

Measuring Sample Path Causal Influences with Relative Entropy

Gabriel Schamberg, *Member, IEEE*, and Todd P. Coleman, *Senior Member, IEEE*,

Abstract—We present a sample path dependent measure of causal influence between time series. The proposed causal measure is a random sequence, a realization of which enables identification of specific patterns that give rise to high levels of causal influence. We show that these patterns cannot be identified by existing measures such as directed information (DI). We demonstrate how sequential prediction theory may be leveraged to estimate the proposed causal measure and introduce a notion of regret for assessing the performance of such estimators. We prove a finite sample bound on this regret that is determined by the worst case regret of the sequential predictors used in the estimator. Justification for the proposed measure is provided through a series of examples, simulations, and application to stock market data. Within the context of estimating DI, we show that, because joint Markovicity of a pair of processes does not imply the marginal Markovicity of individual processes, commonly used plug-in estimators of DI will be biased for a large subset of jointly Markov processes. We introduce a notion of DI with “stale history”, which can be combined with a plug-in estimator to upper and lower bound the DI when marginal Markovicity does not hold.

Index Terms—Causality, Relative Entropy, Sequential Prediction, Granger, Markov Process.

I. INTRODUCTION

The need to identify and quantify causal relationships between two or more time series from purely observational data is ubiquitous across academic disciplines. Recent examples include economists seeking to understand the directional interdependencies between foreign stock indices [1], [2] and neuroscientists seeking to identify directed networks that explain neural spiking activity [3], [4].

Building upon the the ideas of Wiener [5], Granger [6] proposed the following perspective on causal influence: *We say that a time series Y is “causing” X if we can better predict X given the past of all information than given the past of all information in the universe excluding Y .* While Granger’s original treatment only considered linear Gaussian

This work was supported in part by the Center for Science of Information (CSoI), NSF Science and Technology Center, under grant agreement CCF-0939370; ARO MURI award under contract ARO-W911NF-15-1-0479; NIH award 1R01MH110514; and NSF award IIS-1522125. The material in this paper was presented in part at the 2018 IEEE International Symposium on Information Theory, Vail, CO, and in part at the 2019 IEEE International Symposium on Information Theory, Paris, FR.

G. Schamberg is with the Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA (e-mail: gabes@mit.edu).

T.P. Coleman is with the Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA (e-mail: tpcoleman@ucsd.edu).

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

regression models, his proposed definition applies in general and is here collectively termed *Granger causality* (GC). The inclusion of “all information in the universe” in GC serves to avoid the effects of confounding, i.e. to avoid incorrectly inferring that Y influences X when in reality both X and Y are influenced by a third process, Z . It is important to note that GC lacks mention of *interventions*, a concept that is central to well-accepted notions of causal influence popularized by Pearl [7], [8]. In [9], Eichler and Didelez develop a framework that formalizes interventions in the context of GC, enabling the distinction between scenarios where changing the value of Y (by means of an intervention) results in a change in the value of X , and those where Y merely aids in the *prediction* of X . Absent this formal analysis, GC is better viewed as a measure of predictive utility. Nevertheless, it continues to be a popular tool (for example [10]), and is the focus of this paper. Thus, we use the terms “cause,” “causal effect,” etc. within the context of Granger’s perspective unless otherwise stated.

More modern information theoretic interpretations of Granger’s perspective on causality include directed information (DI) [11]–[13] and transfer entropy (TE) [14], which is equivalent to GC for Gaussian autoregressive processes [15]. Justification for use of the DI for characterizing directional dependencies between processes was given in [16], where it was shown that, under mild assumptions, the DI graph is equivalent to the so-called minimal generative model graph. It was further shown in [17] that the DI graph can be viewed as a generalization of linear dynamical graphs. As a result, the directional dependencies encoded by DI are well equipped to identify the presence or absence of a causal link under Granger’s perspective in the general non-linear and non-Gaussian settings.

Interestingly, both GC and DI are determined entirely by the underlying probabilistic model (i.e. joint distribution) of the random processes in question. It is clear that once the model is determined, these methods provide no ability to distinguish between varying levels of causal influence that may be associated with *specific realizations* of those processes. As a result, GC and DI are only well suited to answer causal questions that are concerned with average influences between processes. Examples of this style of question include “*Does dieting affect body weight?*” and “*Does the Dow Jones stock index influence Hang Seng stock index?*”. Symbolically, we represent this question as Q1: “*Does Y^{i-1} cause X_i ?*”, where the superscript represents and the collection of samples up to time $i - 1$ and capital letters are used to represent random variables and processes.

A natural next question to ask is how the aforementioned

measures may be adapted to be sample path dependent. In particular, one might pose the question Q2: “*Did y^{i-1} cause x_i ?*”, where the lowercase letters now represent specific realizations of the processes X and Y . Examples of these questions would be “*Did eating salad cause me to lose weight?*” and “*Did the dip in the price of the Dow Jones cause the spike in the price of the Hang Seng?*”. One information theoretic approach to answering Q2 is the substitution of self-information for entropy wherever entropy appears in the definition of DI [18]. The issue is that the resulting “local” extension of DI may take on negative values, and it is unclear how these values should be interpreted with regard to the presence/absence of a causal link. As a result, causal measures that use the self information have not seen widespread adoption. While this may appear to be a result of a particular methodology, it is in fact a fundamental challenge with Q2 arising from the handling of *counterfactuals*. This challenge relates to what Holland [19] referred to as the “fundamental problem of causal influence,” namely that we cannot observe the value that X_i would take under two realizations of Y^{i-1} , i.e. the true realization y^{i-1} and some counterfactual realization \tilde{y}^{i-1} . A popular approach to dealing with counterfactuals is structural equation models (SEMs). Using an SEM, one can estimate the “noise” that gave rise to an outcome x_i and infer the \tilde{x}_i that would have occurred had y^{i-1} been \tilde{y}^{i-1} . The interested reader is referred to [7] and [20] for more details on SEMs.

While it is clear that Q1 lacks the resolution to identify specific points on a sample path for which a large causal influence is elicited, Q2 introduces the added challenge of counterfactuals and thus there is no clear approach within the GC framework. This observation motivates our proposed question of study, Q3: “*Does y^{i-1} cause X_i ?*”. In other words, we seek to identify the causal effect that particular values of Y have on *the distribution* of the subsequent sample of X . Examples of this include “*Which diets are most informative about weight loss outcomes?*” and “*When does the Dow Jones have the greatest effect on the Hang Seng?*”. To answer this question, we build on the work of [21] and [22] in the development of a sample path dependent measure of causal influence.

Such a measure will necessarily capture dynamic changes in causal influence between processes. The means by which causal influences vary with time is two-fold. First, it is clear that when the joint distribution of the collection of processes is non-stationary, there will be variations in time with respect to their causal interactions. Second, we note that stationary processes may exhibit time-varying causal phenomena when certain realizations of a process have a greater level of influence than others (see Section III-A1). The latter cannot be captured by GC and DI, which are determined entirely by the joint distribution and thus will only change when the distribution changes. Furthermore, since estimating GC and DI requires taking a time-average, capturing dynamic changes resulting from non-stationarities necessitates approximating an expectation using a sliding window. The sample path dependent measure, on the other hand, captures both types of temporal dynamics: estimates of the sample path measure can be obtained for any processes for which we can have reliable

sequential prediction algorithms.

In developing techniques for estimating the proposed measure, we have identified a challenge in estimating information theoretic measures of causal influence that has been commonly overlooked in the literature. While it is well understood that a collection of jointly Markov processes does not necessarily exhibit Markovicity for *subsets* of processes, the implications of this on information theoretic causal measures are not well studied. An analogous statement with regard to finite order autoregressive processes and the biasing effect this has on estimates of GC was studied in [23], but this work has yet to be adopted in the information theory community. It comes as no surprise that the issues with GC estimators identified in [23] may be extended to DI estimators. Thus, a characterization of when estimators of DI are unbiased and a means of addressing the bias when it arises are lacking. As such, we build upon our earlier work [24] to address both of these unmet needs in Section IV-A in an effort to establish an understanding of when one can expect to obtain unbiased estimates of information theoretic causal measures.

The contributions of this paper may be summarized as follows:

- A methodology for assessing causal influences between time series in a sample-path specific and time-varying manner, by answering the question “Does y^{i-1} cause X_i ?”. This is particularly relevant when there are infrequent events which exhibit large causal influences, which would be “averaged out” using any causal measure (e.g. GC and DI) which takes an average over all sample paths.
- A framework using sequential prediction for estimating the dynamic causal measure with associated upper bounds on the worst case “causality regret”.
- A characterization of when unbiased estimates of DI can be obtained using concepts from causal graphical models and a novel methodology for bounding the DI when unbiased estimates cannot be obtained.
- Demonstration of the causal measure’s value through application to simulated and real data.

The remainder of this paper is organized as follows: following a brief overview of notation, Section II provides a technical summary of related work. In Section III we define the measure, present key properties, and provide justification for the measure through several examples. Section IV provides a framework for estimating the measure. Section V demonstrates the measure on simulated and real data. Finally, Section VI contains a discussion of the results and opportunities for future work.

A. Notation and Definitions

Let X , Y , and Z denote discrete finite-alphabet random processes, unless otherwise specified. We denote processes at a given time point with a subscript and denote the space of values they may take with caligraphic letters, i.e. $X_i \in \mathcal{X}$. Without loss of generality, let $\mathcal{X} = \{0, 1, 2, \dots, |\mathcal{X}| - 1\}$. A temporal range of a process is denoted by a subscript and superscript, i.e. $X_i^n = (X_i, X_{i+1}, \dots, X_n)$, and we define $X^n \triangleq X_1^n$. Realizations of processes are given by lowercase

letters. Probability mass functions (pmfs) are equivalently referred to as “distributions” and are denoted by f . These distributions are characterized by a subscript, which is often omitted when context allows. For example $p_{X_i}(x_i) \equiv p(x_i)$ gives the distribution of a single time point of X , $p_{X^n, Y^n}(x^n, y^n) \equiv p(x^n, y^n)$ gives the joint distribution of X and Y , and $p_{X_i|X^{i-1}}(x_i | x^{i-1}) \equiv p(x_i | x^{i-1})$ gives the conditional distribution of X at a single time conditioned on the past of X . Lastly, we define the *causally conditional distribution with lag k* as:

$$p(x^n || y^{n-k}) \triangleq \prod_{i=1}^n p(x_i | x^{i-1}, y^{i-k}). \quad (1)$$

Note that the standard interpretation of the causal¹ conditioning (as in [25]) is recovered by letting $k = 0$.

We will briefly review some information theoretic quantities that are used frequently throughout the paper. The entropy is given by:

$$H(X^n) = \sum_{x^n} p(x^n) \log \frac{1}{p(x^n)}$$

where it is implied that the sum is over all $x^n \in \mathcal{X}^n$ and the logarithm is base two (as are all logarithms throughout). The conditional entropy is given by:

$$H(X^n | Y^n) = \sum_{x^n, y^n} p(x^n, y^n) \log \frac{1}{p(x^n | y^n)}$$

The causally conditional entropy is given by substituting the causally conditional distribution for the conditional distribution:

$$H(X^n || Y^{n-k}) = \sum_{x^n, y^n} p(x^n, y^n) \log \frac{1}{p(x^n || y^{n-k})}$$

For any of the above defined variants of entropy, the corresponding entropy *rates* are given by:

$$\bar{H}(X) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X^n) \quad (2)$$

$$\bar{H}(X | Y) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X^n | Y^n) \quad (3)$$

$$\bar{H}^{(k)}(X || Y) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X^n || Y^{n-k}) \quad (4)$$

It should be noted that the entropy rates may not exist for all processes.

The conditional mutual information is given by:

$$I(X^n; Y^n | Z^n) = H(X^n | Z^n) - H(X^n | Y^n, Z^n)$$

with the (unconditional) mutual information $I(X^n; Y^n)$ being obtained by removing Z^n everywhere it appears in the above equation. Finally, the relative entropy or KL-divergence between two distributions p_{X_i} and p'_{X_i} is given by:

$$D(p_{X_i} || p'_{X_i}) = \sum_{x_i} p(x_i) \log \frac{p(x_i)}{p'(x_i)}.$$

¹The term “causal” is overloaded, as it is used here in the control theoretic sense strictly to mean “non-anticipative.”

II. RELATED WORK

We now provide a brief summary of three key concepts in the measurement of causal influence across time series, namely Granger causality (GC) [6], directed information (DI) [12], [13], and causal strength (CS) [26]. While some key points will be presented here, a comprehensive summary of the relationships between GC and DI may be found in [27].

A. Granger Causality

While Granger’s perspective on causality underlies most modern studies in causality between time series, his original treatment was limited to linear Gaussian AR models [6]. For clarity, we will here present the case with scalar time series. Formally, define the three real-valued random processes $(X_i, Y_i, Z_i : i \geq 1)$. As in Granger’s original treatment, we let Z^n represent all the information in the universe in order to avoid the effects of confounding. Next, define two models of X_i :

$$X_i = \sum_{j=1}^d a_j X_{i-j} + b_j Y_{i-j} + c_j Z_{i-j} + U_i \quad (5)$$

$$X_i = \sum_{j=1}^d d_j X_{i-j} + e_j Z_{i-j} + V_i \quad (6)$$

where $a_j, b_j, c_j, d_j, e_j \in \mathbb{R}$ are the model parameters and $U_i \sim \mathcal{N}(0, \sigma_U^2)$ and $V_i \sim \mathcal{N}(0, \sigma_V^2)$. We see that the class of models given by (6) is a subset of the models given by (5) where the next X_i does not depend on past Y^{i-1} . Thus, a non-negative measure of the extent of causal influence of Y on X may be defined by:

$$G_{Y \rightarrow X} \triangleq \ln \frac{\sigma_V^2}{\sigma_U^2} \quad (7)$$

The limitations of Granger causality extend considerably beyond the restriction to linear models (see [23] for a comprehensive summary). Of particular interest is the fact that if a VAR process is of finite order, subsets of the process will in general be infinite order. While it is possible to redefine the model in (6) to be infinite order, this creates obvious challenges in attempting to estimate Granger causality. Considering this issue is not addressed by the subsequent existing methods, we will revisit this issue in Section IV-A.

B. Directed Information

The concept directed information was first introduced under the name transinformation by Marko in 1973 [12] in the context of bidirectional communication theory. It was later revisited in 1990 by Massey [13] who defined the directed information from a sequence Y^n to X^n as:

$$I(Y^n \rightarrow X^n) = \sum_{i=1}^n I(Y^i; X_i | X^{i-1}) \quad (8)$$

$$= H(X^n) - H(X^n || Y^n) \quad (9)$$

Unless otherwise specified, we assume that there are no instantaneous causations, i.e. that X_i and Y_i are conditionally independent given the past X^{i-1} and Y^{i-1} . Should one want

to allow for instantaneous causations, the proposed methods may be trivially extended to accommodate. As such, we will primarily consider the reverse DI [28]:

$$I(Y^{n-1} \rightarrow X^n) = \sum_{i=1}^n I(Y^{i-1}; X_i | X^{i-1}) \quad (10)$$

$$= H(X^n) - H(X^n || Y^{n-1}) \quad (11)$$

noting that under the assumption of no instantaneous causation, $I(Y^n \rightarrow X^n) = I(Y^{n-1} \rightarrow X^n)$.

Given that the DI is given by a sum over time, one may be interested in how causal relationships are exhibited on average. This can be accomplished through use of the directed information rate [25], given by:

$$\bar{I}(Y \rightarrow X) = \lim_{n \rightarrow \infty} \frac{1}{n} I(Y^{n-1} \rightarrow X^n) \quad (12)$$

$$= \bar{H}(X) - \bar{H}^{(1)}(X || Y) \quad (13)$$

Lastly, we note that in order to avoid confounding, we may include the side information Z^n to get the following definitions of causally conditioned DI and causally conditioned DI rate:

$$I(Y^{n-1} \rightarrow X^n || Z^{n-1}) = \sum_{i=1}^n I(Y^i; X_i | X^{i-1}, Z^{i-1}) \quad (14)$$

$$\bar{I}(Y \rightarrow X || Z) = \lim_{n \rightarrow \infty} \frac{1}{n} I(Y^{n-1} \rightarrow X^n || Z^{n-1}) \quad (15)$$

C. Causal Strength

In [26], Janzing et al. propose an axiomatic measure of *causal strength* (CS) based on a set of postulates that they propose should be satisfied by a causal measure. Furthermore, they present numerous examples to illustrate where Granger causality and directed information do not give results consistent with intuition. While this measure was proposed to measure influences in general causal graphs, it has a clear interpretation in the context of measuring causal influences between two time series. In particular, for measuring the CS from Y to X , begin by considering the generalization of the two models utilized by GC in (5) and (6) to arbitrary probability distributions $p(X_i | X^{i-1}, Y^{i-1}, Z^{i-1})$ and $p(X_i | X^{i-1}, Z^{i-1})$. Next, note that the second distribution has the following factorization when summing over all possible pasts of Y :

$$p(X_i | X^{i-1}, Z^{i-1}) = \sum_{y^{i-1}} p(X_i | X^{i-1}, y^{i-1}, Z^{i-1}) p(y^{i-1} | X^{i-1}, Z^{i-1})$$

The first term in the sum may be viewed as measuring the *direct* effects of the pasts of X , Y , and Z , on the distribution of X_i . The second term, however, is in some sense measuring the *indirect* effects of the pasts of X and Z on X_i in that they affect the distribution of X_i through their effect on the distribution of Y^{i-1} . Thus, the key idea behind CS is the introduction of the “post-cutting” distribution, where the conditional distribution found in the second term is replaced with a marginal distribution (see Section 4.1 of [26] for a

formal definition). As a result, the (time series) CS from Y to X with side information Z is given by:

$$\mathcal{C}_{Y \rightarrow X} \triangleq E [D(p_{X_i | X^{i-1}, Y^{i-1}, Z^{i-1}} || \tilde{p}_{X_i | X^{i-1}, Z^{i-1}})] \quad (16)$$

where the expectation is taken with respect to $p_{X^{i-1}, Y^{i-1}, Z^{i-1}}$ and the post-cutting distribution is defined as:

$$\tilde{p}_{X_i | X^{i-1}, Z^{i-1}}(X_i) \triangleq \sum_{y^{i-1}} p(X_i | X^{i-1}, y^{i-1}, Z^{i-1}) p(y^{i-1}) \quad (17)$$

The post-cutting distribution is designed to ensure that the extent to which Y has a causal effect on X depends only upon Y and other *direct* causes of X (see P2 in [26]). In the context of measuring causal influences between time series, this can be seen as correcting for scenarios in which X may be very well *predicted* by its own past while not being *caused* by its own past. This scenario arises in models like the one depicted in the center of Figure 1. In such a scenario, it is possible to have $I(Y^{n-1} \rightarrow X^n) = 0$ despite the fact that Y_{i-1} is, in some sense, the sole cause of X_i . The details of this example are made clear in Section III-A2.

By presenting an axiomatic framework for measuring causal influences, Janzing et al. provide a robust justification CS. With that said, we note that like GC and DI, CS is determined solely by the underlying probabilistic model. As such, it may be the preferred technique for addressing Q1, but it does not represent how different realizations may give rise to different levels of causal influence.

D. Self-Information Measures

All of the aforementioned techniques involve taking an expectation over the histories of the time series in question, and are thus well suited to address Q1. In order to address Q2, a notion of locality may be introduced through use of *self-information*. For a given realization x of a random variable $X \sim p_X$, the self-information is given by $h(x) \triangleq -\log p_X(x)$ and represents the amount of surprise associated with that realization. By replacing entropy with self-information, and its conditional form $h(x | y) \triangleq -\log p_{X|Y}(x | y)$, a local version of DI and its conditional extension may be obtained (see Table 1 in [29] for other so-called “local measures”). As an example, we note that for a given pair of realizations x^i and y^{i-1} , a “directed information density” (using the language of [30]) may be given by:

$$i(y^{n-1} \rightarrow x^n) = \sum_{i=1}^n \log \frac{p(x_i | x^{i-1}, y^{i-1})}{p(x_i | x^{i-1})} \quad (18)$$

While this indeed creates a sample path measure of causality whose expectation is DI, it is clear it may take on negative values. Such a scenario occurs when the knowledge that $Y^{i-1} = y^{i-1}$ makes the observation of $X_i = x_i$ less likely to have occurred. While self-information measures are a good candidate for beginning to address Q2 given their dependence upon realizations, the potential for negative values creates difficulty in trying to obtain an easily interpretable answer in all cases.

E. Time-Varying Causal Measures

A popular extension of GC style causal measures is application to time-varying scenarios [31], [32]. In order to adapt existing methods to these types of scenarios, it is necessary to evaluate them over stretches of time for which there is stationarity. As such, estimation in this scenario necessitates some sort of sliding window technique in order to approximate an expectation, giving rise to a trade-off between sensitivity to dynamic changes and accuracy. Despite being concerned with time-varying causal influences, these approaches are still ultimately attempts to answer Q1 in that the quantity being estimated is determined solely by the underlying joint distribution. The temporal variability that is measured by these approaches is a result only of potential non-stationarities. This is fundamentally different from the question we are asking, which is concerned with the dynamic causal influences that are associated with a particular realization of a process that may or may not be stationary.

III. A SAMPLE PATH MEASURE OF CAUSAL INFLUENCE

We begin by considering the scenario where, having observed $(x^{i-1}, y^{i-1}, z^{i-1})$, we wish to determine the causal influence that y^{i-1} has on the next observation of X_i . Define the *restricted* (denoted (r)) and *complete* (denoted (c)) histories as:

$$\begin{aligned}\mathcal{H}_i^{(r)} &\triangleq \{x_1, \dots, x_{i-1}\} \cup \{z_1, \dots, z_{i-1}\} \\ \mathcal{H}_i^{(c)} &\triangleq \mathcal{H}_i^{(r)} \cup \{y_1, \dots, y_{i-1}\}\end{aligned}$$

The current time samples of side information from the histories (i.e. y_i and z_i) are intentionally omitted, as we assume that there is no instantaneous coupling. We next define the restricted and complete conditional distributions as:

$$\begin{aligned}p_{X_i}^{(r)}(x_i) &\triangleq p_{X_i}(x_i | \mathcal{H}_i^{(r)}) \\ p_{X_i}^{(c)}(x_i) &\triangleq p_{X_i}(x_i | \mathcal{H}_i^{(c)}).\end{aligned}$$

Using these distributions, the sample path causal measure from Y to X in the presence of side information Z at time i is defined by:

$$C_{Y \rightarrow X}(\mathcal{H}_i^{(c)}) \triangleq D(p_{X_i}^{(c)} || p_{X_i}^{(r)}) \quad (19)$$

For ease of notation, we may refer to the causal measure at time i simply as $C_{Y \rightarrow X}(i)$.

The proposed causal measure has an interesting relationship to the directed information. To illustrate this, consider the conditional mutual information term that appears in the sum in (10), along with two equivalent representations:

$$\begin{aligned}I(Y^{i-1}; X_i | X^{i-1}) \\ = H(X_i | X^{i-1}) - H(X_i | X^{i-1}, Y^{i-1})\end{aligned} \quad (20)$$

$$= E_{X^i, Y^{i-1}} \left[\log \frac{p(X_i | X^{i-1}, Y^{i-1})}{p(X_i | X^{i-1})} \right] \quad (21)$$

These equivalent definitions of directed information yield two interpretations. While (20) considers the reduction in uncertainty obtained by conditioning on Y^{i-1} , (21) considers the change in the distribution resulting from the added conditioning as measured by a log-likelihood ratio. When we wish

to condition on a realization $(X^{i-1}, Y^{i-1}) = (x^{i-1}, y^{i-1})$, these representations are no longer equivalent:

$$H(X_i | x^{i-1}) - H(X_i | x^{i-1}, y^{i-1}) \quad (22)$$

$$\neq E_{X_i} \left[\log \frac{p(X_i | x^{i-1}, y^{i-1})}{p(X_i | x^{i-1})} \middle| x^{i-1}, y^{i-1} \right] \quad (23)$$

The representation given by (23) is chosen to be the sample path causal measure and is indeed equivalent to the proposed measure in (19). This choice is made clear by noting two properties of (22). First, we note that (22) may be negative. Second, for particular realizations of x^{i-1} and y^{i-1} , we may have that conditioning on y^{i-1} drastically shifts the distribution of X_i while only mildly affecting the conditional entropy, yielding a value of nearly zero for a scenario when there is a clear causal influence. We note that the difference between definitions of DI that is induced by conditioning on a realization is acknowledged in [28], where four unique estimators of DI are proposed based on these various equivalent definitions of DI. While these estimators converge to the same result in the estimation of DI, the different perspectives yield different results for the question we are addressing and thus their implications must be considered.

As a result of the added conditioning, the proposed measure is a *random variable* that takes on a value for each possible history and may be related to the directed information as follows:

Proposition 1. *In the absence of instantaneous influences, the sum of the expectation over sample paths of the proposed causal measure is the directed information:*

$$\sum_{i=1}^n E[C_{Y \rightarrow X}(\mathcal{H}_i^{(c)})] = I(Y^{n-1} \rightarrow X^n || Z^{n-1}) \quad (24)$$

See Appendix D-A for a proof of the proposition.

A second key property of the proposed measure is non-negativity (for any history), which follows directly from the properties of the KL-divergence. Furthermore, the measure will take a value of zero if and only if the complete and restricted distributions are equivalent for a given history. As such, the proposed causal measure may take on a large value when the additional condition on y^{i-1} introduces a large amount of *uncertainty* into the distribution of X_i . In such a scenario, we would expect y^{i-1} to have a significant causal influence on X_i even though it is not causing X_i to take on a *specific value*. It is this type of scenario that makes Q2 so difficult to answer in a consistent manner, despite having a clear interpretation in terms of Q3.

Remark 1. *Despite the fact that Granger's perspective on causal influence includes no remarks on the role of interventions, it may be of interest to consider a version of the proposed measure where the value of the influencer is forced by means of an intervention². For example, one might wish to consider:*

$$C_{Y \rightarrow X}^{(i)}(\mathcal{H}_i^{(c)}) \triangleq D(p_{X_i}^{(i)} || p_{X_i}^{(r)}) \quad (25)$$

²We thank an anonymous reviewer for this suggestion.

where the first argument of the divergence is defined as the interventional distribution:

$$p_{X_i}^{(i)}(x_i) \triangleq p(x_i | \mathcal{H}_i^{(r)}, do(Y^{i-1} = y^{i-1})) \quad (26)$$

We have here used the *do*-operator of Pearl [7] to represent the action of forcing Y^{i-1} to take the values y^{i-1} irrespective of the probability with which those values occur. Given that it is often infeasible to perform such interventions in real world scenarios, a large body of causality research is focused on determining when these interventional distributions can be learned from observed data. While providing a general characterization of the scenarios for which $C_{Y \rightarrow X}^{(i)} = C_{Y \rightarrow X}$ is outside the scope of this paper, this equivalence does in fact hold for the three examples in the following section (with a mild technical assumption). This follows intuitively from the depictions in Figure 1 and is shown formally in Appendix A. By contrast, we would not expect this equivalence to hold for the stock market example considered in Section V-B, as discussed in Remark 2.

A. Justification for Measurement of Sample Path Influences

We now present a series of examples that illustrate the value of a sample path causal measure. Graphical representations of the three examples can be seen in Figure 1.

1) *IID Influences*: Let $Y_i \sim \text{Bern}(\epsilon)$ iid for $i = 1, 2, \dots$ and:

$$X_i \sim \begin{cases} \text{Bern}(p_1), & Y_{i-1} = 1 \\ \text{Bern}(p_2), & Y_{i-1} = 0 \end{cases} \quad (27)$$

for $\epsilon, p_1, p_2 \in [0, 1]$. Intuitively, the extent to which Y_{i-1} influences X_i will vary for different values y_{i-1} provided that $p_1 \neq p_2$. In order to compute the causal measure $C_{Y \rightarrow X}(i)$, we first need to find the restricted distribution of X_i given only its own past:

$$\begin{aligned} p_{X_i}^{(r)}(1) &= \mathbb{P}(X_i = 1 | X^{i-1} = x^{i-1}) \\ &= \mathbb{P}(X_i = 1) \\ &= \sum_{y_{i-1} \in \{0,1\}} \mathbb{P}(X_i = 1 | Y_{i-1} = y_{i-1}) \mathbb{P}(Y_{i-1} = y_{i-1}) \\ &= p_1 \epsilon + p_2 (1 - \epsilon). \end{aligned}$$

Noting that $p_{X_i}^{(c)}(1) = p_1$ when $y_{i-1} = 1$ and $p_{X_i}^{(c)}(1) = p_2$ when $y_{i-1} = 0$, the causal measure is given by:

$$C_{Y \rightarrow X}(i) = \begin{cases} D(p_1 || p_1 \epsilon + p_2 (1 - \epsilon)), & y_{i-1} = 1 \\ D(p_2 || p_1 \epsilon + p_2 (1 - \epsilon)), & y_{i-1} = 0 \end{cases}$$

Thus, we see that as $\epsilon \rightarrow 0$,

$$C_{Y \rightarrow X}(i) \rightarrow \begin{cases} D(p_1 || p_2), & y_{i-1} = 1 \\ 0, & y_{i-1} = 0 \end{cases}$$

By contrast, the DI rate is given by taking the expectation of $C_{Y \rightarrow X}(i)$ over possible values Y_{i-1} . Defining $C_{Y \rightarrow X}(i) \triangleq C_{Y \rightarrow X}(y_{i-1})$, we get:

$$\bar{I}(Y \rightarrow X) = C_{Y \rightarrow X}(1) \epsilon + C_{Y \rightarrow X}(0) (1 - \epsilon) \xrightarrow{\epsilon \rightarrow 0} 0$$

As a result, it is clear that the sample paths that occur with lower probability will give rise to a greater causal measure than

those that occur with higher probability; however, as a result of their lesser probability, these infrequent, highly influential events will have little influence in the computation of the DI rate.

We further note that while it is tempting to invoke ‘‘conditioning reduces entropy’’ to conclude that $C_{Y \rightarrow X}(i) > 0$ represents a *reduction* in uncertainty that is obtained by including the past y^{i-1} in the prediction of X_i , this is not the case. To make this clear, assign values $p_1 \approx 0.5$ and $p_2 \approx 1$ in (27) and again let ϵ approach zero. In such a scenario, we find that:

$$\begin{aligned} p_{X_i}^{(r)}(1) &\approx 1 \\ p_{X_i}^{(c)}(1) &\approx \begin{cases} 1, & y_{i-1} = 1 \text{ (w.p. } 1 - \epsilon) \\ 0.5, & y_{i-1} = 0 \text{ (w.p. } \epsilon) \end{cases} \end{aligned}$$

As such, it is clear that by additionally conditioning on $y_{i-1} = 0$, there is a considerable increase in uncertainty. Thus, while it is certainly true that $H(X_i | X^{i-1}) \leq H(X_i | X^{i-1}, Y^{i-1})$, there are scenarios in which a particular realization of Y^{i-1} may *cause uncertainty* in X_i . Revisiting Q2, it is not clear how to answer the extent to which the event $\{Y_{i-1} = 0\}$ causes any particular outcome $\{X_i = x_i\}$, because all possible outcomes are equally likely. On the other hand, if we consider Q3, it is quite clear that the event $\{Y_{i-1} = 1\}$ has significant influence on X_i and that this is reflected by the proposed measure.

2) *Perturbed Cross Copying*: We next consider a scenario where two processes repeatedly swap values. This example was originally posed in [33] and modified to include noise in [26]. Formally, the processes may be defined as:

$$X_i = \begin{cases} Y_{i-1}, & \text{w.p. } 1 - \epsilon \\ Y_{i-1} \oplus 1, & \text{w.p. } \epsilon \end{cases} \quad Y_i = \begin{cases} X_{i-1}, & \text{w.p. } 1 - \epsilon \\ X_{i-1} \oplus 1, & \text{w.p. } \epsilon \end{cases} \quad (28)$$

where $X_i, Y_i \in \{0, 1\}$ for all i and \oplus is the XOR operator. We again consider the limiting case where ϵ is taken to approach zero. As is shown in [26], the DI rate approaches zero as $\epsilon \rightarrow 0$. This results from the fact that for very small ϵ , Y_{i-1} on average contains virtually no information about X_i that is not contained in X_{i-2} .

Janzing et al. [26] note that because X_i and X_{i-2} are *independent* given Y_{i-1} , Y_{i-1} should, in some sense, be fully responsible for the information that is known about X_i . As a result, for this example their proposed causal strength measures the average reduction in uncertainty obtained by conditioning on Y_{i-1} versus conditioning on *nothing at all*, i.e. $\mathfrak{C}_{Y \rightarrow X} = D(\epsilon || 0.5) \rightarrow 1$ as $\epsilon \rightarrow 0$ (under the assumption the X and Y are initiated by fair coin tosses).

Next, we consider our proposed sample path measure. First, we note that the complete distribution of X_i depends only upon y_{i-1} and the restricted distribution depends only upon x_{i-2} . Explicitly, we get the following distributions:

$$\begin{aligned} p_{X_i}^{(c)}(x_i) &= \begin{cases} 1 - \epsilon & x_i = y_{i-1} \\ \epsilon & x_i \neq y_{i-1} \end{cases} \\ p_{X_i}^{(r)}(x_i) &= \begin{cases} \epsilon^2 + (1 - \epsilon)^2 & x_i = x_{i-2} \\ 2\epsilon(1 - \epsilon) & x_i \neq x_{i-2} \end{cases} \end{aligned}$$

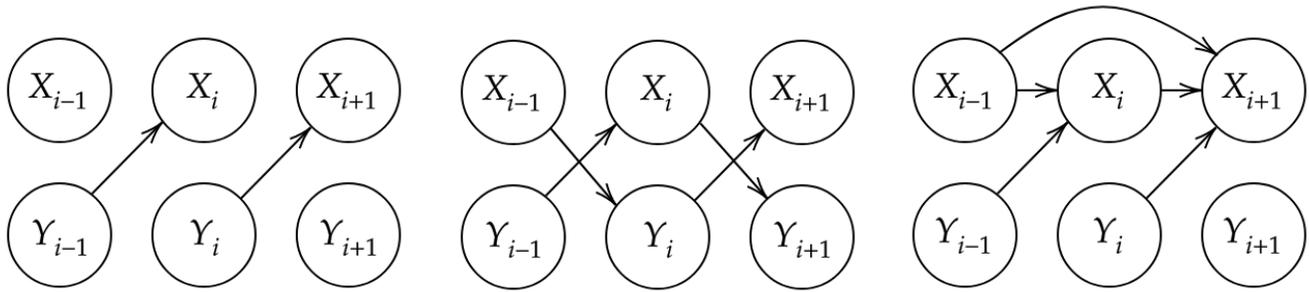


Fig. 1. Graphical representation of the IID influences (left), perturbed cross copying (center), and horse betting (right) examples

As a result, we see that for a given complete history $\mathcal{H}_i^{(c)} = \{x_{i-2}, y_{i-1}\}$ we get:

$$C_{Y \rightarrow X}(\mathcal{H}_i^{(c)}) = \begin{cases} D(\epsilon \parallel 2\epsilon(1-\epsilon)), & x_{i-2} = y_{i-1} \\ D(\epsilon \parallel \epsilon^2 + (1-\epsilon)^2), & x_{i-2} \neq y_{i-1} \end{cases}$$

Thus, we see that as $\epsilon \rightarrow 0$, $C_{Y \rightarrow X} \rightarrow 0$ if $x_{i-2} = y_{i-1}$ and $C_{Y \rightarrow X} \rightarrow \infty$ otherwise.

A comparison of the three measures makes clear that each provides a slightly different perspective. DI rate is loyal to the Granger's perspective in that it captures how, as $\epsilon \rightarrow 0$, Y_{i-1} contains less and less information about X_i that is *not already known*. As a result $\bar{I}(Y \rightarrow X)$ is strictly decreasing for decreasing ϵ . Causal strength, on the other hand, is loyal to the causal Markov condition in the sense that it restricts consideration to only the immediate parents of the node in question (see P2 in Section 2 of [26]). As such, decreasing ϵ yields a smaller level of uncertainty in X_i conditioned on Y_{i-1} , and therefore the causal strength is strictly increasing for decreasing ϵ . The proposed measure lies somewhere in between the two in that it simultaneously captures the decrease and increase in effect of Y on X as ϵ shrinks. Deciding which perspective is "correct" is a philosophical question that must be answered on a problem-by-problem basis. In any case, the proposed measure provides an interesting perspective that, to our knowledge, has not been considered in the literature.

3) *Horse Betting*: Consider the problem of horse race gambling with side information as presented in Section III-A of [34] (with minor adjustments to notation). At each time i the gambler bets all of their wealth based on the past winners $X^{i-1} \in [M]^{i-1}$ and side information Y^{i-1} . As a result, the gambler's wealth at time i , denoted $w(X^i, Y^{i-1})$, is a function of the winning horses and side information up to that time. Lastly, the amount of money that is won for betting on the winning horse is given by the odds $o(X_i | X^{i-1})$, and the portion of wealth bet on each horse is given by $b(X_i | X^{i-1}, Y^{i-1}) \geq 0$ with $\sum_x b(x | X^{i-1}, Y^{i-1}) = 1$. Thus, the evolution of the wealth can be described recursively as:

$$\begin{aligned} w(X^i, Y^{i-1}) \\ = b(X_i | X^{i-1}, Y^{i-1})o(X_i | X^{i-1})w(X^{i-1}, Y^{i-2}) \end{aligned}$$

Finally, the expected growth rate of the wealth is defined as $\frac{1}{n} E[\log w(X^n, Y^{n-1})]$.

It is shown in [34] that the betting strategy that maximizes the expected growth rate is given by distributing bets according to the conditional distribution of X_i given all available information:

$$b^*(X_i | X^{i-1}, Y^{i-1}) = p(X_i | X^{i-1}, Y^{i-1}).$$

Similarly, we can define a restricted betting strategy $b(X_i | X^{i-1})$ where the side information is not available (and optimal strategy $b^*(X_i | X^{i-1}) = p(X_i | X^{i-1})$). The wealth that is obtained under that strategy is then given by:

$$w(X^i) = b(X_i | X^{i-1})o(X_i | X^{i-1})w(X^{i-1})$$

Letting $w^*(X^i, Y^{i-1})$ and $w^*(X^i)$ represent the wealth resulting from using the optimal strategies, it is further shown in [34] that the increase in growth rate resulting from including side information in the betting strategy is given by:

$$\frac{1}{n} E[\log w^*(X^n, Y^{n-1}) - \log w^*(X^n)] = \frac{1}{n} I(Y^{n-1} \rightarrow X^n) \quad (29)$$

It should be noted that the result in (29) holds for any choice of odds $o(X_i | X^{i-1})$. Thus, we proceed by making the mild assumption that the odds chosen by the racetrack are such that, for any past sequence of winners x^{i-1} , the gambler optimally betting without side information is expected to lose money on round i :

$$E[\log b^*(X_i | X^{i-1})o(X_i | X^{i-1}) | x^{i-1}] = \log \delta < 0 \quad (30)$$

for some $0 < \delta < 1$. We define the above equation as the conditional expected growth rate for race i (without side information). As a consequence, this implies a negative expected growth rate for the gambler's wealth without side information:

$$\begin{aligned} E[\log w^*(X^n)] \\ = E[\log b^*(X_n | X^{n-1})o(X_n | X^{n-1})] + E[\log w^*(X^{n-1})] \\ = \sum_{i=1}^n E[\log b^*(X_i | X^{i-1})o(X_i | X^{i-1})] + \log w_0 \\ = n \log \delta < 0 \end{aligned}$$

where the initial wealth w_0 is assumed, without loss of generality, to be 1.

It follows that a gambler with access to side information ought to gamble only if their expected growth rate is greater

than zero. Applying this condition to (29), a gambler with side info can expect to win money if:

$$\frac{1}{n} I(Y^{n-1} \rightarrow X^n) > -\log \delta \quad (31)$$

Thus, when equipped with the DI, a gambler will decide either to visit the racetrack and bet on every race or to stay at home. It turns out, however, that the gambler may be doing themselves a disservice by staying home any time that (31) does not hold. To see this, suppose that before race i the gambler has witnessed winners x^{i-1} and side information y^{i-1} , and wishes to gamble if they expect to make money on the current race. Such a scenario occurs when the conditional expected growth rate for round i is positive:

$$E[\log b^*(X_i | X^{i-1}, Y^{i-1}) o(X_i | X^{i-1}) | x^{i-1}, y^{i-1}] > 0 \quad (32)$$

Combining (32) with the rate for round i in (30), the condition for which the gambler should place a bet becomes:

$$\begin{aligned} & E[\log b^*(X_i | X^{i-1}, Y^{i-1}) - \log b^*(X_i | X^{i-1}) | x^{i-1}, y^{i-1}] \\ &= \sum_{x_i} p(x_i | x^{i-1}, y^{i-1}) \log \frac{b^*(x_i | x^{i-1}, y^{i-1})}{b^*(x_i | x^{i-1})} \\ &= \sum_{x_i} p(x_i | x^{i-1}, y^{i-1}) \log \frac{p(x_i | x^{i-1}, y^{i-1})}{p(x_i | x^{i-1})} \\ &= C_{Y \rightarrow X}(x^{i-1}, y^{i-1}) \\ &> -\log \delta \end{aligned}$$

Thus we can see that while the DI represents the *time averaged* expected increase in wealth growth rate resulting from side information, the proposed measure gives the *per round* expected increase. It is important to note that with problems in communication theory, low probability events may indeed be of little concern, and thus the DI may be the correct technique with which to analyze the relationship between Y and X . In the case of betting and the applications discussed in Section V-B, we note that there may be great interest in how the two time series interact for specific realizations, even if those realizations are rare.

IV. ESTIMATING THE CAUSAL MEASURE

An estimate of the causal measure can be obtained by simply estimating the complete and restricted distributions and then computing the KL divergence between the two at each time. Such an estimator allows us to leverage results from the field of sequential prediction [35]. The sequential prediction problem formulation we consider is as follows: for each round $i \in \{1, \dots, n\}$, having observed some history \mathcal{H}_i , a learner selects a probability assignment $\hat{p}_i \in \mathcal{P}$, where \mathcal{P} is the space of probability distributions over \mathcal{X} . Once \hat{p}_i is chosen, x_i is revealed and a loss $l(\hat{p}_i, x_i)$ is incurred by the learner, where the loss function $l : \mathcal{X} \rightarrow \mathbb{R}$ is chosen to be the self-information loss given by $l(p, x) = -\log p(x)$.

The performance of sequential predictors may be assessed using a notion of *regret* with respect to a reference class of probability distributions $\tilde{\mathcal{P}} \subset \mathcal{P}$. For a given round i and reference distribution $\tilde{p}_i \in \tilde{\mathcal{P}}$, the learner's regret is:

$$r(\hat{p}_i, \tilde{p}_i, x_i) = l(\hat{p}_i, x_i) - l(\tilde{p}_i, x_i) \quad (33)$$

In many cases the performance of sequential predictors will be measured by the worst case regret, given by:

$$R_n(\tilde{\mathcal{P}}_n) = \sup_{x^n \in \mathcal{X}^n} \sum_{i=1}^n l(\hat{p}_i, x_i) - \inf_{\tilde{p} \in \tilde{\mathcal{P}}_n} \sum_{i=1}^n l(\tilde{p}_i, x_i) \quad (34)$$

$$\triangleq \sup_{x^n \in \mathcal{X}^n} \sum_{i=1}^n r(\hat{p}_i, f_i^*, x_i) \quad (35)$$

where $p_i^* \in \tilde{\mathcal{P}}$ is defined as the distribution from the reference class with the smallest cumulative loss up to time n , i.e. the \tilde{p}_i for which R_n is largest. We also define $p^* \in \tilde{\mathcal{P}}_n \subset \mathcal{P}^n$ to be the cumulative loss minimizing *joint* distribution, noting that the reference class of joint distributions $\tilde{\mathcal{P}}_n$ is not necessarily equal to $\tilde{\mathcal{P}}^n$ (i.e. $\tilde{\mathcal{P}} \times \tilde{\mathcal{P}} \times \dots$), as often times there may be a constraint on the selection of the best reference distribution that is imposed in order to establish bounds. In the absence of any restrictions, the reference distributions may be selected at each time such that $p_i^*(x_i) = 1$, resulting in zero cumulative loss for any sequence x^n . Thus, sequential prediction problems impose restrictions on the reference distributions with which to compare predictor performance [35]. For example, one may assume stationarity by enforcing $p_1^* = p_2^* = \dots = p_n^*$ or assume that $p_i^* = p_{i+1}^*$ for all but some small number of indices. For various learning algorithms (i.e. strategies for selecting \hat{p}_i given \mathcal{H}_i) and reference classes $\tilde{\mathcal{P}}_n$, these bounds on the worst case regret are defined as a function of the sequence length n :

$$R_n(\tilde{\mathcal{P}}_n) \leq M(n) \quad (36)$$

It follows naturally that an estimator for our causal measure can be constructed by building two sequential predictors. The restricted predictor $\hat{p}_{X_i}^{(r)}$ computed at each round using $\mathcal{H}_i^{(r)}$, and the complete predictor $\hat{p}_{X_i}^{(c)}$ computed at each round using $\mathcal{H}_i^{(c)}$. It then follows that each of these predictors will have an associated worst case regret, given by $R_n^{(r)}(\tilde{\mathcal{P}}_n^{(r)})$ and $R_n^{(c)}(\tilde{\mathcal{P}}_n^{(c)})$, where $\tilde{\mathcal{P}}_n^{(r)}$ and $\tilde{\mathcal{P}}_n^{(c)}$ represent the restricted and complete reference classes. Using these sequential predictors, we define our estimated causal influence from Y to X at time i as:

$$\hat{C}_{Y \rightarrow X}(i) = D(\hat{p}_{X_i}^{(c)} \parallel \hat{p}_{X_i}^{(r)}) \quad (37)$$

It should be noted that when averaged over time, this estimator becomes a universal estimator of the directed information rate for certain predictors and classes of signals [28].

To assess the performance of an estimate of the causal measure, we define a notion of causality regret:

$$CR(n) \triangleq \sum_{i=1}^n \left| \hat{C}_{Y \rightarrow X}(i) - C_{Y \rightarrow X}^*(i) \right| \quad (38)$$

where we define:

$$C_{Y \rightarrow X}^*(i) = D(p_{X_i}^{(c)*} \parallel p_{X_i}^{(r)*}) \quad (39)$$

with $p_{X_i}^{(c)*} \in \tilde{\mathcal{P}}^{(c)}$ and $p_{X_i}^{(r)*} \in \tilde{\mathcal{P}}^{(r)}$ defined as the loss minimizing distributions from the complete and restricted reference classes. We note that with this notion of causal regret, the estimated causal measure is being compared against the best estimate of the causal measure from within a reference

class. As such, we limit our consideration to the scenario in which the reference classes are sufficiently representative of the true sequences to produce a desirable $C_{Y \rightarrow X}^*$ (i.e. $C_{Y \rightarrow X}^*(i) \approx C_{Y \rightarrow X}(i)$ for all i).

We now present the necessary assumptions for proving a finite sample bound on the estimates of causality regret.

Assumption 1. For sequential predictors $\hat{p}_{X_i}^{(c)}$ and $\hat{p}_{X_i}^{(r)}$ and observations $(x^n, y^n, z^n) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n$, we assume that $\hat{p}_{X_i}^{(c)}$ and $\hat{p}_{X_i}^{(r)}$ are absolutely continuous with respect to each other, i.e.:

$$\sup_{x \in \mathcal{X}} \left| \log \frac{\hat{p}_{X_i}^{(c)}(x)}{\hat{p}_{X_i}^{(r)}(x)} \right| < \infty \quad i = 1, \dots, n \quad (40)$$

Clearly, the above assumption will be satisfied for any sequential prediction algorithm that does not assign zero probability to any outcomes.

Assumption 2. For loss minimizing distributions $p_{X_i}^{(c)*} \in \tilde{\mathcal{P}}^{(c)}$ and $p_{X_i}^{(r)*} \in \tilde{\mathcal{P}}^{(r)}$, restricted sequential predictor $\hat{p}_{X_i}^{(r)}$, and observations $(x^n, y^n, z^n) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n$:

$$\sum_{i=1}^n \left| E_{p_{X_i}^{(c)*}} \left[r(\hat{p}_{X_i}^{(r)}, p_{X_i}^{(r)*}, X_i) \right] \right| \leq M^{(r)}(n) \quad (41)$$

While it is understood that the expected regret is in general bounded by worst case regret, Assumption 2 requires that the reference classes are sufficiently rich that the expected regret is not too large in *absolute value*. This is necessary in bounding the causality regret because unlike the regret defined by (34), $CR(n)$ increases when the estimated distributions outperform the regret minimizing distributions.

We now present our main theoretical result, a finite sample bound on the causality regret under Assumptions 1 and 2:

Theorem 1. Let the worst case regret for the predictors $\hat{p}_{X_i}^{(r)}$ and $\hat{p}_{X_i}^{(c)}$ be bounded by $R_n^{(r)}(\tilde{\mathcal{P}}_n^{(r)}) \leq M^{(r)}(n)$ and $R_n^{(c)}(\tilde{\mathcal{P}}_n^{(c)}) \leq M^{(c)}(n)$, respectively. Then, for any collection of observations $(x^n, y^n, z^n) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n$ satisfying Assumptions 1 and 2, we have:

$$CR(n) \leq M^{(c)}(n) + M^{(r)}(n) + \frac{\|\vec{c}_n\|_2}{\sqrt{2}} \sqrt{M^{(c)}(n)}. \quad (42)$$

where $\vec{c}_n = [c_1, \dots, c_n]$ is a vector with elements:

$$c_i = \sum_{x \in \mathcal{X}} \left| \log \frac{\hat{p}_{X_i}^{(c)}(x)}{\hat{p}_{X_i}^{(r)}(x)} \right| \quad (43)$$

A proof of the theorem may be found in Appendix E-A. We note that because each c_i depends solely on the estimated complete and restricted distributions, a finite sample bound may be computed at each point in time. If we make the additional assumption that the absolute log ratio of our complete and restricted predictors is bounded:

$$\sup_{x \in \mathcal{X}} \left| \log \frac{\hat{p}_{X_i}^{(c)}(x)}{\hat{p}_{X_i}^{(r)}(x)} \right| \leq L \quad i = 1, 2, \dots \quad (44)$$

then we can simplify the bound by observing that:

$$\|\vec{c}_n\|_2 \leq L |\mathcal{X}| \sqrt{n}. \quad (45)$$

When such a scenario holds, we can make use of the following Corollary to Theorem 1 regarding the asymptotic behavior of the causality regret:

Corollary 1. Let the worst case regret for the predictors $\hat{p}_{X_i}^{(r)}$ and $\hat{p}_{X_i}^{(c)}$ be sublinear in n and the absolute log ratio of the complete and restricted sequential predictors be bounded as in (44). Then, under Assumptions 1 and 2, for any collection of observations $(x^n, y^n, z^n) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n$, the causality regret will be sublinear in n :

$$\lim_{n \rightarrow \infty} \frac{1}{n} CR(n) = 0 \quad (46)$$

Lastly, we note that in the special case where the true complete and restricted distributions are in the reference classes (i.e. $p_{X_i}^{(r)} \in \tilde{\mathcal{P}}^{(r)}$ and $p_{X_i}^{(c)} \in \tilde{\mathcal{P}}^{(c)}$), then under an appropriately modified Assumption 2 with $p_{X_i}^{(c)}$ and $p_{X_i}^{(r)}$ substituted for $p_{X_i}^{(c)*}$ and $p_{X_i}^{(r)*}$, we have that:

$$\sum_{i=1}^n \left| \hat{C}_{Y \rightarrow X}(i) - C_{Y \rightarrow X}(i) \right| \leq CR(n). \quad (47)$$

While in practice it is not expected that we would know whether or not the true underlying distribution is in a particular class of reference distributions, this observation will be used in performing simulations in Section V-A.

A. Addressing Infinite Order Restricted Models

It is clear that the proposed causality regret only serves as a meaningful metric of estimation accuracy insofar as the reference class optimal causal measure $C_{Y \rightarrow X}^*$ serves as a useful proxy for the true causal measure $C_{Y \rightarrow X}$. This consideration is not unique to the proposed causal measure. In an extensive analysis of problems encountered when using Granger causality, [23] describes a bias-variance tradeoff that results from the fact that subsets of VAR models will in general be of infinite order even if the complete VAR model is finite order. In the context of our estimation framework, this tradeoff lies in the selection of reference classes $\tilde{\mathcal{P}}_n^{(r)}$ and $\tilde{\mathcal{P}}_n^{(c)}$, which need to be rich enough to yield sufficiently good $C_{Y \rightarrow X}^*$ but not so rich that there do not exist sequential prediction methods for which low cumulative regret may be achieved. This issue appears in numerous forms throughout the DI literature. In particular, in order to reliably estimate the DI, it is typically assumed that in addition to all processes being jointly d -Markov, the subset of processes that does not include the directed information source is itself d -Markov [28], [32], [36]–[38]. In the present context, this equates to assuming that (i) X , Y , and Z are jointly d -Markov and (ii) X is “conditionally d -Markov given Z ,” i.e. $p(X_i | X^{i-1}, Z^{i-1}) = p(X_i | X_{i-d}^{i-1}, Z_{i-d}^{i-1})$. While these assumptions are widely utilized to establish performance guarantees of DI estimators, their implications on the nature of the relationship between X and Y is absent in the literature. Thus, we seek to identify *when* stationary jointly Markov processes X , Y , and Z are such that X is conditionally Markov given Z . It should be noted that, by letting $Z = \emptyset$, the following results can be extended to the standard setting consisting of only X and Y .

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 dependencies in collections of random variables using a directed acyclic graph (DAG) $\mathcal{G} = (V, E)$, where $V = \{V_1, \dots, V_m\}$ is a set of random variables (equivalently nodes or vertices) and $E \subset V \times V$ is a set of directed edges that it do not contain any cycles [39]. 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:

$$p(V) = \prod_{i=1}^m p(V_i | \mathcal{P}_i). \quad (48)$$

If this factorization holds for a given p and \mathcal{G} , we say \mathcal{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 \mathcal{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 \mathcal{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 [39], [40]. When a graph \mathcal{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 \mathcal{G} [39].

Algorithm 1 d-Separation [41]

Input: DAG $\mathcal{G} = (V, E)$ and disjoint sets $A, B, C \subset V$

- 1: Create a subgraph containing only nodes in A, B , or C or with a directed path to A, B , 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
-

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:

$$p(X^n, Y^n, Z^n) = \prod_{i=1}^n p(X_i, Y_i, Z_i | X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}). \quad (49)$$

Assuming that there are no instantaneous influences 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 presence of instantaneous influences, we cannot construct a *unique* Bayesian network representation of Markov processes without making alternative assumptions. This is similar reasoning to that of [38], where the absence of instantaneous influences is used to establish the equivalence between DI graphs and minimal generative model graphs. Under this assumption, we can further simplify (49) as:

$$p(X^n, Y^n, Z^n) = \prod_{i=1}^n \prod_{S \in \{X_i, Y_i, Z_i\}} p(S | X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}). \quad (50)$$

Comparing (48) and (50), 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 , or Z_i) with parents $\mathcal{P}_{X_i}, \mathcal{P}_{Y_i}, \mathcal{P}_{Z_i} \subseteq \{X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}\}$. Given that there may be multiple valid Bayesian networks for a particular distribution, we note that X_i, Y_i , and Z_i may not be conditionally dependent on the entire set $\{X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}\}$. 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 $k = 1, \dots, d$ only if:

$$I(S_{i-k}; S'_i | \{X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}\} \setminus S_{i-k}) > 0. \quad (51)$$

Using the Bayesian network construction given by (51), 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 $(X_{i-l}^{i-1}, Z_{i-l}^{i-1})$:

Theorem 2. *Let (X, Y, Z) be a collection of jointly stationary d-Markov processes such that there are no instantaneous influences. If $I(Y^n \rightarrow X^n || Z^n) = 0$, then X is conditionally d-Markov given Z . If $I(Y^n \rightarrow X^n || Z^n) > 0$, X is conditionally Markov given Z of order $2d$ or less if:*

$$I(Y_j; Y_k | X^i, Z^i) = 0 \quad \forall j \leq k \leq i \quad (52)$$

If $I(Y^n \rightarrow X^n || Z^n) > 0$ but (52) is not satisfied, there will not exist any positive integer l such that $(X_{i-l}^{i-1}, Z_{i-l}^{i-1})$ d-separates X_i from (X^{i-l-1}, Z^{i-l-1}) in the Bayesian network generated according to (51).

A proof of the theorem can be found in Appendix E-B. The implication of this theorem is that the desired d-separation criteria only occurs when no two time points of Y directly influence each other and X_i is only causally influenced by a single Y_j for some $j \leq i$. In particular we note that this excludes jointly stationary d-Markov processes aside from the special case where $p(Y_i | X_{i-d}^{i-1}, Y_{i-d}^{i-1}) = p(Y_i | X_{i-d}^{i-1})$ and $p(X_i | X_{i-d}^{i-1}, Y_{i-d}^{i-1}) = p(X_i | X_{i-d}^{i-1}, Y_{i-\tau})$ for some $0 \leq \tau \leq d$.

Theorem 2 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 (52)), 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 [39] and citations therein), a formal characterization within the present context is desired.

Theorem 3. *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 $(X_{i-l}^{i-1}, Z_{i-l}^{i-1})$ does not d -separate X_i from (X^{i-l-1}, Z^{i-l-1}) in the Bayesian network constructed by (51) has Lebesgue measure zero with respect to \mathbb{R}^N .*

A proof of the theorem is given in Appendix E-C.

This observation is not to detract from the literature on the estimation and application of DI, but rather to extend the concerns expressed in [23] to the information theoretic analogues of Granger causality. The issue arises not from the estimators themselves, but rather from the properties of d -separation in causal models. When X is not marginally Markov, any plug-in estimator that requires an estimate of $p_{X_i}^{(r)}$ will necessitate approximation of an infinite order model and therefore introduce a bias-variance tradeoff.

The most relevant previous work in this area was in [42], where the performance of plug-in estimators of DI were analyzed. It was noted in this work that in scenarios where X is not marginally Markov, the finite order approximation is nevertheless interesting, in particular because it serves as an *upper bound* for the DI. Moreover, it is shown that this approximate measure may be reliably estimated even in scenarios where X is not marginally Markov.

We now propose a second alternative to DI for which reliable estimates may be obtained when X is not marginally Markov. The proposed measure is complementary to that proposed in [42] in that it serves as a *lower bound* for the DI. In order to avoid dependence on an infinite past, we consider conditioning on a “stale” history:

Theorem 4. *Let $(X, Y) \sim p$ be a jointly stationary irreducible aperiodic finite-alphabet d -Markov process. For a fixed k , define $\tilde{X}_i \triangleq (X_i, Y_{i-k+1})$. Then \tilde{X} is a jointly stationary irreducible aperiodic $(d+k)$ -Markov process and the following equality holds:*

$$p(X_i | X^{i-1}, Y^{i-k}) = p(X_i | X_{i-k-d}^{i-1}, Y_{i-k-d}^{i-k}). \quad (53)$$

The above theorem states that so long as the distribution of X_i is conditioned upon its own past and any d consecutive samples of Y , then it is independent of all X and Y that precede those samples of Y . The proof of the theorem may be found in Appendix E.

It follows that by conditioning on some stale past of Y , a proxy for DI may be obtained. Formally, define the partial history with lag (or staleness) k to be:

$$\mathcal{H}_i^{(k)} \triangleq \mathcal{H}_i^{(r)} \cup \{y_1, \dots, y_{i-k}\}. \quad (54)$$

Similarly, define the partial conditional distribution:

$$p_{X_i}^{(k)}(x_i) \triangleq p(x_i | \mathcal{H}_i^{(k)}). \quad (55)$$

We note that the partial conditional distribution is a generalization of the complete and restricted distributions in that $p_{X_i}^{(1)} = p_{X_i}^{(c)}$ and $p_{X_i}^{(i)} = p_{X_i}^{(r)}$. Finally, we define the partial causal measure with lag k to be:

$$C_{Y \rightarrow X}^{(k)}(\mathcal{H}_i^{(c)}) \triangleq D(p_{X_i}^{(c)} || p_{X_i}^{(k)}) \quad (56)$$

This sample path dependent measure may be related to the model dependent measures by defining a partial DI (PDI) and PDI rate:

$$I(Y_{n-k}^{n-1} \rightarrow X^n) \triangleq \sum_{i=1}^n E \left[C_{Y \rightarrow X}^{(k)}(\mathcal{H}_i^{(c)}) \right] \quad (57)$$

$$\bar{I}_P^{(k)}(Y \rightarrow X) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n I(Y_{n-k}^{n-1} \rightarrow X^n) \quad (58)$$

Much like the DI (rate), the PDI (rate) can be represented as a difference of entropies (rates):

$$I(Y_{n-k}^{n-1} \rightarrow X^n) = H(X^n || Y^{n-k-1}) - H(X^n || Y^{n-1}) \quad (59)$$

$$\bar{I}_P^{(k)}(Y \rightarrow X) = \bar{H}^{(k-1)}(X || Y) - \bar{H}^{(1)}(X || Y) \quad (60)$$

where we have replaced the first entropy terms on the right hand side of (11) and (13) with a lagged causally conditioned entropy.

The partial causal measures of (56), (57), and (58) have straightforward interpretations in relationship to their complete counterparts. In particular, we note that these measures can be viewed as measuring the causal effect of the *recent* past of Y on X . Additionally, the PDI serves as a lower bound for DI. Therefore, effectively estimating the PDI for scenarios in which we do not have a universal estimator of the DI may be of great interest. The estimators of [28] can be extended to be universal estimators of the PDI rate:

Theorem 5. *Let $(X, Y) \sim p$ be a jointly stationary irreducible aperiodic finite-alphabet Markov process of order d or less. Let $\hat{p}_{X_i}^{(c)}$ be a depth- d CTW estimate of $p_{X_i}^{(c)}$ with access to $\mathcal{H}_i^{(c)}$ and $\hat{p}_{X_i}^{(k)}$ be a depth- $(d+k)$ CTW estimate of $p_{X_i}^{(k)}$ with access to $\mathcal{H}_i^{(k)}$. Then:*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n D(\hat{p}_{X_i}^{(c)} || \hat{p}_{X_i}^{(k)}) = \bar{I}_P^{(k)}(Y \rightarrow X) \quad p_{X^n, Y^n} - a.s. \quad (61)$$

The proof of the theorem may be found in Appendix E. We note that this theorem is analogous with Theorem 3 from [28], where here we have removed any assumptions about how X behaves marginally. It should also be noted that here we are only considering one of four estimators proposed by [28]. Presumably, similar results could be obtained for the other estimators, though this is not the primary concern of this work.

As a final point, we note that by combining the PDI and the work of [42], we can reliably estimate upper and lower bounds of DI for jointly Markov processes regardless of whether or not the sub-processes are themselves marginally Markov:

Proposition 2. Let (X, Y) be a jointly stationary Markov process of order d . Define the truncated DI (TDI) rate of order k as:

$$\bar{I}_T^{(k)}(Y \rightarrow X) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n I(X_i; Y_{i-k}^{i-1} | X_{i-k}^{i-1}). \quad (62)$$

Then for $k_1 \geq 1$ and $k_2 \geq d$, we have:

$$\bar{I}_P^{(k_1)}(Y \rightarrow X) \leq \bar{I}(Y \rightarrow X) \leq \bar{I}_T^{(k_2)}(Y \rightarrow X) \quad (63)$$

with both bounds becoming equalities as k_1 and k_2 approach infinity.

The proof of the proposition is given in Appendix D-B. It should be noted that, in a scenario where both joint and marginal Markovicity hold, the TDI and DI rates are equivalent. We have here simply renamed the TDI to emphasize that, in many cases, they are in fact *not* equivalent. The implication of this proposition is that, for the large class of jointly Markov processes for which the sub-processes are not marginally Markov (as characterized by Theorems 2 and 3), we can avoid the bias-variance tradeoff resulting from using truncated approximations of infinite-order models by reliably estimating *bounds* on DI rather than the DI itself. In particular, the quantity that is estimated under the assumption that X is marginally Markov (given by (62)) provides a lower bound, and the PDI resulting from conditioning on a stale history provides an upper bound. Furthermore, both of these quantities can be reliably estimated, as demonstrated by [42] and Theorem 5.

V. RESULTS

A. Stationary Markov Processes

We begin by demonstrating estimation of the causal measure for simulated stationary first order Markov processes using a context tree weighting (CTW) sequential prediction algorithm [43]. For the purpose of estimating either the complete conditional distribution $p_{X_i}^{(c)}$, or partial conditional distribution $p_{X_i}^{(k)}$, we utilize a CTW with side information as in [44]. In order to evaluate the causality regret of a given estimator of the causal measure, we need worst case regret bounds on the sequential predictors utilized in the estimator.

Lemma 1 ([28], [45]). Let \hat{p} be a depth- d CTW probability assignment of a stationary finite-alphabet Markov process $X \sim p$ of order less than or equal to d . Then the worst case regret is bounded as:

$$\sup_{x^n} \log \frac{p(x^n)}{\hat{p}(x^n)} \leq \frac{(|\mathcal{X}| - 1)L}{2} \log \frac{n}{L} + L \left(\frac{|\mathcal{X}|}{|\mathcal{X}| - 1} + \log |\mathcal{X}| \right) - \frac{1}{|\mathcal{X}| - 1} \quad (64)$$

where L is the number of leaves in the CTW (equivalently, the number of states of X).

Next we note that this bound may be extended to the case where X is given access to causal side information Y :

Proposition 3. Let \hat{p} be a depth- d CTW probability assignment of X causally conditioned on Y , where $(X, Y) \sim p$ is a pair

of jointly stationary finite-alphabet process of order less than or equal to d . Then the worst case regret is bounded as:

$$\sup_{x^n, y^n} \log \frac{p(x^n || y^n)}{\hat{p}(x^n || y^n)} \leq \frac{(|\mathcal{X}| - 1)L}{2} \log \frac{n}{L} + L(|\mathcal{X}| - 1) + S \quad (65)$$

where L is the number of leaves in the CTW, S is the total number of nodes in the CTW.

A proof of the proposition is provided in Appendix D-C. Using the above lemma and proposition we can compute the values of the causality regret bounds by using (64) and (65) for $M^{(r)}(n)$ and $M^{(c)}(n)$, respectively, in (42). In the following sections we compare the causal regret bound with the true estimation accuracy for three scenarios.

1) *Independent Processes:* Let X and Y be independent ternary processes (i.e. $X_i, Y_i \in \{0, 1, 2\}$ for all i), with each process being stationary first order Markov. As such, the processes are completely characterized by the probabilities $p(x_i | x_{i-1})$ and $p(y_i | y_{i-1})$ for $x_i, x_{i-1}, y_i, y_{i-1} \in \{0, 1, 2\}$. Given the independence of X and Y , we have that for all $i = 1, 2, \dots$, $C_{Y \rightarrow X}(i) = C_{X \rightarrow Y}(i) = 0$.

Figure 2 shows the estimate of the causal measure over time for $n = 10000$ samples. We can see that the estimated causal measure in both directions quickly becomes very small at all times. The true causal measure is not shown because it is always zero. In the bottom panel of the figure we see the normalized causal regret with respect to the true causal measure as in (47), which in this case is given by the running average of the estimated causal measure. Additionally, we show the causal regret bounds, which are computed using $|\mathcal{X}| = 3$, $L = 3$ in (64), and $L = 9$ and $S = 10$ in (65).

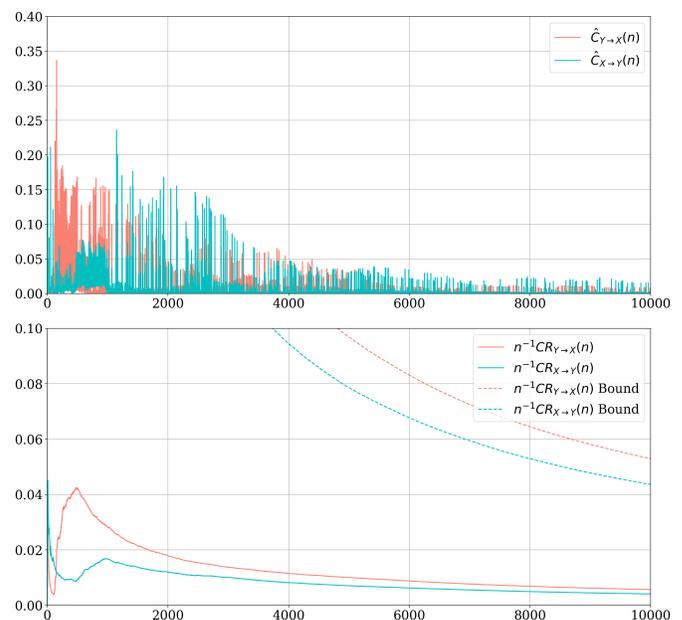


Fig. 2. Top - Estimates of causal measure in each direction for independent processes. Bottom - Normalized cumulative absolute error of estimates (solid) and normalized causality regret bounds (dashed).

2) *Unidirectional IID Influence*: For the second scenario we consider a pair of processes wherein each Y_i is independent and identically distributed and X_i is dependent only upon X_{i-1} and Y_{i-1} . As such, it is clear that, in addition to X and Y being jointly first-order Markov, X is marginally first-order Markov. While the marginal Markovicity is immediately clear in this case, we point out that these processes do indeed satisfy the condition of Theorem 2 for any parameterization of the probabilities in question.

Figure 3 shows the true causal measure $C_{Y \rightarrow X}$ alongside the estimates $\hat{C}_{Y \rightarrow X}$ and $\hat{C}_{X \rightarrow Y}$. For clarity, only the last 100 time points are shown. We can see that in this time window the estimated $\hat{C}_{Y \rightarrow X}$ tracks the true causal measure $C_{Y \rightarrow X}$ very well, and the estimated $\hat{C}_{X \rightarrow Y}$ has converged to 0 as desired. In the bottom panel we see that, because the causal measure $\hat{C}_{X \rightarrow Y}(i) < \hat{C}_{Y \rightarrow X}(i)$, c_i in equation (43) is much smaller in the $X \rightarrow Y$ direction and thus the causal regret bound is considerably tighter. This is consistent with the result obtained in [42] that plug-in estimators of the DI rate will converge at a faster rate if the DI rate is zero.

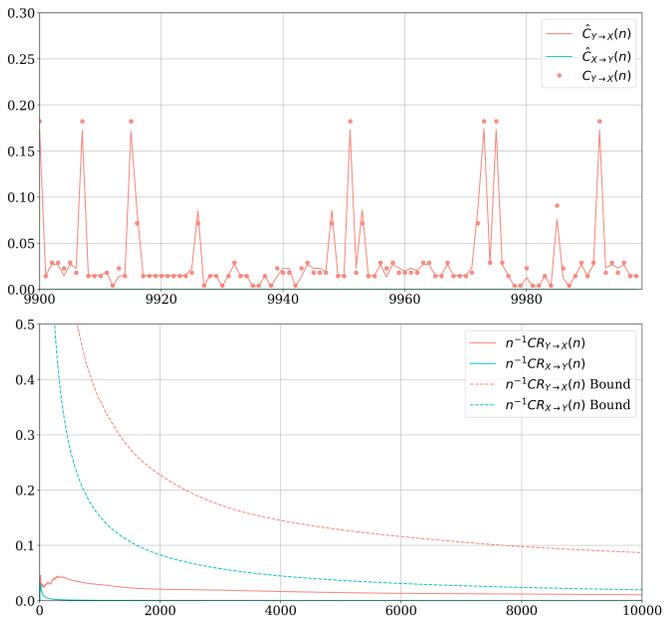


Fig. 3. Top - Estimates of causal measure in each direction for unidirectional influences. Bottom - Normalized cumulative absolute error of estimate (solid) and normalized causality regret bounds (dashed).

3) *Bidirectional Influences*: Lastly, we consider the scenario where X and Y mutually influence each other. Specifically, let each X_i and Y_i be independently influenced by X_{i-1} and Y_{i-1} such that the processes are fully characterized by the probabilities $p(x_i | x_{i-1}, y_{i-1})$ and $p(y_i | x_{i-1}, y_{i-1})$.

Figure 4 shows the true and estimated causal measures in both directions. The bottom panel shows the cumulative absolute error alongside the causal regret bounds. We note that here we have extended the time horizon to $n = 50000$ to illustrate that the estimators exhibit bias resulting from the fact that X is not marginally Markov. As a result, it is important to note that the true restricted distribution $p_{X_i}^{(r)}$ will not be in the reference class of restricted distributions $\tilde{\mathcal{P}}^{(r)}$

and we can expect the causality regret bound to be lower than the cumulative absolute error as $n \rightarrow \infty$. Due to the non-Markovicity of X , computing the true restricted distribution at each time becomes increasingly challenging. To address this, we derive a recursive updating algorithm for efficiently computing the true causal measure $C_{Y \rightarrow X}(i)$ in such settings. Details can be found in Appendix B.

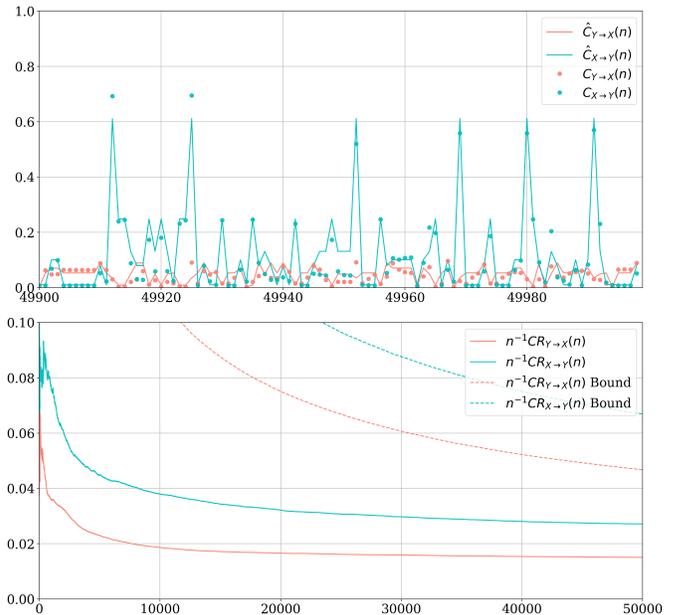


Fig. 4. Top - Estimates of causal measure in each direction for bidirectional influences. Bottom - Normalized cumulative absolute error of estimate (solid) and normalized causality regret bounds (dashed).

To address the estimation bias seen in Figure 4, we consider the partial causal measure $C_{Y \rightarrow X}^{(k)}(i)$ defined by (56). Figure 5 shows an estimate of the partial causal measure on the same sequence considered in Figure 4 with a staleness of $k = 1$. The bottom panel of Figure 5 depicts the cumulative absolute error and the causal regret bounds. While the worst case regret for the complete predictor $M^{(c)}(n)$ remains the same as in the previous examples, the regret of the partial predictor is computed using equation (65) with $L = 27$ (3 values for x_{i-1} times 9 possible values for (x_{i-2}, y_{i-2})) and $N = 31$ (27 leaf nodes, 3 depth-1 nodes, and 1 root node).

We can see in Figure 5 that due to the increased number of nodes in the CTW estimate of the partial distribution, the normalized absolute error decreases more slowly at the beginning. Regardless, the estimate of the partial causal measure does not exhibit the same behavior of converging on a biased estimate. We see the error continues to decrease throughout the entire sequence. Moreover, a visual comparison of the true and estimated measures makes clear that the estimate is unbiased.

Given that the same sequences were used in generating Figures 4 and 5, we can compare the values of the complete causal measure with the partial causal measure. It is clear that while there is considerable agreement on the positions of the spikes in causal influence, the strengths vary. While it is true that the partial causal measure will be smaller than the complete causal measure in expectation (i.e. partial DI is

less than DI), there are times where the stale history y^{i-k} is misleading about the recent history y_{i-k+1}^{i-1} , and thus we sometimes see that the partial causal measure is larger than the complete causal measure on a given sample path. Lastly, we note that because the true partial distribution is in the class of reference partial distributions, the causality regret bounds will bound the cumulative absolute error.

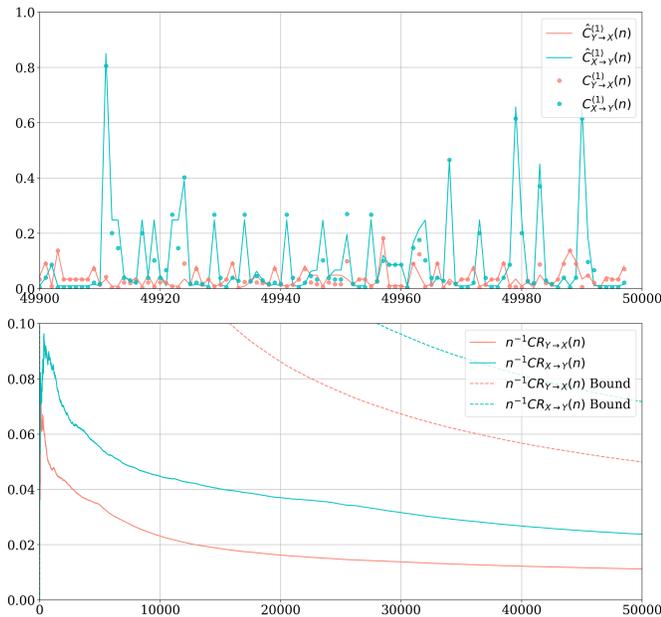


Fig. 5. Top - Estimates of the partial causal measure with $k = 1$ in each direction for bidirectional influences. Bottom - Normalized cumulative absolute error of estimates (solid) and normalized causality regret bounds (dashed).

B. Stock Market Indices

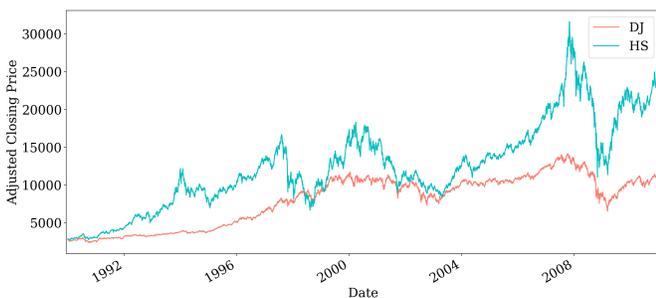


Fig. 6. The Dow Jones (DJ) Industrial Average and Hang Seng (HS) indices.

We now demonstrate the use of the sample path causal measure on historical stock market data from the Dow Jones (DJ) Industrial Average index on the New York Stock Exchange (NYSE) and the Hang Seng (HS) index on the Shanghai Stock Exchange (SSE), as in [28]. In [28] it was shown that the DJ index had a greater influence on the HS index than vice versa by measuring the DI between the sequences of daily changes in adjusted closing price. Here we consider the same dataset, shown in Figure 6.

The data was downloaded from Yahoo Finance. Given that the NYSE and SSE are closed on different holidays, missing values were interpolated on days where one was open and the other was not (weekends and shared holidays were not interpolated). We next consider the inter-day percentage change in adjusted closing price and quantize it to a ternary sequence with a value of 0 indicating a drop by more than 0.8%, 2 indicating a rise by more than 0.8%, and 1 representing no significant change. In computing the influence of HS on DJ, the HS data is shifted forward by one day. This is due to the fact that on each day, the NYSE closes before the SSE opens, as noted in [28]. Thus, the HS is affected by the same day DJ, while the DJ is affected by the previous day HS.

Figure 7 shows the estimate of the causal measure for different previous day states in each direction using depth-1 CTW predictors in addition to the estimated DI (dashed line). Values from the final 3 years of the data are shown in the box plot, with the fairly tight error bars showing that the CTW estimators have mostly converged (we would not expect complete convergence due to the non-stationarity of stock market data).

There are numerous noteworthy points in Figure 7. First we note that, as a result of averaging, the DI is never *equal* to the causal measure in the DJ to HS direction. In particular we note that when DJ does not change, it has virtually no effect on the distribution of HS. Furthermore, most of the time DJ does not change. On the other hand, on the rare occasion that DJ went down and HS went up on the previous day (5.1% of days), the causal measure is almost 4X the DI. Similarly, on the 4.2% of days where DJ goes up and HS goes down, the causal measure from HS to DJ is roughly 10X the DI. When considering this type of data, the added value of Q3 over Q1 becomes very clear. If one has access to what has already happened (i.e. the previous day state), then why take an expectation over the past?

Remark 2. *It is crucially important to make clear the notion of causality that is considered in this context. There is no doubt that there are confounding factors (i.e. factors that affect both DJ and HS) that would decrease the measured influence if included in the model. That having been said, it is clear that there is information contained in the DJ that provides us with an improved ability to predict the next day's HS, a finding which may certainly be of use for applications outside of classical "causal inference". In order to make claims of causal influence, careful attention needs to be paid to potential causes that are not considered in the model, thus requiring domain expertise. As such, in scenarios where one cannot confidently rule out the potential presence of confounding factors, the proposed measure may be more accurately viewed as a measure of increased predictability (as in [46]).*

VI. DISCUSSION

The concepts presented in this paper can be distilled to three primary contributions. First, we have introduced a need for measuring causal influences between random processes that depend on the sample paths of those processes. We have shown that in both simple thought experiments and real stock market

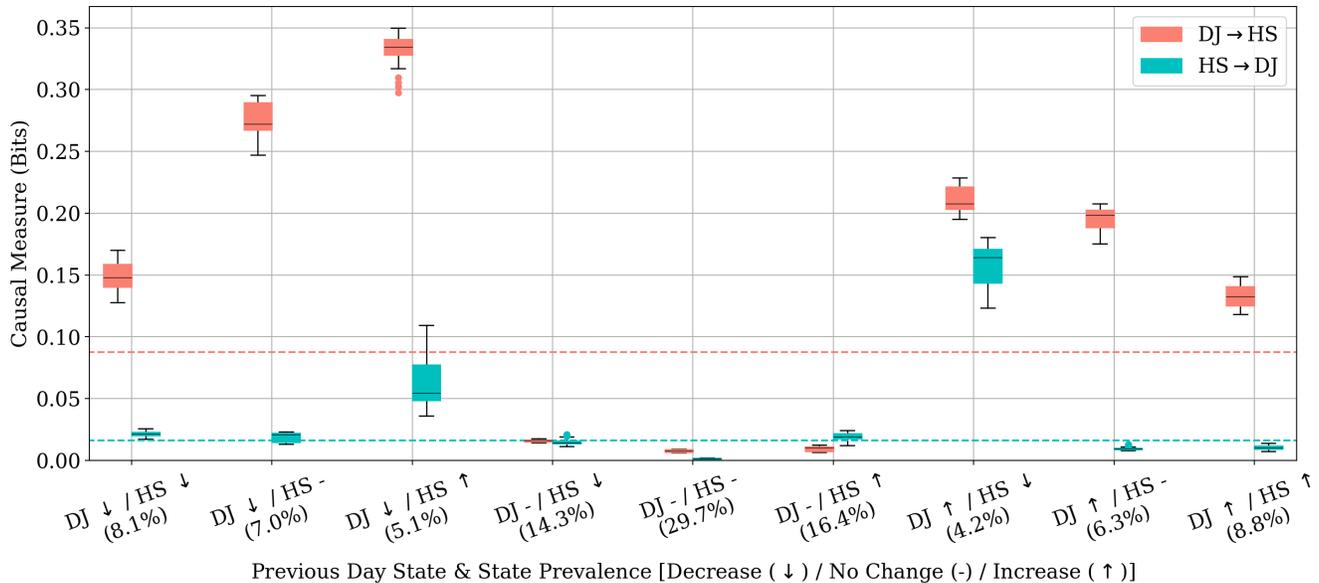


Fig. 7. Causal measure between stock indices for different previous day states from 2008 to 2011. Below each of the 9 possible previous day states we include the percentage of days in which that state occurs. The dashed lines represent the DI estimate.

data, there exist sample path dependent causal influences that may occur infrequently, and are thus not captured by average measures such as GC and DI. Second, we have proposed a measure for identifying these influences. We have shown that this measure gives results consistent with intuitions in a number of examples. Furthermore, we proved finite sample bounds on the performance of an estimator of our proposed measure. Third, we have presented a characterization of when it is possible to reliably estimate both the proposed measure and the DI. This characterization utilizes the connections between DI, GC, and causal graphical models to demonstrate that in many common cases, we should not expect to be able to get unbiased estimates of DI, which to our knowledge has been repeatedly overlooked throughout the information theory literature. To address this issue we introduce a notion of partial DI and a corresponding sample path dependent partial causal measure.

There are numerous directions for continued research in this area. Further leveraging the tools from causal graphical models can enable a better understanding of the circumstances in which we can estimate measures of causal influence reliably. Furthermore, the tools from this field are necessary to distinguish between true causal influences and measures of improved predictability. Additionally, extending the three questions proposed in the introduction to the case of general causal graphs is of great interest. In particular, how can we measure how different realizations of groups of random variables affect another group of random variables in a given directed graph? We believe the philosophy presented in this paper may be used to address this question.

It is important to note the present work is built upon the restrictive assumption that all processes are observed. There has been considerable recent interest in estimating causal in-

fluences and graph structures when only a subset of processes may be observed [47]–[49]. As such, there is an opportunity to study how these results may be applied to inferring dynamic causal influences that are dependent upon *realizations* of a subset of processes.

Another line of future work is further investigation of the significance of partial directed information developed in Section IV-A and its application in quantifying information leakage for coupled systems with delayed information [50], providing fundamental performance limitations of closed-loop systems [51] subject to delay constraints, or in characterizing rate-performance tradeoffs [52] for network control problems with non-classical information structures [53], [54] pertaining to information and delay constraints.

A final area for future work is the demonstration of how the causal measure can provide added value in decision making. A promising avenue lies in the use of the causal measure for aiding in *time-varying model selection*. Take, for example, the stock market example in Section V-B. It is shown in [55] that using DI for model selection can yield improvements in the systemic risk. A natural extension of this would be to use the sample path causal measure to create a collection of models that are dependent upon the current “state” of the stock market. This would enable minimizing the number of estimated parameters while ensuring that opportunities for leveraging directed influences are not overlooked.

APPENDIX A EQUIVALENCE OF INTERVENTIONAL AND NON-INTERVENTIONAL CAUSAL MEASURE IN SECTION III-A

First, we introduce the *causal model* defined by Pearl [7, Definition 2.2.2], which consists of a causal structure (i.e. a

DAG) and a set of functions defining a probability distribution over each node in the DAG. For the three examples in section III-A, we have that the causal structure is given by Figure 1. For the first two examples, the functions are given by equations (27) and (28), respectively. For the third example of horse betting, we assume that for each i , $X_i = f_i(X^{i-1}, Y_{i-1}, U_i)$ and $Y_i = V_i$, where f_i is some collection of functions, U is a collection of iid random variables (independent of X and Y), and V is a collection of iid random variables (independent of X , Y , and U). The key element of this assumption is that the winner of the i^{th} race, X_i , is functionally dependent on the side information Y_{i-1} , meaning that changing the side information could change the winner. Without this technicality, the example would not constitute a causal model in the sense of [7, Definition 2.2.2]. Once these causal models are established, showing the equivalence between the interventional and non-interventional measures discussed in Remark 1 can easily be shown using the second rule of the so-called *do*-calculus [56, Theorem 3]. Specifically, showing that $p(x_i | x^{i-1}, do(y_{i-1})) = p(x_i | x^{i-1}, y_{i-1})$ amounts to showing that X_i and Y_{i-1} are d -separated by X^{i-1} in an augmented DAG where the outgoing arrows from Y_{i-1} have been removed. This holds trivially in all three DAGs in Figure 1 because removing the outgoing arrows from Y_{i-1} results in there being *no* path connecting Y_{i-1} to X_i in the augmented DAG.

APPENDIX B

COMPUTING TRUE CAUSAL MEASURE WITH HIDDEN MARKOV MODELS

In order to compute the true causal measure, it is necessary to compute the true restricted distribution. As discussed in Section IV-A, the restricted distribution is, in general, non-Markov. As such, it is desirable to have an efficient method for computing the true restricted distribution $p(y_i | y^{i-1})$. Here we derive update equations for recursively computing $p(y_i | y^{i-1})$. The proposed updating scheme is a generalization of the well known recursive method for evaluating the likelihood of a process under a standard hidden Markov model where the likelihood is given by $p(y_i | x_i)$ and the one-step prediction distribution is given by $p(x_i | x_{i-1})$ [57, Ch. 9].

First, assume X and Y are jointly first order Markov as in Section V-A and decompose the restricted distribution as the product of “likelihood” and “prior” terms:

$$\begin{aligned} p(y_{i+1} | y^i) &= \sum_{x_i} p(y_{i+1}, x_i | y^i) \\ &= \sum_{x_i} \underbrace{p(y_{i+1} | x_i, y_i)}_{\text{Likelihood}} \underbrace{p(x_i | y^i)}_{\text{Prior}} \end{aligned}$$

where we note that only the prior term has a long-term dependence on the past. The prior may be further decomposed into the sum of products of “one-step prediction” and “posterior” terms:

$$\begin{aligned} p(x_i | y^i) &= \sum_{x_{i-1}} p(x_i, x_{i-1} | y^i) \\ &= \sum_{x_{i-1}} \underbrace{p(x_i | x_{i-1}, y_{i-1})}_{\text{One-Step Prediction}} \underbrace{p(x_{i-1} | y^i)}_{\text{Posterior}} \end{aligned}$$

where now only the posterior has a long-term dependence on the past. Lastly, we can use Bayes’ Rule to show that the posterior depends only on the previous likelihood evaluated at the newly observed y_i and the previous prior:

$$\begin{aligned} p(x_{i-1} | y^i) &= \frac{p(y_i | x_{i-1}, y^{i-1})p(x_{i-1} | y^{i-1})}{\sum_{\tilde{x}_{i-1}} (y_i | \tilde{x}_{i-1}, y^{i-1})p(\tilde{x}_{i-1} | y^{i-1})} \\ &= \frac{p(y_i | x_{i-1}, y_{i-1})p(x_{i-1} | y^{i-1})}{\sum_{\tilde{x}_{i-1}} (y_i | \tilde{x}_{i-1}, y_{i-1})p(\tilde{x}_{i-1} | y^{i-1})} \end{aligned}$$

Thus, the restricted distribution can be computed in a recursive manner. To initialize the algorithm, define $y_0 = x_0 = \emptyset$, $p(\emptyset | \cdot) = 1$, and starting distributions $p(\cdot | \emptyset) = p(\cdot)$.

APPENDIX C USEFUL LEMMAS

We first show that the cumulative KL divergence from the best reference distribution to the predicted distribution is less than the predictor’s worst-case regret.

Lemma 2. *For a sequential predictor \hat{p}_i with worst case regret $M(n)$, a collection observations (x^n, y^n, z^n) , and any distribution from the reference class $p \in \tilde{\mathcal{P}}_n$:*

$$\sum_{i=1}^n D(p_i || \hat{p}_i) \leq M(n) \quad (66)$$

Proof.

$$\begin{aligned} \sum_{i=1}^n D(p_i || \hat{p}_i) &= \sum_{i=1}^n \sum_{x \in \mathcal{X}} p_i(x) \log \frac{p_i(x)}{\hat{p}_i(x)} \\ &\leq \sum_{i=1}^n \left[\sup_{x \in \mathcal{X}} \log \frac{p_i(x)}{\hat{p}_i(x)} \right] \sum_{x \in \mathcal{X}} p_i(x) \\ &= \sum_{i=1}^n \sup_{x \in \mathcal{X}} r(\hat{p}_i, p_i, x) \\ &\leq \sup_{x^n \in \mathcal{X}^n} \sum_{i=1}^n r(\hat{p}_i, p_i, x_i) \\ &\leq \sup_{x^n \in \mathcal{X}^n} \sup_{p \in \tilde{\mathcal{P}}_n} \sum_{i=1}^n r(\hat{p}_i, p_i, x_i) \\ &\leq M(n) \end{aligned}$$

□

Next, we bound the cumulative difference in expectation of a bounded function between the best reference distribution and sequential predictor.

Lemma 3. *For a sequential predictor \hat{p}_i with worst case regret $M(n) \geq 1$, a collection observations (x^n, y^n, z^n) , cumulative loss minimizing distribution p_i^* , and a collection of functions $g_i : \mathcal{X} \rightarrow \mathbb{R}$ for $i = 1, \dots, n$:*

$$\sum_{i=1}^n |E_{p_i^*}[g_i(X)] - E_{\hat{p}_i}[g_i(X)]| \leq \frac{\|\tilde{c}_n\|_2}{\sqrt{2}} \sqrt{M(n)} \quad (67)$$

where $\vec{c}_n = [c_1, \dots, c_n]$ is a vector with elements:

$$c_i = \sum_{x \in \mathcal{X}} |g_i(x)| \quad (68)$$

Proof.

$$\begin{aligned} & \sum_{i=1}^n |E_{p_i^*}[g_i(X)] - E_{\hat{p}_i}[g_i(X)]| \\ &= \sum_{i=1}^n \left| \sum_{x \in \mathcal{X}} [p_i^*(x) - \hat{p}_i(x)] g_i(x) \right| \\ &\leq \sum_{i=1}^n \sum_{x \in \mathcal{X}} |p_i^*(x) - \hat{p}_i(x)| |g_i(x)| \end{aligned} \quad (69)$$

$$\leq \sum_{i=1}^n \left[\sum_{x \in \mathcal{X}} |p_i^*(x) - \hat{p}_i(x)| \right] \left[\sum_{x \in \mathcal{X}} |g_i(x)| \right] \quad (70)$$

$$\begin{aligned} &\leq \sum_{i=1}^n \sqrt{\frac{1}{2} D(p_i^* \parallel \hat{p}_i)} \sum_{x \in \mathcal{X}} |g_i(x)| \\ &= \frac{1}{\sqrt{2}} \sum_{i=1}^n c_i \sqrt{D(p_i^* \parallel \hat{p}_i)} \end{aligned} \quad (71)$$

where (69) uses the triangle inequality, (70) follows from both terms of the sum being positive, and (71) uses Pinsker's inequality. Focusing on the sum, we define a vector $\vec{v}_n = [v_1, \dots, v_n]$ such that $v_i = \sqrt{D(p_i^* \parallel \hat{p}_i)}$ for $i = 1, \dots, n$:

$$\sum_{i=1}^n c_i \sqrt{D(p_i^* \parallel \hat{p}_i)} = |\vec{c}_n \cdot \vec{v}_n| \quad (72)$$

$$\leq \|\vec{c}\|_2 \|\vec{v}\|_2 \quad (73)$$

$$\begin{aligned} &= \|\vec{c}\|_2 \left(\sum_{i=1}^n D(p_i^* \parallel \hat{p}_i) \right)^{\frac{1}{2}} \\ &\leq \|\vec{c}\|_2 \sqrt{M(n)} \end{aligned} \quad (74)$$

where (72) follows from the fact that $c_i \geq 0$ and $v_i \geq 0$ for all i , (73) uses the Cauchy-Schwarz inequality and (74) uses Lemma 2 and the assumption that $M(n) \geq 1$. \square

APPENDIX D PROOF OF PROPOSITIONS

A. Proof of Proposition 1

Using the definition of the causal measure, we get that the left hand side of (24) is:

$$\begin{aligned} & \sum_{\mathcal{H}_i^{(c)}} p(\mathcal{H}_i^{(c)}) D(p_{X_i}^{(c)} \parallel p_{X_i}^{(r)}) \\ &= \sum_{\mathcal{H}_i^{(c)}} p(\mathcal{H}_i^{(c)}) \sum_{x_i} p_{X_i}^{(c)}(x_i) \log \frac{p_{X_i}^{(c)}(x_i)}{p_{X_i}^{(r)}(x_i)} \\ &= \sum_{\mathcal{H}_i^{(c)}} p(\mathcal{H}_i^{(c)}) \sum_{x_i} p(x_i \mid \mathcal{H}_i^{(c)}) \log \frac{p(x_i \mid \mathcal{H}_i^{(c)})}{p(x_i \mid \mathcal{H}_i^{(r)})} \\ &= \sum_{\mathcal{H}_i^{(c)}, x_i} p(\mathcal{H}_i^{(c)}, x_i) \log \frac{p(x_i \mid \mathcal{H}_i^{(c)})}{p(x_i \mid \mathcal{H}_i^{(r)})} \end{aligned}$$

$$\begin{aligned} &= E_{X_i, Y^{i-1}, Z^{i-1}} \left[\log \frac{p(X_i \mid X^{i-1}, Y^{i-1}, Z^{i-1})}{p(X_i \mid X^{i-1}, Z^{i-1})} \right] \\ &= I(Y^{n-1} \rightarrow X^n \parallel Z^{n-1}) \end{aligned}$$

\square

B. Proof of Proposition 2

First we note that the lower bound holds simply by applying conditioning reduces entropy to the first term of the PDI rate as defined in (60). Formally, we have:

$$\begin{aligned} & \bar{I}_P^{(k)}(Y \rightarrow X) \\ &= \bar{H}^{(k-1)}(X \parallel Y) - \bar{H}^{(1)}(X \parallel Y) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} [H(X^n \parallel Y^{n-k-1}) - H(X^n \parallel Y^{n-1})] \\ &\leq \lim_{n \rightarrow \infty} \frac{1}{n} [H(X^n) - H(X^n \parallel Y^{n-1})] \\ &= \bar{H}(X) - \bar{H}^{(1)}(X \parallel Y) = \bar{I}(Y \rightarrow X) \end{aligned}$$

Next, for the TDI upper bound, we note that for $k_2 = d$, the bound was proven in [42]. For completeness we extend the proof for $k_2 \geq d$:

$$\begin{aligned} & \bar{I}_T^{(k)}(Y \rightarrow X) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n I(X_i; Y_{i-k_2}^{i-1} \mid X_{i-k_2}^{i-1}) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(X_i \mid X_{i-k_2}^{i-1}) - H(X_i \mid X_{i-k_2}^{i-1}, Y_{i-k_2}^{i-1}) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(X_i \mid X_{i-k_2}^{i-1}) - H(X_i \mid X_{i-d}^{i-1}, Y_{i-d}^{i-1}) \\ &\geq \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(X_i \mid X^{i-1}) - H(X_i \mid X_{i-d}^{i-1}, Y_{i-d}^{i-1}) \\ &= \bar{H}(X) - \bar{H}^{(1)}(X \parallel Y) = \bar{I}(Y \rightarrow X) \end{aligned}$$

as was to be shown. \square

C. Proof of Proposition 3

As in the statement of the proposition, let L be the number of leaves in the CTW and N be the total number of nodes in the tree. Define $p_e(x^n)$ to be the Dirichlet estimator introduced in [58], otherwise known as the KT-estimator. Then it is known that the worst case regret is given by [45]:

$$\sup_{x_n} \log \frac{p(x^n)}{p_e(x^n)} \leq \frac{|\mathcal{X}| - 1}{2} \log n + |\mathcal{X}| - 1 \quad (75)$$

We next define the KT-tree estimator with side information $p_c(x^n \parallel y^n)$ as the estimator where, for each possible "context" $(x_{i-d}^{i-1}, y_{i-d}^{i-1})$, a separate instance of a KT-estimator is maintained. Letting $\mathcal{L} \triangleq \{(x_{i-d}^{i-1}, y_{i-d}^{i-1}) : (x_{i-d}^{i-1}, y_{i-d}^{i-1}) \in \mathcal{X}^d \times \mathcal{Y}^d\}$ be the set of contexts (i.e. leaf nodes), we have that $|\mathcal{L}| = L$. Defining $p_e^{(l)}(x^n \parallel y^n) \triangleq p_e^{(l)}(x^{(l)})$ to be the KT-estimator that assigns probabilities to $x^{(l)} \triangleq \{x_i : (x_{i-d}^{i-1}, y_{i-d}^{i-1}) = l\}$ for $l \in \mathcal{L}$ with $|x_i^{(l)}| \triangleq n_l$, we can

derive the worst case regret of the KT-tree estimator with side information as follows:

$$\begin{aligned} \sup_{x^n, y^n} \log \frac{p(x^n)}{p_c(x^n)} &= \sup_{x^n, y^n} \log \frac{p(x^n \parallel y^n)}{\prod_{l \in \mathcal{L}} p_e^{(l)}(x^n \parallel y^n)} \\ &= \sup_{x^n, y^n} \log \prod_{l \in \mathcal{L}} \frac{p(x^{(l)})}{p_e^{(l)}(x^{(l)})} \\ &= \sup_{x^n, y^n} \sum_{l \in \mathcal{L}} \log \frac{p(x^{(l)})}{p_e^{(l)}(x^{(l)})} \\ &\leq \sum_{l \in \mathcal{L}} \left(\frac{|\mathcal{X}| - 1}{2} \log n_l + |\mathcal{X}| - 1 \right) \quad (76) \end{aligned}$$

$$\begin{aligned} &= \frac{L(|\mathcal{X}| - 1)}{2} \sum_{l \in \mathcal{L}} \frac{1}{L} \log n_l + L(|\mathcal{X}| - 1) \\ &\leq \frac{L(|\mathcal{X}| - 1)}{2} \log \sum_{l \in \mathcal{L}} \frac{n_l}{L} + L(|\mathcal{X}| - 1) \quad (77) \end{aligned}$$

$$= \frac{L(|\mathcal{X}| - 1)}{2} \log \frac{n}{L} + L(|\mathcal{X}| - 1) \quad (78)$$

where (76) follows from the bound in (75), (77) follows from Jensen's inequality, and (78) follows from the fact that $\sum_l n_l = n$. We now define the set of all nodes to be $\mathcal{S} \triangleq \{(x_{i-k}^{i-1}, y_{i-k}^{i-1}) : (x_{i-k}^{i-1}, y_{i-k}^{i-1}) \in \mathcal{X}^k \times \mathcal{Y}^k, k = 1, \dots, d\}$, with $|\mathcal{S}| = S$. Then, we can define a context tree by letting defining a probability $p_w^{(s)}(x^n \parallel y^n)$ for each node $s \in \mathcal{S}$ as follows:

$$p_w^{(s)}(x^n \parallel y^n) = \begin{cases} \frac{1}{2} p_e^{(s)}(x^n \parallel y^n) + \frac{1}{2} \prod_{s' \in \mathcal{X} \times \mathcal{Y}} p_w^{(s's)}(x^n \parallel y^n) & s \notin \mathcal{L} \\ p_e^{(s)}(x^n \parallel y^n) & s \in \mathcal{L} \end{cases} \quad (79)$$

where $s's = (x_{i-k-1}^{i-1}, y_{i-k-1}^{i-1}) \in \mathcal{X}^{k+1} \times \mathcal{Y}^{k+1}$ represents a child node of $s = (x_{i-k}^{i-1}, y_{i-k}^{i-1})$ with $s' = (x_{i-k-1}^{i-1}, y_{i-k-1}^{i-1})$. Letting λ be the root node of the tree (i.e. $s\lambda = s$), the CTW probability assignment is given by $p_w(x^n \parallel y^n) \triangleq f_w^{(\lambda)}(x^n \parallel y^n)$. This probability assignment may be recursively lower-bounded as:

$$\begin{aligned} p_w(x^n \parallel y^n) &= \frac{1}{2} p_e^{(\lambda)}(x^n \parallel y^n) + \frac{1}{2} \prod_{s \in \mathcal{X} \times \mathcal{Y}} p_w^{(s)}(x^n \parallel y^n) \\ &\geq \frac{1}{2} \prod_{s \in \mathcal{X} \times \mathcal{Y}} p_w^{(s)}(x^n \parallel y^n) \\ &\geq \frac{1}{2} \prod_{s \in \mathcal{X} \times \mathcal{Y}} \frac{1}{2} \prod_{s' \in \mathcal{X} \times \mathcal{Y}} p_w^{(s's)}(x^n \parallel y^n) \\ &\geq \dots \\ &\geq \frac{1}{2^S} \prod_{l \in \mathcal{L}} p_w^{(l)}(x^n \parallel y^n) \\ &= \frac{1}{2^S} \prod_{l \in \mathcal{L}} p_e^{(l)}(x^n \parallel y^n) \\ &= \frac{1}{2^S} p_c^{(l)}(x^n \parallel y^n). \end{aligned}$$

Finally, we can consider the log-likelihood ratio of true probability and the CTW probability in order to obtain a bound on the worst case regret of the CTW:

$$\begin{aligned} \sup_{x^n, y^n} \log \frac{p(x^n \parallel y^n)}{p_w(x^n \parallel y^n)} &\leq S + \log \frac{p(x^n \parallel y^n)}{p_c(x^n \parallel y^n)} \\ &\leq S + \frac{L(|\mathcal{X}| - 1)}{2} \log \frac{n}{L} + L(|\mathcal{X}| - 1) \end{aligned}$$

as was to be shown. \square

APPENDIX E PROOF OF THEOREMS

A. Proof of Theorem 1

We begin by defining the functions:

$$\hat{g}_i(X) \triangleq \log \frac{\hat{p}_{X_i}^{(c)}(X)}{\hat{p}_{X_i}^{(r)}(X)} \quad g_i^*(X) \triangleq \log \frac{p_{X_i}^{(c)*}(X)}{p_{X_i}^{(r)*}(X)}.$$

Using the definition of the causal measure and KL-divergence:

$$\begin{aligned} &\sum_{i=1}^n \left| \hat{C}_{Y \rightarrow X}(i) - C_{Y \rightarrow X}^*(i) \right| \\ &\quad - \left| E_{p_{X_i}^{(c)*}} [\hat{g}_i(X)] - E_{\hat{p}_{X_i}^{(c)}} [\hat{g}_i(X)] \right| \quad (80) \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^n \left| E_{p_{X_i}^{(c)*}} [g_i^*(X)] - E_{\hat{p}_{X_i}^{(c)}} [\hat{g}_i(X)] \right| \\ &\quad - \left| E_{p_{X_i}^{(c)*}} [\hat{g}_i(X)] - E_{\hat{p}_{X_i}^{(c)}} [\hat{g}_i(X)] \right| \\ &\leq \sum_{i=1}^n \left| \left| E_{p_{X_i}^{(c)*}} [g_i^*(X)] - E_{\hat{p}_{X_i}^{(c)}} [\hat{g}_i(X)] \right| \right. \\ &\quad \left. - \left| E_{p_{X_i}^{(c)*}} [\hat{g}_i(X)] - E_{\hat{p}_{X_i}^{(c)}} [\hat{g}_i(X)] \right| \right| \quad (81) \end{aligned}$$

$$\begin{aligned} &\leq \sum_{i=1}^n \left| \left| E_{p_{X_i}^{(c)*}} [g_i^*(X)] - E_{\hat{p}_{X_i}^{(c)}} [\hat{g}_i(X)] \right| \right. \\ &\quad \left. - E_{p_{X_i}^{(c)*}} [\hat{g}_i(X)] + E_{\hat{p}_{X_i}^{(c)}} [\hat{g}_i(X)] \right| \quad (82) \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^n \left| E_{p_{X_i}^{(c)*}} [g_i^*(X) - \hat{g}_i(X)] \right| \\ &= \sum_{i=1}^n \left| E_{p_{X_i}^{(c)*}} \left[\log \frac{p_{X_i}^{(c)*}(X)}{\hat{p}_{X_i}^{(c)}(X)} - \log \frac{p_{X_i}^{(r)*}(X)}{\hat{p}_{X_i}^{(r)}(X)} \right] \right| \\ &\leq \sum_{i=1}^n \left| D(p_{X_i}^{(c)*} \parallel \hat{p}_{X_i}^{(c)}) \right| + \left| E_{p_{X_i}^{(c)*}} \left[\log \frac{p_{X_i}^{(r)*}(X)}{\hat{p}_{X_i}^{(r)}(X)} \right] \right| \quad (83) \\ &\leq M^{(c)}(n) + M^{(r)}(n) \quad (84) \end{aligned}$$

where (81) follows from the properties of absolute value, (82) follows from the reverse triangle inequality, (83) follows from the triangle inequality, and (84) follows from non-negativity of the KL-divergence, Lemma 2, and Assumption 2. Moving the second term of (80) to the other side of the inequality yields:

$$\sum_{i=1}^n \left| \hat{C}_{Y \rightarrow X}(i) - C_{Y \rightarrow X}^*(i) \right|$$

$$\begin{aligned} &\leq M^{(c)}(n) + M^{(r)}(n) + \sum_{i=1}^n \left| E_{p_{X_i}^{(c)*}} [\hat{g}_i(X)] - E_{\hat{p}_{X_i}^{(c)}} [\hat{g}_i(X)] \right| \\ &\leq M^{(c)}(n) + M^{(r)}(n) + \frac{\|\tilde{c}_n\|_2}{\sqrt{2}} \sqrt{M^{(c)}(n)} \end{aligned} \quad (85)$$

where (85) follows from Lemma 3. This concludes the proof. \square

B. Proof of Theorem 2

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

$$\begin{aligned} &p(X_i | X^{i-1}, Z^{i-1}) \\ &= \sum_{y_{i-d}^{i-1}} p(X_i | X^{i-1}, y_{i-d}^{i-1}, Z^{i-1}) \prod_{j=i-d}^{i-1} p(y_j | X^{i-1}, Z^{i-1}) \end{aligned} \quad (86)$$

$$= \sum_{y_{i-d}^{i-1}} p(X_i | X_{i-d}^{i-1}, y_{i-d}^{i-1}, Z_{i-d}^{i-1}) \prod_{j=i-d}^{i-1} p(y_j | X_{j-d}^{i-1}, Z_{j-d}^{i-1}) \quad (87)$$

$$= \sum_{y_{i-d}^{i-1}} p(X_i | X_{i-2d}^{i-1}, y_{i-d}^{i-1}, Z_{i-2d}^{i-1}) \prod_{j=i-d}^{i-1} p(y_j | X_{i-2d}^{i-1}, Z_{i-2d}^{i-1}) \quad (88)$$

where (86) follows from the chain rule and that y_{i-d}^{i-1} are conditionally independent given (X^{i-1}, Z^{i-1}) , (87) follows from the joint Markovicity of X and Y and the conditional independence of y_{i-d}^{i-1} , and (88) 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 $I(Y_j; Y_k | X^i, Z^i) > 0$, then there is no positive integer l such that $(X_{i-l}^{i-1}, Z_{i-l}^{i-1})$ d-separates (X^{i-l-1}, Z^{i-l-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 $\tau_1 = k - j$, then by the joint stationarity of (X, Y, Z) , every Y_i will be connected to $Y_{i-\tau_1}$ at the end of step 4. Furthermore, we know that $I(Y^n \rightarrow X^n | Z^n) > 0$ 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 $Y_{i-\tau_2}, Y_{i-\tau_2-\tau_1}, Y_{i-\tau_2-2\tau_1}, \dots$. Thus, for any $l \geq 1$ this path can be followed r steps such that $r\tau_1 > d$. Then we know that $Y_{i-\tau_2-r\tau_1}$ is connected via an undirected edge to $X_{i-\tau_2-r\tau_1+\tau_2} = X_{i-r\tau_1}$. Recalling that in step 3 of the d-separation algorithm, $(X_{i-l}^{i-1}, Z_{i-l}^{i-1})$ have been removed from the graph, we note that since $i - r\tau_1 < i - l$, $X_{i-r\tau_1}$

is in the graph. Thus, there is an undirected path connecting $X_{r\tau_1} \in X^{i-l-1}$ and X_i , which implies that $(X_{i-l}^{i-1}, Z_{i-l}^{i-1})$ does not d-separate (X^{i-l-1}, Z^{i-l-1}) and X_i for any l . \square

C. Proof of Theorem 3

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

$$p(x_i | x_{i-l-1}^{i-1}, z_{i-l-1}^{i-1}) = p(x_i | \tilde{x}_{i-l-1}^{i-1}, \tilde{z}_{i-l-1}^{i-1}) \quad (89)$$

where we define $\tilde{x}_{i-l-d}^{i-1} \triangleq \{x_{i-l}^{i-1}, x'_{i-l-1}\}$ and $\tilde{z}_{i-l-1}^{i-1} \triangleq \{z_{i-l}^{i-1}, z'_{i-l-1}\}$. Define $\theta_{X,Y,Z}^{S_i} \triangleq p(S_i | X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1})$ for $S \in \{X, Y, Z\}$ and $\theta \triangleq \{\theta_b^a : a \in \mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}, b \in \mathcal{X}^d \times \mathcal{Y}^d \times \mathcal{Z}^d\}$. We will demonstrate that the equation given by (89) amounts to solving a polynomial function of the parameters θ . It is shown in [59] 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 (89), we see that:

$$\begin{aligned} &p(x_i | x_{i-l-1}^{i-1}, z_{i-l-1}^{i-1}) \quad (90) \\ &= \sum_{y_{i-l-1}^{i-1}} \theta_{x,y,z}^{x_i} p(y_{i-l-1}^{i-1} | x_{i-l-1}^{i-1}, z_{i-l-1}^{i-1}) \\ &= \sum_{y_{i-l-1}^{i-1}} \theta_{x,y,z}^{x_i} \frac{p(x_{i-l-1}^{i-1}, y_{i-l-1}^{i-1}, z_{i-l-1}^{i-1})}{p(x_{i-l-1}^{i-1}, z_{i-l-1}^{i-1})} \\ &= \frac{\sum_{y_{i-l-1}^{i-1}} \theta_{x,y,z}^{x_i} \pi(x_{i-l-1}, y_{i-l-1}, z_{i-l-1}) \prod_{j=1}^l \theta_{x,y,z}^{(x,y,z)_{i-j}}}{\sum_{\tilde{y}_{i-l-1}^{i-1}} \pi(x_{i-l-1}, \tilde{y}_{i-l-1}, z_{i-l-1}) \prod_{j=1}^l \theta_{x,\tilde{y},z}^{(x,\tilde{y},z)_{i-j}}} \end{aligned} \quad (91)$$

where $\pi : |\mathcal{X}| \times |\mathcal{Y}| \times |\mathcal{Z}| \rightarrow [0, 1]$ is the invariant distribution and $\theta_{x,y,z}^{(x,y,z)_i} \triangleq \theta_{x,y,z}^{x_i} \theta_{x,y,z}^{y_i} \theta_{x,y,z}^{z_i}$. Next, define a matrix $A \in \mathbb{R}^{|\mathcal{X}||\mathcal{Y}||\mathcal{Z}| \times |\mathcal{X}||\mathcal{Y}||\mathcal{Z}|}$ containing the transition probabilities, i.e. $A_{j,k} = \theta_{R_j}^{R_k}$ where R is some enumeration over the $|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|$ possible values taken by (X, Y, Z) . Then we can represent π in vector form $\pi \in [0, 1]^{|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|}$ as the unique solution to $\pi = \pi A$. Let $\tilde{\pi}$ be an arbitrary vector satisfying $(A^T - I)\tilde{\pi} = 0$, and note that for any $\tilde{\pi}$ there is a constant C such that $C\tilde{\pi} = \pi$. Such a vector $\tilde{\pi}$ can be found by performing Gauss-Jordan elimination on $(A^T - I)$, and as a result, each element $\tilde{\pi}_j$ can be written as fractions of polynomial functions of θ . Replacing π with $C\tilde{\pi}$ in its functional form $\tilde{\pi} : |\mathcal{X}| \times |\mathcal{Y}| \times |\mathcal{Z}| \rightarrow \mathbb{R}$ in (91) we see that C cancels in the numerator and denominator and thus each side of (89) can be written entirely as fractions of polynomial functions of θ . Next, repeat the process on the right hand side of (89) with \tilde{x}_{i-l-d}^{i-1} and \tilde{z}_{i-l-d}^{i-1} . Then, for any term that appears as a fraction, we can multiply both sides of (89) by the denominator and repeat until (89) is a polynomial function of θ . Finally, we note that the polynomial given by (89) is trivial only if every process is a solution which can be shown not to be the case by a straightforward counterexample. \square

D. Proof of Theorem 4

We will first show that \tilde{X} is $(d+k)$ -Markov, i.e. $\tilde{X}_i \perp \tilde{X}^{i-k-d-1} | \tilde{X}_{i-k-d}^{i-1}$. Note that the distribution of \tilde{X}_i given \tilde{X}^{i-1} may be written as:

$$\begin{aligned}
 & p(X_i, Y_{i-k+1} | X^{i-1}, Y^{i-k}) \\
 &= \sum_{y_{i-k+2}^{i-1}} p(X_i, Y_{i-k+1}, y_{i-k+2}^{i-1} | X^{i-1}, Y^{i-k}) \\
 &\triangleq \sum_{y_{i-k+2}^{i-1}} p(X_i, \tilde{Y}_{i-k+1}^{i-1} | X^{i-1}, Y^{i-k}) \\
 &= \sum_{y_{i-k+2}^{i-1}} p(X_i | X^{i-1}, \tilde{Y}^{i-1}) p(\tilde{Y}_{i-k+1}^{i-1} | X^{i-1}, Y^{i-k}) \\
 &= \sum_{y_{i-k+2}^{i-1}} p(X_i | X_{i-k-d}^{i-1}, \tilde{Y}_{i-k-d}^{i-1}) p(\tilde{Y}_{i-k+1}^{i-1} | X_{i-k-d}^{i-1}, Y_{i-k-d}^{i-k}) \\
 &= \sum_{y_{i-k+2}^{i-1}} p(X_i, \tilde{Y}_{i-k+1}^{i-1} | X_{i-k-d}^{i-1}, Y_{i-k-d}^{i-k}) \\
 &= p(X_i, Y_{i-k+1} | X_{i-k-d}^{i-1}, Y_{i-k-d}^{i-k})
 \end{aligned} \tag{92}$$

where, for ease of notation, we have defined $\tilde{Y}_j = y_j$ if $i-k+2 \leq j \leq i-1$ and $\tilde{Y}_j = Y_j$ otherwise, and (93) follows from the joint Markovicity of X and Y and:

$$\begin{aligned}
 p(\tilde{Y}_{i-k+1}^{i-1} | X^{i-1}, Y^{i-k}) &= \prod_{j=i-k+1}^{i-1} p(\tilde{Y}_j | X^{i-1}, \tilde{Y}^{j-1}) \\
 &= \prod_{j=i-k+1}^{i-1} p(\tilde{Y}_j | X_{i-k-d}^{i-1}, \tilde{Y}_{i-k-d}^{j-1}) \\
 &= p(\tilde{Y}_{i-k+1}^{i-1} | X_{i-k-d}^{i-1}, Y_{i-k-d}^{i-k})
 \end{aligned}$$

where we define $\tilde{Y}_a^b = \emptyset$ when $b < a$. This proves the Markovicity of \tilde{X} . To get the equality given by (53), we simply take the sum over Y_{i-k+1} in the above equations.

Next, we will show that \tilde{X} is irreducible. We note that the possible states of \tilde{X} may be a subset of the possible states of (X, Y) , i.e. $\tilde{\mathcal{X}} \subset \mathcal{X} \times \mathcal{Y}$. Each state $\tilde{x} \in \tilde{\mathcal{X}}$ occurs as a result of visiting a state (x_{i-k+1}, y_{i-k+1}) followed by (x_i, y_i) after $k-1$ steps. Given that (X, Y) is irreducible, every state $(x_{i-k+1}, y_{i-k+1}) \in \mathcal{X} \times \mathcal{Y}$ can be visited from any state $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. As a result, every state in $\tilde{\mathcal{X}}$ can be visited from any other state $\tilde{x}' \in \tilde{\mathcal{X}}$. Therefore, \tilde{X} is irreducible.

Lastly, we will show that if (X, Y) is aperiodic then \tilde{X} is also aperiodic. Note that for any state $\tilde{x}_i = (X_i = a, Y_{i-k+1} = b) \in \tilde{\mathcal{X}}$, we know there exists $c \in \mathcal{X}$ and $d \in \mathcal{Y}$ such that $p(X_{i-k+1} = c, Y_{i-k+1} = b, X_i = a, Y_i = d) > 0$. By the aperiodicity we know that the greatest common divisor of the set of τ such that:

$$p(X_{i-k+1} = c, Y_{i-k+1} = b, X_{i-k+1+\tau} = c, Y_{i-k+1+\tau} = b) > 0$$

is one. As a result, the same is true of τ such that:

$$\begin{aligned}
 & 0 < p(X_{i-k+1} = c, Y_{i-k+1} = b, X_i = a, Y_i = d, \\
 & \quad X_{i-k+1+\tau} = c, Y_{i-k+1+\tau} = b, X_{i+\tau} = a, Y_{i+\tau} = d) \\
 & \leq p(Y_{i-k+1} = b, Y_{i-k+1+\tau} = b, X_i = a, X_{i+\tau} = a)
 \end{aligned}$$

$$= p(\tilde{x}_i, \tilde{x}_{i+\tau})$$

implying that \tilde{X} is aperiodic. \square

E. Proof of Theorem 5

Let the estimate of the partial DI rate be given by:

$$\hat{I}_{P,n}^{(k)}(Y \rightarrow X) \triangleq \frac{1}{n} \sum_{i=1}^n D(\hat{p}_{X_i}^{(c)} || \hat{p}_{X_i}^{(k)}). \tag{94}$$

Then the theorem states that $\hat{I}_{P,n}^{(k)}$ converges to $\bar{I}_P^{(k)}$ almost surely. Following the proof of Theorem 3 in [28], decompose the estimate as:

$$\begin{aligned}
 \hat{I}_n^{(k)}(Y \rightarrow X) &= \frac{1}{n} \sum_{i=1}^n \sum_{x_i} \hat{p}_{X_i}^{(c)}(x_i) \log \frac{1}{\hat{p}_{X_i}^{(k)}(x_i)} \\
 &\quad - \frac{1}{n} \sum_{i=1}^n \sum_{x_i} \hat{p}_{X_i}^{(c)}(x_i) \log \frac{1}{\hat{p}_{X_i}^{(c)}(x_i)}.
 \end{aligned}$$

It was shown in [28] that the second term on the right hand side of the above equation converges to $\bar{H}^{(1)}(X || Y)$ almost surely. Next, define the quantity:

$$F_n^{(k)} \triangleq \frac{1}{n} \sum_{i=1}^n \sum_{x_i} \hat{p}_{X_i}^{(c)}(x_i) \log \frac{1}{\hat{p}_{X_i}^{(k)}(x_i)}. \tag{95}$$

Then it remains to be shown that $F_n^{(k)}$ converges to $\bar{H}^{(k)}(X || Y)$ almost surely. Next, define $R_n^{(k)}$ and $S_n^{(k)}$ as:

$$\begin{aligned}
 R_n^{(k)} &\triangleq \frac{1}{n} \sum_{i=1}^n \left[\sum_{x_i} p_{X_i}^{(c)}(x_i) \log p_{X_i}^{(k)}(x_i) - \hat{p}_{X_i}^{(c)}(x_i) \log \hat{p}_{X_i}^{(k)}(x_i) \right] \\
 S_n^{(k)} &\triangleq - \frac{1}{n} \sum_{i=1}^n \sum_{x_i} p_{X_i}^{(c)}(x_i) \log p_{X_i}^{(k)}(x_i) - \bar{H}^{(k)}(X || Y)
 \end{aligned}$$

and note that $F_n^{(k)} - \bar{H}^{(k)}(X || Y) = R_n^{(k)} + S_n^{(k)}$. As such, all that remains to be shown is that $R_n^{(k)}$ and $S_n^{(k)}$ converge to zero almost surely. It is shown in Lemma 2 of [28] that the CTW probability assignment $\hat{p}_{X_i}^{(c)}(x_i)$ converges to $p_{X_i}^{(c)}(x_i)$ almost surely if (X, Y) is a stationary irreducible aperiodic finite-alphabet Markov process. We showed in Theorem 4 that this condition implies that the process \tilde{X} with $\tilde{X}_i \triangleq (X_i, Y_{i-k+1})$ is also a stationary aperiodic finite-alphabet Markov process and thus $\hat{p}_{X_i}^{(k)}(x_i)$ converges to $p_{X_i}^{(k)}(x_i)$ almost surely as well. As a result, we see that the bracketed term in $R_n^{(k)}$ converges to zero almost surely as i tends to infinity. Furthermore, since $R_n^{(k)}$ is the Cesáro mean of the bracketed term, it too converges to zero almost surely.

To show that $S_n^{(k)}$ converges to zero, first define the first term as:

$$\begin{aligned}
 G_i^{(k)} &\triangleq - \sum_{x_i} p_{X_i}^{(c)}(x_i) \log p_{X_i}^{(k)}(x_i) \\
 &= - \sum_{x_i} p(x_i | x^{i-1}, y^{i-1}) \log p(x_i | x^{i-1}, y^{i-k}) \\
 &= - \sum_{x_i} p(x_i | x_{i-k-d}^{i-1}, y_{i-k-d}^{i-1}) \log p(x_i | x_{i-k-d}^{i-1}, y_{i-k-d}^{i-k}) \\
 &\triangleq g(x_{i-k-d}^{i-1}, y_{i-k-d}^{i-1})
 \end{aligned}$$

Then, from Breiman's generalized ergodic theorem [60], it follows that the following equality holds almost surely:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n g(x_{i-k-d}^{-1}, y_{i-k-d}^{-1}) = E[g(X_{-k-d}^{-1}, Y_{-k-d}^{-1})].$$

Finally, using the law of iterated expectation, we note that:

$$\begin{aligned} & E[g(X_{-k-d}^{-1}, Y_{-k-d}^{-1})] \\ &= E \left[- \sum_{x_0} p(x_0 | X_{-k-d}^{-1}, Y_{-k-d}^{-1}) \right. \\ & \quad \left. \log p(x_0 | X_{-k-d}^{-1}, Y_{-k-d}^{-1}) \right] \\ &= E \left[E \left[- \sum_{x_0} f(x_0 | X_{-k-d}^{-1}, Y_{-k-d}^{-1}) \right. \right. \\ & \quad \left. \left. \log p(x_0 | X_{-k-d}^{-1}, Y_{-k-d}^{-1}) \middle| X_{-k-d}^{-1}, Y_{-k-d}^{-1} \right] \right] \\ &= E \left[- \sum_{x_0} p(x_0 | X_{-k-d}^{-1}, Y_{-k-d}^{-1}) \right. \\ & \quad \left. \log p(x_0 | X_{-k-d}^{-1}, Y_{-k-d}^{-1}) \right] \\ &= \bar{H}^{(k)}(X || Y) \end{aligned}$$

Thus, we conclude that:

$$\lim_{n \rightarrow \infty} S_n^{(k)} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n G_i^{(k)} - \bar{H}^{(k)}(X || Y) = 0 \text{ } p - a.s.$$

as was to be shown. \square

ACKNOWLEDGMENT

The authors would like to thank Will Chapman for insightful conversations and the anonymous reviewers for their suggestions.

REFERENCES

- [1] L. S. Junior, A. Mullokandov, and D. Y. Kenett, "Dependency relations among international stock market indices," *Journal of Risk and Financial Management*, vol. 8, no. 2, pp. 227–265, 2015.
- [2] T. Dimpfl and F. J. Peter, "The impact of the financial crisis on transatlantic information flows: An intraday analysis," *Journal of International Financial Markets, Institutions and Money*, vol. 31, pp. 1–13, 2014.
- [3] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A Granger causality measure for point process models of ensemble neural spiking activity," *PLoS computational biology*, vol. 7, no. 3, e1001110, 2011.
- [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," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.
- [5] N. Wiener, "The theory of prediction," *Modern mathematics for engineers*, 1956.
- [6] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, 1969.
- [7] J. Pearl, *Causality*. Cambridge university press, 2009.
- [8] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic Books, 2018.
- [9] M. Eichler and V. Didelez, "On granger causality and the effect of interventions in time series," *Lifetime data analysis*, vol. 16, no. 1, pp. 3–32, 2010.
- [10] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger causality analysis in neuroscience and neuroimaging," *Journal of Neuroscience*, vol. 35, no. 8, pp. 3293–3297, 2015.
- [11] H. Marko, "Die theorie der bidirektionalen kommunikation und ihre anwendung auf die nachrichtenübermittlung zwischen menschen (subjektive information)," *Kybernetik*, vol. 3, no. 3, pp. 128–136, 1966.
- [12] —, "The bidirectional communication theory—a generalization of information theory," *IEEE Transactions on communications*, 1973.
- [13] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.*, 1990.
- [14] T. Schreiber, "Measuring information transfer," *Physical review letters*, 2000.
- [15] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Physical review letters*, 2009.
- [16] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Equivalence between minimal generative model graphs and directed information graphs," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, IEEE, 2011, pp. 293–297.
- [17] J. Etesami and N. Kiyavash, "Directed information graphs: A generalization of linear dynamical graphs," in *American Control Conference (ACC), 2014*, IEEE, 2014, pp. 2563–2568.
- [18] J. T. Lizier, "Measuring the dynamics of information processing on a local scale in time and space," in *Directed information measures in neuroscience*, Springer, 2014, pp. 161–193.
- [19] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.
- [20] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.
- [21] S. Kim, C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Dynamic and succinct statistical analysis of neuroscience data," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 683–698, 2014.
- [22] G. Schamberg and T. P. Coleman, "A sample path measure of causal influence," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 2251–2255. DOI: 10.1109/ISIT.2018.8437627.
- [23] P. A. Stokes and P. L. Purdon, "A study of problems encountered in Granger causality analysis from a neuroscience perspective," *Proceedings of the National*

- Academy of Sciences*, vol. 114, no. 34, E7063–E7072, 2017.
- [24] G. Schamberg and T. P. Coleman, “On the bias of directed information estimators,” *arXiv preprint arXiv:1902.00580*, 2019.
- [25] G. Kramer, “Directed information for channels with feedback,” PhD thesis, 1998.
- [26] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf, “Quantifying causal influences,” *The Annals of Statistics*, pp. 2324–2358, 2013.
- [27] P.-O. Amblard and O. J. Michel, “The relation between Granger causality and directed information theory: A review,” *Entropy*, vol. 15, no. 1, pp. 113–143, 2012.
- [28] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, “Universal estimation of directed information,” *IEEE Transactions on Information Theory*, 2013.
- [29] J. T. Lizier, “JIDT: An information-theoretic toolkit for studying the dynamics of complex systems,” *arXiv preprint arXiv:1408.3270*, 2014.
- [30] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.
- [31] A. Sheikhattar, S. Miran, J. Liu, J. B. Fritz, S. A. Shamma, P. O. Kanold, and B. Babadi, “Extracting neuronal functional network dynamics via adaptive Granger causality analysis,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 17, E3869–E3878, 2018.
- [32] B. Oselio and A. Hero, “Dynamic reconstruction of influence graphs with adaptive directed information,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, IEEE, 2017.
- [33] N. Ay and D. Polani, “Information flows in causal networks,” *Advances in complex systems*, vol. 11, no. 01, pp. 17–41, 2008.
- [34] H. H. Permuter, Y.-H. Kim, and T. Weissman, “Interpretations of directed information in portfolio theory, data compression, and hypothesis testing,” *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3248–3259, 2011.
- [35] N. Merhav and M. Feder, “Universal prediction,” *IEEE Transactions on Information Theory*, 1998.
- [36] Y. Murin, “K-NN estimation of directed information,” *arXiv preprint arXiv:1711.08516*, 2017.
- [37] Y. Murin, J. Kim, and A. Goldsmith, “Tracking epileptic seizure activity via information theoretic graphs,” in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, IEEE, 2016, pp. 583–587.
- [38] C. J. Quinn, N. Kiyavash, and T. P. Coleman, “Directed information graphs,” *IEEE Transactions on information theory*, 2015.
- [39] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson, *Causation, prediction, and search*. MIT press, 2000.
- [40] C. Meek, “Strong completeness and faithfulness in Bayesian networks,” in *Proc. of the Eleventh Conf. on Uncertainty in Artificial Intelligence*, 1995.
- [41] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer, “Independence properties of directed Markov fields,” *Networks*, vol. 20, no. 5, pp. 491–505, 1990.
- [42] I. Kontoyiannis and M. Skoularidou, “Estimating the directed information and testing for causality,” *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6053–6067, 2016.
- [43] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, “The context-tree weighting method: Basic properties,” *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.
- [44] H. Cai, S. R. Kulkarni, and S. Verdú, “A universal lossless compressor with side information based on context tree weighting,” in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, IEEE, 2005, pp. 2340–2344.
- [45] T. J. Tjalkens, Y. M. Shtarkov, and F. M. Willems, “Sequential weighting algorithms for multialphabet sources,” in *6th Joint Swedish-Russian Int. Worksh. Inform. Theory*, Citeseer, 1993.
- [46] R. Kleeman, “Measuring dynamical prediction utility using relative entropy,” *Journal of the atmospheric sciences*, vol. 59, no. 13, pp. 2057–2072, 2002.
- [47] P. Geiger, K. Zhang, B. Schoelkopf, M. Gong, and D. Janzing, “Causal inference by identification of vector autoregressive processes with hidden components,” in *International Conference on Machine Learning*, 2015, pp. 1917–1925.
- [48] J. Etesami, N. Kiyavash, and T. Coleman, “Learning minimal latent directed information polytrees,” *Neural computation*, vol. 28, no. 9, pp. 1723–1768, 2016.
- [49] V. Matta and A. H. Sayed, “Consistent tomography under partial observations over adaptive networks,” *IEEE Transactions on Information Theory*, 2018.
- [50] S. K. Gorantla, S. Kadloor, N. Kiyavash, T. P. Coleman, I. S. Moskowitz, and M. H. Kang, “Characterizing the efficacy of the NRL network pump in mitigating covert timing channels,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 64–75, 2012.
- [51] N. C. Martins and M. A. Dahleh, “Feedback control in the presence of noisy channels: “Bode-like” fundamental limitations of performance,” *IEEE Transactions on Automatic Control*, vol. 53, no. 7, pp. 1604–1615, 2008.
- [52] T. Tanaka, P. M. Esfahani, and S. K. Mitter, “LQG control with minimum directed information: Semidefinite programming approach,” *IEEE Transactions on Automatic Control*, vol. 63, no. 1, pp. 37–52, 2018.
- [53] S. Gorantla and T. Coleman, “Information-theoretic viewpoints on optimal causal coding-decoding problems,” *arXiv preprint arXiv:1102.0250*, 2011.
- [54] A. A. Kulkarni and T. P. Coleman, “An optimizer’s approach to stochastic control problems with nonclassical information structures,” *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 937–949, 2015.
- [55] J. Etesami, A. Habibnia, and N. Kiyavash, “Econometric modeling of systemic risk: Going beyond pairwise comparison and allowing for nonlinearity,” *Systemic*

- Risk Centre, The London School of Economics and Political Science, 2017.
- [56] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.
- [57] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson London, 2014, vol. 3.
- [58] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.
- [59] M. Okamoto, "Distinctness of the eigenvalues of a quadratic form in a multivariate sample," *The Annals of Statistics*, 1973.
- [60] L. Breiman, "The individual ergodic theorem of information theory," *The Annals of Mathematical Statistics*, vol. 28, no. 3, pp. 809–811, 1957.

Gabriel Schamberg (S'15–M'19) received his B.S. in computer engineering in 2012, M.S. in electrical engineering in 2016, and Ph.D. in electrical engineering in 2019, all from the University of California, San Diego (UCSD). From 2012 to 2014 he worked as a software developer for NKI Engineering in San Diego, California. In 2014 he began his graduate studies and was awarded the Jacobs Fellowship from the Department of Electrical and Computer Engineering at UCSD. He is currently a Picower Institute Postdoctoral Fellow in the Neuroscience Statistics Research Laboratory at the Massachusetts Institute of Technology.

Todd P. Coleman (S'01–M'05–SM'11) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2005. He was a Postdoctoral Scholar in neuroscience at MIT and Massachusetts General Hospital (MGH) during the 2005-2006 academic year. He was an Assistant Professor in Electrical and Computer Engineering as well as Neuroscience at the University of Illinois from 2006 to 2011. He is currently a Professor in Bioengineering and Principal Investigator of the Neural Interaction Laboratory at the University of California San Diego, La Jolla, CA, USA. His research interests lie at the intersection of applied mathematics, bio-electronics, neuroscience, and medicine. He was named a Gilbreth Lecturer for the National Academy of Engineering in 2015 and a Fellow of the American Institute for Medical and Biological Engineering in 2019.