UNIVERSITY OF CALIFORNIA SAN DIEGO

Information Theoretic Measures and Estimators of Specific Causal Influences

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Communication Theory and Systems)

by

Gabriel Schamberg

Committee in charge:

Professor Todd P. Coleman, Chair
Professor Young-Han Kim, Co-Chair
Professor Andrea Chiba
Professor Piya Pal
Professor Lara Rangel
Professor Ramesh Rao

2019

The dissertation of Gabriel Schamberg is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2019

DEDICATION

To grandpa Richard.

EPIGRAPH

It turns out that an eerie type of chaos can lurk just behind a facade of order - and yet, deep inside the chaos lurks an even eerier type of order.

—Douglas R. Hofstadter

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The need to identify and measure causal influences is ubiquitous across academic disciplines. This ubiquity results in the field of causality being one of great breadth in both the questions being asked by epidemiologists, philosophers, economists, climate scientists, physicists etc., and the types of solutions that these researchers propose with help from statisticians, machine learners, and information theorists. As a result, the field of causality has many corners. These corners can differ not only on the best way to measure and estimate causal influences, but on the best way to define causality. To make matters more complicated, these corners can disagree on whether or not causality *exists*. Even if one assumes causality's existence as a starting point, it is easy to end up in the middle of a debate over the right way to characterize its existence.

My introduction to causality was through the lens of Granger causality. In 1969, the economist Clive Granger proposed the following definition: "We say that Y_t is causing X_t if we are better able to predict X_t using all available information than if the information apart from Y_t had been used." Not only is this definition rather intuitive, it helped win Granger the Nobel Prize in Economics in 2003. Granger causality remains a highly popular area of research and is now applied frequently in neuroscience and climate science. The search term "Granger Causality" yields 6,850 results on Google Scholar between January 1, 2019 and August 1, 2019. As such, a researcher can reside comfortably within the confines of Granger causality, as I did for some time.

Enter Judea Pearl, author of "Causality" [92], pioneer of the so-called *Causal Revolution*, and Turing Award winner. Pearl's viewpoint cannot be summarized as concisely as Granger's, but a primary focus of Pearl's work is to provide the mathematical tools for working with *interventions* in causality. The value of interventions is best understood through an example. We know that a barometer reading will correlate with the air pressure. This correlation tells us that knowing the air pressure will convey some information about the value displayed by the barometer. Given that correlation is symmetric, it also tells us that knowing the barometer reading provides us with information about the true air pressure. When performing interventions, however, this symmetry

is broken. To see this, imagine forcing a specific air pressure to occur – regardless of circumstance (for example, the altitude at which this experiment is being conducted); this forcing of air pressure will affect the barometer reading. On the other hand, if we directly manipulated the barometer, forcing it to display a particular air pressure, this would have no effect on the true air pressure. One of Pearl's major contributions is the *do*-calculus, which is part of a larger mathematical formalism defining the nature of interventions using graphical models.

The idea of interventions reflects intuition, and while it is central to Pearl's perspective, it is absent from Granger's. It thus comes as no surprise when Pearl presents his take on Granger causality (in no uncertain terms) in the first chapter of his recent book "The Book of Why: The New Science of Cause and Effect" [94]:

> [I]n their effort to mathematize the concept of causation – itself a laudable idea – philosophers were too quick to commit to the only uncertainty-handling language they knew, the language of probability. They have for the most part gotten over this blunder, but unfortunately similar ideas are being pursued in econometrics even now, under names like "Granger causality" and "vector autocorrelation."

In other words, Pearl believes that the study of cause and effect requires specialized mathematical tools, such as the *do*-calculus. His book contains a very compelling argument in support of this belief, and those interested in causality are encouraged to read it.

As a causality researcher, it can be challenging to reconcile the difference in these perspectives. On one hand, Pearl's work clearly elucidates the limitations of Granger's perspective; on the other hand, Granger's concise definition gives rise to a very straightforward, user-friendly framework that yields interpretable results. As such, it is my belief that there is value to the vast body of work built upon Granger's ideas and that it is unfair to dismiss Granger causality entirely on the grounds of its inability to address all questions of a causal nature. At the same time, it is irresponsible to present Granger causality in a vacuum without making clear the aspects of causality that it fails to identify.

The brief summary above does not come close to completely characterizing Pearl's perspective, the shortcomings of Granger causality, or the breadth of ideas within the causality

community. Nevertheless, I feel that it is important to acknowledge these points in order to ensure that the contributions that follow are not taken out of context. This dissertation, titled "Information Theoretic Measures and Estimators of Specific Causal Influences," is not an attempt to provide universally applicable measures and estimators of causal influences, nor is it an attempt to say that information theory provides the "right" tools for studying causality. Rather, the goal of this dissertation is to present a series of contributions along my path from Granger to Pearl. In doing so, I will build upon preexisting notions of causality within the information theory community in order to (i) maximize their utility should they be deemed appropriate for a given problem and (ii) attempt to provide some context for how these notions coincide with the broader causality landscape.

ACKNOWLEDGEMENTS

I first want to thank my advisor, Professor Todd Coleman. Not only have you been an excellent academic mentor, you have always supported my commitment to work life balance. Thank you for helping me to have a well rounded and healthy graduate school experience. I would additionally like to thank my co-advisor Professor Young-Han Kim and my collaborators Professors Demba Ba, Lara Rangel, and Andrea Chiba for insightful discussions and continued support over the years.

I want to thank my many colleagues in the Neural Interaction Lab who have all played a central role in making the pursuit of a PhD an enjoyable experience. I am especially grateful for the friendship and guidance given to me by Armen Gharibans and Diego Mesa. Outside of the lab I have many friends to thank for a variety of reasons, including (but not limited to): Mark for sharing my love of Hare Krishna Wednesdays, Will for helping me to combine working and surfing, Praveen for having a like mind, Jesse and Billy for enduring my practice presentations, and Garrick and Blake for inspiring me to do a PhD in the first place.

Many thanks are owed to my brother, Zack, and my parents, Mom and Dad. I am so fortunate to receive your unconditional love, support, and understanding. It makes all the difference. Finally, to my partner, Ashley – Everybody seemed to think that our long distance relationship would be the hardest part of the PhD. Thanks for helping me prove them wrong. I'm so grateful to have you in my life and am so happy that we have made it to where we are today.

Chapter 2, in full, is a reprint of the material as it appears in IEEE Transactions on Signal Processing 2018. Schamberg, Gabriel; Ba, Demba; Coleman, Todd, IEEE, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is a reprint of the material as it appears in the proceedings of the IEEE International Symposium on Information Theory 2019. Schamberg, Gabriel; Coleman, Todd, IEEE, 2019. Chapter 3, in part, is currently under review for publication of the material. Schamberg, Gabriel; Coleman, Todd. The dissertation author was the primary investigator and

author of these materials.

Chapter 4, in part, is currently under review for publication of the material. Schamberg, Gabriel; Coleman, Todd. The dissertation author was the primary investigator and author of this material.

Chapter 5, in part, is currently under review for publication of the material. Schamberg, Gabriel; Coleman, Todd. Chapter 5, in part, is currently being prepared for submission for publication of the material. Schamberg, Gabriel; Chapman, William; Coleman, Todd. The dissertation author was the primary investigator and author of these materials.

VITA

2012	B. S. in Computer Engineering *cum laude*, University of California, San Diego
2016	M. S. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego
2016	Teaching Assistant, University of California, San Diego
2018	Instructor of Record, University of California, San Diego
2019	Ph. D. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego

PUBLICATIONS

G. Schamberg, W. Chapman, and T. P. Coleman, "Direct and Indirect Effects – An Information Theoretic Perspective," In Preparation.

G. Schamberg and T. P. Coleman, "Information Theoretic Measures of Direct and Indirect Effects," Under Review.

G. Schamberg and T. P. Coleman, "Measuring Sample Path Causal Influences with Relative Entropy," Under Review.

A. Allegra, A. Gharibans, G. Schamberg, D. Kunkel, and T. P. Coleman, "Bayesian Inverse Methods for Spatiotemporal Characterization of Gastric Electrical Activity from Cutaneous Multi-Electrode Recording," Under Review.

G. Schamberg and T. P. Coleman, "On the Bias of Directed Information Estimators," IEEE International Symposium on Information Theory (ISIT), July 2019.

G. Schamberg and T. P. Coleman, "Quantifying Context-Dependent Causal Influences," NeurIPS Workshop On Causal Learning, December 2018.

G. Schamberg and T. P. Coleman, "A Sample Path Measure of Causal Influence," IEEE International Symposium on Information Theory (ISIT), June 2018.

G. Schamberg, D. Ba, and T. P. Coleman, "A Modularized Efficient Framework for Non-Markov Time Series Estimation," IEEE Transactions on Signal Processing, Volume 66, Issue 12, June 2018.

G. Schamberg, M. Wagner, D. Ba, and T. P. Coleman, "Efficient Low-Rank Spectrotemporal Decomposition using ADMM," IEEE Statistical Signal Processing Workshop, June 2016.

ABSTRACT OF THE DISSERTATION

Information Theoretic Measures and Estimators of Specific Causal Influences

by

Gabriel Schamberg

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California San Diego, 2019

Professor Todd P. Coleman, Chair
Professor Young-Han Kim, Co-Chair

The need to measure causal influences between random variables or processes in complex networks arises throughout academic disciplines. In four parts, we here develop techniques for measuring and estimating causal influences using tools from information theory, with the explicit goal of providing context for how information theoretic perspectives on causal influence fit within the vast and interdisciplinary body of work studying causality. Throughout the dissertation, we demonstrate the utility of the proposed methods with applications to physiologic, economic, and climatological datasets.

Beginning with a focus on time series, we present a modularized approach to finding the

maximum a posteriori estimate of a latent time series that obeys a dynamic stochastic model and is observed through noisy measurements. We specifically consider modern signal processing problems with non-Markov signal dynamics (e.g., group sparsity) and/or non-Gaussian measurement models (e.g., point process observation models used in neuroscience). Importantly, this framework can be leveraged in the estimation of the latent parameters specifying the probability distribution of a time series, which is a fundamental step in the estimation of causal influences between time series.

Second, we study the conditions under which directed information, a popular information theoretic notion of causal influence between time series, can be estimated without bias. While the assumptions made by estimators of directed information are often presented explicitly, a characterization of when we can expect these assumptions to hold is lacking. Using the concept of d-separation from Bayesian networks, we present sufficient and almost everywhere necessary conditions for which proposed estimators can be implemented without bias. We further introduce a notion of partial directed information, which can be used to bound the bias under a milder set of assumptions.

Third, we present a sample path dependent measure of causal influence between time series. The proposed measure is a random sequence, a realization of which enables identification of specific patterns that give rise to high levels of causal influence. 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 which we subsequently bound.

Finally, we extend our focus to general causal graphs and show that information theoretic measures of causal influence are fundamentally different from mainstream (e.g. statistical) notions in that they (1) compare distributions over the effect rather than values of the effect and (2) are defined with respect to random variables representing a cause rather than specific values of a cause. We leverage perspectives from the statistical causality literature to present a novel information theoretic framework for measuring direct, indirect, and total causal effects in natural

complex networks. In addition to endowing information theoretic approaches with an enhanced "resolution," the proposed framework uniquely elucidates the relationship between the information theoretic and statistical perspectives on causality.

Chapter 1

Introduction

We consider the problem of understanding the causal influences between elements of a network of interacting processes or variables. We view this problem as having three distinct aspects. First, *measurement* involves defining a quantitative measure of causal influence that satisfies desirable properties and behaves well in examples where the joint distributions of the interacting elements are fully known. Given that there is no single measure that will work in every problem setting, we present multiple measures that employ a variety of perspectives and offer different benefits in terms of ease of estimation and interpretability. Second, *estimation* addresses the practical difficulties associated with learning the value of a causal measure when the true distributions are unknown. This aspect involves developing estimation algorithms, proving performance guarantees, and evaluating the efficacy of estimators through simulations. Third, the *application* of the estimation techniques to data is used to demonstrate the value of the newly defined measures in addressing real world questions.

Consider Reichenbach's common cause principle [102], which states that if X and Y are correlated, then either (i) X causes Y, (ii) Y causes X, or (iii) X and Y share a common cause. The field of *structure identification* (to use the language of Peters et al. [97]) seeks to identify which of these three explanations correctly characterizes the correlation between X

and Y. It is important to note that the aspects of causality considered in this dissertation do not include structure identification. We avoid the need to identify the underlying structure in one of two ways, depending on the problem setting. In the context of measuring influences between time series (Chapters 3 and 4), we make the assumption that a cause will precede its effect. Thus, when discussing the causal effects between two *processes* X and Y, we are utilizing the passage of time to determine the direction of an arrow and a time-lagged correlation[1] to determine the presence/absence of an arrow. This assumption does not alone resolve issues around distinguishing a causal effect from a shared common cause – this requires much stricter assumptions, which are discussed in detail. In the context of measuring influences between random variables, we assume that a graphical representation of the causal structure is provided in the form of a directed acyclic graph. While this may at first seem like a strong assumption, it may be reasonable to obtain these graphs through domain expertise or common sense in certain settings. For example, in the climatological example considered in Chapter 5, climate scientists provide the knowledge that that the phase of the El Niño-Southern Oscillation may influence temperatures, but not the other way around. Similarly, common sense tells us that one may expect the time of year to influence temperatures, but not the other way around.

As we explore different problem settings, we employ different notions of causal influence. Broadly speaking, this dissertation can be seen as beginning with a perspective akin to Clive Granger (see [45, 111]) and transitioning to a perspective akin to Judea Pearl (see [92, 94, 91]). Along this path, we consistently utilize and build upon the relevant tools from information theory, such as directed information [79, 80, 82], transfer entropy [110], causal strength [54], information flow [10], and local information measures [74]. With the introduction of these different aspects of the literature, it is important to be aware of the limitations of the proposed measures and estimators, to understand the assumptions that are made, and to take seriously the implications of the assumptions not holding. In particular, we note that there are some very compelling caveats

[1] We here use "correlation" to mean a symmetric measure of dependence such as the mutual information, for example.

to Granger's perspective on causal influence that are discussed briefly in the Preface and in detail in the following chapters.

The substance of this dissertation begins with Chapter 2. While this chapter does not focus directly on the topic of causality, it introduces a method that can be leveraged in the estimation of causal influences between time series. Specifically, we present a compartmentalized approach to finding the maximum a-posteriori (MAP) estimate of a latent time series that obeys a dynamic stochastic model and is observed through noisy measurements. We focus primarily on modern signal processing problems with non-Markov signal dynamics (e.g. group sparsity) and/or non-Gaussian measurement models (e.g. point process observation models used in neuroscience). Through the use of auxiliary variables in the MAP estimation problem, we show that a consensus formulation of the alternating direction method of multipliers (ADMM) [16] enables iteratively computing separate estimates based on the likelihood and prior and subsequently "averaging" them in an appropriate sense using a Kalman smoother. We further show that this estimation procedure converges to the true MAP estimate under mild log-concavity assumptions. As such, this approach can be applied to a broad class of problem settings and only requires modular adjustments when interchanging various aspects of the statistical model. Within the context of causality, we note that all measures of causal influence discussed in this dissertation require knowledge of the underlying joint distribution of the data. Thus, when the distribution is unknown, it must be estimated. In scenarios where the parameters specifying such a joint distribution are known to vary with time, we can treat those parameters as a latent time series and use the proposed framework to obtain an estimate. This chapter is presented as a reprint of [109], where, in addition to developing the methodology, we present two example applications involving (i) group-sparsity priors, within the context of electrophysiologic specrotemporal estimation, and (ii) non-Gaussian measurement models, within the context of dynamic analyses of learning with neural spiking and behavioral observations.

In Chapter 3, we shift focus to the estimation of directed information (DI), which can

be viewed as a generalization of Granger's definition of causality. The primary focus of this chapter is the characterization of scenarios for which the assumptions that are typically required by estimators of DI can be expected to hold. This area of study was motivated by the observation that, in general, subsets of finite order autoregressive processes are themselves autoregressive processes of *infinite* order. This observation has been discussed at length in the Granger causality community, as it has serious implications for the estimation of Granger causality. Given the relationship between Granger causality and DI, it comes as no surprise that estimators of DI make an analogous set of assumptions with regard to the Markovicity of collections of processes and subsets of processes. This chapter marks the beginning of a transition toward Pearl's school of thought in that we approach the problem from the perspective of Bayesian networks. While we utilize Bayesian networks strictly as a framework for identifying conditional independence relationships via the d-separation criterion (i.e. without utilizing any causal language), Bayesian networks are a foundational element of Pearl's graphical modeling approach to causality. Using this Bayesian network perspective, we demonstrate sufficient conditions for which collections of processes will satisfy the desired Markovicity assumption. We further demonstrate that these conditions are in fact necessary with the exception of a zero measure set of parameters defining the processes. Given this strictness of the identified condition, we propose alternative measures to the DI that can be estimated under a milder set of assumptions and can be used to bound the true DI from above and below, with both the upper and lower bounds approach the true DI as the model order grows to infinity. Using these augmented notions of DI, we run simulations in order to assess the extent to which estimating DI in the absence of the necessary assumptions results in biases. This chapter is composed largely of contributions initially presented in [107].

While Chapter 3 introduces augmented notions of DI in order to relax the assumptions necessary for estimating the DI, Chapter 4 introduces a measure aimed at increasing interpretability. Specifically, we begin by making the observation that Granger causality, directed information, transfer entropy, etc. are all *process level* measures in that they provide a single value for a

given joint distribution defining the processes in question. We propose that, even in a setting where two processes are jointly stationary, it is reasonable to expect that *certain values* of a process can give rise to larger causal influences than others. As such, we define a measure that is itself a random sequence whose expected sum is the DI. Through a series of examples, we illustrate how our sample path dependent measure of causal influence between time series can uncover specific patterns that give rise to large causal influence. We further leverage results from sequential prediction theory in order to develop estimators of the proposed measure and to prove finite sample bounds on the performance of these estimators. Through application to stock market data, we demonstrate that there is reason to believe that the behaviors exhibited in the examples are not limited to thought experiments. This chapter represents a step toward Pearl's perspective in the sense that it defines a measure of influence for every possible value of a cause, irrespective of the probability with which those values occurs. This property is reminiscent of the notion of intervention, which is based around forcing a cause to take a particular value as a means of bypassing the distribution the cause would normally obey. Nevertheless, Chapter 4 lacks a full treatment of the nature of interventions in the sample path dependent measure, and thus remains more closely aligned with Granger's perspective. The ideas in this chapter were originally presented in [107] and later refined in [106].

This dissertation concludes with Chapter 5, where we extend the perspective described in Chapter 4 to general directed acyclic graphs in the development of an information theoretic framework for measuring total, direct, and indirect causal effects. Whereas the previous chapter sought to define a version of DI that was dependent on a sample path, Chapter 5 uses a value dependent notion of mutual information as a starting point. Value dependent versions of mutual information have appeared in a variety of settings throughout the literature, including experimental design [71, 26] and neural stimulus response [29]. We specifically utilize the so-called specific mutual information $I(x;Y)$ as a foundation for measuring the influence of a specific x on a random variable Y. Given that this is, to our knowledge, the first application of specific mutual

information in the context of causality, we refer to the resulting measures as measures of *specific causal influence*. When moving beyond simple two-node DAGs, we define our measures of specific causal influence with respect to an intervention on the variable representing a cause. After defining multiple notions of specific causal influence, we present a collection of theoretical results relating the proposed measures to existing information theoretic measures such as causal strength and information flow, as well as a set of conditions under which the proposed measures can be estimated from observational data. We further provide three examples that illustrate notions of causal influence that are uniquely described by our specific causal measures. Finally, the chapter concludes with an in-depth case study that applies the proposed framework to a large climatological dataset. Chapter 5 marks the end of the transition from Granger to Pearl. Nevertheless, information theoretic measures of causal influence possess some very fundamental differences from their mainstream statistical counterparts. As such, a central goal of this final chapter is to elucidate the nature of these differences with the hopes of communicating the strengths and weaknesses of information theory in the context of causality.

Chapter 2

A Modularized Framework for Non-Markov Time Series Estimation

2.1 Introduction

We consider the problem of estimating a latent time series based on an underlying dynamic model and noisy measurements. Such a problem appears in a variety settings, including (but certainly not limited to) tracking [7], medical imaging [95], and video denoising [31]. Given the broad applicability of this problem formulation, the underlying models that are used inevitably become increasingly complex.

Certain scenarios are well studied, such as the case of a linear system with Gaussian noise, where it is well known that the maximum *a-posteriori* (MAP) point estimate can be obtained using a Kalman smoother (KS) [58]. When introducing non-linearities, alternatives include the extended Kalman filter (EKF), which relies on linear approximations, as well as the unscented Kalman filter (UKF) [123] and Particle Filter (PF) [27], which use sample based techniques. While the EKF and UKF are well suited for a broad class of problems, they are not well suited for models with non-Gaussian noise. This is problematic for the increasingly popular problem

of incorporating sparsity inducing models to latent signal estimation. These problems include exploiting sparsity in the underlying signal [121, 5, 129, 21] in addition to exploiting sparsity in the signal dynamics [11, 20, 6]. While some of these methods utilize ℓ_1-regularization to enforce sparsity at a local level and enable causal prediction, there is often knowledge of global structures, such as those favored by the group lasso [47], that dictate a need for batch-wise estimation. In such cases, the desired estimation problem deviates from the classical state estimation problem in that the underlying signal is no longer Markov. In such a scenario, there is no clear extension to the EKF, UKF, or PF that may be utilized to address the non-Markovicity of the underlying signal.

The broad scope of the problem in question dictates a need for a systematic approach to latent time series estimation for a variety of measurement models and system models. Furthermore, a solution framework that can compartmentalize these two models facilitates interchangeability and allows new regularization techniques to be easily incorporated to an estimation procedure.

We develop a framework using the alternating direction method of multipliers (ADMM) [16] that, under mild (i.e. log-concavity) assumptions, yields the MAP estimate for problems with non-Markov latent variables and/or nonlinear observations. While ADMM has been utilized to decompose specific dynamic systems into simpler subproblems [105, 6], our approach applies to arbitrary log-concave dynamic models. In particular, we utilize auxiliary variables to enable a solution involving iterative updates to three modules, one that pertains to the measurement model, another that pertains to the prior distribution on the latent signal, and a third that is a Kalman smoother. As such, our framework enables various sparsity models to be easily applied to the signal and/or dynamics with adjustments only required to the corresponding module. We demonstrate implementation of the framework in two distinct applications, namely latent state estimation and spectrotemporal estimation. We show that in the case of state estimation, our method outperforms a fixed interval smoother and particle filter for two state-space models coupled with non-Gaussian observations. In the case of spectrotemporal estimation, we demonstrate the

efficacy of our method when using non-Markov priors. The proposed method yields an intuitive approach to latent process estimation with iterative use of a Kalman smoother in tandem with standard convex optimization techniques. We provide a mathematical justification for the intuition by proving that our approach guarantees convergence to the MAP solution under the same relatively mild conditions that apply to general ADMM approaches. Finally, we provide software to enable the reader to reproduce the results of this chapter and to easily apply the framework to novel models. Our contributions may be summarized as follows:

- We present an efficient iterative solution framework for latent time series estimation with a guarantee of convergence to the MAP estimate under mild log-concavity assumptions.
- In the presence of non-Linear, non-Gaussian measurement models, our method does not require a Gaussian approximation, unlike KS variants, and is more efficient than Sequential Monte Carlo (SMC) methods.
- Our framework accommodates non-Markov signals despite there being no clear method for adapting EKF, UKF, and SMC methods for such a scenario, particularly when the prior applies to highly non-linear functions of the latent process, such as a singular value decomposition.
- Through the use of auxiliary variables, the ADMM solution to our reformulated MAP estimation problem is modular, with the observation and system models in disjoint modules that are unified by a Kalman smoother.

The chapter is structured as follows: Section 2.2 provides the general formulation of the problem we are solving in addition to a brief review of relevant work solving specific instances of the problem. Section 2.3 details a novel systematic approach for solving the MAP estimation problem in its general form. Section 2.4 demonstrates the capabilities of the framework through implementation on two existing problems. Section 2.5 concludes the chapter with a discussion of the results and future work.

2.2 Preliminaries

2.2.1 Notation

While it is intended that the notation is presented unambiguously, we here present some notational conventions. Bold letters are used to represent vectors and matrices, whereas non-bold letters represent scalars. Subscripts are used for indexing scalar elements of a vector, or columns of a matrix. A double subscript is used to specify scalar elements of a matrix. For example, x_n gives the nth element of a vector $\mathbf{x}$, $\mathbf{x}_n$ gives the nth column of a matrix $\mathbf{x}$, and $x_{n,m}$ gives the mth row of the nth column of a matrix $\mathbf{x}$. Capital/lowercase letter pairs represent either random variable/realization pairs or total count/index pairs. For example, we may have that $\mathbf{x}_n$ gives a specific value of the random vector $\mathbf{X}_n$, which is the nth column of a random matrix $\mathbf{X}$ with N columns in total. We let f and p denote probability density functions (pdfs) and probability mass functions (pmfs), respectively. Various joint and conditional pdfs and pmfs are made clear by their subscripts. For example, the pdf of X given $Y = y$ is $f_{X|Y}(\cdot|y)$. We let $\mathbb{R}$ denote the space of real numbers, $\mathbb{R}_+$ denote the non-negative reals, $\mathbb{R}^{A\times B}$ denote the space of A by B real valued matrices, and $\mathbb{R}^{AB}$ denote the space of real valued vectors of length A times B.

2.2.2 Problem Setup

Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces and N be the length of time series pertaining to the latent process $\mathbf{X} \in \mathcal{X}^N$ and observed process $\mathbf{Y} \in \mathcal{Y}^N$. Unless otherwise specified, we assume $\mathcal{X} = \mathbb{R}^K$ and $\mathcal{Y} = \mathbb{R}^P$ where K is the dimension of the latent process at any time, and P is the dimension of the observation process at any time. As such, $\mathbf{X} \in \mathbb{R}^{K\times N}$ is the latent time series we wish to estimate and $\mathbf{Y} \in \mathbb{R}^{P\times N}$ is the collection of noisy observations. Furthermore, assume that

these observations are conditionally independent given the underlying time series:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \prod_{n=1}^{N} f_{\mathbf{Y}_n|\mathbf{X}_n}(\mathbf{y}_n \mid \mathbf{x}_n) \tag{2.1}$$

where $f_{\mathbf{Y}|\mathbf{X}}$ is the likelihood of the entire collection of observations given the entire latent time series and $f_{\mathbf{Y}_n|\mathbf{X}_n}$ is the likelihood of a single observation given the corresponding element of the latent time series.

Next, define the latent signal's dynamics (or system behavior) in terms of $\mathbf{W} \in \mathbb{R}^{K\times N}$ for which:

$$\mathbf{W}_n = \begin{cases} \mathbf{X}_1 & \text{n=1} \\ \mathbf{X}_n - \mathbf{D}\mathbf{X}_{n-1} & \text{n = 2,...,N} \end{cases},$$

where $\mathbf{D} \in \mathbb{R}^{K\times K}$ is a transition matrix and $\mathbf{W}_n \in \mathbb{R}^K$ and $\mathbf{X}_n \in \mathbb{R}^K$ represent the nth columns of $\mathbf{W}$ and $\mathbf{X}$, respectively. For compactness we write this as $\mathbf{W} = \mathcal{A}(\mathbf{X})$, where $\mathcal{A}$ represents a linear operator that is fully defined by $\mathbf{D}$. We assume that $\mathbf{W}$ is distributed according to a known prior pdf $f_{\mathbf{W}}(\mathbf{w})$. Note that this framework includes, for the special case of $\mathbf{W}_n = \mathbf{X}_n - \mathbf{X}_{n-1}$ and $\mathbf{W}_n \sim \mathcal{N}(\mu_n, \Sigma_n)$ are independent Gaussian random vectors for $n = 2,\ldots,N$, the well-studied scenario in which the underlying time series $\mathbf{X}$ is a Gauss-Markov process.

Here, we consider the problem of finding the maximum a posteriori estimate:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \; -\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) - \log f_{\mathbf{X}}(\mathbf{x}) \tag{2.2}$$

where $-\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x})$ is the negative log-likelihood and $-\log f_{\mathbf{X}}(\mathbf{x})$ is the negative log-prior. We note that because $\mathbf{W}$ is a linear function of $\mathbf{X}$, we have $f_{\mathbf{X}}(\mathbf{x}) \propto f_{\mathbf{W}}(\mathcal{A}(\mathbf{x}))$. This relationship indicates that knowing a prior on either $\mathbf{X}$ or $\mathbf{W}$ induces a prior on the other. Thus, we can

equivalently rewrite our problem as:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \; -\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) - \log f_{\mathbf{W}}(\mathcal{A}(\mathbf{x})) \tag{2.3}$$

$$= \underset{\mathbf{x}}{\operatorname{argmin}} \, L(\mathbf{y} \mid \mathbf{x}) + \beta\phi(\mathcal{A}(\mathbf{x})) \tag{2.4}$$

with $\beta \in \mathbb{R}_+$ and where we define the measurement model $L : \mathbb{R}^{N\times K} \to \mathbb{R}$ and system model $\phi : \mathbb{R}^{N\times K} \to \mathbb{R}$ as:

$$L(\mathbf{y} \mid \mathbf{x}) \coloneqq -\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) \tag{2.5}$$

$$\phi(\mathbf{w}) \coloneqq -\frac{\log f_{\mathbf{W}}(\mathbf{w})}{\beta}. \tag{2.6}$$

The inclusion of β in (2.4) is to facilitate the cases when the system model is only known up to a proportionality constant or when ϕ is a regularizer used to exploit a desired dynamic characteristic of the latent signal (as opposed to representing the *true* distribution of $\mathbf{W}$). In either of these cases β is interpreted as a tuning parameter used to control the extent to which the system model is weighted (as in λ throughout [47]).

Throughout this chapter, we will interchangeably use the names log-likelihood and measurement model in reference to L, and log-prior, system model, and dynamic model in reference to ϕ. Due to the assumption that observations are conditionally independent given the state variables, the measurement model can be decomposed into a sum over N measurements, each depending on the state variable at a single time instance:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \left(\sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) \right) + \beta\phi(\mathcal{A}(\mathbf{x})) \tag{2.7}$$

where $L_n(\mathbf{y}_n \mid \mathbf{x}_n) \coloneqq -\log f_{\mathbf{Y}_n|\mathbf{X}_n}(\mathbf{y}_n \mid \mathbf{x}_n)$. It should be noted that the problem presented in (2.7) is made difficult by the second term. In particular, imposing a prior on the *differences* of the underlying time series prevents separability across the N time points. Furthermore, by allowing

for non-Markov models, it is possible to have models that do not allow the second term to be separated into terms each containing only $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$ for each $n = 1 \ldots N$. In the following section, we present a framework for efficiently solving problems in the form of (2.7) for a broad class of measurement models L and system models ϕ.

2.2.3 Related Work

Works related to our proposed method include both the investigation of new algorithms for estimating latent time series and the creation/application of new time series models. Notably, the Kalman smoother [58] and its variants [123, 27] provide structured approaches to estimating latent signals in a subset of problems with dynamical system models and noisy measurements. While the Kalman smoother is MAP optimal for the very specific case of a linear system with Gaussian noise, its non-linear variants do not guarantee optimality and do not offer solutions for a comprehensive class of measurement and system models. In particular, there has been growing interest in models exploiting the sparsity of states and/or dynamics of signals [121, 5, 129, 11, 20, 6], which in many cases do not lend themselves to solutions via the existing Kalman smoother variants.

Table 2.1: Examples of common models. For the multiple modalities case, we define $\mathbf{y} = (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)})$ to be a J-tuple of simultaneous and conditionally independent observations, each with its own dimensionality and associated measurement model $L^{(j)}$.

Measurement Models - $L(\mathbf{y} \mid \mathbf{x})$	
Linear Gaussian (LG)	$\sum_{n=1}^{N} \|\|\mathbf{y}_n - \mathbf{A}\mathbf{x}_n + \mathbf{b}\|\|_2^2 \triangleq LG(\mathbf{x})$
Sparse LG	$LG(\mathbf{x}) + \|\|\mathbf{x}\|\|_1$
Group sparse LG	$LG(\mathbf{x}) + \sum_{k=1}^{K} \left(\sum_{n=1}^{N} x_{k,n}^2 \right)^{\frac{1}{2}}$
Multiple Modalities	$\sum_{j=1}^{J} L^{(j)}(\mathbf{y}^{(j)} \mid \mathbf{x})$
System Models - $\phi(\mathbf{w})$	
LG	$\sum_{n=1}^{N} \|\|\mathbf{C}\mathbf{w}_n - \mathbf{d}\|\|_2^2$
Sparse	$\|\|\mathbf{w}\|\|_1$
Group sparse	$\sum_{k=1}^{K} \left(\sum_{n=1}^{N} w_{k,n}^2 \right)^{\frac{1}{2}}$

For such sparsity-inducing models, existing causal estimators are often heuristic extensions of the Kalman filter, such as ℓ_1-regularized Kalman filter updates [20] and tracking a belief of the support set [121]. Causal estimation is made particularly challenging for the models that are non-Markov in nature. As such, the aforementioned causal estimators lack performance guarantees. Existing batchwise solutions utilize a Kalman smoother to solve the updates for a particular iterative algorithm, such as IRLS for group sparse dynamics [11] and ADMM for group sparse states [6]. In the latter example, their non-consensus formulation of ADMM is reliant upon the choice of a Gaussian system model.

In addition to the Kalman smoother variants, sample based methods such as Markov chain Monte Carlo (MCMC) and SMC are viable options for latent time series estimation. While these methods can accommodate non-linear and non-Gaussian models [41] and can simultaneously estimate the state and model parameters [19, 4], they are often computationally prohibitive. Furthermore, these methods do not have a straightforward extension to non-Markov and non-linear priors such as the ℓ_1/ℓ_2 and nuclear norm priors (see Remark 2).

Here we propose a generalized framework for obtaining the MAP estimate in many of the aforementioned problems in a batchwise manner. Tables 2.1 and 2.2 show the models used in some of these problems and serve to illustrate the primary contribution of our framework, namely that for a given problem, the solution is modular in that the choice of measurement model can be made independently of the system model without requiring a complete rederivation of the solution.

2.3 Modular MAP Estimation Framework

The alternating direction method of multipliers (ADMM) allows large global problems to be decomposed into smaller subproblems whose solutions can be coordinated to achieve the global solution. ADMM offers an iterative solution of the dual problem that has the decomposability of

Table 2.2: Examples of measurement model/system model pairings in previous works.

	Measurement Model	**System Model**
Kalman Smoother [58]	LG	LG
State Space Model of Learning [24]	Non-linear/ multiple modalities	Gaussian
Spectrotemporal Pursuit [11]	LG	Group sparse
Lasso-Kalman Smoother [6]	Group sparse	LG
Sparse States and Sparse Innovations [20]	Sparse	Sparse

dual descent in addition to the convergence guarantees of the method of multipliers, which hold under fairly mild conditions. While the details of dual optimization and ADMM are omitted here, they can be found in [15] and [16], respectively.

We begin by reformulating (2.7) to create separability in the objective function by including $\mathbf{w}$ as an optimization variable and introducing a constraint to preserve the relationship between $\mathbf{x}$ and $\mathbf{w}$:

$$\begin{aligned} (\hat{\mathbf{x}}, \hat{\mathbf{w}}) = \underset{\mathbf{x},\mathbf{w}}{\operatorname{argmin}} \quad & \sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \beta\phi(\mathbf{w}) \\ s.t. \quad & \mathbf{w} = \mathcal{A}(\mathbf{x}). \end{aligned} \tag{2.8}$$

The optimization problem given by (2.8) can be solved using ADMM, and would yield a solution that enables the measurement model and penalty function to be addressed in independent subproblems. However, when using the above formulation, the update equations yielded by the ADMM algorithm would require one of the aforementioned approximate or sample-based methods for non-Gaussian measurement models (see Remark 1).

We use a variant of ADMM known as consensus ADMM and construct a modular solution framework shown in Fig. 2.1 that only requires making local adjustments to the solution when modifying the measurement model (L), penalty function (ϕ), or transition model ($\mathcal{A}$). This is accomplished by introducing an auxiliary variable $\mathbf{z} \in \mathbb{R}^{K \times N}$ to achieve separability (of $\mathbf{x}$ and $\mathbf{w}$)

in the constraints as well as the objective function:

$$\begin{aligned}(\hat{\mathbf{x}},\hat{\mathbf{w}},\hat{\mathbf{z}}) = \underset{\mathbf{x},\mathbf{w},\mathbf{z}}{\operatorname{argmin}} \quad & \sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \beta\phi(\mathbf{w}) \\ s.t. \quad & \mathbf{x} = \mathbf{z} \\ & \mathbf{w} = \mathcal{A}(\mathbf{z}).\end{aligned} \tag{2.9}$$

The optimization problem given by (2.9) is termed the consensus formulation, and $\mathbf{z}$ the consensus variable. By introducing this variable, our iterative updates with respect to the measurement model and penalty function are not only independent of each other, but are also independent of the transition model determined by $\mathcal{A}$.

The first step in solving (2.9) using ADMM requires generating the augmented Lagrangian:

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{x},\mathbf{w},\mathbf{z},\lambda,\alpha) = & \sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \beta\phi(\mathbf{w}) \\ & + \langle\lambda, \mathbf{x}-\mathbf{z}\rangle + \langle\alpha, \mathbf{w}-\mathcal{A}(\mathbf{z})\rangle + \frac{\rho}{2}||\mathbf{x}-\mathbf{z}||_F^2 + \frac{\rho}{2}||\mathbf{w}-\mathcal{A}(\mathbf{z})||_F^2\end{aligned} \tag{2.10}$$

where $\lambda \in \mathbb{R}^{K\times N}$ and $\alpha \in \mathbb{R}^{K\times N}$ are Lagrange multipliers, $\langle\cdot,\cdot\rangle$ is the Frobenius inner product, $||\cdot||_F$ is the matrix Frobenius norm, and $\rho \in \mathbb{R}_+$ is the penalty parameter for the augmented Lagrangian. Note that in the case where $\rho = 0$, the augmented Lagrangian is equivalent to the standard (unaugmented) Lagrangian.

Given the augmented Lagrangian, the ADMM solution is obtained by iteratively alternating between minimization with respect to the primal variables ($\mathbf{x}$, $\mathbf{w}$ and $\mathbf{z}$) and performing gradient ascent on the Lagrange multipliers. These iterations represent a trade off between finding a solution that minimizes the cost function in (2.9) while ensuring that the Lagrange multipliers are such that the dual function of (2.9) is increasing in i and thus ensuring the constraints are satisfied. Letting $\mathbf{x}^{(i)}$ represent the estimate of $\mathbf{x}$ after i iterations (similarly for $\mathbf{w}^{(i)}$, $\mathbf{z}^{(i)}$, $\lambda^{(i)}$, and

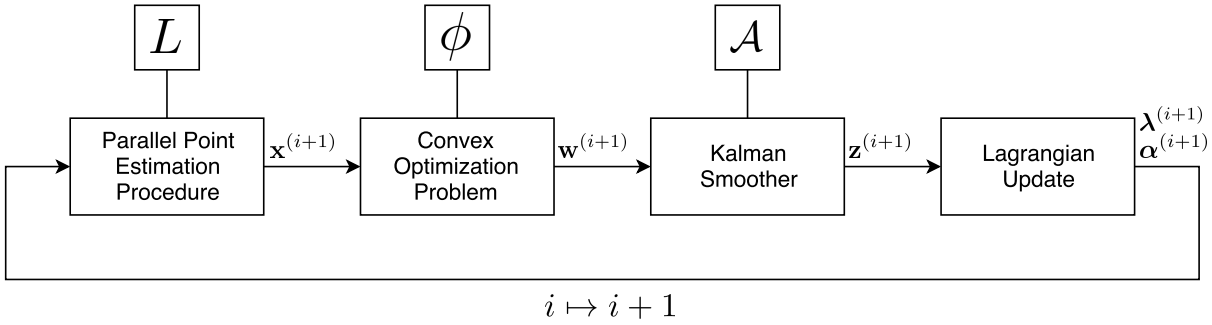

Figure 2.1: Block diagram of the modular MAP estimation framework illustrates how the selection of L, ϕ, and $\mathcal{A}$ affects independent parts of the estimation procedure.

$\alpha^{(i)}$), each iteration of ADMM is composed of the following updates [16, Sec. 3.1]:

$$\begin{aligned}
\mathbf{x}^{(i+1)} &= \operatorname*{argmin}_{\mathbf{x}} \mathcal{L}_\rho(\mathbf{x}, \mathbf{w}^{(i)}, \mathbf{z}^{(i)}, \lambda^{(i)}, \alpha^{(i)}) \\
\mathbf{w}^{(i+1)} &= \operatorname*{argmin}_{\mathbf{w}} \mathcal{L}_\rho(\mathbf{x}^{(i+1)}, \mathbf{w}, \mathbf{z}^{(i)}, \lambda^{(i)}, \alpha^{(i)}) \\
\mathbf{z}^{(i+1)} &= \operatorname*{argmin}_{\mathbf{z}} \mathcal{L}_\rho(\mathbf{x}^{(i+1)}, \mathbf{w}^{(i+1)}, \mathbf{z}, \lambda^{(i)}, \alpha^{(i)}) \\
\lambda^{(i+1)} &= \lambda^{(i)} + \rho(\mathbf{x}^{(i+1)} - \mathbf{z}^{(i+1)}) \\
\alpha^{(i+1)} &= \alpha^{(i)} + \rho(\mathbf{w}^{(i+1)} - \mathcal{A}(\mathbf{z}^{(i+1)})).
\end{aligned} \tag{2.11}$$

By fixing all but one variable in each update, the objective functions can be simplified by dropping the terms in (2.10) that do not contain the optimization variable for the corresponding update. As a result, when updating with respect to the measurement model L and the system model ϕ, we only need to consider the model corresponding to that update and an ℓ_2-norm proximal operator [89] that ensures the update is moving in the appropriate direction to achieve a global consensus. This inclusion of the proximal operators in the augmented Lagrangian enables the use of ADMM with non-smooth objective functions [89, Sec 4.4]. Then, updating of the consensus variable involves "centering" it such that it gives equal representation to our current estimates based on the measurements and our estimates based on the system dynamics. In this sense, our ADMM framework yields a mathematical justification for a very intuitive approach,

namely, iteratively finding the best estimate based on measurements, finding the best estimate based on dynamics, and "averaging" the two in the appropriate sense. This viewpoint will be made clearer in the following sections where we detail the specific update equations.

2.3.1 Measurement Model Update

When updating with respect to the measurement model, only terms containing $\mathbf{x}$ in the augmented Lagrangian must be considered. To simplify notation, we will consider the scaled form of the update equations [16, Sec. 3.1.1], which can be obtained by combining the appropriate linear and quadratic terms in (2.10) by completing the square:

$$\mathbf{x}^{(i+1)} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \frac{\rho}{2}||\mathbf{x} - \tilde{\mathbf{x}}^{(i)}||_F^2 \tag{2.12}$$

where $\tilde{\mathbf{x}}^{(i)} := \mathbf{z}^{(i)} - \lambda^{(i)}/\rho$ is fixed within the scope of this update. Details for deriving the scaled form of the update can be found in Appendix A.1. Given that the squared Frobenius norm can be decomposed to the sum of squared ℓ_2 norms, we note that the measurement model update is separable over n, meaning that we can solve for $\mathbf{x}_n^{(i+1)}$ for each $n = 1, \ldots, N$ independently:

$$\mathbf{x}_n^{(i+1)} = \underset{\mathbf{x}_n}{\operatorname{argmin}} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \frac{\rho}{2}||\mathbf{x}_n - \tilde{\mathbf{x}}_n^{(i)}||_2^2 \tag{2.13}$$

where $\tilde{\mathbf{x}}_n^{(i)} := \mathbf{z}_n^{(i)} - \lambda_n^{(i)}/\rho$.

Remark 1. *Note that the ability to separate each of the N updates is a result of the inclusion of the consensus variable. Excluding this variable would require that the dynamics be considered in the update of the measurement model:*

$$\mathbf{x}^{(i+1)} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \frac{\rho}{2}||\mathbf{x}_n - \mathbf{D}\mathbf{x}_{n-1} - \tilde{\mathbf{x}}_n'^{(i)}||_2^2$$

where $\tilde{\mathbf{x}}_n'^{(i)} := \mathbf{w}_n^{(i)} - \gamma^{(i)}/\rho$, $\mathbf{x}_0 := 0$, *and* γ *represents the single Lagrange multiplier that would be*

required in solving (2.8) *using ADMM. Requiring that the dynamics of the underlying time series be included in the measurement model update prohibits solving for* $\mathbf{x}_n^{(i)}$ *independently across N. Thus, using ADMM in this fashion does not offer any simplifications over traditional approaches for non-Gaussian measurement models. As such, incorporation of the consensus variable not only enables faster processing by allowing each update to be parallelized across N, but it allows the framework to be applied in a straightforward, non-approximate manner to a broad class of measurement models.*

It should be noted that while we assume conditional independence of the observations given the latent time series, one can revert to the update in (2.12) for the case when the observations are correlated. In this case the ability to parallelize across n is lost, but the ability to ignore system dynamics is preserved (i.e. the optimization problem in (2.12) still does not depend on ϕ).

2.3.2 System Model Update

In the system model update, only terms in (2.10) that contain $\mathbf{w}$ must be included. Again, we consider the scaled form:

$$\mathbf{w}^{(i+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\ \beta\phi(\mathbf{w}) + \frac{\rho}{2}||\mathbf{w} - \tilde{\mathbf{w}}^{(i)}||_F^2 \tag{2.14}$$

where $\tilde{\mathbf{w}}^{(i)} := \mathcal{A}(\mathbf{z}^{(i)}) - \alpha^{(i)}/\rho$. In this form we can clearly interpret the system model update as finding a new collection of latent variable transitions $\mathbf{w}^{(i+1)}$ that is both representative of our system model ϕ and proximal to the appropriately scaled current consensus on the transitions $\tilde{\mathbf{w}}^{(i)}$.

The key observation is that this framework imposes no restrictions as to whether or not our underlying signal is Markov. In the case where the signal is indeed Markov, then $\mathbf{w}_n^{(i+1)}$ would be updated independently over n, but in general we do not assume this is the case. This provides the ability to impose batch-level structures on the dynamics of the signal. Furthermore, we note

that the nature of the proximal operator enables closed form solutions when ϕ is chosen to be a number of common sparsity inducing priors. In particular, because the proximal operator is not multiplying $\mathbf{w}$ by a non-orthonormal matrix, the ℓ_1, group sparse, and nuclear norm priors all offer soft-thresholding solutions [118]. Furthermore, we note that for a fixed K, the complexity of the soft-thresholding solutions for the ℓ_1 and group sparse priors scale linearly with N per iteration. The nuclear norm prior, however, requires a singular value decomposition (SVD), and thus scales quadratically with N per iteration [42]. Similarly, for a fixed N, the same scaling factors apply to K. It should be noted however, that if increasing N and K, the complexity of the SVD will scale quadratically with $\max\{K,N\}$ and cubically with $\min\{K,N\}$.

2.3.3 Consensus Update

Updating the consensus variable depends on neither the measurement model nor the system model. We can think of this step as averaging our current estimates of our signal based on measurements $\mathbf{x}^{(i+1)}$ and based on dynamics $\mathbf{w}^{(i+1)}$:

$$\mathbf{z}^{(i+1)} = \underset{\mathbf{z}}{\operatorname{argmin}} \, ||\mathbf{z} - \tilde{\mathbf{z}}_{\mathbf{x}}^{(i)}||_F^2 + ||\mathcal{A}(\mathbf{z}) - \tilde{\mathbf{z}}_{\mathbf{w}}^{(i)}||_F^2 \tag{2.15}$$

where $\tilde{\mathbf{z}}_{\mathbf{x}}^{(i)} := \mathbf{x}^{(i+1)} + \lambda^{(i)}/\rho$ and $\tilde{\mathbf{z}}_{\mathbf{w}}^{(i)} = \mathbf{w}^{(i+1)} + \alpha^{(i)}/\rho$. Note that given the nature of the linear operator $\mathcal{A}$, (2.15) can always be solved efficiently using a Kalman smoother.

This step clarifies the notion of "averaging" the current estimates $\mathbf{x}^{(i+1)}$ and $\mathbf{w}^{(i+1)}$. By framing our problem from a consensus ADMM perspective, we can carve out various elements of the model and delegate them to independent updates. Then, given the nature of the relationship between the signal $\mathbf{x}$ and the dynamics $\mathbf{w}$, establishing consensus between the two estimates is a Kalman smoothing problem *regardless* of the measurement and system models. This is a result of the use of ℓ_2-norms in the augmented Lagrangian, which can be thought of as representing Gaussian noise with identity covariance. In other words, at each iteration i, the consensus update

is a Kalman smoothing problem where each of our measurements are given by $\tilde{\mathbf{z}}_{\mathbf{x}}^{(i)}$ and each of our predictions are given by $\tilde{\mathbf{z}}_{\mathbf{w}}^{(i)}$. In this sense, the consensus update gives equal weight to the current iterates of our measurement and system estimates. This follows from the fact that the log-likelihood and log-prior have their own uncertainty terms that dictate how far the updates $\mathbf{x}^{(i+1)}$ and $\mathbf{w}^{(i+1)}$ can deviate from the consensus in their respective updates, namely measurement noise and the tuning parameter β. We note that because both terms in (2.15) can be thought of as representing Gaussian noise with identity covariance and the transition model $\mathcal{A}$ is invariant over iterations i, all matrix inversions required by the Kalman smoother can be precomputed. As a result, each iteration requires on the order of N matrix multiplications.

2.3.4 Convergence

Next we consider the practical and theoretical convergence of the proposed framework. To begin, we present the optimality conditions and the means with which we can in practice implement convergence checks. The derivations are omitted, as they closely follow Section 3.3 of [16]. The optimality conditions for the proposed framework are given by:

$$\left.\begin{aligned} 0 &= \hat{\mathbf{x}} - \hat{\mathbf{z}} \\ 0 &= \hat{\mathbf{w}} - \mathcal{A}(\hat{\mathbf{z}}) \end{aligned}\right\} \textit{Primal Feasibility}$$
$$\left.\begin{aligned} 0 &\in \tfrac{\partial}{\partial \hat{\mathbf{x}}} L(\mathbf{y} \mid \hat{\mathbf{x}}) + \hat{\lambda} \\ 0 &\in \tfrac{\partial}{\partial \hat{\mathbf{w}}} \beta\phi(\hat{\mathbf{w}}) + \hat{\alpha} \\ 0 &= \hat{\lambda} + \mathcal{A}(\hat{\alpha}) \end{aligned}\right\} \textit{Dual Feasibility} \tag{2.16}$$

where $\partial/\partial\cdot$ is the subgradient operator (or gradient when defined, in which case $\in$ becomes an equality). The primal feasibility conditions ensure that our $\hat{\mathbf{z}}$ preserves the desired relationship between $\hat{\mathbf{x}}$ and $\hat{\mathbf{w}}$, and the dual feasibility conditions serve the purpose of ensuring that the optimal Lagrange multipliers are such that $\hat{\mathbf{x}}$ and $\hat{\mathbf{w}}$ jointly minimize L and ϕ.

Using these optimality conditions, we can derive the primal and dual residuals:

$$\left.\begin{aligned} r_1^{(i)} &= \mathbf{x}^{(i)} - \mathbf{z}^{(i)} \\ r_2^{(i)} &= \mathbf{w}^{(i)} - \mathcal{A}(\mathbf{z}^{(i)}) \end{aligned}\right\} \textit{Primal Residuals}$$
$$\left.\begin{aligned} s_1^{(i)} &= \rho\mathcal{A}(\mathbf{w}^{(i)} - \mathbf{w}^{(i-1)}) \\ s_2^{(i)} &= \rho(\mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}) \end{aligned}\right\} \textit{Dual Residuals} \tag{2.17}$$

where primal feasibility is achieved when $r_j^{(i)} = 0$ and dual feasibility is achieved when $s_j^{(i)} = 0$ for all $j \in \{1,2\}$. In practice, we declare the algorithm converged when $||r_j^{(i)}||_F \leq \varepsilon_j^{pri}$ and $||s_j^{(i)}||_F \leq \varepsilon_j^{dual}$ for all $j \in \{1,2\}$, with the thresholds given by:

$$\begin{aligned} \varepsilon_1^{pri} &= \varepsilon^{rel} \max\{||\mathbf{x}^{(i)}||_F, ||\mathbf{z}^{(i)}||_F\} + \varepsilon^{abs}\sqrt{KN} \\ \varepsilon_2^{pri} &= \varepsilon^{rel} \max\{||\mathbf{w}^{(i)}||_F, ||\mathcal{A}(\mathbf{z}^{(i)})||_F\} + \varepsilon^{abs}\sqrt{KN} \\ \varepsilon_1^{dual} &= \varepsilon^{rel}||\lambda^{(i)}||_F + \varepsilon^{abs}\sqrt{KN} \\ \varepsilon_2^{dual} &= \varepsilon^{rel}||\alpha^{(i)}||_F + \varepsilon^{abs}\sqrt{KN} \end{aligned} \tag{2.18}$$

where ε^{rel} (relative tolerance) and ε^{abs} (absolute tolerance) are small positive parameters.

In general, ADMM does not guarantee convergence for more than two optimization variables [22]. As such, it is not immediately clear that our ADMM framework would guarantee convergence given that it optimizes over $\mathbf{x}$, $\mathbf{w}$, and $\mathbf{z}$. As it turns out, for the particular version of consensus ADMM that we are proposing, we can guarantee convergence under the same mild conditions required in standard ADMM.

Theorem 1. *Given an observation* $\mathbf{y}$*, when* $L(\mathbf{y} \mid \cdot)$ *and* $\phi(\cdot)$ *are closed, proper, and convex functions, the ADMM algorithm given by* (2.10) *and* (2.11) *converges to the solution of* (2.9)*, i.e.* $(\mathbf{x}^{(i)}, \mathbf{w}^{(i)}, \mathbf{z}^{(i)}) \to (\hat{\mathbf{x}}, \hat{\mathbf{w}}, \hat{\mathbf{z}})$ *as* $i \to \infty$*.*

The proof of Theorem 1 is based on a consensus ADMM formulation presented in section 5 of [32] and is given in detail in Appendix A.2.

2.4 Applications

2.4.1 State-Space Model of Learning

We begin by demonstrating how the ADMM framework can be applied to a problem with a highly non-linear multimodal measurement model. In the state-space model of learning [113], the system model is a traditional state-space Gauss-Markov process, where the state represents an unobservable cognitive state that represents a subject's ability to perform a task over time. The corresponding measurement model provides a statistical relationship between the underlying state and the observed task performance for a given trial.

We define $\mathbf{X} \in \mathbb{R}^{1\times N}$ to be the cognitive state (with $K = 1$), where N represents the number of trials conducted. The system model is given by:

$$X_n = \kappa X_{n-1} + \gamma + V_n \tag{2.19}$$

where $\kappa \in [0,1]$ is a forgetting factor, $\gamma \in \mathbb{R}_+$ is a positive bias that represents a tendency for the cognitive state to increase with time, and $V_n \sim \mathcal{N}(0, \sigma_V^2)$ is noise in the system model.

Using the state-space model of learning pertaining with multiple behavioral and neurophysiological measures, we assume that each of the N trials has an associated binary success/failure outcome, a reaction time, and neural spiking behavior. As such, each observation is given by a triplet $\mathbf{Y}_n = (B_n, R_n, \mathbf{S}_n) \in \{0,1\} \times \mathbb{R} \times \{0,1\}^J$, where B_n is a binary random variable indicating whether or not the trial was completed successfully, R_n is the log of the subject's reaction time to complete the task, and $\mathbf{S}_n$ is a length J point process that indicates whether or not there was neural spiking activity in each discrete Δt time window.

Each of the three observation modalities is associated with an appropriate statistical model.

First, the binary success/failure outcomes obey a Bernoulli probability model:

$$\mathbb{P}(B_n = b_n \mid X_n = x_n) = p_n^{b_n}(1-p_n)^{1-b_n} \tag{2.20}$$

where p_n is given by a logistic function that maps the cognitive state between 0 and 1:

$$p_n = \frac{\exp(\nu + \eta x_n)}{1 + \exp(\nu + \eta x_n)} \tag{2.21}$$

where $\nu, \eta \in \mathbb{R}$ are model parameters.

Next, the reaction time obeys a log-normal probability model, with:

$$R_n \sim \mathcal{N}(\psi + \omega X_n, \sigma_R^2) \tag{2.22}$$

where $\psi \in \mathbb{R}$ is the estimated initial log reaction time, $\omega \in \mathbb{R}_-$ is negative to ensure that the reaction time tends to decrease with an increasing cognitive state and σ_R^2 represents the level of stochasticity in the relationship between the cognitive state and reaction time.

Lastly, the neural spiking activity is modeled as a point process (as in equation 2.6 of [23]), with the negative log-probability of a given set of spikes given by:

$$-\log \mathbb{P}(\mathbf{S}_n = \mathbf{s}_n \mid X_n = x_n) = \sum_{j=1}^{J} -\log(\Lambda_{n,j}) s_{n,j} + \Lambda_{n,j}\Delta t \tag{2.23}$$

where $s_{n,j} \in \{0,1\}$ is the j^{th} bit of $\mathbf{s}_n$ and $\log \Lambda$ is the conditional intensity function, given by a generalized linear model [120]:

$$\log \Lambda_{n,j} = \xi + a x_n + \sum_{m=1}^{M} c_m s_{n,j-m} \tag{2.24}$$

where $\xi \in \mathbb{R}$ gives a base intensity level, $a \in \mathbb{R}$ determines the effect of the cognitive state on the spiking intensity, and $\mathbf{c} = (c_1, \ldots, c_M) \in \mathbb{R}^M$ accounts for the refractory period in neural spiking,

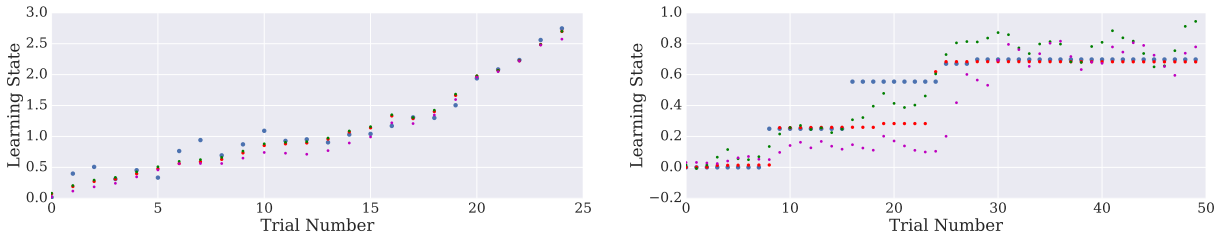

Figure 2.2: Sample realization (blue) for Gaussian state-space model (left) and sparse-variation state-space model (right), along with the estimates using ADMM (red), FIS (green), and SMC (purple). While the Gaussian states are well estimated by all three methods, the ADMM approach utilizing the ℓ_1 prior yields the only estimate that captures the piecewise constant nature of the sparse-variation states.

i.e. the fact that it is unlikely to see spiking activity in neighboring bins. The point process model given by (2.23) represents a discrete approximation of the negative log-likelihood for an inhomogeneous Poisson process where the rate in trial n and time j is $\Lambda_{n,j}$.

Next we adapt the state-space model of learning to the ADMM framework. We begin by considering the negative log-likelihood of the observations given the underlying cognitive state. We note that not only are the observations temporally conditionally independent given a sequence of cognitive states, but each of the three observations within a trial is conditionally independent given the cognitive state corresponding with that trial:

$$L(\mathbf{y} \mid \mathbf{x}) = \sum_{n=1}^{N} L_n(\mathbf{y}_n \mid x_n) = \sum_{n=1}^{N} L_{B_n}(b_n \mid x_n) + L_{R_n}(r_n \mid x_n) + L_{\mathbf{S}_n}(\mathbf{s}_n \mid x_n) \tag{2.25}$$

where the negative log-likelihoods $L_{B_n} := -\log p_{B_n|X_n}$, $L_{R_n} := -\log f_{R_n|X_n}$, and $L_{\mathbf{S}_n} := -\log p_{\mathbf{S}_n|X_n}$ are defined to be the negative log of the appropriate pdf/pmf corresponding with the respective observations. It is important to note that L is indeed convex. Considering this is not immediately obvious, it is shown in Appendix A.4.

Next we consider the system model. By defining $W_n = X_n - \kappa X_{n-1} = \gamma + V_n$ with $W_0 = X_0$, we get that $W_n \sim \mathcal{N}(\gamma, \sigma_V^2)$, i.e. each W_n is distributed iid Gaussian. Thus, our negative log-prior

is given by:

$$\begin{aligned}\phi(\mathbf{w}) &= -\log \prod_{n=1}^{N} \mathcal{N}(w_n; \gamma, \sigma_V^2) \\ &\propto \sum_{n=1}^{N} \frac{(w_n - \gamma)^2}{2\sigma_V^2}\end{aligned} \tag{2.26}$$

where $\mathcal{N}(x; \mu, \sigma^2)$ gives the value of a normal distribution with mean μ and variance σ^2 evaluated at x. Additionally, under this definition of $\mathbf{W}$ we get that the transition matrix $\mathbf{D}$ is in fact just a scalar, namely $\kappa \in \mathbb{R}$.

Plugging L, ϕ, and $\mathcal{A}$ into equations (2.12), (2.14), and (2.15), we obtain the update equations for solving the state-space model of learning problem. Beginning with the measurement model update, as a result of its separability across trials, each update decomposes into N univariate convex minimization problems. As such, these N problems can be solved in parallel using a convex solver such as CVX [25]. For the system model update, we note that because (2.14) is separable over $n = 1, \ldots, N$, the update is reduced to n quadratic minimizations that can be solved in closed form. Given that the density for $\mathbf{W}$ is assumed to be fully known, we set the tuning parameter $\beta = 1$. The details of these updates can be found in Appendix A.3.

We demonstrate the state-space model of learning solution on simulated data with $N = 25$, using parameters from section V-A of [24]. The proposed method is compared with the fixed-interval smoother (FIS) detailed in [24] and a sequential Monte Carlo (SMC) method. In particular, we develop a particle smoother using the forward-filtering backward-sampling technique with systematic resampling at each step [30]. For the ADMM method, we set $\rho = 30$ and limit the procedure to 25 iterations, i.e. $\hat{\mathbf{x}} := \mathbf{x}^{(25)}$. For the SMC method, we use 100 particles. In Table 2.3 we look at the average root-mean-square error (RMSE) and average runtime for each method over 50 trials, where for a given realization $\mathbf{x}$ and a given estimate $\hat{\mathbf{x}}$, $\text{RMSE}(\hat{\mathbf{x}}) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 / \sqrt{N}$. We note that the proposed method is both most efficient and most accurate in the RMSE sense. While the SMC method would presumably benefit from a larger number of particles, we see that even with limited samples, it is very computationally intensive. While the difference in RMSE

is negligible across all 3 methods, it is worth noting that each method obtains a fundamentally different estimate. To be specific, the proposed method gives the MAP estimate in the limit of large iterations, while the other methods yield conditional expectations of the states given the entire observation sequence. In the case of the FIS, the estimate is the conditional expectation under a Gaussian approximation of the posterior. The SMC method, on the other hand, yields the true conditional expectation in the limit of large particle count.

It should be noted that in the case of a Gaussian state space, the problem formulations given by (2.8) and (2.9) are nearly equivalent. In particular, it is possible to omit the consensus variable and modify the constraint such that $\mathbf{W} = \mathbf{X}$. In such a scenario, the measurement model update would remain the same and the system model update would be solvable with a Kalman smoother. Thus, we further demonstrate the utility of our method by considering a second state-space model with sparse variations where such an approach is not possible. We simulate a state-space model with sparse variations by defining $X_n = X_{n-1} + V_n$ with V_n obeying a commonly used sparsity inducing mixture model [128]:

$$V_n = \begin{cases} 0 & \text{w.p. } p \\ \sigma U_n & \text{w.p. } 1-p \end{cases} \tag{2.27}$$

where $p \in [0,1]$ is a probability, $\sigma \in \mathbb{R}_+$ is a positive constant, and we define $U_n \sim \chi_2^2$ as i.i.d. Chi-Squared random variables with two degrees of freedom. This model represents a scenario supported by neurophysiological findings [69, 12] wherein infrequent, discontinuous changes in neural activity arise.

We again conduct 50 trials, setting $N = 50$, $p = 0.9$, and $\sigma = 0.1$, and estimate the state using ADMM, FIS, and SMC approaches. For the ADMM approach, we note that the true system model is no longer log-concave, so we instead use a sparsity inducing ℓ_1 regularizer, i.e. we define $\phi(\mathbf{w}) = \beta ||\mathbf{w}||_1$. As such, we set $\beta = 15$, noting that is no longer determined by the model

and must be treated as a tuning parameter. The resulting system model update is given by:

$$\mathbf{w}^{(i+1)} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{\rho}{2} \left|\left| \tilde{\mathbf{w}}^{(i)} - \mathbf{w} \right|\right|_2^2 + \beta \left|\left|\mathbf{w}\right|\right|_1 .$$

This problem is known as the LASSO problem and may be efficiently solved by applying a soft threshold operation to $\mathbf{w}^{(i)}$ at each iteration [117].

Given the model mismatch, we observe that the proposed method takes longer to converge on a desirable estimate, and thus increase the maximum number of iterations to 75. For the FIS, given that there is no systematic approach to obtain an estimate with sparse variations, we again utilize a Gaussian approximation, with the noise at each step being modeled by a Gaussian distribution with zero-mean and variance $\text{Var}(U_n) = 4\sigma^2$. The SMC method is given the benefit of using the true underlying system model when generating samples on the forward pass. However, when performing the backward pass on sample x_n^i with respect to a fixed $\hat{x}_{n+1}$, we get that when $x_n^i > \hat{x}_{n+1}$, the likelihood $f_{X_{n+1},\mathbf{Y}|X_n}(\hat{x}_{n+1}, \mathbf{y} \mid x_n^i) = 0$, causing the smoother to continually lower $\hat{x}_n$ for $n = N, N-1, \ldots, 1$ until the smoother fails (i.e. $x_k^i > \hat{x}_{k+1}$ for all i for some $k \in \{1, \ldots, N\}$). As such, we only utilize the forward pass particle filter. Referring to Table 2.3 for results, we note that the proposed method again outperforms the other methods in the RMSE sense. From a computational perspective, the 3X increase in iterations causes the ADMM approach to take slightly longer than the FIS, though both remain significantly more efficient than the SMC method.

Table 2.3: Performance metrics for the proposed method (ADMM), fixed-interval smoother (FIS), and sequential Monte Carlo (SMC) averaged over 50 trials with the Gaussian state-space model given by (2.19) and the state-space model with sparse variations given by (2.27).

	Gaussian State		**Sparse Variations**	
	RMSE	Run Time (s)	RMSE	Run Time (s)
ADMM	**0.165**	**1.8**	**0.141**	7.0
FIS	0.168	2.6	0.181	**5.2**
SMC	0.188	53.5	0.186	105.7

2.4.2 Spectrotemporal Pursuit

Next we demonstrate application of the ADMM framework to the method of spectrotemporal pursuit, originally presented in [11]. Spectrotemporal pursuit formulates the problem of estimating time varying frequency coefficients as a compressive sensing problem. We define $\mathbf{Y} \in \mathbb{R}^{P \times N}$ to be a matrix version of an observed time series of length PN, where each column of $\mathbf{Y}$ gives a length P window of the time series. Next, we define $\mathbf{X} \in \mathbb{R}^{K \times N}$ to be a matrix of frequency coefficients, with each column $\mathbf{X}_n \in \mathbb{R}^K$ representing the frequency coefficients corresponding with the time window $\mathbf{Y}_n \in \mathbb{R}^P$. By defining $\mathbf{X}$ to be real valued, it is implied that the frequency coefficients are in rectangular form, and thus a frequency resolution of $K/2$ is achieved. Using this representation, we define the quadratic measurement model:

$$L(\mathbf{y} \mid \mathbf{x}) = \sum_{n=1}^{N} ||\mathbf{y}_n - \mathbf{F}_n \mathbf{x}_n||_2^2 \tag{2.28}$$

where $\mathbf{F}_n \in \mathbb{R}^{P \times K}$ is an inverse Fourier matrix, i.e. $(\mathbf{F}_n)_{p,k} := \cos(2\pi((n-1)P+p)\frac{k-1}{K}$ and $(\mathbf{F}_n)_{p,k+\frac{K}{2}} := \sin(2\pi((n-1)P+p)\frac{k-1+K/2}{K}$ for $p = 1,\ldots,P$ and $k = 1,\ldots,K/2$. In this sense we can view the spectrotemporal estimation problem as a traditional linear measurement with Gaussian noise problem. As such, it is well defined when $P \geq K$, which is consistent with the well known fact that the number of frequency coefficients associated with a time series can not exceed the number of samples.

The method of spectrotemporal pursuit removes this constraint by introducing a sparsity inducing prior on the frequency coefficients, paralleling the approaches in compressive sensing used to estimate the coefficients underlying a system with an underdetermined set of observations. In particular, spectrotemporal pursuit imposes a group-sparsity prior on the first differences of the frequency coefficients. Letting $\mathbf{W}_n = \mathbf{X}_n - \mathbf{X}_{n-1}$ (i.e. D is the identity matrix), we define the

system model:

$$\phi(\mathbf{w}) = \sum_{k=1}^{K} \left(\sum_{n=1}^{N} w_{k,n}^2 \right)^{\frac{1}{2}}. \tag{2.29}$$

We can view this function as the ℓ_1-norm of a vector whose entries are the ℓ_2-norms of the rows of the argument. As such, $\phi(\mathbf{w})$ is small when only a small number of the rows of $\mathbf{w}$ are non-zero. Furthermore, the rows that are non-zero should have a small ℓ_2-norm. Application of this function to the *differences* of the frequency coefficients over time ensures that throughout a given time series, most frequency coefficients do not vary, and those that do vary are varying smoothly. This time-frequency characterization is known to occur in certain biological time-series. Thus, spectrotemporal pursuit utilizes this knowledge to obtain significantly denoised spectrotemporal estimates while avoiding the time/frequency resolution trade-off without necessitating a sliding window approach. This is again reminiscent of compressive sensing, which makes strong claims regarding the recoverability of a set of coefficients with underdetermined measurements so long as the coefficients are sufficiently sparse.

The spectrotemporal pursuit solution initially proposed in [11] is an iteratively reweighted least squares (IRLS) algorithm. While the IRLS algorithm is also exact and offers convergence guarantees, it requires inversion of $N \times N$ and $K \times K$ matrices N times per iteration of the algorithm. Furthermore, design of the state-covariance matrix obfuscates the problem and requires careful thought when modifying the system model.

The proposed ADMM framework yields a straightforward solution to the spectrotemporal pursuit problem. First, plugging L into equation (2.12) yields:

$$\begin{aligned} \mathbf{x}_n^{(i+1)} &= \underset{\mathbf{x}_n}{\operatorname{argmin}} ||\mathbf{y}_n - \mathbf{F}_n \mathbf{x}_n||_2^2 + \frac{\rho}{2} ||\mathbf{x}_n - \tilde{\mathbf{x}}_n^{(i)}||_2^2 \\ &= \underset{\mathbf{x}_n}{\operatorname{argmin}} ||\mathbf{x}_n + \mathbf{C}_n \mathbf{b}_n^{(i)}||_{\mathbf{C}_n}^2 \\ &= -\mathbf{C}_n \mathbf{b}_n^{(i)} \end{aligned} \tag{2.30}$$

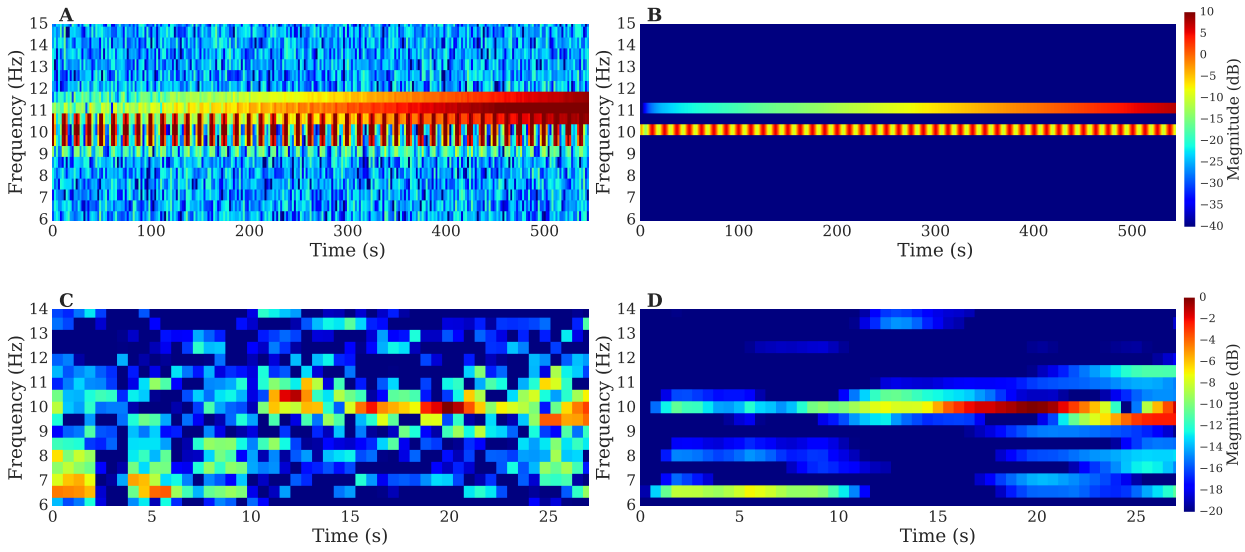

Figure 2.3: Spectrotemporal decompositions for simulated time series given by (2.32) (**A/B**) and single channel EEG recording (**C/D**). **A**: Traditional spectrogram with $NFFT = 2f_s$, no overlap, and Hanning window. **B**: Spectrotemporal pursuit estimate with $K = 2f_s$, $P = K/8$. **C**: Traditional spectrogram with $NFFT = 1024$, 75% overlap and Hanning window. **D**: Low-Rank Spectrotemporal Decomposition with $K = 1024$ and $P = K/4$.

where $\mathbf{C}_n := (\mathbf{F}_n^T\mathbf{F}_n + \frac{\rho}{2}\mathbf{I})^{-1}$ and $\mathbf{b}_n^{(i)} := -\frac{1}{2}(\mathbf{F}_n^T\mathbf{y}_n + \rho\tilde{\mathbf{x}}_n^{(i)})$. We note that when $P < K$, $\mathbf{F}_n^T\mathbf{F}_n$ is rank deficient and it is our choice of ρ that ensures the update is well formed. Also, it is important to note that each $\mathbf{C}_n$ for $n = 1,\ldots,N$ can be computed once at initialization, as they do not change throughout iterations.

Next, placing the group-sparsity prior in equation (2.14) shows that the system model update is given by a standard group-lasso problem:

$$\mathbf{w}^{(i+1)} = \underset{\mathbf{w}}{\text{argmin}}||\tilde{\mathbf{w}}^{(i)} - \mathbf{w}||_2^2 + \frac{2\beta}{\rho}\sum_{k=1}^{K}\left(\sum_{n=1}^{N} w_{k,n}^2\right)^{\frac{1}{2}}. \tag{2.31}$$

Furthermore, this special case with an orthonormal regressor matrix (i.e. the identity) yields a closed form solution, namely a row-wise shrinkage operator applied to $\tilde{\mathbf{w}}^{(i)}$ [118]. The shrinkage amount is proportional to the tuning parameter β, with larger β yielding a smaller number of non-zero rows in $\mathbf{w}$.

We demonstrate the ADMM solution for spectrotemporal pursuit on a simulated example recreated from the original paper [11]. Let $\tilde{\mathbf{y}} \in \mathbb{R}^M$ be the vectorized version of $\mathbf{y}$ with $M = NP$ and $\mathbf{y}_n = [\tilde{y}_{(n-1)P+1}, \tilde{y}_{(n-1)P+2}, \ldots, \tilde{y}_{nP}]^T$ for $n = 1, \ldots, N$. Then, we consider the signal:

$$\tilde{y}_m = 10\cos^8(2\pi f_0 m)\sin(2\pi f_1 m) + 10\exp\left(4\frac{m-M}{M}\right)\cos(2\pi f_2 m) + v_m \tag{2.32}$$

where $f_0 = 0.04$ Hz, $f_1 = 10$ Hz, $f_0 = 11$ Hz, and $v_m \sim \mathcal{N}(0,1)$ iid for $m = 1, \ldots, M$. Letting the sampling frequency be $f_s = 125$ Hz and $M = 7500$ gives a simulated time-series 600 seconds in duration. We note that $\mathbf{y}$ contains a sparse number of active frequency components, and the frequency components that are active are modulated over time in a smooth fashion. Additionally, the active frequency components f_1 and f_2 are chosen to be in neighboring frequencies, creating an increased difficulty when trying to distinguish their respective contributions.

The top row of Fig. 2.3 shows time-frequency estimates of the simulated time-series using traditional methods and spectrotemporal pursuit. First, we observe that the standard spectrogram (Fig. 2.3A) suffers from significant spectral leakage and is unable to clearly distinguish between the 10 Hz and 11 Hz frequency components. For the spectrotemporal pursuit estimate (Fig. 2.3B) we select $P < K$, meaning that the number of samples in each time window is less than the number of frequency bins. As such, we are effectively increasing the temporal resolution while still maintaining the spectral resolution without the use of overlapping windows. Because this would in general be an underdetermined problem, the group-sparsity prior is needed to ensure the problem has a unique solution. In addition to increased temporal resolution, we witness that spectrotemporal pursuit enables the contributions from f_1 and f_2 to be clearly distinguishable. Further benefits of this approach to spectrotemporal decompositions are given in detail in [11]. Here, we are proposing an algorithm that offers improvements in efficiency, modularity, and interpretability. In particular, we witness a roughly $10\times$ speedup per iteration on the same size data when using the ADMM framework rather than IRLS.

To further illustrate the modularity of the proposed framework, we next demonstrate that we can utilize an entirely different system model with a minor adjustment to a single update. Specifically, we consider a low-rank spectrotemporal decomposition (LRSD) which substitutes the nuclear norm for the group sparsity prior [108]. As such, the LRSD estimate is obtained by substituting the system model update given by (2.31) with:

$$\mathbf{w}^{(i+1)} = \underset{\mathbf{w}}{\operatorname{argmin}} ||\tilde{\mathbf{w}}^{(i)} - \mathbf{w}||_F^2 + \beta ||\mathbf{w}||_* \tag{2.33}$$

where $||\cdot||_*$ is the nuclear norm, given by the sum of the singular values of the argument. Conveniently, this update is known as the matrix lasso and yields a straightforward solution via singular value soft thresholding [75]. By making a simple adjustment to the means by which $\mathbf{w}^{(i)}$ is updated, we are able to obtain an entirely different spectrotemporal decomposition.

This point is illustrated by the bottom row of Fig. 2.3 where we demonstrate the LRSD on human single-channel EEG data using adhesive flexible sensors [59]. The data in question contains a 30-second recording in which the subject's eyes are closed at the 10 second mark, at which point we would expect to see increased energy in the alpha band (10-12 Hz). The change point nature of the recording suggests that the group sparsity prior on the dynamics, which enforces smoothness across time, is ill-suited for this recording, and the traditional spectrogram (Fig. 2.3C) suffers significantly from noise. By not explicitly enforcing smoothness in time, the low-rank enforcing nuclear norm prior (Fig. 2.3D) accommodates the change point and is able to significantly suppress activity outside of the alpha band. Similarly to the spectrotemporal pursuit example, we are able to set $P < K$ and achieve equivalent temporal resolution to the spectrogram without utilizing overlapping windows or sacrificing spectral resolution.

Remark 2. *Comparisons with other methods are intentionally omitted in this section given that there is no systematic application to these non-Markov problem formulation. While the original problem proposed in equation* (2.7) *does not lend itself to an obvious solution for the discussed*

non-Markov models, the consensus ADMM formulation given by (2.11) *may be solved in a straightforward manner. In particular, we note that the EKF and UKF have no clear extensions for non-Markov scenarios and the and the use of sampling based methods for such models would require drawing samples of group-sparse or low-rank matrices.*

2.5 Discussion

We have presented a unified framework for solving a broad class of dynamic modeling problems. The proposed method can be applied to systems with non-linear measurements and/or non-Markov dynamics. As demonstrated on two applications, our framework can be applied in a straightforward manner to acquire efficient solutions to problems that may otherwise require complex or approximate solutions. Furthermore, we have shown that this algorithm will converge on the true MAP estimate of the latent signal in the limit of large iterations. With this provably accurate algorithm comes a mathematical justification for an intuitive approach to dynamic time-series estimation, namely iteratively computing estimates based on the measurement model and system model and then averaging them in the appropriate sense.

There are a number of extensions to this framework still to be explored. The most glaring shortcomings are the inability to conduct the estimation procedure causally and the necessity to know model parameters a priori. Regarding the former, we note the use of homotopy schemes for causal estimation that gradually incorporate new observations into the solution [9, 8]. Additionally, there has been recent research investigating algorithms for performing ADMM in an online fashion [124, 72] that could potentially be leveraged by our framework. To address the latter, expectation-maximization (EM) techniques can be built into the ADMM iterations in order to estimate model parameters jointly with the desired latent time-series. In that regard, the E-step, which requires sampling from the posterior distribution, is typically the bottleneck. To address that, Langevin based methods and stochastic gradient descent methods can be used to

efficiently sample from the posterior distribution [78]. Identifying sufficient conditions on mixing times for generating approximately i.i.d. posterior samples for the M-step could be the subject of future in-depth work. We note that while there exist sample based methods for estimating model parameters [19, Sec. IV], these methods can be computationally prohibitive as witnessed in Table 2.3.

Lastly, we note that there is considerable interest in state-space estimation where the observations or system are subject to noise from heavy-tailed distributions such as the Student's t or Cauchy distributions [1, 52], which are not log-concave. Recent literature has shown that ADMM can be shown to converge under even milder conditions than those assumed by Theorem 1 [125, 77]. Given that both the Student's t and Cauchy distributions are log-quasi-concave, continuous, and possess a single local maximum, we could reasonably expect convergence of our framework to the MAP estimate in such a scenario. This topic provides interesting opportunities for future experimental and theoretical work.

2.6 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in IEEE Transactions on Signal Processing 2018. Schamberg, Gabriel; Ba, Demba; Coleman, Todd, IEEE, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Identifying and Addressing Bias in Directed Information Estimators

3.1 Introduction

The directed information (DI) is a popular measure of asymmetric relationships between two stochastic processes. Since its origination in 1966 [79, 80] and its reemergence in 1990 [82], the DI has been increasingly pervasive throughout science and engineering disciplines. When using the DI to study the inter-process relationships exhibited by real data, i.e. when the true underlying joint statistics are unknown, it is necessary to utilize DI estimation techniques. DI estimators have been studied extensively in the literature using a variety of approaches, including sequential estimation via context tree weighting (CTW) [55], maximum likelihood estimation of generalized linear models for DI between point processes [99], k-NN estimation [86], and plug-in estimation [63]. With the exception of [63], when estimating the DI from Y to X, all of these estimators work under the assumptions that (i) X and Y are jointly stationary ergodic Markov processes and (ii) X is itself a jointly stationary ergodic Markov process of the same order. For the plug-in estimator studied in [63], it is noted that when assumption (ii) does not hold, the quantity

being estimated is in fact not the DI, but rather an upper bound for the DI. Despite the common adoption of assumptions (i) and (ii), the conditions under which they hold and the implications when they do not are not well studied. Our present work seeks to fill this gap in order to ensure that the estimation of DI across scientific disciplines can be conducted in a manner such that the results are reliable.

Relevant discussions regarding the issues surrounding assumption (ii) have been held in the literature on Granger causality (GC) [45]. GC can be viewed as a special case of DI where the processes in question obey a vector autoregressive (VAR) model with Gaussian noise. It is noted in the GC literature that subsets of finite-order VAR processes are in general infinite order autoregressive processes [115]. Thus, estimating a "restricted" model (i.e. one where the candidate influencer is hidden) from data requires estimating a truncated model and induces a bias-variance tradeoff. For the linear Gaussian case, this issue can be avoided by computing the restricted model directly from the full model using the Yule-Walker equations [13]. Unfortunately, there is no clear extension of this approach for arbitrary Markov processes, and other techniques are required.

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

The contributions of this chapter are summarized as follows:

- For networks of interacting processes, we provide sufficient conditions for which the conditional independencies needed to obtain unbiased estimates of the directed information hold.

- We show that these conditions are also necessary with the exception of a set of parameters with Lebesgue measure zero.

- We present a bound for the estimation bias that can be estimated reliably under mild conditions.
- To understand the magnitude of the biases in question, we compute the proposed bound for simulated processes in a variety of problem settings.

3.2 Preliminaries

3.2.1 Notation and Basic Definitions

Let X, Y, and Z denote discrete finite-alphabet random processes, unless otherwise specified, where, at any time i, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$, and $Z_i \in \mathcal{Z}$. Without loss of generality, Z may represent a collection of processes $(Z^{(1)}, \ldots, Z^{(m)}) \in \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_m \triangleq \mathcal{Z}$. 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. $X_i \in \mathcal{X}$. A temporal range of a process is denoted by a subscript and superscript, i.e. $X_i^n = (X_i, X_{i+1}, \ldots, X_n)$, and we define $X^n \triangleq X_1^n$. Realizations of processes are given by lowercase letters. When a process is Markov of order d we will refer to it as d-Markov. Probability mass functions (pmfs) are equivalently referred to as "distributions" and are denoted by p. 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 $p_{X_i|X^{i-1}}(x_i \mid x^{i-1}) \equiv p(x_i \mid x^{i-1})$ gives the conditional distribution of X at a single time conditioned on the past of X. We define the *causally conditional distribution with lag k* as:

$$p(x^n \mid\mid y^{n-k}) \triangleq \prod_{i=1}^{n} p(x_i \mid x^{i-1}, y^{i-k}). \tag{3.1}$$

Note that the standard interpretation of the causal[1] conditioning (as in [65]) is recovered by letting

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

$k = 0$.

We will briefly review some information theoretic quantities that are used frequently throughout this chapter. 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:

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

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

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

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

$$\bar{H}(X) = \lim_{n \to \infty} \frac{1}{n} H(X^n) \tag{3.2}$$

$$\bar{H}(X \mid Y) = \lim_{n \to \infty} \frac{1}{n} H(X^n \mid Y^n) \tag{3.3}$$

$$\bar{H}^{(k)}(X \mid\mid Y) = \lim_{n \to \infty} \frac{1}{n} H(X^n \mid\mid Y^{n-k}) \tag{3.4}$$

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

The conditional mutual information is given by:

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

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:

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

3.2.2 Directed Information

The directed information from a sequence Y^n to X^n was defined by Massey [82] as:

$$I(Y^n \to X^n) = \sum_{i=1}^{n} I(Y^i; X_i \mid X^{i-1}) \tag{3.5}$$

$$= H(X^n) - H(X^n \,||\, Y^n) \tag{3.6}$$

When ignoring the instantaneous relationship between X_i and Y_i, the reverse DI [55] may be used:

$$I(Y^{n-1} \to X^n) = \sum_{i=1}^{n} I(Y^{i-1}; X_i \mid X^{i-1}) \tag{3.7}$$

$$= H(X^n) - H(X^n \,||\, Y^{n-1}) \tag{3.8}$$

Note that under the assumption that X_i and Y_i are conditionally independent given their pasts, we have $I(Y^n \to X^n) = I(Y^{n-1} \to X^n)$. Given that the DI is given by a sum over time, one may be interested in a process level measure of DI. This can be accomplished through use of the directed information rate [65], given by:

$$\bar{I}(Y \to X) = \lim_{n\to\infty} \frac{1}{n} I(Y^{n-1} \to X^n) \tag{3.9}$$

$$= \bar{H}(X) - \bar{H}^{(1)}(X \,||\, Y) \tag{3.10}$$

When measuring the amount of unique directed information from Y to X that is not contained in

a third process, Z, one may use the causally conditional DI from Y to X given Z, defined as:

$$I(Y^n \to X^n \,||\, Z^n) = \sum_{i=1}^{n} I(X_i; Y^i \mid X^{i-1}, Z^i) \tag{3.11}$$

$$= \sum_{i=1}^{n} H(X_i \mid X^{i-1}, Z^i) - H(X_i \mid X^{i-1}, Y^i, Z^i) \tag{3.12}$$

and the associated causally conditional DI rate (when it exists) as:

$$\bar{I}(Y \to X \,||\, Z) = \lim_{n\to\infty} \frac{1}{n} I(Y^n \to X^n \,||\, Z^n). \tag{3.13}$$

If we assume that (i) (X,Y,Z) are jointly d-Markov, i.e. that $p(X_i \mid X^{i-1}, Y^i, Z^i) = p(X_i \mid X_{i-d}^{i-1}, Y_{i-d}^i, Z_{i-d}^i)$, the second entropy term in (3.12) can be simplified to $H(X_i \mid X_{i-d}^{i-1}, Y_{i-d}^i, Z_{i-d}^i)$. If the further assumption is made that (ii) $X_i \perp (X^{i-d-1}, Z^{i-d-1}) \mid (X_{i-d}^{i-1}, Z_{i-d}^i)$ (henceforth "X is conditionally d-Markov given Z"), then the first entropy term can be simplified as $H(X_i \mid X_{i-d}^{i-1}, Z_{i-d}^i)$. Once this assumption is made, it is clear that the DI can be estimated from data by splitting a stream (X^n, Y^n, Z^n) into a collection of samples $\{(X_{i-d}^i, Y_{i-d}^i, Z_{i-d}^i)\}_{i=1}^n$ and estimating the appropriate distributions using a variety of methods [86, 63, 99, 55]. The goal of this work is to understand when we can expect both of these assumptions to hold and what the consequences are of assuming they both hold when in fact only the the first holds. It should be noted that while we consider only a network of processes and the causally conditional DI as above, all of the results demonstrated in the following sections still apply when $Z = \emptyset$ and the standard DI for a pair of processes is used.

3.2.3 Bayesian Networks

To understand the conditions under which the desired independence relationships hold, we can leverage tools from Bayesian networks, which can be used to represent conditional independencies in collections of random variables using a directed acyclic graph (DAG) $\mathcal{G} =$

(V,E), where $V = \{V_1, \ldots, V_m\}$ is a set of random variables (equivalently nodes or vertices) and $E \subset V \times V$ is a set of directed edges that it do not contain any cycles [114]. The parent set of a node V_i in a DAG is defined as the set of nodes with arrows going into V_i, $\mathcal{P}_i \triangleq \{V_j : (V_j \to V_i) \in E\}$. The defining characteristic of a Bayesian network representation of a joint distribution over the nodes $V \sim p$ is the ability to factorize the distribution as:

$$p(V) = \prod_{i=1}^{m} p(V_i \mid \mathcal{P}_i). \tag{3.14}$$

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 [114, 84]. 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 $\mathcal{G}$ [114].

Algorithm 1 d-Separation [68]

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

3.3 Characterizing Conditionally Markov Processes

3.3.1 Network Representation of Markov Processes

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

$$p(X^n, Y^n, Z^n) = \prod_{i=1}^{n} p(X_i, Y_i, Z_i \mid X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}). \tag{3.15}$$

We next make the additional assumption (A1) that X_i, Y_i, and Z_i are pairwise conditionally independent given the past $\{X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}\}$. This assumption facilitates construction of a Bayesian network, as we can rely on the arrow of time to determine the direction of arrows in the network. In the absence of (A1), we cannot construct a *unique* Bayesian network representation of Markov processes without making alternative assumptions (the details of which will be discussed in future work). This is similar reasoning to that of [101], where (A1) is used for establishing the equivalence between DI graphs and minimal generative model graphs. Under (A1), we can

further simplify (3.15) as:

$$p(X^n, Y^n, Z^n) = \prod_{i=1}^{n} \prod_{S \in \{X_i, Y_i, Z_i\}} p(S \mid X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}). \tag{3.16}$$

Comparing (3.14) and (3.16), 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 $S_{i-k} \to S_i'$ for $S, S' \in \{X, Y, Z\}$ and $k = 1, \ldots, d$ only if:

$$I(S_{i-k}; S_i' \mid \{X_{i-d}^{i-1}, Y_{i-d}^{i-1}, Z_{i-d}^{i-1}\} \setminus S_{i-k}) > 0. \tag{3.17}$$

3.3.2 Necessary and Sufficient Conditions for d-Separation

Using the Bayesian network construction given by (3.17), 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 satisfying (A1). If $I(Y^n \to X^n \mid\mid Z^n) = 0$ then X is conditionally d-Markov given Z. If $I(Y^n \to X^n \mid\mid Z^n) > 0$, X is conditionally Markov given Z of order $2d$ or less if:*

$$I(Y_j; Y_k \mid X^i, Z^i) = 0 \; \forall \, j \leq k \leq i \tag{3.18}$$

If $I(Y^n \to X^n \mid\mid Z^n) > 0$ but (3.18) 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 (3.17).

A proof of the theorem can be found in Appendix B.1.

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 $p(Y_i \mid X_{i-d}^{i-1}, Y_{i-d}^{i-1}) = p(Y_i \mid X_{i-d}^{i-1})$ and $p(X_i \mid X_{i-d}^{i-1}, Y_{i-d}^{i-1}) = p(X_i \mid X_{i-d}^{i-1}, Y_{i-\tau})$ for some $0 \leq \tau \leq d$. This theorem has particularly strong implications for processes with bidirectional influences, as summarized by the following corollary:

Corollary 1. *Let X and Y be a pair of jointly d-Markov processes with bidirectional influences given (without loss of generality) by $I(Y^{n-1} \to X^n) > 0$ and $I(X^n \to Y^n) > 0$. Then X and Y will be marginally Markov if there exist integers $\tau_1 > 0$ and $\tau_2 \geq 0$ such that for all i:*

$$I(X_i; X^{i-1}, Y^{i-1} \setminus Y_{i-\tau_1} \mid Y_{i-\tau_1}) = 0 \tag{3.19}$$

$$I(Y_i; Y^{i-1}, X^{i-1} \setminus X_{i-\tau_2} \mid X_{i-\tau_2}) = 0 \tag{3.20}$$

where $Y^{i-1} \setminus Y_{i-\tau_1} \triangleq \{Y_1, \ldots, Y_i\} \setminus \{Y_{i-\tau_1}\}$ and $X^{i-1} \setminus X_{i-\tau_2}$ is defined similarly. Furthermore, for any distribution not satisfying (3.19) *and* (3.20)*, there will not exist any positive integer l such that X_{i-l}^{i-1} d-separates X_i from X^{i-l-1} or Y_{i-l}^{i-1} d-separates Y_i from Y^{i-l-1}.*

The above corollary follows directly from the application of Theorem 2 to both X and Y. It may be interpreted as stating that the only scenario in which joint Markovicity implies marginal Markovicity in pairs of processes with bidirectional influence is when each process at any given time is independent of everything that has happened up to that point when conditioned on a single sample from the other process. We note that even the most basic sensible feedback communication system does not fit this model. While it is reasonable to assume that each channel output Y_i is dependent solely upon the channel input X_i, no sensible communication scheme would then have

the next transmission X_{i+1} depend solely upon the feedback Y_i (i.e. be independent of all previous transmissions X^i).

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 (3.18)), 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 [114] and citations therein), a formal characterization within the present context is desired.

3.3.3 Completeness of d-Separation

For a DAG $\mathcal{G} = (V, E)$, define $\Gamma_{\mathcal{G}} \subset \mathbb{R}^M$ to represent the set of *all* discrete distributions $p(V)$ such that the $\mathcal{G}$ is a Bayesian network for p. Further define $\Gamma^u_{\mathcal{G}} \subset \Gamma_{\mathcal{G}}$ to be the subset of those distributions that are unfaithful to $\mathcal{G}$. Then, it was shown in [84] the $\Gamma^u_{\mathcal{G}}$ has Lebesgue measure zero with respect to $\mathbb{R}^M$, where M is the number of parameters needed to specify the joint distribution p. Unfortunately, this result cannot be directly applied to our problem. To see why, let $\Theta_{\mathcal{G}} \subset \mathbb{R}^N$ represent the set of parameters defining discrete jointly stationary d-Markov processes satisfying (A1) for which $\mathcal{G}$ gives the Bayesian network constructed according (3.17). Defining $\theta^{S_i}_{X,Y,Z} \triangleq p(S_i \mid X^{i-1}_{i-d}, Y^{i-1}_{i-d}, Z^{i-1}_{i-d})$ for $S \in \{X, Y, Z\}$ and $\theta \triangleq \{\theta^a_b : a \in \mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}, b \in \mathcal{X}^d \times \mathcal{Y}^d \times \mathcal{Z}^d\}$, we can see that there are $N \triangleq (|\mathcal{X}| + |\mathcal{Y}| + |\mathcal{Z}| - 3)|\mathcal{X}|^d|\mathcal{Y}|^d|\mathcal{Z}|^d$ many parameters uniquely defining such a process. Next define $\Theta^u_{\mathcal{G}} \subset \Theta_{\mathcal{G}}$ to be the subset of parameters such that the induced distribution p is unfaithful to $\mathcal{G}$. Then, it is clear that, due to the stationarity constraint, $N << M$, and the Lebesgue measure of $\Gamma^u_{\mathcal{G}}$ with respect to $\mathbb{R}^M$ does not tell us what the Lebesgue measure of $\Theta^u_{\mathcal{G}}$ is with respect to $\mathbb{R}^N$. Returning to the question at hand, we seek to know when we can expect X to be conditionally d-Markov given Z despite the conditional independence not being implied by d-separation. Using a similar technique to [84], the following theorem states

that, when $d = 1$, the set of such parameters has Lebesgue measure zero:

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* (3.17) *has Lebesgue measure zero with respect to $\mathbb{R}^N$.*

A proof of the theorem is given in Appendix B.2.

3.4 Quantifying Estimation Bias

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

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

$$I_T^{(k)}(Y^n \to X^n \,||\, Z^n) \triangleq \sum_{i=1}^{n} I(X_i; Y_{i-k}^{i} \mid X_{i-k}^{i-1}, Z_{i-k}^{i}) \tag{3.21}$$

The TDI in its unconditional form is discussed in [63] in the context of plug-in estimators of DI. Should both Markovicity and conditional Markovicity hold for a collection of processes, then the TDI and the DI are equivalent. However, having shown that conditional Markovicity is unlikely to hold, we here name the TDI to emphasize that it is a fundamentally different measure from the traditional DI.

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

$$I_P^{(k)}(Y^n \to X^n \,||\, Z^n) \triangleq \sum_{i=1}^{n} I(X_i; Y_{i-k}^{i} \mid X^{i-1}, Y^{i-k-1}, Z^{i}) \tag{3.22}$$

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

$$I(X_i;Y_{i-k}^i \mid X^{i-1},Y^{i-k-1},Z^i) = H(X_i \mid X_{i-k-d}^{i-1},Y_{i-k-d}^{i-k-1},Z_{i-k-d}^i) - H(X_i \mid X_{i-d}^{i-1},Y_{i-d}^i,Z_{i-d}^i)$$

Thus, estimators of DI can be extended to estimate the PDI without the additional requirement of conditional Markovicity. This idea is formalized in the next chapter and the details of such an estimator utilizing the context tree weighting predictors are provided. Defining the TDI and PDI rates $\bar{I}_T^{(k)}$ and $\bar{I}_P^{(k)}$ to be the normalized limits analagous with the DI rate given by (3.13), we are able to bound the DI rate from above and below as follows:

Theorem 4. *Let (X,Y,Z) be jointly stationary d-Markov. For $k_1 \geq 1$ and $k_2 \geq d$, the causally conditional PDI and TDI rates bound the DI rate as:*

$$\bar{I}_P^{(k_1)}(Y \to X \mid\mid Z) \leq \bar{I}(Y \to X \mid\mid Z) \leq \bar{I}_T^{(k_2)}(Y \to X \mid\mid Z) \tag{3.23}$$

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

A proof of the theorem can be found in Appendix B.3.

3.5 Simulations

In the above sections we have demonstrated that while one cannot reasonably expect to data to satisfy the necessary assumptions for obtaining unbiased estimates of DI, the TDI and PDI can be used to provide upper and lower bounds for the true DI. A natural next question is, how significant is the difference between PDI and TDI? To address this question, we simulate a pair of jointly stationary Markov discrete processes in four settings, each characterized by a

particular simplification of the generative distribution $p(X_i, Y_i \mid X^{i-1}, Y^{i-1})$:

$$p(X_i \mid Y_{i-1})p(Y_i \mid Y_{i-1}) \tag{S1}$$

$$p(X_i \mid X_{i-1}, Y_{i-1})p(Y_i \mid Y_{i-1}) \tag{S2}$$

$$p(X_i \mid X_{i-1}, Y_{i-1})p(Y_i \mid X_{i-1}, Y_{i-1}) \tag{S3}$$

$$p(X_i \mid X_{i-2}^{i-1}, Y_{i-2}^{i-1})p(Y_i \mid X_{i-2}^{i-1}, Y_{i-2}^{i-1}) \tag{S4}$$

For each of these graphical structures, we conducted 100 experiments with $|\mathcal{X}| = |\mathcal{Y}| = 4$ for (S1)-(S3) and $|\mathcal{X}| = |\mathcal{Y}| = 3$ for (S4). In each experiment, the parameters were sample as independent exponential random variables and then appropriately normalized, yielding parameters drawn uniformly from the probability simplex [28]. Using the sampled parameters, sequences (x^n, y^n) were generated with $n = 300000$ large enough to ensure that accurate estimates of the TDI and PDI could be obtained. $\bar{I}_T^{(k)}(Y \to X)$ and $\bar{I}_P^{(k)}(Y \to X)$ were estimated using CTW estimators in the style of $\hat{I}_3$ in [55] for $k = d$, $d+1$, and $d+2$. Figure 3.1 shows boxplots representing $\hat{\bar{I}}_T^{(k)}(Y \to X) - \hat{\bar{I}}(Y \to X)$ and $\hat{\bar{I}}_P^{(k)}(Y \to X) - \hat{\bar{I}}(Y \to X)$ for varying values of k along with the mean (across trials) DI rate, which was determined by the value converged upon by the TDI and PDI[2]. We can see that while the bound on the bias quickly converges to zero as k increases, there are many examples in every setting when $k = d$ for which the bound on the bias is rather large relative to the mean DI rate. Furthermore, we can see that as the structures get more complex, the bound on the bias tends to be larger. This suggests that while (S4) is not covered by Theorem 3, alternative proof techniques may exist for demonstrating that the results hold for $d > 1$. Thus, when working with real data, it may be prudent to use the TDI and PDI to upper *and* lower bound the DI rate rather than simply relying on the TDI as a proxy for DI.

[2]Code can be found in the following repository: `https://github.com/gabeschamberg/directed_info_bias`.

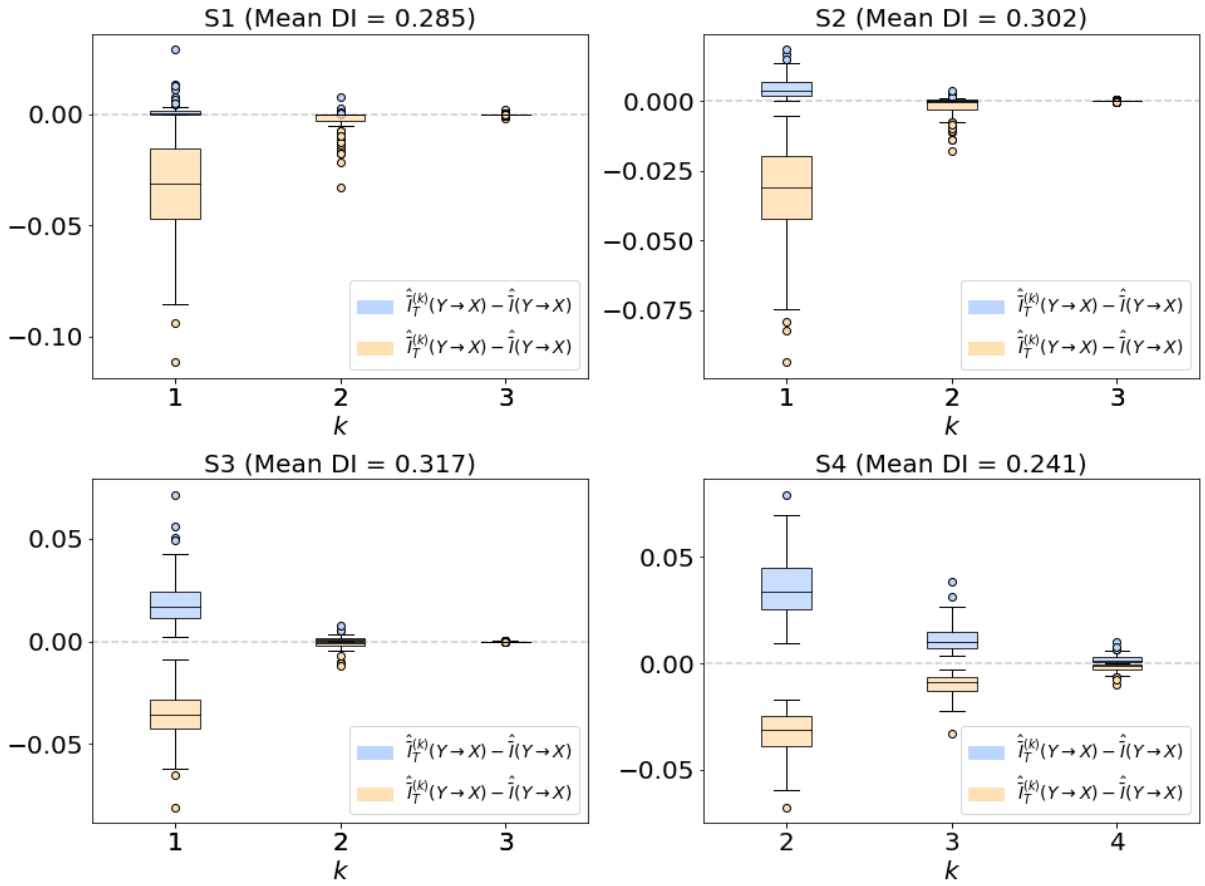

Figure 3.1: Difference between TDI and DI (blue) and PDI and DI (orange) for different values of k (x-axis) under different process structures (panels).

3.6 Acknowledgements

Chapter 3, in part, is a reprint of the material as it appears in the proceedings of the IEEE International Symposium on Information Theory 2019. Schamberg, Gabriel; Coleman, Todd, IEEE, 2019. Chapter 3, in part, is currently under review for publication of the material. Schamberg, Gabriel; Coleman, Todd. The dissertation author was the primary investigator and author of these materials.

Chapter 4

Measuring Sample Path Causal Influences with Relative Entropy

4.1 Introduction

Building upon the the ideas of Wiener [126], Granger [45] 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 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 [92, 94]. In [34], 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 [111]), 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) [79, 80, 82] and transfer entropy (TE) [110], which is equivalent to GC for Gaussian autoregressive processes [14]. Justification for use of the DI for characterizing directional dependencies between processes was given in [100], 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 [35] 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 [74]. 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 [50] 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 [92] and [97] 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 [60] and [107] 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 4.3.1). 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 [115], 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 [115] 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 address both of these unmet needs in Section 4.4.1 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 chapter 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".

- Demonstration of the causal measure's value through application to simulated and real data.

The remainder of this chapter is organized as follows: following a brief overview of notation, Section 4.2 provides a technical summary of related work. In Section 4.3 we define the measure, present key properties, and provide justification for the measure through several examples. Section 4.4 provides a framework for estimating the measure. Section 4.5 demonstrates the measure on simulated and real data. Finally, Section 4.6 contains a discussion of the results and opportunities for future work.

4.2 Related Work

In discussions of causal inference, it is important to differentiate between the *deterministic* and *stochastic* settings as well as the *interventional* and *observational* settings. With regard to the former, we limit consideration strictly to the stochastic setting. In other words, This restriction is necessary when utilizing Granger's perspective, as comparing qualities of *prediction* in stochastic settings is our main interest. With regard to the latter, we focus on the observational setting, wherein the potential causes (i.e. y^{i-1} in Q3) may not be controlled or perturbed, in comparison to the causal intervention calculus pioneered by Pearl [92].

We now provide a brief summary of three key concepts in the measurement of causal influence across time series, namely Granger causality (GC) [45], directed information (DI) [80, 82], and causal strength (CS) [54].

4.2.1 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 [45]. 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:

$$X_i = \sum_{j=1}^{d} a_j X_{i-j} + b_j Y_{i-j} + c_j Z_{i-j} + U_i \tag{4.1}$$

$$X_i = \sum_{j=1}^{d} d_j X_{i-j} + e_j Z_{i-j} + V_i \tag{4.2}$$

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 (4.2) is a subset of the models given by (4.1) 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:

$$G_{Y \to X} \triangleq \ln \frac{\sigma_V^2}{\sigma_U^2} \tag{4.3}$$

The limitations of Granger causality extend considerably beyond the restriction to linear models (see [115] 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 (4.2) 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 4.4.1.

4.2.2 Directed Information

Recall the causally conditional DI and causally conditional DI rate from the previous chapter:

$$I(Y^n \to X^n \,||\, Z^n) = \sum_{i=1}^{n} I(Y^i; X_i \mid X^{i-1}, Z^i) \tag{4.4}$$

$$\bar{I}(Y \to X \,||\, Z) = \lim_{n\to\infty} \frac{1}{n} I(Y^n \to X^n \,||\, Z^n) \tag{4.5}$$

Unless otherwise stated, we make the assumption 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}. This assumption is consistent with Granger's original presentation, as we can see in (4.1) and (4.2) that neither Y_i nor Z_i are included in either model prediction X_i. In such a setting, we have that the causally conditional DI (rate) can be written as:

$$I(Y^{n-1} \to X^n \,||\, Z^{n-1}) = \sum_{i=1}^{n} I(Y^i; X_i \mid X^{i-1}, Z^{i-1}) \tag{4.6}$$

$$\bar{I}(Y \to X \,||\, Z) = \lim_{n\to\infty} \frac{1}{n} I(Y^{n-1} \to X^n \,||\, Z^{n-1}) \tag{4.7}$$

It follows intuitively that DI serves as a generalization of GC in that the causally conditional DI provides a measure of how much *unique* information is shared between the past of Y and the present of X, where we use the word unique to indicate that this information is not also contained in the past of X or Z. This connection is further elucidated when rewriting the conditional mutual information term in (4.6) as:

$$I(Y^i; X_i \mid X^{i-1}, Z^{i-1}) = \mathbb{E}\left[\log \frac{p(X_i \mid X^{i-1}, Y^{i-1}, Z^{i-1})}{p(X_i \mid X^{i-1}, Z^{i-1})}\right]$$

In words, the conditional mutual information can be rewritten as the difference in the expected log loss of two predictors of X_i, one that has access to the past of Y, and one that does not. This

framing of DI provides intuition for the connection between DI and GC. An extensive exposition of the precise nature of this relationship can be found in [3].

4.2.3 Causal Strength

In [54], 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 (4.1) and (4.2) to arbitrary probability distributions $p(X_i \mid X^{i-1}, Y^{i-1}, Z^{i-1})$ and $p(X_i \mid X^{i-1}, Z^{i-1})$. Next, note that the second distribution has the following factorization when summing over all possible pasts of Y:

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

Here we can

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

$$\mathfrak{C}_{Y \to X} \triangleq \mathbb{E}\left[D(p_{X_i|X^{i-1},Y^{i-1},Z^{i-1}} \mid\mid \tilde{p}_{X_i|X^{i-1},Z^{i-1}}) \right] \tag{4.8}$$

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:

$$\tilde{p}_{X_i|X^{i-1},Z^{i-1}}(X_i) \triangleq \sum_{y^{i-1}} p(X_i \mid X^{i-1}, y^{i-1}, Z^{i-1}) p(y^{i-1}) \tag{4.9}$$

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 [54]). 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 4.1. In such a scenario, it is possible to have $I(Y^{n-1} \to X^n) = 0$ 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 4.3.1.

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.

4.2.4 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 $h(x \mid y) \triangleq -\log p_{X|Y}(x \mid y)$, a local version of DI and its conditional extension may be obtained (see Table 1 in [73] 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 [46]) may be given by:

$$i(y^{n-1} \to x^n) = \sum_{i=1}^{n} \log \frac{p(x_i \mid x^{i-1}, y^{i-1})}{p(x_i \mid x^{i-1})} \tag{4.10}$$

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.

4.2.5 Time-Varying Causal Measures

A popular extension of GC style causal measures is application to time-varying scenarios [112, 88]. 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.

4.3 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:

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

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:

$$p_{X_i}^{(r)}(x_i) \triangleq p_{X_i}(x_i \mid \mathcal{H}_i^{(r)})$$
$$p_{X_i}^{(c)}(x_i) \triangleq p_{X_i}(x_i \mid \mathcal{H}_i^{(c)}).$$

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:

$$C_{Y \to X}(\mathcal{H}_i^{(c)}) \triangleq D(p_{X_i}^{(c)} \,||\, p_{X_i}^{(r)}) \tag{4.11}$$

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

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

(3.7), along with two equivalent representations:

$$I(Y^{i-1};X_i \mid X^{i-1}) = H(X_i \mid X^{i-1}) - H(X_i \mid X^{i-1},Y^{i-1}) \tag{4.12}$$

$$= \mathbb{E}_{X^i,Y^{i-1}}\left[\log \frac{p(X_i \mid X^{i-1},Y^{i-1})}{p(X_i \mid X^{i-1})}\right] \tag{4.13}$$

These equivalent definitions of directed information yield two interpretations. While (4.12) considers the reduction in uncertainty obtained by conditioning on Y^{i-1}, (4.13) 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:

$$H(X_i \mid x^{i-1}) - H(X_i \mid x^{i-1},y^{i-1}) \tag{4.14}$$

$$\neq$$

$$\mathbb{E}_{X_i}\left[\log \frac{p(X_i \mid x^{i-1},y^{i-1})}{p(X_i \mid x^{i-1})}\middle| x^{i-1},y^{i-1}\right] \tag{4.15}$$

The representation given by (4.15) is chosen to be the sample path causal measure and is indeed equivalent to the proposed measure in (4.11). This choice is made clear by noting two properties of (4.14). First, we note that (4.14) 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 [55], 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:*

$$\sum_{i=1}^{n} \mathbb{E}[C_{Y\to X}(\mathcal{H}_i^{(c)})] = I(Y^{n-1} \to X^n \,||\, Z^{n-1}) \tag{4.16}$$

See Appendix C.4 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 3. *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 intervention. For example, one might wish to consider:*

$$C_{Y\to X}^{(i)}(\mathcal{H}_i^{(c)}) \triangleq D(p_{X_i}^{(i)} \,||\, p_{X_i}^{(r)}) \tag{4.17}$$

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

$$p_{X_i}^{(i)}(x_i) \triangleq p(x_i \mid \mathcal{H}_i^{(r)}, do(Y^{i-1} = y^{i-1})) \tag{4.18}$$

We have here used the do-operator of Pearl [92] to represent the action of forcing Y^{i-1} to take

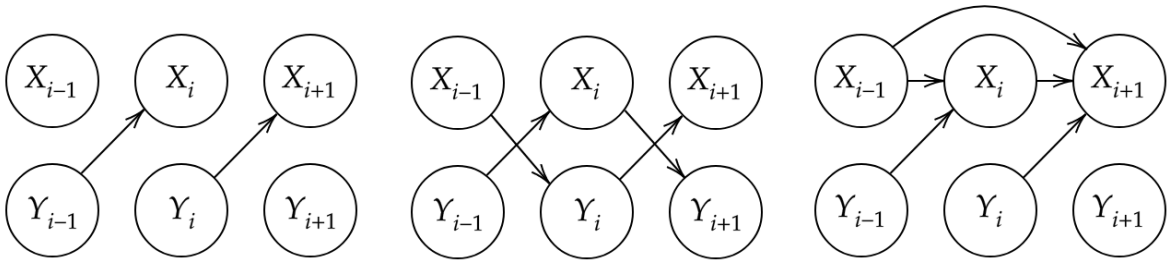

Figure 4.1: Graphical representation of the IID influences (left), perturbed cross copying (center), and horse betting (right) examples.

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 $C^{(i)}_{Y\to X} = C_{Y\to X}$ is outside the scope of the present discussion, 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 4.1 and is shown formally in Appendix C.1. By contrast, we would not expect this equivalence to hold for the stock market example considered in Section 4.5.2, as discussed in Remark 4. The use of the do-operator in this type of causal measure will be revisited in the following chapter where we extend the present discussion to be applied to general graphs.

4.3.1 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 4.1.

IID Influences

Let $Y_i \sim Bern(\varepsilon)$ iid for $i = 1, 2, \ldots$ and:

$$X_i \sim \begin{cases} \text{Bern}(p_1), & Y_{i-1} = 1 \\ \text{Bern}(p_2), & Y_{i-1} = 0 \end{cases} \tag{4.19}$$

for $\varepsilon, p_1, p_2 \in [0,1]$. 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 $C_{Y \to X}(i)$, we first need to find the restricted distribution of X_i given only its own past:

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

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:

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

Thus, we see that as $\varepsilon \to 0$,

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

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

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

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 $C_{Y\to X}(i) > 0$ 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 (4.19) and again let ε approach zero. In such a scenario, we find that:

$$p^{(r)}_{X_i}(1) \approx 1 \qquad p^{(c)}_{X_i}(1) \approx \begin{cases} 1, & y_{i-1} = 1 \ (w.p.\ 1-\varepsilon) \\ 0.5, & y_{i-1} = 0 \ (w.p.\ \varepsilon) \end{cases}$$

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 $H(X_i \mid X^{i-1}) \leq H(X_i \mid X^{i-1}, Y^{i-1})$, 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.

Perturbed Cross Copying

We next consider a scenario where two processes repeatedly swap values. This example was originally posed in [10] and modified to include noise in [54]. Formally, the processes may be defined as:

$$X_i = \begin{cases} Y_{i-1}, \ w.p. \ 1-\varepsilon \\ Y_{i-1} \oplus 1, \ w.p. \ \varepsilon \end{cases} \qquad Y_i = \begin{cases} X_{i-1}, \ w.p. \ 1-\varepsilon \\ X_{i-1} \oplus 1, \ w.p. \ \varepsilon \end{cases} \tag{4.20}$$

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 [54], the DI rate approaches zero as $\varepsilon \to 0$. 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. [54] 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. $\mathfrak{C}_{Y\to X} = D(\varepsilon \,||\, 0.5) \to 1$ as $\varepsilon \to 0$ (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:

$$p^{(c)}_{X_i}(x_i) = \begin{cases} 1-\varepsilon, & x_i = y_{i-1} \\ \varepsilon, & x_i \neq y_{i-1} \end{cases} \qquad p^{(r)}_{X_i}(x_i) = \begin{cases} \varepsilon^2 + (1-\varepsilon)^2, & x_i = x_{i-2} \\ 2\varepsilon(1-\varepsilon), & x_i \neq x_{i-2} \end{cases}$$

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

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

Thus, we see that as $\varepsilon \to 0$, $C_{Y\to X} \to 0$ if $x_{i-2} = y_{i-1}$ and $C_{Y\to X} \to \infty$ 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 $\varepsilon \to 0$, Y_{i-1} contains less and less information about X_i that is *not already known*. As a result $\bar{I}(Y \to X)$ 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 [54]). 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.

Horse Betting

Consider the problem of horse race gambling with side information as presented in Section III-A of [96] (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 $o(X_i \mid X^{i-1})$, and the portion of wealth bet on each horse is given by

$b(X_i \mid X^{i-1}, Y^{i-1}) \geq 0$ with $\sum_x b(x \mid X^{i-1}, Y^{i-1}) = 1$. Thus, the evolution of the wealth can be described recursively as:

$$w(X^i, Y^{i-1}) = b(X_i \mid X^{i-1}, Y^{i-1}) o(X_i \mid X^{i-1}) w(X^{i-1}, Y^{i-2})$$

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 [96] 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:

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

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

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

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 [96] that the increase in growth rate resulting from including side information in the betting strategy is given by:

$$\frac{1}{n} \mathbb{E}\left[\log w^*(X^n, Y^{n-1}) - \log w^*(X^n)\right] = \frac{1}{n} I(Y^{n-1} \to X^n) \tag{4.21}$$

It should be noted that the result in (4.21) holds for any choice of odds $o(X_i \mid X^{i-1})$. 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:

$$\mathbb{E}[\log b^*(X_i \mid X^{i-1}) o(X_i \mid X^{i-1}) \mid x^{i-1}] = \log \delta < 0 \tag{4.22}$$

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:

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

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 (4.21), a gambler with side info can expect to win money if:

$$\frac{1}{n} I(Y^{n-1} \to X^n) > -\log \delta \tag{4.23}$$

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 (4.23) 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:

$$E[\log b^*(X_i \mid X^{i-1}, Y^{i-1}) o(X_i \mid X^{i-1}) \mid x^{i-1}, y^{i-1}] > 0 \tag{4.24}$$

Combining (4.24) with the rate for round i in (4.22), the condition for which the gambler should place a bet becomes:

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

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 4.5.2, we note that there may be great interest in how the two time series interact for specific realizations, even if those realizations are rare.

4.4 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 [85]. The sequential prediction problem formulation we consider is as follows: for each round $i \in \{1, \ldots, n\}$, 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 $l : \mathcal{X} \to \mathbb{R}$ 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:

$$r(\hat{p}_i, \tilde{p}_i, x_i) = l(\hat{p}_i, x_i) - l(\tilde{p}_i, x_i) \tag{4.25}$$

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

$$R_n(\tilde{\mathcal{P}}_n) = \sup_{x^n \in \mathcal{X}^n} \sum_{i=1}^{n} l(\hat{p}_i, x_i) - \inf_{\tilde{p} \in \tilde{\mathcal{P}}_n} \sum_{i=1}^{n} l(\tilde{p}_i, x_i) \tag{4.26}$$

$$\triangleq \sup_{x^n \in \mathcal{X}^n} \sum_{i=1}^{n} r(\hat{p}_i, f_i^*, x_i) \tag{4.27}$$

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 [85]. 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:

$$R_n(\tilde{\mathcal{P}}_n) \leq M(n) \tag{4.28}$$

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:

$$\hat{C}_{Y\to X}(i) = D(\hat{p}_{X_i}^{(c)} \,||\, \hat{p}_{X_i}^{(r)}) \tag{4.29}$$

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 [55].

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

$$CR(n) \triangleq \sum_{i=1}^{n} \left| \hat{C}_{Y\to X}(i) - C^*_{Y\to X}(i) \right| \tag{4.30}$$

where we define:

$$C^*_{Y\to X}(i) = D(p_{X_i}^{(c)*} \,||\, p_{X_i}^{(r)*}) \tag{4.31}$$

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

$$\sup_{x \in \mathcal{X}} \left| \log \frac{\hat{p}^{(c)}_{X_i}(x)}{\hat{p}^{(r)}_{X_i}(x)} \right| < \infty \;\; i = 1, \ldots, n \tag{4.32}$$

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

$$\sum_{i=1}^{n} \left| \mathbb{E}_{p^{(c)*}_{X_i}} \left[r(\hat{p}^{(r)}_{X_i}, p^{(r)*}_{X_i}, X_i) \right] \right| \leq M^{(r)}(n) \tag{4.33}$$

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 (4.26), $CR(n)$ *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 5. *Let the worst case regret for the predictors $\hat{p}^{(r)}_{X_i}$ and $\hat{p}^{(c)}_{X_i}$ be bounded by $R^{(r)}_n(\tilde{\mathcal{P}}^{(r)}_n) \leq M^{(r)}(n)$ and $R^{(c)}_n(\tilde{\mathcal{P}}^{(c)}_n) \leq M^{(c)}(n)$, respectively. Then, for any 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:*

$$CR(n) \leq M^{(c)}(n) + M^{(r)}(n) + \frac{||\vec{c}_n||_2}{\sqrt{2}} \sqrt{M^{(c)}(n)}. \tag{4.34}$$

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

$$c_i = \sum_{x \in \mathcal{X}} \left| \log \frac{\hat{p}^{(c)}_{X_i}(x)}{\hat{p}^{(r)}_{X_i}(x)} \right| \tag{4.35}$$

A proof of the theorem may be found in Appendix C.5. 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:

$$\sup_{x \in \mathcal{X}} \left| \log \frac{\hat{p}^{(c)}_{X_i}(x)}{\hat{p}^{(r)}_{X_i}(x)} \right| \leq L \;\; i = 1, 2, \ldots \tag{4.36}$$

then we can simplify the bound by observing that:

$$||\vec{c}_n||_2 \leq L|\mathcal{X}|\sqrt{n}. \tag{4.37}$$

When such a scenario holds, we can make use of the following Corollary to Theorem 5 regarding the asymptotic behavior of the causality regret:

Corollary 2. *Let the worst case regret for the predictors* $\hat{p}^{(r)}_{X_i}$ *and* $\hat{p}^{(c)}_{X_i}$ *be sublinear in n and the absolute log ratio of the complete and restricted sequential predictors be bounded as in* (4.36)*. 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:*

$$\lim_{n \to \infty} \frac{1}{n} CR(n) = 0 \tag{4.38}$$

Lastly, we note that in the special case where the true complete and restricted distributions are in the reference classes (i.e. $p^{(r)}_{X_i} \in \tilde{\mathcal{P}}^{(r)}$ and $p^{(c)}_{X_i} \in \tilde{\mathcal{P}}^{(c)}$), then under an appropriately modified

Assumption 2 with $p^{(c)}_{X_i}$ and $p^{(r)}_{X_i}$ substituted for $p^{(c)*}_{X_i}$ and $p^{(r)*}_{X_i}$, we have that:

$$\sum_{i=1}^{n} \left| \hat{C}_{Y\to X}(i) - C_{Y\to X}(i) \right| \leq CR(n). \tag{4.39}$$

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 4.5.1.

4.4.1 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 $C^*_{Y\to X}$ serves as a useful proxy for the true causal measure $C_{Y\to X}$. This consideration is not unique to the proposed causal measure. In an extensive analysis of problems encountered when using Granger causality, [115] 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 $\tilde{\mathcal{P}}^{(r)}_n$ and $\tilde{\mathcal{P}}^{(c)}_n$, which need to be rich enough to yield sufficiently good $C^*_{Y\to X}$ but not so rich that there do not exist sequential prediction methods for which low cumulative regret may be achieved.

This issue is addressed by building upon the notion of partial directed information (PDI) introduced in the previous chapter. Specifically, we note that by giving a restricted predictor access to a "stale" history, then the desired Markov properties hold:

Theorem 6. *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:*

$$p(X_i \mid X^{i-1}, Y^{i-k}) = p(X_i \mid X^{i-1}_{i-k-d}, Y^{i-k}_{i-k-d}). \tag{4.40}$$

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 C.5.

Define the partial history with lag (or staleness) k to be:

$$\mathcal{H}_i^{(k)} \triangleq \mathcal{H}_i^{(r)} \cup \{y_1, \ldots, y_{i-k}\}. \tag{4.41}$$

Similarly, define the partial conditional distribution:

$$p_{X_i}^{(k)}(x_i) \triangleq p(x_i \mid \mathcal{H}_i^{(k)}). \tag{4.42}$$

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 can define a partial causal measure with lag k to be:

$$C_{Y \to X}^{(k)}(\mathcal{H}_i^{(c)}) \triangleq D(p_{X_i}^{(c)} \mid\mid p_{X_i}^{(k)}) \tag{4.43}$$

This sample path dependent measure can be related to the PDI introduced in the previous section by noting that:

$$I_P^{(k)}(Y^n \to X^n) = \sum_{i=1}^{n} \mathbb{E}\left[C_{Y \to X}^{(k)}(\mathcal{H}_i^{(c)})\right] \tag{4.44}$$

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

$$I_P^{(k)}(Y^n \to X^n) = H(X^n \mid\mid Y^{n-k-1}) - H(X^n \mid\mid Y^{n-1}) \tag{4.45}$$

$$\bar{I}_P^{(k)}(Y \to X) = \bar{H}^{(k-1)}(X \mid\mid Y) - \bar{H}^{(1)}(X \mid\mid Y) \tag{4.46}$$

where we have replaced the first entropy terms on the right hand side of (3.8) and (3.10) with a lagged causally conditioned entropy.

These partial measures have straightforward interpretations in relationship to their com-

plete counterparts. In particular, we note that these measures can be viewed as measuring the causal effect of the *recent* past of Y on X. 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 [55] can be extended to be universal estimators of the PDI rate:

Theorem 7. *Let $(X,Y) \sim p$ be a jointly stationary irreducible aperiodic finite-alphabet Markov process of order d or less. Let $\hat{p}^{(c)}_{X_i}$ be a depth-d CTW estimate of $p^{(c)}_{X_i}$ with access to $\mathcal{H}_i^{(c)}$ and $\hat{p}^{(k)}_{X_i}$ be a depth-$(d+k)$ CTW estimate of $p^{(k)}_{X_i}$ with access to $\mathcal{H}_i^{(k)}$. Then:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} D(\hat{p}^{(c)}_{X_i} \,||\, \hat{p}^{(k)}_{X_i}) = \bar{I}_P^{(k)}(Y \to X) \;\; p_{X^n,Y^n} - a.s. \tag{4.47}$$

The proof of the theorem may be found in Appendix C.5. We note that this theorem is analogous with Theorem 3 from [55], 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 [55]. Presumably, similar results could be obtained for the other estimators, though this is not the primary concern of this work.

4.5 Results

4.5.1 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 [127]. For the purpose of estimating either the complete conditional distribution $p^{(c)}_{X_i}$, or partial conditional distribution $p^{(k)}_{X_i}$, we utilize a CTW with side information as in [18]. 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 ([119, 55])**.** *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:

$$\sup_{x^n} \log \frac{p(x^n)}{\hat{p}(x^n)} \leq \frac{(|\mathcal{X}|-1)L}{2} \log \frac{n}{L} + L\left(\frac{|\mathcal{X}|}{|\mathcal{X}|-1} + \log|\mathcal{X}|\right) - \frac{1}{|\mathcal{X}|-1} \tag{4.48}$$

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 2. *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:*

$$\sup_{x^n, y^n} \log \frac{p(x^n \,||\, y^n)}{\hat{p}(x^n \,||\, y^n)} \leq \frac{(|\mathcal{X}|-1)L}{2} \log \frac{n}{L} + L(|\mathcal{X}|-1) + S \tag{4.49}$$

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 C.4. Using the above lemma and proposition we can compute the values of the causality regret bounds by using (4.48) and (4.49) for $M^{(r)}(n)$ and $M^{(c)}(n)$, respectively, in (4.34). In the following sections we compare the causal regret bound with the true estimation accuracy for three scenarios.

Independent Processes

Let X and Y be independent ternary processes, with each process being stationary first order Markov. As such, the processes are completely characterized by the probabilities $p(x_i \mid x_{i-1})$ and $p(y_i \mid y_{i-1})$ 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 $i = 1,2,\ldots, C_{Y\to X}(i) = C_{X\to Y}(i) = 0$.

Figure 4.2 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 (4.39), 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 (4.48), and $L = 9$ and $S = 10$ in (4.49).

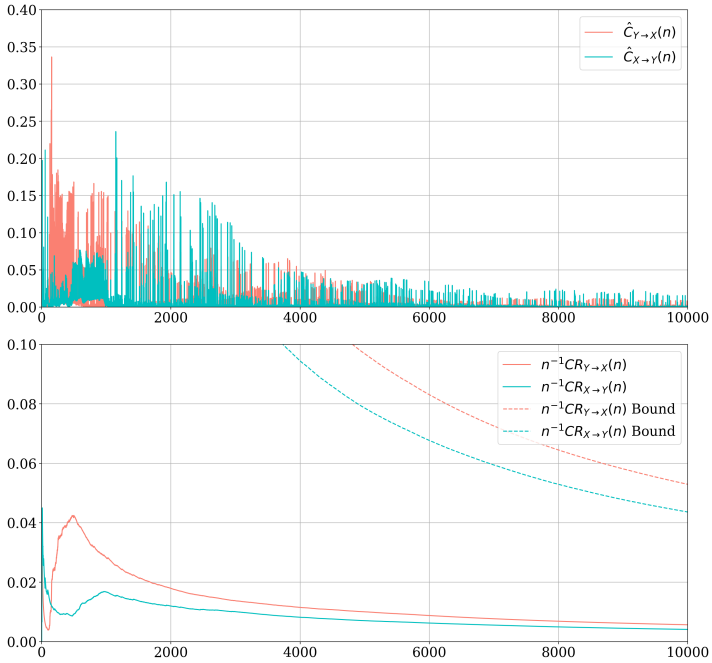

Figure 4.2: Top - Estimates of causal measure in each direction for independent processes. Bottom - Normalized cumulative absolute error of estimates (solid) and normalized causality regret bounds (dashed).

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 conditions described in Chapter 3 for any parameterization of the probabilities in question.

Figure 4.3 shows the true causal measure $C_{Y\to X}$ alongside the estimates $\hat{C}_{Y\to X}$ and $\hat{C}_{X\to Y}$. For clarity, only the last 100 time points are shown. We can see that in this time window the estimated $\hat{C}_{Y\to X}$ tracks the true causal measure $C_{Y\to X}$ very well, and the estimated $\hat{C}_{X\to Y}$ has converged to 0 as desired. In the bottom panel we see that, because the causal measure $\hat{C}_{X\to Y}(i) < \hat{C}_{Y\to X}(i)$, c_i in equation (4.35) is much smaller in the $X \to Y$ direction and thus the causal regret bound is considerably tighter. This is consistent with the result obtained in [63] that plug-in estimators of the DI rate will converge at a faster rate if the DI rate is zero.

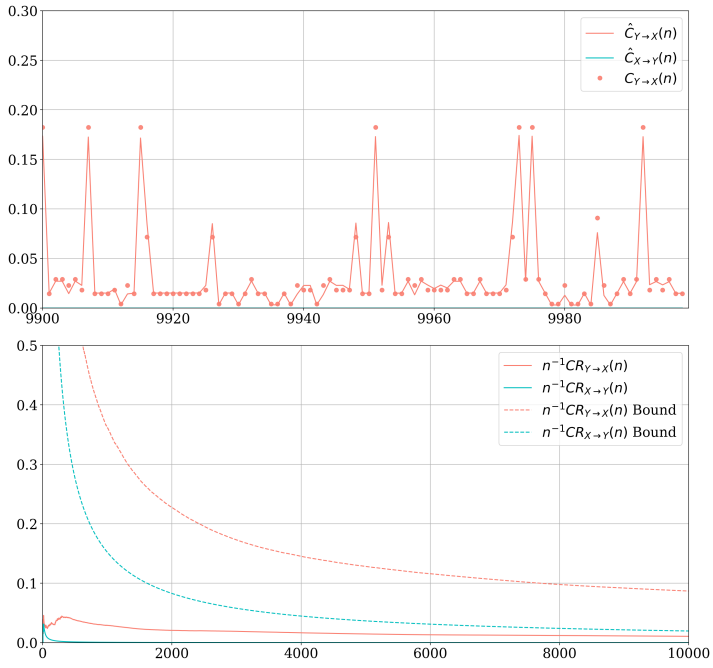

Figure 4.3: Top - Estimates of causal measure in each direction for unidirectional influences. Bottom - Normalized cumulative absolute error of estimate (solid) and normalized causality regret bounds (dashed).

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 $p(x_i \mid x_{i-1}, y_{i-1})$ and $p(y_i \mid x_{i-1}, y_{i-1})$.

Figure 4.4 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 $n \to \infty$. 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 $C_{Y \to X}(i)$ such settings. Details can be found in Appendix C.2.

To address the estimation bias seen in Figure 4.4, we consider the partial causal measure $C_{Y \to X}^{(k)}(i)$ defined by (4.43). Figure 4.5 shows an estimate of the partial causal measure on the same sequence considered in Figure 4.4 with a staleness of $k = 1$. The bottom panel of Figure 4.5 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 (4.49) 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 4.5 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

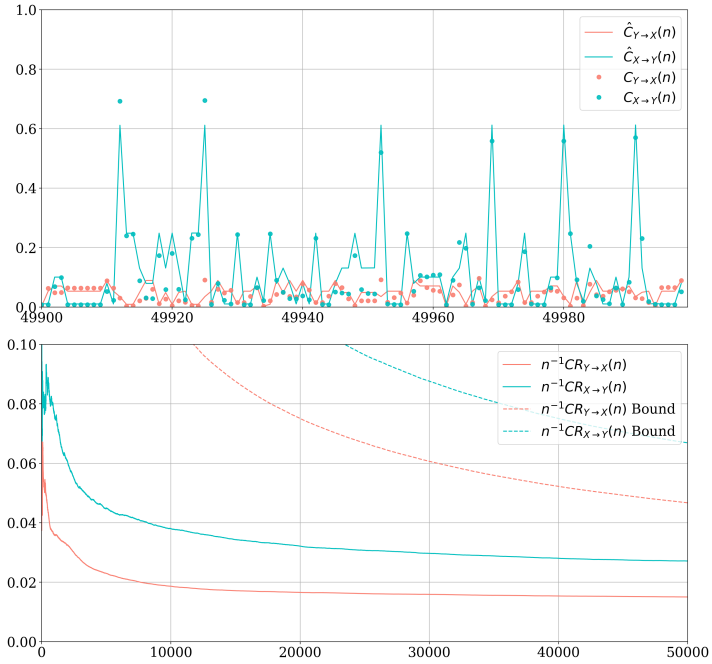

Figure 4.4: Top - Estimates of causal measure in each direction for bidirectional influences. Bottom - Normalized cumulative absolute error of estimate (solid) and normalized causality regret bounds (dashed).

estimate is unbiased.

Given that the same sequences were used in generating Figures 4.4 and 4.5, 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.

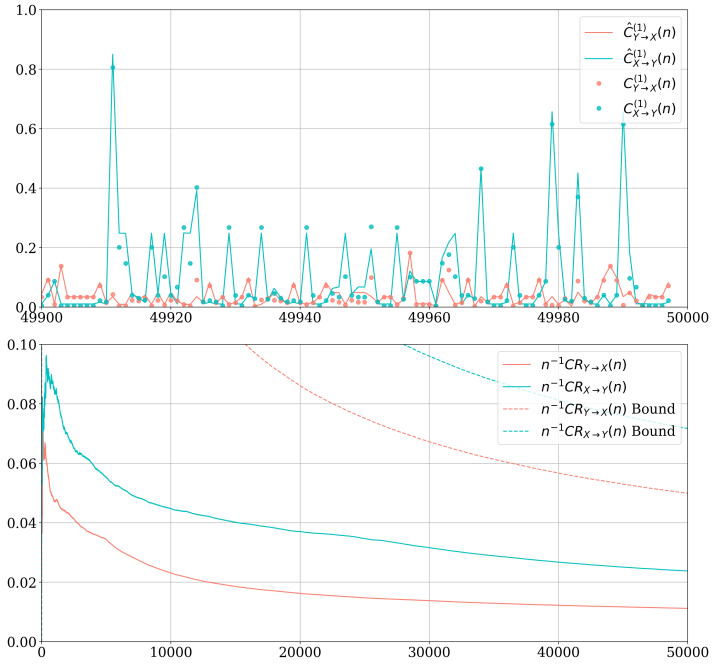

Figure 4.5: Top - Estimates of the partial causal measure with $k = 1$ in each direction for bidirectional influences. Bottom - Normalized cumulative absolute error of estimates (solid) and normalized causality regret bounds (dashed).

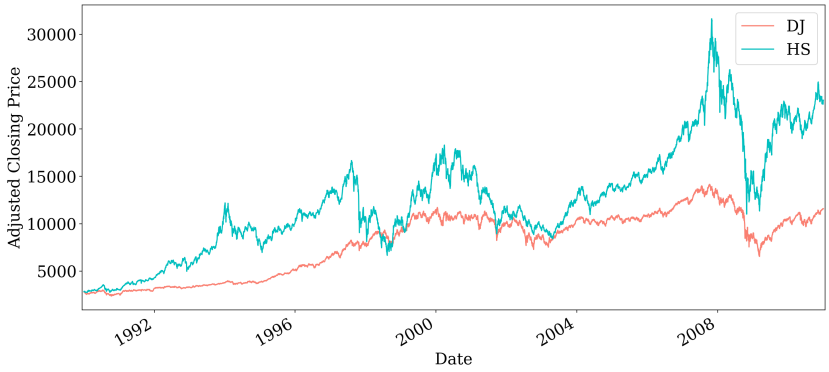

Figure 4.6: The Dow Jones (DJ) Industrial Average and Hang Seng (HS) indices.

4.5.2 Stock Market 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 [55]. In [55] it was

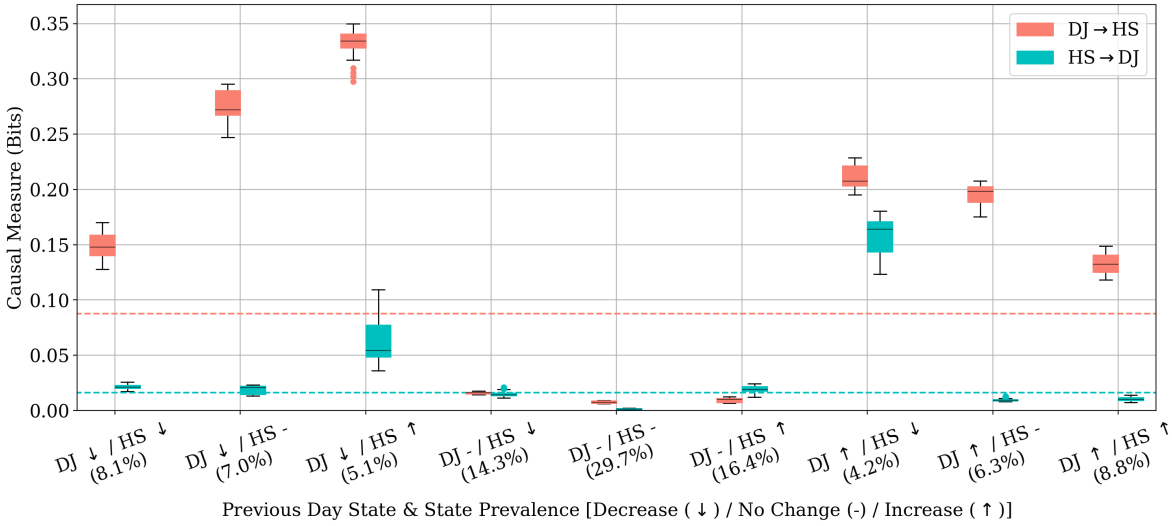

Figure 4.7: Causal measure between stock indices for different previous day states from 2008 to 2011. Below each of the 9 possible previous day states we include the percentage of days in which that state occurs. The dashed lines represent the DI estimate.

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 4.6.

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 [55]. Thus, the HS is affected by the same day DJ, while the DJ is affected by the previous day HS.

Figure 4.7 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 4.7. 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 4. *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 [61]).*

4.6 Discussion

The concepts presented in this chapter can be distilled to two 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 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.

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 variables affect another group of random variables in a given directed graph? We believe the philosophy presented in this chapter 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 influences and graph structures when only a subset of processes may be observed [40, 36, 83]. 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 4.4.1 and its application in quantifying information leakage for

coupled systems with delayed information [44], providing fundamental performance limitations of closed-loop systems [81] subject to delay constraints, or in characterizing rate-performance tradeoffs [116] for network control problems with non-classical information structures [43, 67] 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 4.5.2. It is shown in [37] 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.

4.7 Acknowledgements

Chapter 4, in part, is currently under review for publication of the material. Schamberg, Gabriel; Coleman, Todd. The dissertation author was the primary investigator and author of this material.

Chapter 5

An Information Theoretic Perspective on Direct and Indirect Effects

5.1 Introduction

Consider a directed acyclic graph (DAG), where nodes represent random variables and edges represent a direct causal influence between two variables. We here discuss the problem of *quantifying* these causal influences. Considerable attention has been paid to this problem in a variety of communities; for the sake of exposition, we here coarsely categorize methods as either statistical (i.e. those summarized by [93]) or information theoretic (IT) [54, 10, 38, 98]. When viewed from an applications perspective, these two approaches are quite different. Statistical approaches are common in epidemiology and medicine, whereas IT methods appear in the study of complex natural systems, for example climatological [49] or neuroscientific [60]. While this disparity makes sense given the fundamentally different perspectives employed by the two approaches, the difference in perspectives is not well presented in the development of IT methodologies.

To illustrate the philosophical differences between these two approaches, consider a

simple example with a two node graph $X \to Y$, where $X \in \{0,1\}$ represents whether or not an individual has won the lottery and $Y \in \mathbb{R}$ represents that individual's average monthly spending (assume for clarity that there are no confounding factors). A statistical measure such as the average causal effect (ACE) [51, 92] would seek to answer the question "What is the effect of *winning* the lottery on spending?" by comparing the average spending of lottery winners ($X = 1$) against the average spending of lottery non-winners ($X = 0$): $\mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0]$. We would of course expect this to be quite large. It is important to note that the ACE is defined irrespective of the marginal distribution of X, meaning that the probability with which x occurs has no bearing on the effect of x on Y. An IT approach addresses a subtly different question: "What is the effect of the lottery on spending?" In other words, an IT measure considers the effect of the random variable representing whether or not one wins the lottery on spending. Specifically, the effect of X on Y would be given by the mutual information (MI), $I(X;Y)$ (see [54, Sec. 2, P1]). Using a simple IT inequality, we get $I(X;Y) \leq H(X) \approx 0$. In words, because so few people win the lottery, an IT measure indicates that the lottery has a negligible effect on spending. In other words, the statistical measures consider the effect of a *specific cause*, whereas IT measures consider the effect at a *systemic level*.

A second difference is that, whereas statistical approaches measure causal effects on the *value* of an outcome, IT approaches measure the causal effect on the *distribution* of an outcome. Each of these comes with benefits and drawbacks. With statistical approaches, the units are preserved (in the previous example, the units of the ACE are dollars). While IT measures yield the less interpretable unit of bits, they are able capture more complex causal effects, for instance the effect that a variable has on the variance of another. Acknowledging this difference helps to understand the disparity between the applications of statistical and IT measures. When evaluating the causal link between smoking and cancer, the number of bits of information shared by the smoking and cancer variables is not of great use. However, when studying the nature of complex natural networks, it may be desirable to use a measure that can capture higher order causal effects.

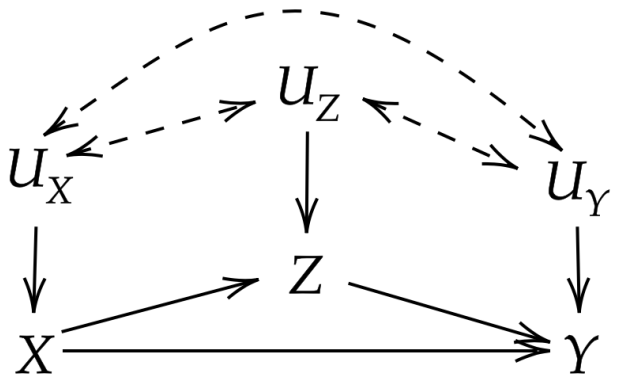

Figure 5.1: DAG $\mathcal{G}$ representing a mediation model

In the present work we seek to resolve the first difference by endowing IT measures with the ability to measure *specific* causal effects. Furthermore, we show that existing IT measures of causal influences are ill-equipped for distinguishing direct and indirect effects. Following a parallel storyline to that of Pearl [91], we will provide measures of the total, (natural and controlled) direct, and natural indirect effects. We will show that these measures do not fundamentally change the underlying IT perspective on causality, but enable obtaining "higher resolution" measures of causal influence. In doing so, we will provide increased clarity to the aforementioned differences between IT and statistical causal measures.

5.2 Preliminaries

5.2.1 Notation and Problem Setup

Throughout this chapter we will be developing techniques for measuring the causal influence of $X \in \mathcal{X}$ upon $Y \in \mathcal{Y}$ in the presence of a mediating variable $Z \in \mathcal{Z}$ using the DAG $\mathcal{G}$ depicted in Figure 5.1. Without loss of generality, Z may represent a collection $(Z_1, Z_2, \ldots, Z_k) \in \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_k = \mathcal{Z}$ of all mediating variables. We define the parent sets as $PA_X = U_X$, $PA_Z = \{X\} \cup U_Z$, and $PA_Y = \{X, Z\} \cup U_Y$, where dashed double headed arrows are used to indicate

unknown dependencies between $U_X \in \mathcal{U}_X$, $U_Y \in \mathcal{U}_Y$, and $U_Z \in \mathcal{U}_Z$ (including the possibility of $U_S \cap U_T \neq \emptyset$ for $S,T \in \{X,Y,Z\}$). We may use the shorthand $U = U_X \cup U_Y \cup U_Z \in \mathcal{U}$. For simplicity, we will assume that all variables are discrete with arbitrary finite supports, though extending the proposed methods to continuous or mixed random variables is straightforward. In general, we let p be the probability mass function (pmf) for all variables in the graph (i.e. $X,Y,Z,U \sim p$), capital letters represent random variables, and lowercase letters represent their realizations. For example, $p(x \mid pa_X)$ gives the marginal probability of the event $X = x$ given that its parents took on values pa_X. We further assume that p satisfies the causal Markov condition with respect to $\mathcal{G}$ [92] , with $p(x,y,z,u) = p(u)p(x \mid u_X)p(z \mid x,u_Z)p(y \mid x,z,u_Y)$. We use a hat to indicate the *do*-operator, which represents taking the action of forcing a variable to assume a particular value by means of intervention. For example, $p(y \mid \hat{z}) = p(y \mid do(Z = z))$ gives the probability of y given that Z is forced to take the value z, irrespective of the probability with which that value occurs. When working with distributions utilizing the *do*-operator, a set of rules known as the *do*-calculus can be used to identify if and how the interventional distributions correspond to observational distributions that do not utilize the interventions. While the reader is referred to [92, Sec. 3.4] for the complete *do*-calculus, we provide a description of the rule which enables swapping interventions for observations in Appendix D.1.

We conclude this section with definitions of information theoretic quantities that will play a central role. The entropy of a random variable Y and conditional entropy of Y given X are given by $H(Y) = -\sum_y p(y) \log p(y)$ and $H(Y \mid X) = -\sum_{x,y} p(x,y) \log p(y \mid x)$. It is worth noting that the conditional entropy yields the *expected* uncertainty of Y given X, and is not to be confused with $H(Y \mid X = x) = -\sum_y p(y) \log p(y \mid x)$, which gives the uncertainty of Y when conditioning on a *particular value* of X. For two distributions p and q over $\mathcal{Y}$, the KL-divergence from p to q represents the excess number of bits needed to represent Y if the distribution is assumed to be q when it is in fact p, and is given by $D(p(Y) \mid\mid q(Y)) = \sum_y p(y) \log {}^{p(y)}/_{q(y)}$. The KL-divergence is zero if and only if $p(y) = q(y)$ for all y, and is deemed infinite if there exists a y such that $p(y) > 0$

and $q(y) = 0$. We use $Bern(\alpha)$ to represent the distribution of a Bernoulli random variable with parameter $0 < \alpha < 1$. For the KL divergence between two Bernoulli random variables with parameters α and β, we will use the shorthand $D(\alpha \,||\, \beta)$. Finally, the mutual information (MI) between X and Y is given by $I(X;Y) = H(Y) - H(Y \mid X) = \mathbb{E}[D(p(Y \mid X) \,||\, p(Y))]$, where the expectation is taken with respect to X. These equivalent definitions of MI give rise to two interpretations: (i) the average reduction in uncertainty in Y obtained by conditioning on X and (ii) the average increased ability to predict Y resulting from conditioning on X. It is worth noting that (barring some technical details), these definitions can be applied to continuous valued random variables by substituting integrals for sums and probability density functions for pmfs.

5.2.2 Direct and Indirect Effects

Building upon the work of Robins and Greenland [103], Pearl [91] formalized definitions of direct and indirect effects in the context of graphical models. Such a distinction is useful in disentangling the mechanisms via which causal influences occur[1]. A canonical example is presented by [48], wherein a birth control pill is suspected of directly increasing the likelihood of thrombosis in women, while simultaneously reducing thrombosis through its prevention of pregnancy (which is positively linked to thrombosis). In each of Pearl's definitions, the magnitude of the causal effect is specified for a specific value x and is measured with respect to a reference (or baseline) value x^*. The simplest of these measures is the total effect (TE) of $X = x$ on Y given by $\mathbb{E}[Y \mid \hat{x}] - \mathbb{E}[Y \mid \hat{x}^*]$. The TE yields the answer to a very concise causal question, namely "How much would we expect the value of Y to change if we were to change X from x^* to x?" As indicated by the name, the TE does not distinguish effects that x has on Y directly from those that occur via a mediating variable Z. As such, Pearl proceeds to define the controlled direct effect (CDE) of x on Y with mediator z as $\mathbb{E}[Y \mid \hat{x}^*, \hat{z}] - \mathbb{E}[Y \mid \hat{x}, \hat{z}]$. Once again, this measure addresses a

[1]The present discussion only scratches the surface of the extensive field of mediation analysis. For a more in depth summary the reader is referred to [93, Sec. 5.1] and citations therein.

clear causal question: "How much would we expect the value of Y to change if we were to change X from x^* to x, but kept Z at a fixed value z?" While this is an intuitive notion of direct effect, it is important to note that it requires the intervention $do(Z = z)$. Given that it may be of interest to know the direct effect that occurs when the mediating variable is *not* controlled for, Pearl defines the *natural* direct effect (NDE) as $\mathbb{E}[Y \mid \hat{x}, Z_{x^*}] - \mathbb{E}[Y \mid \hat{x}^*]$, where Z_{x^*} is the value Z would have taken had X been x^*. This notion of simultaneously assigning a value to X and allowing Z to take the value it would under a different X is central to the measurement of indirect effects. As such, Pearl defines the natural indirect effect (NIE) as $\mathbb{E}[Y \mid \hat{x}^*] - \mathbb{E}[Y \mid \hat{x}^*, Z_x]$. In words, the natural indirect effect represents the expected change in Y resulting from changing Z from the value it would take under x to the value it takes under x^* while leaving X fixed at x^*. Next we will show how this systematic decomposition of causal effects is absent from the IT literature.

5.2.3 Information Theoretic Notions of Causal Influence

While there is a considerable body of work on the development of IT techniques for measuring causal influence, we here focus on information flow [10] and causal strength [54].

Information Flow

Drawing on the relationship between mutual information and statistical dependence, Ay and Polani [10] define an IT notion of *causal independence*, which unlike mutual information, is directed. Their definitions rely heavily on the post-interventional distribution, which dictates a truncated factorization of a joint distribution in the presence of interventions. We start by considering the information flow (IF) from X to Y, which is defined as:

$$I(X \to Y) \triangleq \sum_x p(x) \sum_y p(y \mid \hat{x}) \log \frac{p(y \mid \hat{x})}{\sum_{x'} p(x') p(y \mid \hat{x}')} \tag{5.1}$$

We can see that if all the hats are removed from the above equation, then the standard mutual information is recovered. By using these post-interventional distributions, however, all "upstream" dependencies of X are ignored, and thus any relationship between X and Y resulting from confounding variables is removed. Ay and Polani also define a conditional version of IF. Using the mediation model in Figure 5.1, let V be some subset of remaining variables in the graph, i.e. $V \subseteq U \cup \{Z\}$. The IF from X to Y imposing V is then given by:

$$I(X \to Y \mid \hat{V}) \triangleq \sum_{v} p(v) \sum_{x} p(x \mid \hat{v}) \sum_{y} p(y \mid \hat{x}, \hat{v}) \log \frac{p(y \mid \hat{x}, \hat{v})}{\sum_{x'} p(x' \mid \hat{v}) p(y \mid \hat{x}', \hat{v})} \tag{5.2}$$

Noting that V always appears as an intervention, the conditional IF can be interpreted as representing the IF from X to Y when the value of V is controlled. The IF can be extended to measure the flow to and from *sets* of nodes, though at present we only consider the flow from X to Y. IF is not to be confused with the directed information of Massey [82], which does not employ any interventional methods and is only used in the context of time series.

Within the IF framework, we can treat $I(X \to Y)$ as a measure of the total effect of X on Y and $I(X \to Y \mid \hat{Z})$ as a measure of controlled direct effect. While these measures are intuitively analogous to the measures in [91], it is difficult to formalize the nature of this analogy because we cannot formulate IF measures as the answer to a concise causal question similar to those of the previous section. Furthermore, because the conditional version of IF represents *controlling* a set of variables, IF offers no way to measure the *natural* direct and indirect effects proposed by Pearl.

Causal Strength

The causal strength (CS) measure proposed by Janzing et al. [54] takes a slightly different approach in that it measures the strength of specific edges in a DAG. We call this an "edge-centric" perspective, in contrast with the "node-centric" perspective used by IF. To motivate the definition of CS, the authors propose a collection of five postulates that they argue ought to be satisfied by

measures of CS. Janzing et al. acknowledge that their postulates need not apply to all reasonable measures of causal influence; as such, any present criticisms of CS can be attributed to differences in the problem formulation. The postulates are briefly summarized here, and the reader is referred to [54] for more thorough definitions: **(P0)** If the CS of an arrow is zero, then that arrow should be able to be removed from the DAG without breaking the causal Markov condition. **(P1)** If the entire DAG is given by $X \to Y$, then the CS is $I(X;Y)$. **(P2)** The strength of an arrow $X \to Y$ should be defined locally, i.e. it should depend only upon the distributions $p(y \mid pa_Y)$ and $p(pa_Y)$. **(P3)** The CS of an arrow $X \to Y$ should be at least the conditional mutual information $I(X;Y \mid PA_Y \setminus \{X\})$. **(P4)** If the CS of a set of edges is zero, then the CS of all subsets of those edges should be zero.

Janzing et al. [54] proceed to propose a measure of CS that satisfies these postulates. Central to their CS measure is the post-cutting distribution. Formally, let $V = \{V_1, \ldots, V_n\}$ be the nodes in a graph, PA_j^S be the subset of parents of V_j for which $V_i \to V_j \in S$, and $PA_j^{\bar{S}} = PA_j \setminus PA_j^S$. Then the post-cutting distribution is given by:

$$p_S(v_1, \ldots, v_n) = \prod_j \left[\sum_{pa_j^S} p(v_j \mid pa_j^{\bar{S}}, pa_j^S) \left(\prod_{v \in pa_j^S} p(v) \right) \right] \tag{5.3}$$

We can see that the post-cutting distribution factorizes much like the joint distribution p– however, for nodes at the receiving end of an edge in S, they are fed the *marginal distribution* of the node at the other end, rather than the actual value of that node. Using the post-cutting distribution, the CS of a set of edges S is then given by $\mathfrak{C}_S = D(p \mid\mid p_S)$, and thus provides a measure of how much excess information is needed to accommodate the severed edges. Once again, the reader is referred to [54] for further intuition.

Consider CS in the context of the mediation model in Figure 5.1, i.e. $D(p(X,Y,Z,U) \mid\mid p_S(X,Y,Z,U))$ for some set of edges $S \subseteq \{X \to Y, X \to Z, Z \to Y\}$. Within the constraints of the CS framework, one might seek to measure the total, direct, and indirect effects as the strength of the edge sets $S_{TE} = \{X \to Y, X \to Z, Z \to Y\}$, $S_{DE} = \{X \to Y\}$, and $S_{IE} = \{X \to Z, Z \to Y\}$,

respectively. To see why this is insufficient, consider an extreme case of the birth control pill example above, where the indirect and direct effects of X on Y are perfectly complementary such that for all $x_1, x_2 \in \mathcal{X}$ and $y \in \mathcal{Y}$, $p(y \mid \hat{x}_1) = p(y \mid \hat{x}_2)$. Any reasonable measure of total effect will conclude that no value of x has an effect on Y – however, note that from postulate **(P4)**, the total effect (as we have defined it in the CS framework) must be non-zero if either the direct or indirect effect is non-zero. A similar example can be constructed for the insufficiency of CS as a measure of indirect effects by having the effect of X on Z be canceled out by the effect of Z on Y. Finally, we note that CS is similar to IF in that it does not yield a clear causal question for which it gives the answer. This is perhaps justified by the decision to define a set of formal postulates that are used to link the properties of CS with our intuitions – however, given that causal influences are likely to be measured in order to obtain a better understanding of the system under study, we find it to be of great practical use to pair causal measures with an easily interpretable causal question for which the measure provides an answer. We will now show that this can be achieved by considering the causal effect of specific values of x.

5.3 Novel Information Theoretic Causal Measures

The observation that the MI $I(X;Y)$ does not capture how different values of x may contain different amounts of information about Y has been made in a variety of contexts throughout the literature, including experimental design [71, 26], neural stimulus response [29], measuring surprise [53], and most recently, distinguishing between information transfer and information copying [62]. Central to each of these works is the development of a notion of MI for a *specific value* of x, i.e. $I(x;Y)$. There is, however, no inherent $I(x;Y)$ implied by the definition of $I(X;Y)$ – to see this, we use the notation of [29] and provide two candidate definitions of $I(x;Y)$ based on

the two definitions of $I(X;Y)$ provided in Section 5.2.1:

$$I_1(x;Y) = D(p(Y \mid X = x) \mid\mid p(Y)) \tag{5.4}$$

$$I_2(x;Y) = H(Y \mid X = x) - H(Y) \tag{5.5}$$

It is well understood that, in general, $I_1(x;Y) \neq I_2(x;Y)$. This is clear to see by simply noting that, for any joint distribution $X,Y \sim p$, $I_1(x;Y) \geq 0$ for all x, whereas it is possible to have $I_2(x;Y) < 0$. In words, the knowledge of a specific value of x will only provide us with a more accurate distribution of Y ($I_1 \geq 0$), though it is possible for this distribution to have a greater entropy than the marginal distribution ($I_2 < 0$). We here use I_1 as a foundation for establishing value specific measures of causal influence, and, using the terminology of [62], refer to it as the *specific mutual information* (SMI). Building upon this language in the present context, we refer to these measures as measuring *specific causal effects*. To our knowledge, the use of SMI in the context of quantifying causal influence is novel – as such, we begin with an informal discussion around the use of SMI for the quantification of causal influence in two-node DAGs, followed by a formal definition of various specific causal effects in a mediation model.

5.3.1 Specific Mutual Information in Two-Node DAGs

Consider a DAG $X \to Y$ with joint distribution over nodes $X,Y \sim p$, and for the sake of exposition, assume there are no confounding variables. In this simple scenario, when considering the effect of X on Y, we can freely exchange interventions for observations (assuming we only consider x s.t. $p(x) > 0$), and thus the average causal effect of x with respect to baseline x^* is given by $\mathbb{E}[Y \mid \hat{x}] - \mathbb{E}[Y \mid \hat{x}^*] = \mathbb{E}[Y \mid x] - \mathbb{E}[Y \mid x^*]$. Once again, this addresses the question of how much the value of Y is expected to change as a result of switching from x^* to x. With regard to the CS and IF methods discussed above, both would quantify the effect of X on Y as the $I(X;Y)$.

Now consider using the SMI $I_1(x;Y)$ as a measure of the specific causal influence of x

upon Y, and note the following observations:

(I) We have the equivalence $I(X;Y) = \mathbb{E}[I_1(X;Y)]$, where the expectation is taken with respect to X. As such, we can think of the specific causal effect as a *random variable*, whose expectation is the mutual information. In doing so, we are able to capture that different values of x may have different *magnitudes* of causal effect on Y, with each of those effects occurring with some probability according to $p(x)$. Moreover, this makes clear that the perspective adopted here is consistent with that of other IT measures.

(II) $I_1(x;Y)$ is non-negative for all $x \in \mathcal{X}$. Whereas a negative ACE has the clear interpretation of x causing a decrease in the expected value of Y, we are measuring influences that x has on *the distribution of* Y. Given that there is no obvious notion of a (potentially negative) difference between distributions, we utilize a definition that results in any causal effect yielding a positive magnitude. This comes at the cost of foregoing the ability to differentiate negative and positive causal influences in the sense of the ACE. This further serves as a partial justification for using I_1, rather than I_2, as a foundation.

(III) The SMI does not require specifying a reference value x^*. Instead, we can view SMI as measuring the causal effect of x as compared with the X that would have occurred naturally. This suggests an intuition for why IT measures commonly appear in the measurement of causal influences in complex natural networks – values of x that are seen as *changing the course of nature* will be assigned a large causal influence. Given that we can (in this setting) exchange observation for intervention, we can view the SMI as comparing the effect of an intervention $\hat{x}$ with a random (i.e. non-atomic) intervention $\hat{X}$ with $X \sim p$ (see [92, 97] for discussions on random interventions).

(IV) The SMI addresses a very clear causal question: "How much different would we expect the distribution of Y to be if, instead of forcing X to take the value x, we let X take on a value naturally?" Stated more compactly: "How much would we expect performing the intervention $do(X = x)$ to change the course of nature for Y?"

(V) We can interpret the SMI as comparing a ground truth distribution conditioned on the actually occurring value x ($p(Y \mid x)$) with a counterfactual distribution wherein nature was allowed to run its course ($p(Y)$). This works well with the interpretation of the KL-divergence as a measure of excess bits resulting from encoding Y using the distribution that is not the true distribution from which Y is sampled. The use of the KL-divergence is further justified in this context by the fact that the logarithmic loss is unique in its ability to capture the benefit of conditioning on X in the prediction of Y [56].

(VI) Finally, we note that $I_1(x;Y) = 0$ if and only if $p(y \mid x) = p(y)$ for all y. By contrast, it is possible to have $I_2(x;Y) = 0$ and $p(y \mid x) \neq p(y)$. To illustrate why this is not desirable, consider the following example:

Example 5.3.1. *Consider a two-node DAG $X \to Y$ with $X \sim Bern(1/7)$, $Y \mid X = 0 \sim Bern(1/10)$, and $Y \mid X = 1 \sim Bern(8/10)$. It is clear that the distribution of Y is highly dependent upon the value of X. Next note that $Y \sim Bern(p_1)$, where:*

$$p_1 = \frac{1}{7} \cdot \frac{8}{10} + \frac{6}{7} \cdot \frac{1}{10} = \frac{2}{10} \tag{5.6}$$

Thus, we can see that $H(Y) = H(Y \mid X = 1)$ and thus $I_2(X = 1;Y) = 0$. On the other hand, we have $I_1(X = 1;Y) = D(8/10 \,||\, 2/10) = 1.2$ bits. This exemplifies how simply measuring differences in entropy is insufficient for capturing causal influences.

We will conclude this section by returning briefly to the lottery example discussed in the introduction, recalling that $X \in \{0,1\}$ represents whether or not an individual has won the lottery and $Y \in \mathbb{R}$ represents that individual's average spending. Given that virtually nobody wins the lottery, we have $p(X = 0) \approx 1$ and thus $p(y) = \sum_x p(y \mid x)p(x) \approx p(y \mid X = 0)$. As such, the specific causal effect of losing the lottery is $I_1(X = 0;Y) \approx 0$. By contrast, $p(y \mid X = 1) \neq p(y)$, and thus the specific causal effect of winning the lottery will be $I_1(X = 1;Y) \gg 0$. Framed in terms of the causal question discussed above in **(IV)**, we would expect forcing someone to win

the lottery to change the course of nature much more than forcing someone to not win the lottery.

5.3.2 Specific Causal Effects in the Mediation Model

Following the process of [91], we here formalize a series of definitions of total/direct/indirect causal influences from an information theoretic perspective. When leaving the comfort of the unconfounded two-node DAG, it is important to incorporate the notion of intervention directly into the definition of the causal measures:

Definition 3. *The specific total effect of x on Y is defined as:*

$$STE(x \to Y) \triangleq D(p(Y \mid \hat{x}) \mid\mid \sum_{x'} p(x')p(Y \mid \hat{x}')) \tag{5.7}$$

With the exception of the interventional notation, the STE is equivalent to the SMI. Note that for a DAG given by $X \to Y$, we will have $STE(x \to Y) = I_1(x;Y)$ but $STE(y \to X) = 0 \neq I_1(y;X) = D(p(X \mid y) \mid\mid p(X))$, where $STE(y \to X)$ represents the specific total effect of y on X.

Next we define the specific controlled direct effect (SCDE) of x on Y. Given that computing the controlled direct effect must be done by means of intervention on Z, we define the SCDE with respect to a specific value z, as it is unclear what distribution over Z should be used if the definition were to take an expectation over *all* possible values of z (see Theorem 9).

Definition 4. *The specific controlled direct effect of x on Y with mediator z is defined as:*

$$SCDE(x \to Y; z) \triangleq D(p(Y \mid \hat{x}, \hat{z}) \mid\mid \sum_{x'} p(x')p(Y \mid \hat{x}', \hat{z})) \tag{5.8}$$

The SCDE measures how much we would expect performing the intervention $do(X = x)$ to change the course of nature given that Z is held fixed at z. As mentioned in Section 5.2.2, computing the controlled direct effect involves intervening upon the mediating variable Z, and thus does not convey the direct effect that occurs naturally from fixing a value of X.

Next, the specific natural direct effect measures the direct effect of x on Y that occurs naturally when the mediator is not controlled for:

Definition 5. *The specific natural direct effect of x on Y is defined as:*

$$SNDE(x \to Y) \triangleq D(p(Y \mid \hat{x}) \mid\mid \sum_{x',z'} p(x')p(z' \mid \hat{x})p(Y \mid \hat{x}', z')) \tag{5.9}$$

It is helpful to dissect the two distributions of Y considered by the SNDE. Both distributions are given by a weighted combination of the distribution of Y conditioned upon intervened values of X and Z. In both cases, the intervened values Z are weighted by the probability with which they would occur under the intervention $\hat{x}$. For the intervened values of X, however, the first distribution uses the "ground truth" value x, whereas the second uses the "naturally occurring" x', weighted according to $p(x')$. Using the same logic, we can define a specific natural indirect effect:

Definition 6. *The specific natural indirect effect of x on Y is defined as:*

$$SNIE(x \to Y) \triangleq D(p(Y \mid \hat{x}) \mid\mid \sum_{x',z'} p(x')p(z' \mid \hat{x}')p(Y \mid \hat{x}, z')) \tag{5.10}$$

Conducting a similar dissection, we see that both arguments of the KL-divergence are given by a weighted average of the conditional distribution Y under the interventions $\hat{x}$ and $\hat{z}'$. The difference is that each value of z' is weighted by its probability with respect to the "ground truth" value x in the first distribution, while each value of z' is weighted by its probability with respect to the "naturally occurring" x' in the second distribution.

Unfortunately, the proposed definitions of SNDE and SNIE yield no obvious inequalities with respect to the STE (for example, $SNDE(x \to Y) + SNIE(x \to Y) \not\leq TE(x \to Y)$ in general). While this is initially unintuitive, it can be justified by the decision to have all causal influences be assigned a non-negative magnitude. As such, we would expect that contradictory indirect and

direct effects could individually have a large magnitude while still resulting in a total effect of zero.

5.3.3 Equivalence Relations

We briefly introduce two theorems relating the proposed specific measures to IF and CS.

Theorem 8. *The expected STE is equivalent to the information flow, i.e.* $\mathbb{E}[STE(X \to Y)] = I(X \to Y)$*, where the expectation is taken with respect to the marginal distribution over* X*.*

A proof is found in Appendix D.3.1. The above theorem shows that the expected STE recovers the standard (unconditional) IF from X to Y and follows directly from the definitions in (5.1) and (5.7). Notably, the expected STE is *not* equivalent to the CS associated with any subset of the arrows in the graph. Next, we show that both IF and CS provide a notion of expected SCDE:

Theorem 9. *The conditional IF is given by the expected value of the SCDE taken with respect to the* marginal *distributions of* X *and* Z*:*

$$I(X \to Y \mid \hat{Z}) = \sum_{x,z} p(x)p(z)SCDE(x \to Y; z) = \mathbb{E}_{p(X)p(Z)}[SCDE(X \to Y; Z)]$$

Furthermore, if the DAG consists of only X*,* Y*, and* Z *(i.e.* $U = \emptyset$*), then the CS of* $X \to Y$ *is given by the expected value of the SCDE taken with respect to the* joint *distribution of* X *and* Z*:*

$$\mathfrak{C}_{X \to Y} = \sum_{x,z} p(x,z)SCDE(x \to Y; z) = \mathbb{E}_{p(X,Z)}[SCDE(X \to Y; Z)]$$

A proof is found in Appendix D.3.2. This theorem clarifies the point made earlier with regard to the value of a measure of *natural* direct effect. In particular, when taking an average with respect to possible control values for the mediator Z, it is not clear what distribution over Z should be used.

5.3.4 Conditional Specific Influences

Even though the above causal measures are defined for specific values of X, they provide a notion of average causal influence in that they are implicitly averaging over all possible covariates U. Given that different values of u may significantly affect the nature of the relationship between x and Y, we define conditional versions of the above definitions for a specific value $U = u$. We here consider the general case where we can only observe a subset of the covariates $\tilde{U} \subseteq U$:

Definition 7. *The conditional STE of x on Y in setting $\tilde{u}$ is defined as:*

$$STE(x \rightarrow Y \mid \tilde{u}) \triangleq D(p(Y \mid \hat{x}, \tilde{u}) \mid\mid \sum_{x'} p(x' \mid \tilde{u}) p(Y \mid \hat{x}', \tilde{u})) \tag{5.11}$$

For the special case where we can observe all relevant covariates, i.e. $\tilde{U} = U$, the conditional STE can be simplified as:

$$STE(x \rightarrow Y \mid u) \triangleq D(p(Y \mid \hat{x}, u_Y, u_Z) \mid\mid \sum_{x'} p(x' \mid u_X) p(Y \mid \hat{x}', u_Y, u_Z)) \tag{5.12}$$

This definition violates the locality postulate **(P2)** of Janzing et al. [54] in that the causal effect of x on Y may be dependent upon how X is affected by *its own parents*. Allowing this is, however, consistent with the perspective that IT measures quantify the deviance from the course of nature in that the value u dictates the current *natural state*. Nevertheless, the terms $p(x' \mid \tilde{u})$ and $p(x' \mid u_X)$ can be replaced with $p(x')$ if one wishes to remain faithful to the locality postulate. The conditional versions of SCDE, SNDE, and SNIE follow very similar logic to that of the STE, and can be found in Appendix D.2.

5.3.5 Identifiability

When U is partially observable or unobservable, the nature of the dependence relationships between U_X, U_Y, and U_Z will dictate the ability to estimate the proposed causal measures from

observational data – more specifically, the ability to determine the interventional distributions given only estimated conditional distributions. This is crucially important given that performing interventions in many complex natural systems is infeasible. The following theorem states when the conditional specific measures can be estimated in the partially observable setting where only $\tilde{U} \subset U$ can be observed:

Theorem 10. *Consider a dataset containing observations of X, Y, Z, and a partial setting $\tilde{U} \subseteq U$. Then, the conditional STE, SNDE, and SNIE, can be estimated from observational data if there exist $\tilde{U}_1, \tilde{U}_2 \subseteq \tilde{U}$ such that the following two conditions hold:* **(1)** $(X \perp\!\!\!\perp Y \mid \tilde{U}_1)_{\mathcal{G}_{\underline{X}}}$ *and* **(2)** $(X \perp\!\!\!\perp Z \mid \tilde{U}_2)_{\mathcal{G}_{\underline{X}}}$*, where $\mathcal{G}_{\underline{X}}$ represents the DAG with all outgoing arrows from X removed, and $(A \perp\!\!\!\perp B \mid C)_{\mathcal{G}}$ represents the d-separation of A and B by C in DAG $\mathcal{G}$.*

The above theorem provides a graphical criteria for which the interventional distributions utilized by the setting specific causal effects may be swapped for conditional distributions. The proof uses a direct application of Pearl's *do*-calculus [92, Theorem 3.4.1], and is provided in Appendix D.3.3. By letting $\tilde{U} = \emptyset$, identifiability conditions for the specific causal effects of Section 5.3.2 are obtained. Similarly, the theorem provides the corollary that the setting specific causal effects may be estimated from observational data when U is fully observable. It is important to note that the above theorem assumes that each conditional distribution can be sufficiently well estimated. Indeed, the "increased resolution" of the proposed measures comes at a cost in that reliable estimation of the proposed measures poses challenges for values of x that occur infrequently. Consider, for example, estimating the distribution in the second argument of the KL-divergence defining the SNDE in (5.9), namely $p(y \mid \hat{x}', z')$. Given that there is a sum over x' and z', it is necessary to know this distribution for *every* pair (x', z'). Thus, when $p(x', z')$ is very small, a significant amount of date will be required to estimate $p(y \mid x', z')$ (and therefore the SNDE) reliably.

5.4 Examples

We now present three examples of notions of causal influence that are unique to specific causal influences.

5.4.1 Chain Reaction

For the first example we will consider a simple chain $X \to Z \to Y$. This can be thought of as a simplified version of the example proposed by Ay and Polani [10] and modified to include noise by Janzing et al. [54, Example 7]. We will consider the simplest case, where a binary message is being passed from X to Z to Y, with the message being flipped by Z and Y with probability ε. We will interpret each variable as representing the message it passes on, i.e. $X = 1$ means "X passes the message 1 to Z." Formally, let $X, Y, Z \in \{0,1\}$ with $X \sim Bern(0.5)$:

$$Z = \begin{cases} X & w.p.\ 1-\varepsilon \\ X \oplus 1 & w.p.\ \varepsilon \end{cases} \qquad Y = \begin{cases} Z & w.p.\ 1-\varepsilon \\ Z \oplus 1 & w.p.\ \varepsilon \end{cases} \tag{5.13}$$

where $\oplus$ is the XOR operation.

Focusing first on the effect of x on Y, we note that because the only path from X to Y is the one through Z, the direct effect is zero and the total and indirect effects are equal. Noting that $Y \sim Bern(0.5)$, $Y \mid do(X=0) \sim Bern(2\varepsilon(1-\varepsilon))$, and $Y \mid do(X=1) \sim Bern(1-2\varepsilon(1-\varepsilon))$, the total effect is the same for both $x \in \{0,1\}$ and is given by:

$$STE(x \to Y) = D(2\varepsilon(1-\varepsilon) \mid\mid 0.5) \xrightarrow[\varepsilon \to 0]{} 1 \tag{5.14}$$

Thus, as the probability of flipping the message approaches zero, Y will be deterministically linked to X, and X resolves the entire one bit of uncertainty associated with Y. Now consider the conditional STE of z on Y for a particular x. We can compute this by comparing the distributions

$p(y \mid x, \hat{z}) = p(y \mid z)$ and $p(y \mid x)$. Given the symmetry of the problem, this will take one of two values depending on whether or not x and z are equal:

$$STE(z \to Y \mid x) = \begin{cases} D(\varepsilon \,||\, 2\varepsilon(1-\varepsilon)) & x = z \\ D(\varepsilon \,||\, \varepsilon^2 + (1-\varepsilon)^2) & x \neq z \end{cases} \xrightarrow[\varepsilon \to 0]{} \begin{cases} 0 & x = z \\ \infty & x \neq z \end{cases}$$

To understand this result, fix ε to be an arbitrarily small number, and we can say with very high confidence that Z will pass on its received message accurately. Thus, when $x = z$, it is, in a sense, unreasonable to endow Z with responsibility for causing the value taken by Y when it is propagating the message in a nearly deterministic manner (note that for any fixed $\varepsilon > 0$ the STE will not be *exactly* zero). In such a case, it is not so much Z that is causing Y, but rather X that initiated a *chain reaction*. On the other hand, in the unlikely occurrence that $x \neq z$, we have that Z does have a causal effect on Y. This scenario can be thought of as Z acting of its own volition in selecting a message to pass to Y.

We acknowledge that the notion of an unbounded causal influence is initially unsettling. When looking closer, however, this property is intuitive. First, we note that for any fixed $\varepsilon > 0$, the STE will be finite. It is only for $\varepsilon = 0$ that the STE could be infinite, but in that case, the setting that results in infinite influence happens *with probability zero*. Thus, in general, an infinite influence could only be achieved through intervention. Furthermore, such an intervention would have to assign a value to a cause that occurs with probability zero, and that cause would in turn have to enable an otherwise impossible effect to have non-zero probability.

As mentioned in Section 5.3.4, this setting specific violates the locality postulate **(P2)** of Janzing et al. [54] in that the effect of z depends on the value of its own parent, x. We do not claim that the perspective taken here is "correct," but merely point out that there can be intuitive justifications for considering the value of the parent of a cause in evaluating the causal effect.

5.4.2 Caused Uncertainty

Consider a 3-node DAG characterized by the connections $X \to Y \leftarrow Z$ and the following (conditional) distributions:

$$X \sim Bern(0.5) \qquad Z \sim Bern(0.1) \qquad Y \mid X,Z \sim \begin{cases} Bern(0.5) & Z = 1 \\ Bern(0.1) & (X,Z) = (0,0) \\ Bern(0.9) & (X,Z) = (1,0) \end{cases}$$

Given that X and Z are both parentless, we can treat interventions on X and Z as observations, and the CS, conditional IF, and conditional mutual information (CMI) are equivalent. In particular, we have that $\mathfrak{C}_{X\to Y} = I(X \to Y \mid \hat{Z}) = I(X;Y \mid Z) \approx 0.48$ and $\mathfrak{C}_{Z\to Y} = I(Z \to Y \mid \hat{X}) = I(Z;Y \mid X) \approx 0.06$. Before considering the specific causal measures, note that characterization of CMI as a difference of conditional entropies as $I(Z;Y \mid X) = H(Y \mid X) - H(Y \mid X,Z)$ provides us with the interpretation of CMI as the reduction in uncertainty of Y resulting from the added conditioning of Z. Of course, as a result of conditioning reduces entropy, this will always be non-negative.

Next we consider $STE(x \to Y \mid z)$ and $STE(z \to Y \mid x)$ for $(x,z) \in \{0,1\}^2$. Given the symmetry of the problem with respect to X, we only need to consider two of the four possible contexts, namely $(x_0,z_0) \triangleq (0,0)$ and $(x_0,z_1) \triangleq (0,1)$. In order to compute the STE for each X and Z to Y for both contexts, we need the following distributions:

$$p(Y \mid x_0,z_0) = Bern(0.1) \qquad p(Y \mid x_0,z_1) = Bern(0.5)$$
$$p(Y \mid z_0) = Bern(0.5) \qquad p(Y \mid z_1) = Bern(0.5)$$
$$p(Y \mid x_0) = Bern(0.14)$$

For a given context, the STE is given by $STE(x \to Y \mid z) = D(p(Y \mid x,z) \mid\mid p(Y \mid z))$ and $STE(z \to$

$Y \mid x) = D(p(Y \mid x,z) \mid\mid p(Y \mid x))$:

$$STE(x \to Y \mid z) \approx \begin{cases} 0.53 & z = 0 \\ 0.00 & z = 1 \end{cases} \qquad STE(z \to Y \mid x) \approx \begin{cases} 0.01 & z = 0 \\ 0.52 & z = 1 \end{cases}$$

The results presented above are intuitive: when $z = 0$, then the value taken by Y is largely determined by X, and the knowledge that $z = 0$ tells us very little about the distribution of Y. On the other hand, when $z = 1$, X has no bearing on the value taken by Y. Thus, in this scenario, it is the value taken by Z that has caused the shift in the distribution of Y, even though Z provides no information with regard to the particular *value* taken by Y. In this sense, we can think of Z as *causing uncertainty* in Y. This scenario makes particularly clear why it makes sense to condition on the cause but take an expectation with respect to the effect – no outcome y could be attributed to being a result of $z = 1$, despite the clear influence that such an event has on the distribution of Y.

5.4.3 Shared Responsibility

Consider a scenario where a collection of n iid variables $X_i \sim Bern(\varepsilon)$ collectively influence a single outcome Y, i.e. $X_i \to Y$ for $i = 1, \ldots, n$. For a given context $\{x_i\}_{i=1}^n$, let k be the number of x_i that are one, i.e. $k = \sum_i x_i$. Then let Y be distributed as:

$$Y \mid X_1, \ldots, X_n \sim Bern\left(\frac{1}{2^K}\right)$$

where $K = \sum_i X_i$ is a random variable. One interpretation of this example is that each X_i is a potential inhibitor of Y. As more inhibitors become activated (i.e. as k grows), the effect of adding another inhibitor diminishes. Since the value taken by K depends on a context, however, this diminishing influence will not be captured by a measure that is not context-dependent.

As with the previous example, the CS, conditional IF, and CMI are equivalent for this problem setting. While there is no simple computation for these measures as a function of ε and n, there are a couple of key points. First, the influence of each of the variables X_i on Y is the same, i.e. $I(X_i;Y \mid X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n) = I(X_1;Y \mid X_2,\ldots,X_n)$ for all $i = 1,\ldots,n$. Second, as $n \to \infty$, the probability of $Y = 1$ goes to zero, and as $\varepsilon \to 0$, the probability of $Y = 1$ goes to one. In either of the limits, the entropy of Y goes to zero and thus so does the causal influence of each X_i as measured by either CMI, conditional IF, or CS.

Now consider a realization $\{x_i\}_{i=1}^n$ and the corresponding $STE(x_1 \to Y \mid x_2,\ldots,x_n)$. While the influence of each x_i on Y will *not* be the same for a given realization, the symmetry of the problem is such that the computation will be performed in the same manner for each x_i. Letting $k_1 \triangleq \sum_{i=2}^n x_i$ be the number of ones excluding x_1, we define the following distributions:

$$p(Y \mid \{x_i\}_{i=1}^n) = p(Y \mid k) = Bern\left(\frac{1}{2^k}\right)$$
$$p(Y \mid \{x_i\}_{i=2}^n) = p(Y \mid k_1) = Bern\left(\frac{\varepsilon}{2^{k_1+1}} + \frac{1-\varepsilon}{2^{k_1}}\right)$$

Then, for a given context, the STE is a function of x_1 and k_1:

$$STE(x_1 \to Y \mid k_1) = D(p(Y \mid k) \mid\mid p(Y \mid k_1)) = \begin{cases} D(\frac{1}{2^{k_1}} \mid\mid \frac{\varepsilon}{2^{k_1+1}} + \frac{1-\varepsilon}{2^{k_1}}) & x_1 = 0 \\ D(\frac{1}{2^{k_1+1}} \mid\mid \frac{\varepsilon}{2^{k_1+1}} + \frac{1-\varepsilon}{2^{k_1}}) & x_1 = 1 \end{cases}$$

In interpreting these results, first assume that ε is small, meaning that for each of the inhibitors, it is unlikely that it will be activated. As a result of this assumption, we have $STE(X_1 = 0 \to Y \mid k_1) < STE(X_1 = 1 \to Y \mid k_1)$, i.e. an inhibitor has a greater influence when it is activated. More interestingly, we note that $STE(x_1 \to Y \mid k_1)$ is strictly decreasing in k_1. This is consistent with the intuition provided above, namely that if a large number of inhibitors are active, then they *share responsibility* and the influence of any single one is negligible. On the other hand, if only

one is activated (i.e. $(x_1, k_1) = (1, 0)$), then in the limit of $\varepsilon \to 0$, its influence will be infinite.

5.5 Case Study – Effect of El Niño-Southern Oscillation on Pacific Northwest Temperature Anomalies

We now present an application of the proposed framework to measuring the specific causal influences of the El Niño-Southern Oscillation (ENSO) on the temperature anomaly signal in the pacific northwest (PNW, latitude: 47°N, longitude: 240°E). ENSO is characterized by the sea surface temperature in the Niño 3.4 region located in the equatorial Pacific (latitude: 5°S-5°N, longitude: 170°W-120°W). The ENSO signal is typically understood by being in one of three phases (or states) – a neutral phase (we will refer to this as $E = 0$) gives rise to a precipitation region centered near longitude 160°E, the El Niño phase ($E = 1$) gives rise to an eastward shifted precipitation region ($\sim$170°W), and the La Niña phase ($E = -1$) gives rise to a westward shifted precipitation region ($\sim$150°E) [2]. Niño and Niña phases can occur with varying intensities during the winter months with a typical return period of two to seven years [70]. When a Niño or Niña phase occurs, the shifted precipitation signal produces large scale atmospheric pressure waves that influence North American land temperatures [122]. We here use the proposed framework to quantify the causal effect of this teleconnection[2], focusing specifically on the temperature in the PNW.

This application is a particularly good fit for the proposed analysis for a number of reasons. First, by utilizing a collection of simulation model runs, we can obtain an immense amount of data. Second, domain expertise can be leveraged to construct causal DAGs prior to performing analysis. For example, it is well known that the ENSO signal influences temperature as opposed to the temperature influencing ENSO. Third, there are well-accepted methods for detrending

[2]Teleconnection is a commonly used term in climate science, defined by Wikipedia as "climate anomalies being related to each other at large distances (typically thousands of kilometers)."

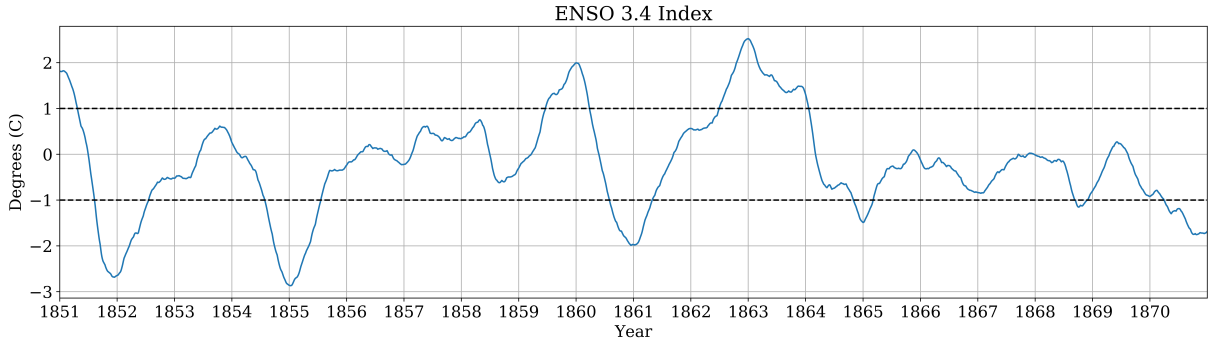

Figure 5.2: ENSO 3.4 index from 1851-1871 along with threshold for determining ENSO phase.

signals, and these methods can be used to control for possible confounding effects. Fourth, it is to be expected that different phases of the ENSO signal will, in some sense, give rise to larger causal effects than other phases. The proposed framework can be used to quantify these differences in a formal sense.

5.5.1 Data and Preprocessing

The analyzed dataset is composed of nine simulated model runs from the National Center for Atmospheric Research's Community Earth System Model, version 2 (CESM2) [39]. This is the gold standard US climate model. Each of the model runs provides an array of daily temperature values spanning the years 1850 to 2015 from which we can compute the ENSO 3.4 index and directly obtain the PNW two-meter temperature. Each of the model runs provides an independent realization of possible evolutions of temperatures that obey the underlying dynamic and thermodynamic equations as encoded by the model. It is important to clarify that the model is not intended for prediction, but rather gives possible atmospheric states for a given set of initial conditions and constraints determined by the selected time period (i.e. CO_2 forcing, solar/lunar cycles, etc.). Both the ENSO index and PNW two-meter temperature signals have the mean and the leading six harmonics of the annual cycle removed, leaving only the anomalous components of the signal. This is standard practice in the analysis of climate data (e.g. [64]). We will henceforth

strictly consider the anomaly signals.

The ENSO index is shown in Figure 5.2, along with a plus/minus one degree threshold for determining the quantized ENSO phase. It is clear that the ENSO signal does not reliably alternate between $E = 1$ and $E = -1$ with a constant period. Furthermore, we can see that the ENSO signal is strongest in or near to January (marked by vertical grid lines). As such, we limit our focus to the months of January, February, and March, as it is not interesting to measure the effect of the ENSO signal in the months where it is not present. We further simplify the problem by quantizing the ENSO index on an annual timescale, i.e. we assign a single value to $E \in \{-1, 0, 1\}$ for January-March of a given year based on the ENSO index value on January 1st of that year. Given that we are estimating the effect of ENSO on temperature, we similarly

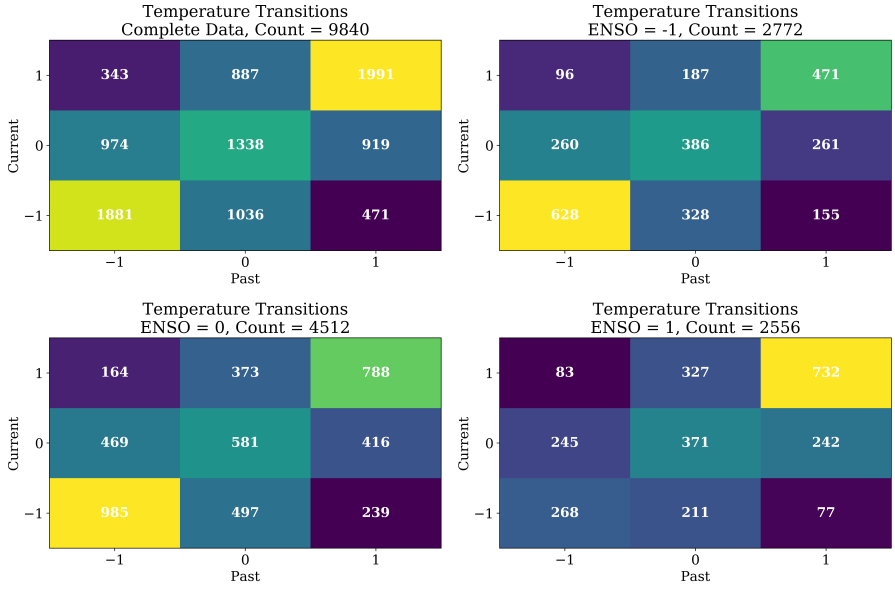

Figure 5.3: Histogram representations of the transitions of temperature averages. Each cell represents the counts of how many times the transition from the past average T_{i-1} to the current average T_i occurs either in the complete dataset (top left) or for specific values of the ENSO signal. The count given in the titles represent the sum of all the cells.

consider the temperature signal only during the months of January, February, and March. Rather than attempting to assess the effect of ENSO on daily temperature anomalies, we choose to focus

on two-week averages. As we will discuss in the next section, this choice also facilitates the causal modeling. As a final processing step, we quantize the temperature anomaly averages to $T \in \{-1, 0, 1\}$. While this quantization does come with an inevitable loss of resolution, it yields the easily understood interpretation of the temperature signal as representing either a cold front, a heat wave, or neutral. We compute the quantization threshold on the entire dataset (i.e. before averaging and before selecting for months) such that one third of days are in each category. The averages are then compared to these thresholds, given by -1.3 and +1.94 degrees. Figure 5.3 gives an impression of how the temperature averages evolve over time with respect to various ENSO phases. Each cell represents the number of counts where a given temperature average $T_{i-1} \in \{-1, 0, 1\}$ was followed by a temperature average $T_i \in \{-1, 0, 1\}$[3]. As we can see from the count in the the top left panel of Figure 5.3, the resultant dataset after selecting for the winter months and taking two-week averages consists of 9840 samples.

5.5.2 Causal Modeling

In order to implement the proposed framework, we first need to formulate a causal DAG representation of the dataset discussed above. As a starting point, consider the DAG on the left side of Figure 5.4, where we let E represent an annual ENSO phase, $T_1, \ldots, T_6$ represent the quantized two-week temperature anomaly averages for January through March (i.e. T_1 averages January 1st through 14th, T_2 averages January 15th through 28th, etc.), and U represents the other factors, such as seasonality and CO_2 forcing. This DAG encodes a number of assumptions. First, it encodes the intuition that seasonality may affect ENSO and the temperature, but not the other way around. Similarly, ENSO will affect the temperature, but not the other way around. The more interesting implicit assumption is that there is a persistence signal in the temperature represented

[3]One might observe that the sum of a row is not necessarily equal to the sum of the corresponding column. This is because the past temperature averages are shifted by two weeks, i.e. they include the last two weeks of December and not the last two weeks of March for each year. Given the number of years and number of model runs, this can result in fairly significant differences.

by the arrow $T_{i-1} \to T_i$. Importantly, we have assumed that this persistence signal is Markov (when conditioned on E and U), i.e. there is no arrow $T_{i-k} \to T_i$ for $k > 1$. This assumption significantly simplifies estimation of the direct and indirect effects of E on T_i, as those require estimating the distribution of T_i for every possible combination of its parents. This serves as a motivation for the decision to consider two-week averages – if we were to simply consider daily temperatures, it is unreasonable to expect that T_i would be independent of T_{i-2} when conditioned on E, U, and T_{i-1}. We next incorporate two assumptions in order to simplify the causal model.

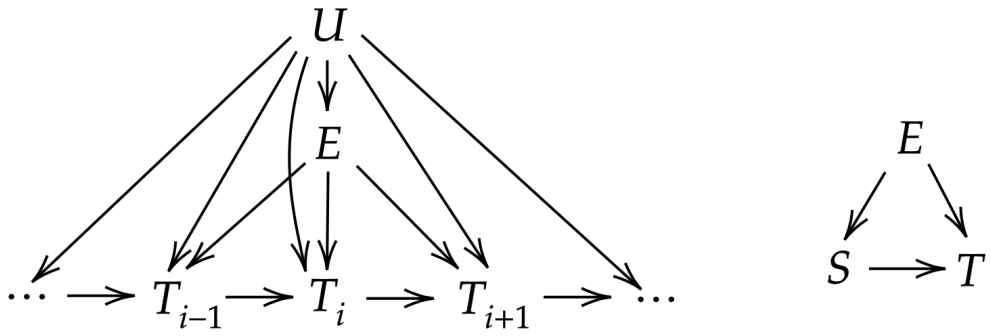

Figure 5.4: Left: Complete DAG representation of climate variables. Right: Simplified DAG after detrending and incorporating the assumption of stationarity.

First, we assume that all the effects of U are removed by the detrending and removal of annual cycle performed in the preprocessing steps. It is to be expected that this assumption will hold for the well known shared causes (such as the aforementioned seasonality and CO_2 forcing), but the possibility of other factors that have effects not captured by the leading six harmonics of the annual cycle is important to note. The second assumption we make is that the distribution of the temperature anomaly averages does not change over time, i.e. that $p(t_i \mid t_{i-1}, e)$ and $p(t_i \mid e)$ are not dependent on i. After making these assumptions, we obtain the simplified DAG on the right of Figure 5.4, where we introduce the new variable S to represent the past temperature anomaly average and T to represent the subsequent temperature average, and note that this perfectly matches the mediation model in Figure 5.1 with $U = \emptyset$. We can think of T as representing T_i and

S as representing either T_{i-1} or the collection $T_1, \ldots, T_{i-1}$. To see why these two interpretations of S are equivalent, consider the SNDE, given by:

$$SNDE(e \to T) = D(p(T \mid \hat{e}) \mid\mid \sum_{e',s'} p(e')p(s' \mid \hat{e}')p(T \mid \hat{e}, s')) \tag{5.15}$$

Now let $T = T_i$ and $S = T_1, \ldots, T_{i-1} \triangleq T_1^{i-1}$, and note that:

$$p(s \mid \hat{e}) = p(t_1^{i-1} \mid \hat{e}) = p(t_{i-1} \mid \hat{e})p(t_1^{i-2} \mid \hat{e}, t_{i-1})$$
$$p(T \mid \hat{e}, s) = p(T \mid \hat{e}, t_1^{i-1}) = p(T \mid \hat{e}, t_{i-1})$$

Plugging these into the second argument in (5.15), we get:

$$\begin{aligned}
\sum_{e',s'} p(e')p(s' \mid \hat{e}')p(T \mid \hat{e}, s') &= \sum_{e', t_1^{i-1\prime}} p(e')p(t'_{i-1} \mid \hat{e}')p(t_1^{i-2\prime} \mid \hat{e}', t'_{i-1})p(T \mid \hat{e}, t'_{i-1}) \\
&= \sum_{e', t'_{i-1}} p(e')p(t'_{i-1} \mid \hat{e}')p(T \mid \hat{e}, t'_{i-1}) \left[\sum_{t_1^{i-2\prime}} p(t_1^{i-2\prime} \mid \hat{e}', t'_{i-1}) \right] \\
&= \sum_{e', t'_{i-1}} p(e')p(t'_{i-1} \mid \hat{e}')p(T \mid \hat{e}, t'_{i-1})
\end{aligned}$$

Given that S appears nowhere in the first argument, we can see that whether $S = T_{i-1}$ or $S = T_1^{i-1}$, the result is the same. The same procedure can be applied to show equivalence for the SNIE. As such, we let $S = T_{i-1}$ and directly use the definitions provided Section 5.3.2 to measure the causal influence of ENSO on temperature. As a result of the assumption that $p(t_i \mid e)$ does not depend on i, we have that $p(t \mid e) = p(s \mid e)$ for $t = s$. It should be noted that for $T = T_1$ (i.e. the average for the first two weeks of January), we define $S = T_0$ to be the average taken over the last two weeks of December. Using this causal model, we define the corresponding dataset from which we estimate the causal influences as $\mathcal{D} = (e_n, s_n, t_n)_{n=1}^{9840}$.

5.5.3 Estimation

Given that there is a large amount of data and a relatively small alphabet size, we utilize plug-in estimators of the proposed measures, where every distribution in question is estimated using a maximum likelihood estimator. Given that E has no parents in the DAG given on the right side of Figure 5.4, we can freely exchange interventions $\hat{e}$ for observations e in the estimation of the effect of e on T. As such, the estimates of the specific effect of ENSO on temperature are given by:

$$\widehat{STE}_{\mathcal{D}}(e \to T) = D(\hat{p}_{\mathcal{D}}(T \mid e) \mid\mid \hat{p}_{\mathcal{D}}(T)) \tag{5.16}$$

$$\widehat{SNDE}_{\mathcal{D}}(e \to T) = D(\hat{p}_{\mathcal{D}}(T \mid e) \mid\mid \textstyle\sum_{e',s'} \hat{p}_{\mathcal{D}}(e')\hat{p}_{\mathcal{D}}(s' \mid e)\hat{p}_{\mathcal{D}}(T \mid e', s')) \tag{5.17}$$

$$\widehat{SNIE}_{\mathcal{D}}(e \to T) = D(\hat{p}_{\mathcal{D}}(T \mid e) \mid\mid \textstyle\sum_{e',s'} \hat{p}_{\mathcal{D}}(e')\hat{p}_{\mathcal{D}}(s' \mid e')\hat{p}_{\mathcal{D}}(T \mid e, s')) \tag{5.18}$$

where $\hat{p}_{\mathcal{D}}$ gives the maximum likelihood estimate of p on the sample $\mathcal{D}$. Specifically, for an arbitrary collection of n samples $\mathcal{C} = (x_i, y_i, z_i)_{i=1}^{n}$ of variables X, Y, Z, the estimate is given by:

$$\hat{p}_{\mathcal{C}}(y) \triangleq \frac{|\{i : y_i = y\}|}{n} \tag{5.19}$$

$$\hat{p}_{\mathcal{C}}(y \mid x) \triangleq \frac{|\{i : x_i = x, y_i = y\}|}{|\{i : x_i = x\}|} \tag{5.20}$$

$$\hat{p}_{\mathcal{C}}(y \mid x, z) \triangleq \frac{|\{i : x_i = x, y_i = y, z_i = z\}|}{|\{i : x_i = x, z_i = z\}|} \tag{5.21}$$

where the $|\{\cdot\}|$ gives the number of elements in the set $\{\cdot\}$.

Next note that the conditional specific total effect of the past temperature average S on the subsequent temperature T conditioned on an ENSO state E is given by:

$$STE(s \to T \mid e) = D(p(T \mid \hat{s}, e) \mid\mid \textstyle\sum_{s'} p(s' \mid e)p(T \mid \hat{s}', e)) \tag{5.22}$$

Letting $X = S$, $Y = T$, $Z = \emptyset$, and $U = E$, it follows from Theorem 10 that we can estimate the

total effect from observational data. Therefore, we use the following plug-in estimator:

$$\widehat{STE}_{\mathcal{D}}(s \to T \mid e) = D(\hat{p}_{\mathcal{D}}(T \mid e,s) \mid\mid \hat{p}_{\mathcal{D}}(T \mid e)) \tag{5.23}$$

By applying these estimators to the complete dataset $\mathcal{D}$, we are able obtain point estimates of the desired measures, which are given by a red $\times$ in the figures in the next section. For ease of notation, we omit $\mathcal{D}$ from the estimates from here on with the understanding that all estimates are performed on $\mathcal{D}$. It is important to note that even though not all estimates will utilize all 9840 samples, Figure 5.3 makes clear there is a considerable amount of samples available for estimating every distribution in question. In particular, we see that:

$$\begin{aligned}
\min_{e,s} |\{i : e_i = e, s_i = s\}| &= |\{i : e_i = 1, s_i = -1\}| \\
&= \sum_t |\{i : t_i = t, e_i = 1, s_i = -1\}| \\
&= 83 + 245 + 268 = 596
\end{aligned}$$

In other words, the distribution estimated on the smallest number of samples is $p(t \mid E = 1, S = -1)$, and this estimate is performed on 596 samples.

In addition to these point estimates, it is desirable to have a means of measuring the significance of the estimated measures. This is particularly important given that the obtained estimates are necessarily non-negative and the units of bits are not as easily interpretable as other units, i.e. temperature. In other words, it is unclear if a causal influence of 0.01 bits (for example) ought to be interpreted as a true influence prior to performing a statistical test. We here pair two approaches – (i) performing a nonparametric bootstrap hypothesis test [76] and (ii) constructing a nonparametric bootstrap confidence interval [33].

The goal of the hypothesis test is to estimate the distribution of the estimated measure under a null hypothesis (H0) and assess the likelihood that our estimate came from such a

distribution. In this case, H0 corresponds to the absence of a causal link, which would result in the true causal measure being equal to zero. The primary challenge to performing this test is the generation of samples from a distribution representative of H0. We accomplish this using a scheme similar to that presented in [54, Example 2] wherein we group the data by one of the three variables (E, S, or T) and shuffle the other two in order to break one of the causal links. For example, when performing the test for the direct effect of E on T, we split the data into three sets: $\{i_{-1} : s_i = -1\}$, $\{i_0 : s_i = 0\}$, and $\{i_1 : s_i = 1\}$. Within each of these sets, we shuffle (i.e. permute) all the samples of E (or T). Because the shuffling occurs within groupings of S, any possible link from E to S and S to T is preserved (and thus so is the indirect effect), but the link between E and T is destroyed. Each of these permutations is then treated as a sample under H0 from which we estimate the SNDE. We perform this shuffling and estimation procedure 100 times and treat the 6th largest estimate as the cutoff threshold for statistical significance. This threshold is given by the upper whisker on the boxplots labeled H0 in the figures in the next section. When performing this test for the indirect effect, we choose to break the link from S to T rather than from E to S in order to preserve the assumption that $p(s \mid e) = p(t \mid e)$ for $s = t$.

Rather than comparing the above threshold to our estimate on the complete dataset (i.e. compare the upper whisker with the red $\times$), we here consider a necessarily stricter test wherein we compare the threshold with the lower end of an estimated confidence interval. In particular, we obtain a straightforward nonparametric bootstrap confidence interval by repeatedly drawing a collection of samples from the empirical distribution of our data and estimating the measure on the new collection of samples. Specifically, let $\mathcal{D}_b^* = (e_{j_b^i}, s_{j_b^i}, t_{j_b^i})_{i=1}^{9840}$ be the bth bootstrap sample, where j_b^i are drawn independently from the uniform distribution over $\{1, 2, \ldots, 9840\}$ and $b = 1, \ldots, 100$. We estimate the causal measure in question on each of the 100 bootstrap samples and, similarly to the hypothesis test, treat the 6th smallest and 6th largest estimates as the lower and upper bounds to our confidence interval. These bounds are given by the whiskers on the boxplots superimposed on the red $\times$ in the figures in the following section.

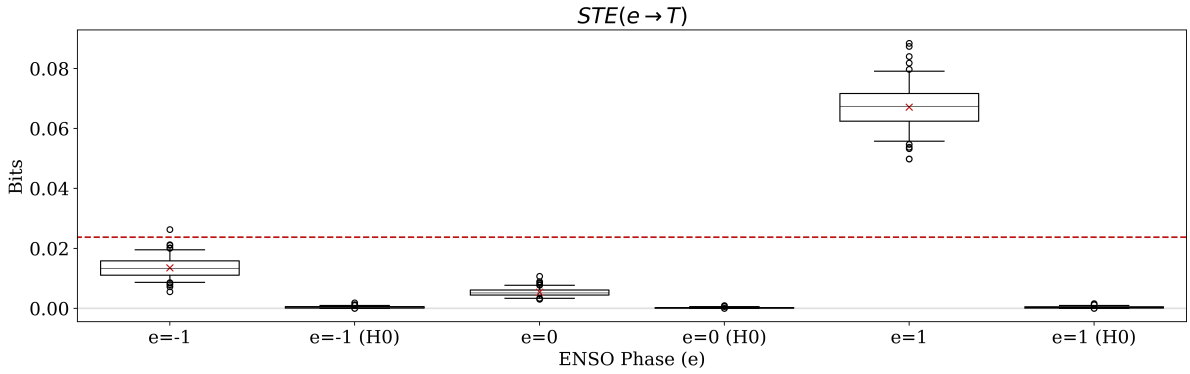

Figure 5.5: Specific total effect of ENSO on temperature anomaly.

5.5.4 Results

We estimate the STE, SNDE, and SNIE of ENSO on temperature and the conditional STE of the past temperature average on the next average. In every case, the measure is estimated on the complete dataset (red $\times$) and compared with the corresponding weighted average, or non-specific, measure (red dashed lines). For the specific measure, we obtain an estimate for each value of the cause, i.e. $e \in \{-1,0,1\}$ or $s \in \{-1,0,1\}$, depending on whether the effect of ENSO or the effect of the past temperature is being estimated. The average measure is then calculated by taking an expectation of the specific measures with respect to $p(e)$ or $p(s \mid e)$. As an example, the red dashed line in Figure 5.5 represents the estimate of $\mathbb{E}_{p(E)}[STE(E \to T)]$, and the three red dashed lines in Figure 5.8 represent the estimates of $\mathbb{E}_{p(S|e)}[STE(S \to T \mid e)]$ for $e \in \{-1,0,1\}$. Each figure also displays two boxplots for each specific measure – the first shows the distribution of the measure estimated on the bootstrap samples and the second shows the distribution of the measure estimated under the null hypothesis that the causal link in question does not exist (denoted "H0").

We begin by considering the total effect of ENSO on temperature shown in Figure 5.5. Given that E is a root node in the DAG representation given on the right of Figure 5.4, we note that STE and SMI are equivalent, i.e. $STE(e \to T) = I_1(e;T)$, and the red dashed line gives an estimate of the mutual information, i.e. $\hat{I}(E;T) = \mathbb{E}_{\hat{p}(E)}[\widehat{STE}(E \to T)]$. This illustrates the value

of considering a specific causal measure – as we can see, the estimated effect of $E = 1$ is roughly three times the effect as estimated by the mutual information. Recall the interpretation of the SMI provided by point **(IV)** in Section 5.3.1, namely that it provides a measure of how much we would expect performing $do(E = e)$ to change the course of nature for T. Under this interpretation, we see that forcing an El Niño year would alter the temperature distribution from what we would expect to occur naturally moreso than forcing a La Niña or neutral year.

While interesting, the perspective of the specific measures as measures of deviation from nature still avoids directly addressing the question of how to interpret a causal influence measured in bits. As mentioned in point **(V)** of Section 5.3.1, we can use the coding theoretic interpretation of the KL-divergence to begin to answer this. In particular, we know that because $T \in \{-1, 0, 1\}$, then the conditional entropy is bounded as $H(T \mid e) \leq \log_2 3 \approx 1.58$ bits. Since $STE(e \to T) = D(p(T \mid e) \mid\mid p(T))$, we can interpret the STE as representing the excess number of bits needed to encode the temperature when we incorrectly assume that E will be drawn according to $p(E)$ (i.e. according to nature), rather than being forced to equal e. As such, it may be of interest to identify the percentage of bits used to encode the naturally occurring T that would have been unnecessary had E been forced to be e. For example, for $E = 1$ we have that:

$$100 \times \left(\frac{\widehat{STE}(E = 1 \to T)}{\widehat{STE}(E = 1 \to T) + \hat{H}(T \mid E = 1)} \right) \approx 100 \times \left(\frac{0.07}{0.07 + 1.53} \right) \approx 4.3\% \tag{5.24}$$

where the estimate of entropy is given by $\hat{H}(T \mid E = 1) \triangleq -\sum_t \hat{p}(t \mid E = 1) \log \hat{p}(t \mid E = 1)$ with $\hat{p}$ as defined in (5.20). In words, if one merely assumes that E will be occur naturally, then the intervention $do(E = 1)$ results in roughly 4.3% of the bits used to encode T being unnecessary had it been known that E would be 1.

We next consider the natural direct and indirect effects shown in Figures 5.6 and 5.7, first noting that both are less than the STE for all values e. This is consistent with the intuition that the direct and indirect effects of ENSO on temperature would not cancel each other out. Intuition is

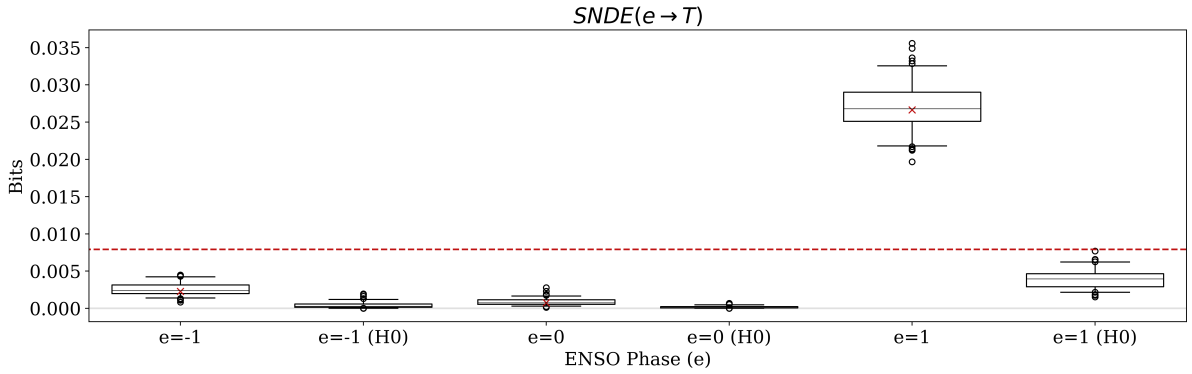

Figure 5.6: Specific natural direct effect of ENSO on temperature anomaly.

also validated by the fact that the SNIE is less than the SNDE for all values. While this need not be the case in general, we make the assumption that S and T are identically distributed given E, and thus we would expect the indirect link $E \to S \to T$ to be weaker than the direct link $E \to T$. As a final point, we note that for both the SNDE and SNIE, only the effect of $E = 1$ passes the proposed statistical significance test. This serves as further justification for the measurement of specific causal influences – when simply measuring average influences with mutual information, causal strength, or information flow, statistical significance testing results in an "all or nothing" result, whereas the present framework enables identifying influences that are significant for only some values of a cause.

We conclude this section with the conditional STE of past on current temperature in a specific ENSO phase, as portrayed by Figure 5.8. We can clearly see that there is a strong persistence in the temperature anomaly signal, i.e. that the past temperature average has a strong effect on the subsequent average, with the largest effect ($STE(S = -1 \to T \mid E = 1)$) being roughly five times that of the effect of $E = 1$. Using the same percentage approach described

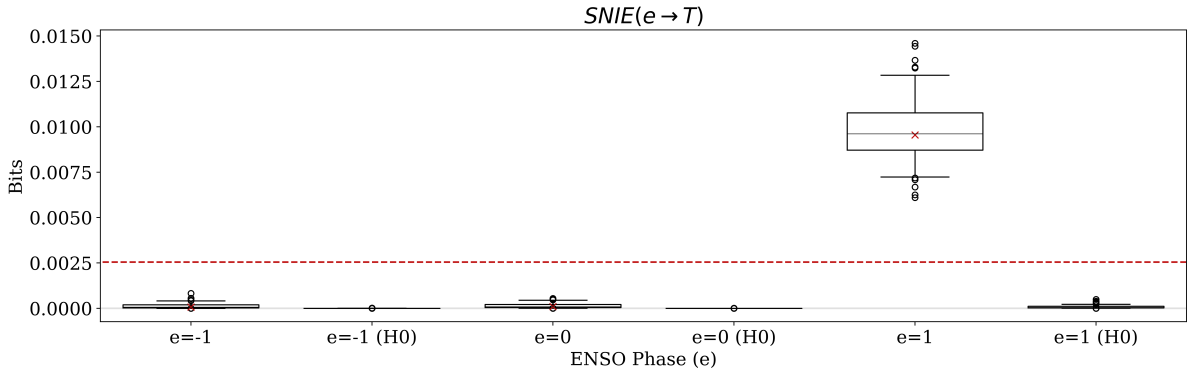

Figure 5.7: Specific natural indirect effect of ENSO on temperature anomaly.

above, we see that:

$$100 \times \left(\frac{\widehat{STE}(S=-1 \to T \mid E=1)}{\widehat{STE}(S=-1 \to T \mid E=1) + \hat{H}(T \mid E=1, S=-1)} \right) \approx 100 \times \left(\frac{0.36}{0.36+1.44} \right) = 20\% \tag{5.25}$$

The fact that the largest effect of S on T occurs when performing the intervention $do(S=-1)$ during an El Niño year can likely be explained by the tendency for El Niño years to give rise to high temperatures. Thus, we would expect that forcing a cold front during an El Niño would alter the course of nature moreso than, say, forcing a heat wave. Furthermore, the second largest effect is seen when $S=1$ and $E=-1$, i.e. when a heat wave is forced during a La Niña year. This result is reminiscent of the example provide in Section 5.4.1, where there is a large causal influence resulting from a broken chain reaction. In this case, since we would expect an El Niño (resp. La Niña) year to assign a higher probability to a heat wave (resp. cold front) that would then persist through the effect of S on T, intervening on S to force a cold front (resp. heat wave) will result in a large deviation from the natural behavior and thus a large causal effect.

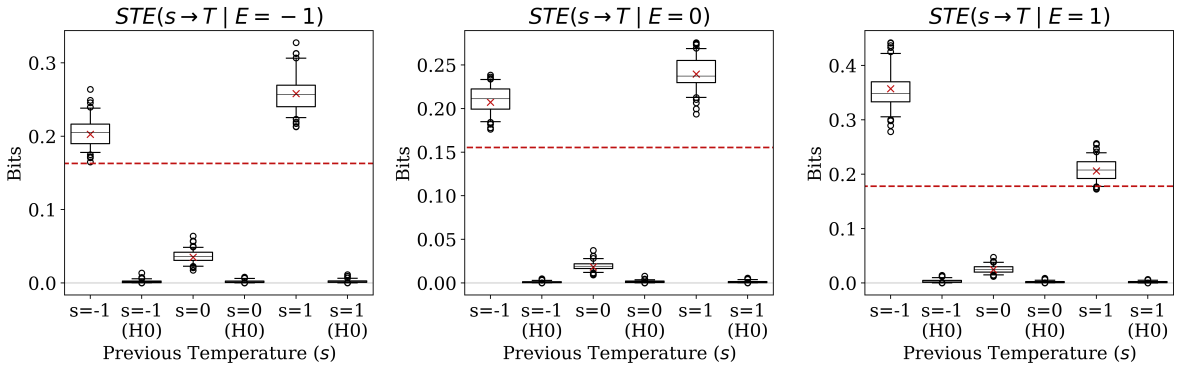

Figure 5.8: Specific total effect of previous temperature anomaly on current temperature anomaly.

5.5.5 Challenges and Caveats

The proposed causal model warrants a number comments. Most notably, any causal interpretation of the results is predicated on the assumption that there are no confounding factors not accounted for in the preprocessing steps. This assumption is less of an issue when measuring the effect of ENSO, where we only need to assume that there is no common cause for E and S or E and T (or rather that there is no backdoor path, to be precise) beyond the seasonality, CO_2 forcing, and any other phenomena captured by the leading six harmonics. When measuring the effect of past temperatures, however, this assumption is a bit more far reaching. For example, we have neglected to consider the temperatures in neighboring regions. Moreover, the explicit nature of the causal effect of S on T is more elusive than that of E on T. While it is reasonable to expect the temperature to have some causal effect in a literal sense (i.e. via the heat equation), it is likely that the estimation procedure is also capturing the effects of temperature related variables. For example, if we additionally included PNW atmospheric pressure waves in the model, we would expect these waves to be a common cause for S and T resulting in a significantly weaker (if not absent) link $S \to T$. As such, the above estimate of $P(s \to T \mid e)$ ought to be viewed as either a measure of predictive utility of the literal temperature, or the causal effect of a "meta variable" representative of the temperature and related quantities that are intervened upon as a whole. In

any case, the present study can serve as a starting point for the development of more intricate causal models representing the relationship between ENSO and temperature.

A second set of challenges arises from the need to estimate the measures for every value of the cause. While these challenges are indeed a fundamental issue with the proposed framework, they provide an opportunity for the development of novel estimation and statistical testing techniques. On one hand, the proposed specific causal measures are necessarily more challenging to estimate than their average counterparts. On the other hand, they necessarily provide more resolution and allow for estimating separate confidence intervals for each element in the analysis. If we are trying to estimate $STE(x \to Y)$ but only have a small number of points in our dataset where $x_i = x$, then we would have very little confidence in our estimate. However, that need not discourage us from having high confidence in an estimate of $STE(x' \to Y)$ for some x' for which we have many samples. That having been said, the proposed estimators and significance test used in the present study lack a formal analysis and leave considerable room for improvement.

As a final discussion point, we return to the comparison of information theoretic and statistical notions of causal influence. Despite having carefully formulated the proposed measures as measures of the extent to which an intervention results in a deviation from the course of nature, the results presented in this section beg the question: *How useful are bits?* As an absolute measure, bits are certainly not very useful considering that a measure in bits will be largely influenced by the number of quantization regions we select. While this can be partially addressed by normalizing as in (5.24) and (5.25), there is no question that the coding theoretic interpretation provided alongside those equations is less intuitive than a measure of, say, the number of degrees warmer we would expect it to be an El Niño year than a La Niña year. Moreover, this intuition gap would be even larger for someone outside of the information theory community, including the majority of the climate scientists for whom these results are intended. This is not to say that the proposed measures are so opaque that they are unusable – in fact, we would argue that

they provide more interpretable notions of causal influence than other information theoretic measures that have experienced some popularity in the literature. Instead, this discussion is merely to maximize the level of intuition that we can associate with the proposed measures while simultaneously acknowledging the limitations of information theoretic measures in terms of interpretability.

5.6 Conclusion

We have sought inspiration from the statistical causality community in order to refine information theoretic measures of causal influence. Specifically, we have developed a series of causal measures that are defined for specific values of the cause in question with the goal of differentiating between total, direct, and indirect effects, and provided conditions under which they can be estimated from observational data. The proposed measures are, at their core, aligned with previous information theoretic measures in that they compare distributions of Y rather than comparing values of Y. As such, they are well-equipped for capturing non-linear, higher order causal effects, although at the cost of foregoing an explanation of the exact nature of the causal effects. Perhaps most importantly, we have elucidated the key insight that information theoretic measures of causal influence can be interpreted as methods for quantifying the magnitude with which an intervention is expected to alter the course of nature. This interpretation stands in stark contrast to that of statistical measures. As such, we hope that a key takeaway will be that information theoretic and statistical notions of causal can provide *complementary* methods in that they yield the answers to fundamentally different causal questions.

5.7 Acknowledgements

Chapter 5, in part, is currently under review for publication of the material. Schamberg, Gabriel; Coleman, Todd. Chapter 5, in part, is currently being prepared for submission for publication of the material. Schamberg, Gabriel; Chapman, William; Coleman, Todd. The dissertation author was the primary investigator and author of these materials.

Appendix A

Appendix to Chapter 2

A.1 Derivation of Scaled Form

We will demonstrate the derivation of the scaled form of the measurement model update only, noting that the derivation for the other updates follows almost identical steps. Consider the original measurement model update:

$$\mathbf{x}^{(i+1)} = \underset{\mathbf{x}}{\operatorname{argmin}} \left(\sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) \right) + \langle \lambda^{(i)}, \mathbf{x} - \mathbf{z}^{(i)} \rangle + \frac{\rho}{2} \left|\left| \mathbf{x} - \mathbf{z}^{(i)} \right|\right|_F^2. \tag{A.1}$$

For ease of notation, the superscript (i) is omitted for the remainder of this appendix. Using the definition of the inner product and Frobenius norm, we can break up the second and third terms across into sums and simplify as follows:

$$\begin{aligned} \mathbf{x}^{(i+1)} &= \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \lambda_n^T(\mathbf{x}_n - \mathbf{z}_n) + \frac{\rho}{2}(\mathbf{x}_n - \mathbf{z}_n)^T(\mathbf{x}_n - \mathbf{z}_n) \\ &= \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \frac{\rho}{2}\mathbf{x}_n^T\mathbf{x}_n + (\lambda_n - \rho\mathbf{z}_n)^T\mathbf{x}_n \\ &= \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{n=1}^{N} \frac{2}{\rho} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \mathbf{x}_n^T\mathbf{x}_n - 2\left(\mathbf{z}_n - \frac{\lambda_n}{\rho}\right)^T \mathbf{x}_n. \end{aligned} \tag{A.2}$$

Defining $\tilde{\mathbf{x}}_n = \mathbf{z}_n - \frac{\lambda_n}{\rho}$ as in Section 2.3.1, we note that $\tilde{\mathbf{x}}_n$ does not depend on $\mathbf{x}$, enabling us to complete the square and simplify as follows:

$$\begin{aligned}
\mathbf{x}_n &= \underset{\mathbf{x}}{\operatorname{argmin}} \quad \sum_{n=1}^{N} \frac{2}{\rho} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \mathbf{x}_n^T \mathbf{x}_n - 2\tilde{\mathbf{x}}_n^T \mathbf{x}_n \\
&= \underset{\mathbf{x}}{\operatorname{argmin}} \quad \sum_{n=1}^{N} \frac{2}{\rho} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + \mathbf{x}_n^T \mathbf{x}_n - 2\tilde{\mathbf{x}}_n^T \mathbf{x}_n + \tilde{\mathbf{x}}_n^T \tilde{\mathbf{x}}_n \\
&= \underset{\mathbf{x}}{\operatorname{argmin}} \quad \sum_{n=1}^{N} \frac{2}{\rho} L_n(\mathbf{y}_n \mid \mathbf{x}_n) + (\tilde{\mathbf{x}}_n - \mathbf{x}_n)^T (\tilde{\mathbf{x}}_n - \mathbf{x}_n) \\
&= \underset{\mathbf{x}}{\operatorname{argmin}} \quad \left(\sum_{n=1}^{N} L_n(\mathbf{y}_n \mid \mathbf{x}_n) \right) + \frac{\rho}{2} ||\mathbf{x} - \tilde{\mathbf{x}}||_F^2 ,
\end{aligned} \tag{A.3}$$

as was to be shown.

A.2 Proof of Theorem 1

Consider the problem in its original form:

$$\begin{aligned}
(\hat{\mathbf{x}}, \hat{\mathbf{w}}) = \underset{\mathbf{x},\mathbf{w}}{\operatorname{argmin}} \quad & L(\mathbf{y} \mid \mathbf{x}) + \beta\phi(\mathbf{w}) \\
s.t. \qquad & \mathbf{w} = \mathcal{A}(\mathbf{x}).
\end{aligned} \tag{A.4}$$

The goal is to show that there is an equivalent two-block ADMM problem whose updates match those given by (2.11). To do so we define the variable $\mathbf{Q} \coloneqq [\mathbf{X}^T, \mathbf{W}^T]^T \in \mathbb{R}^{2K \times N}$ ($\mathbf{X}, \mathbf{W} \in \mathbb{R}^{K \times N}$) and the function $g(\mathbf{Q}) \coloneqq L(\mathbf{y} \mid \mathbf{X}) + \beta\phi(\mathbf{W})$. Next, we define $\mathbf{Z} \coloneqq [\mathbf{Z}_\mathbf{X}^T, \mathbf{Z}_\mathbf{W}^T]^T \in \mathbb{R}^{2K \times N}$ and the function:

$$h(\mathbf{Z}) = \begin{cases} 0 & \mathcal{A}(\mathbf{Z}_\mathbf{X}) = \mathbf{Z}_\mathbf{W} \\ \infty & \mathcal{A}(\mathbf{Z}_\mathbf{X}) \neq \mathbf{Z}_\mathbf{W} \end{cases}. \tag{A.5}$$

Using these newly defined terms, we can write (A.4) equivalently as:

$$\begin{aligned} (\hat{\mathbf{q}}, \hat{\mathbf{z}}) = \underset{\mathbf{q},\mathbf{z}}{\text{argmin}} \quad & g(\mathbf{q}) + h(\mathbf{z}) \\ s.t. \quad & \mathbf{q} - \mathbf{z} = 0. \end{aligned} \tag{A.6}$$

Note that if $\mathbf{q} = [\mathbf{x}^T, \mathbf{w}^T]^T$ is such that $\mathbf{w} \neq \mathcal{A}(\mathbf{x})$ (the constraints in (A.4) are not satisfied) and $\mathbf{z}$ is such that $\mathbf{q} - \mathbf{z} = 0$ (the constraints (A.6) are satisfied), then $h(\mathbf{z}) = \infty$ and $(\mathbf{q}, \mathbf{z})$ are not the minimizers of (A.6). To solve this problem with ADMM, we first find augmented Lagrangian:

$$\mathcal{L}_\rho(\mathbf{q}, \mathbf{z}, \gamma) = g(\mathbf{q}) + h(\mathbf{z}) + \langle \gamma, \mathbf{q} - \mathbf{z} \rangle + \frac{\rho}{2} ||\mathbf{q} - \mathbf{z}||_F^2 \tag{A.7}$$

with Lagrange multiplier $\gamma = [\lambda^T, \alpha^T]^T \in \mathbb{R}^{2K \times N}$ ($\lambda, \alpha \in \mathbb{R}^{K \times N}$). As a result, we get the following update equations:

$$\begin{aligned} \mathbf{q}^{(i+1)} &= \underset{\mathbf{q}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{q}, \mathbf{z}^{(i)}, \gamma^{(i)}) \\ \mathbf{z}^{(i+1)} &= \underset{\mathbf{z}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{q}^{(i+1)}, \mathbf{z}, \gamma^{(i)}) \\ \gamma^{(i+1)} &= \gamma^{(i)} + \rho(\mathbf{q}^{(i+1)} - \mathbf{z}^{(i+1)}). \end{aligned} \tag{A.8}$$

Next we show that the update equations given by (A.8) are equivalent to those given by (2.11).

First, consider the $\mathbf{q}$ update:

$$\begin{aligned} \mathbf{q}^{(i+1)} &= \underset{\mathbf{q}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{q}, \mathbf{z}^{(i)}, \gamma^{(i)}) \\ &= \underset{\mathbf{q}}{\text{argmin}}\ g(\mathbf{q}) + \langle \gamma^{(i)}, \mathbf{q} - \mathbf{z}^{(i)} \rangle + \frac{\rho}{2} ||\mathbf{q} - \mathbf{z}^{(i)}||_F^2 \\ &= \underset{[\mathbf{x}^T, \mathbf{w}^T]^T}{\text{argmin}}\ L(\mathbf{y} \mid \mathbf{x}) + \beta\phi(\mathbf{w}) + \begin{bmatrix} \lambda^{(i)} \\ \alpha^{(i)} \end{bmatrix}^T \left(\begin{bmatrix} \mathbf{x} \\ \mathbf{w} \end{bmatrix} - \begin{bmatrix} \mathbf{z}_\mathbf{x}^{(i)} \\ \mathbf{z}_\mathbf{w}^{(i)} \end{bmatrix} \right) + \left|\left| \begin{bmatrix} \mathbf{x} \\ \mathbf{w} \end{bmatrix} - \begin{bmatrix} \mathbf{z}_\mathbf{x}^{(i)} \\ \mathbf{z}_\mathbf{w}^{(i)} \end{bmatrix} \right|\right|_F^2 \\ &= \begin{bmatrix} \mathbf{x}^{(i+1)} \\ \mathbf{w}^{(i+1)} \end{bmatrix} \end{aligned} \tag{A.9}$$

where $\mathbf{x}^{(i+1)}$ and $\mathbf{w}^{(i+1)}$ are given by:

$$\mathbf{x}^{(i+1)} = \underset{\mathbf{x}}{\operatorname{argmin}}\ L(\mathbf{y} \mid \mathbf{x}) + \langle \lambda^{(i)}, \mathbf{x} - \mathbf{z}_{\mathbf{x}}^{(i)} \rangle + ||\mathbf{x} - \mathbf{z}_{\mathbf{x}}^{(i)}||_F^2 \tag{A.10}$$

$$\mathbf{w}^{(i+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\ \beta\phi(\mathbf{w}) + \langle \alpha^{(i)}, \mathbf{w} - \mathbf{z}_{\mathbf{w}}^{(i)} \rangle + ||\mathbf{w} - \mathbf{z}_{\mathbf{w}}^{(i)}||_F^2 \tag{A.11}$$

and can be found independently of each other.

Next, consider the $\mathbf{z}$ update:

$$\begin{aligned}
\mathbf{z}^{(i+1)} &= \underset{\mathbf{z}}{\operatorname{argmin}} && \mathcal{L}_\rho(\mathbf{q}^{(i+1)}, \mathbf{z}, \gamma^{(i)}) \\
&= \underset{\mathbf{z}}{\operatorname{argmin}} && h(\mathbf{z}) + \langle \gamma^{(i)}, \mathbf{q}^{(i+1)} - \mathbf{z} \rangle + \frac{\rho}{2} ||\mathbf{q}^{(i+1)} - \mathbf{z}||_F^2 \\
&= \underset{[\mathbf{z}_{\mathbf{x}}^T, \mathbf{z}_{\mathbf{w}}^T]^T}{\operatorname{argmin}} && h(\mathbf{z}) + \langle \lambda^{(i)}, \mathbf{x}^{(i+1)} - \mathbf{z}_{\mathbf{x}} \rangle + \frac{\rho}{2} ||\mathbf{x}^{(i+1)} - \mathbf{z}_{\mathbf{x}}||_F^2 + \\
& && \langle \alpha^{(i)}, \mathbf{w}^{(i+1)} - \mathbf{z}_{\mathbf{w}} \rangle + \frac{\rho}{2} ||\mathbf{w}^{(i+1)} - \mathbf{z}_{\mathbf{w}}||_F^2 \\
&= \underset{\mathbf{z}_{\mathbf{x}}}{\operatorname{argmin}} && \langle \lambda^{(i)}, \mathbf{x}^{(i+1)} - \mathbf{z}_{\mathbf{x}} \rangle + \frac{\rho}{2} ||\mathbf{x}^{(i+1)} - \mathbf{z}_{\mathbf{x}}||_F^2 + \\
& && \langle \alpha^{(i)}, \mathbf{w}^{(i+1)} - \mathcal{A}(\mathbf{z}_{\mathbf{x}}) \rangle + \frac{\rho}{2} ||\mathbf{w}^{(i+1)} - \mathcal{A}(\mathbf{z}_{\mathbf{x}})||_F^2 \\
&= && \begin{bmatrix} \mathbf{z}_{\mathbf{x}}^{(i+1)} \\ \mathcal{A}(\mathbf{z}_{\mathbf{x}}^{(i+1)}) \end{bmatrix}
\end{aligned}$$

where $\mathbf{z}_{\mathbf{x}}^{(i+1)}$ is given as the solution to (2.15), i.e. the consensus update for our target problem, and the second to last equality follows from the fact that $h(\mathbf{z})$ is infinite if $\mathbf{z}_{\mathbf{w}} \neq \mathcal{A}(\mathbf{z}_{\mathbf{x}})$, so we can treat the problem as a single variable optimization problem.

Next we can substitute these results into the equations for the $\mathbf{q}$ update to obtain:

$$\mathbf{w}^{(i+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\ \beta\phi(\mathbf{w}) + \langle \alpha^{(i)}, \mathbf{w} - \mathcal{A}(\mathbf{z}_{\mathbf{x}}^{(i)}) \rangle + ||\mathbf{w} - \mathcal{A}(\mathbf{z}_{\mathbf{x}}^{(i)})||_F^2 \tag{A.12}$$

which is the (unscaled) update equation (2.14) for $\mathbf{w}$ in the original formulation, where $\mathbf{z}_{\mathbf{x}}^{(i)}$ in

this formulation corresponds with $\mathbf{z}^{(i)}$ in the original formulation. The $\mathbf{x}$ portion of the $\mathbf{q}$ remains unchanged from (A.10), which is equivalent to the unscaled update equation (2.12) for $\mathbf{x}$ in the original formulation.

Next, we can decompose the matrix multiplication in the same way as above to show that:

$$\gamma^{(i+1)} = \begin{bmatrix} \lambda^{(i+1)} \\ \alpha^{(i+1)} \end{bmatrix} \tag{A.13}$$

where $\lambda^{(i+1)}$ and $\alpha^{(i+1)}$ are given by the original updates in (2.11).

Thus, we have shown that directly solving (A.6) using ADMM yields the proposed updates detailed in the body of the paper. As such, we will show that the ADMM solution to (A.6) is convergent. By assumption, L and ϕ are closed, proper, and convex, and hence, so is their sum g. To show that h is convex, we note that this is an indicator function on the set $H \coloneqq \{(\mathbf{z_X}, \mathbf{z_W}) : \mathcal{A}(\mathbf{z_X}) = \mathbf{z_W}\} \subset \mathbb{R}^2$, thus h is convex if and only if H is convex [104, Ch. 2]. Suppose $\mathbf{z}^1 = [\mathbf{z}_\mathbf{x}^{1\,T}, \mathbf{z}_\mathbf{w}^{1\,T}]^T$ and $\mathbf{z}^2 = [\mathbf{z}_\mathbf{x}^{2\,T}, \mathbf{z}_\mathbf{w}^{2\,T}]^T$ are such that $\mathcal{A}(\mathbf{z}_\mathbf{x}^1) = \mathbf{z}_\mathbf{w}^1$ and $\mathcal{A}(\mathbf{z}_\mathbf{x}^2) = \mathbf{z}_\mathbf{w}^2$, i.e. $\mathbf{z}^1, \mathbf{z}^2 \in H$. Then, if we take a convex combination $\mathbf{z}^\alpha \coloneqq \alpha\mathbf{z}^1 + (1-\alpha)\mathbf{z}^2$ for $\alpha \in [0,1]$, we get:

$$\begin{aligned} \mathbf{z}_\mathbf{w}^\alpha &= \alpha\mathbf{z}_\mathbf{w}^1 + (1-\alpha)\mathbf{z}_\mathbf{w}^2 \\ &= \alpha\mathcal{A}(\mathbf{z}_\mathbf{x}^1) + (1-\alpha)\mathcal{A}(\mathbf{z}_\mathbf{x}^2) \\ &= \mathcal{A}(\alpha\mathbf{z}_\mathbf{x}^1 + (1-\alpha)\mathbf{z}_\mathbf{x}^2) \\ &= \mathcal{A}(\mathbf{z}_\mathbf{x}^\alpha). \end{aligned} \tag{A.14}$$

Thus, we see that $\mathbf{z}^1, \mathbf{z}^2 \in H \implies \mathbf{z}^\alpha \in H$, i.e. H, and therefore h, are convex. It then follows from Section 3.2.1 of [16] that the ADMM solution for (A.6) is convergent, as was to be shown.

■

A.3 State-Space Model of Learning Updates

We begin by deriving expressions for the negative log-likelihoods for each of the observations:

$$\begin{aligned}
L_{B_n}(b_n \mid x_n) &= -\log p_{B_n|X_n}(b_n \mid x_n) \\
&= -\log p_n^{b_n}(1-p_n)^{1-b_n} \\
&= -b_n \log \frac{e^{\nu+\eta x_n}}{1+e^{\nu+\eta x_n}} - (1-b_n)\log \frac{1}{1+e^{\nu+\eta x_n}} \\
&\propto \log\left(1+e^{\nu+\eta x_n}\right) - b_n \eta x_n
\end{aligned}$$

$$\begin{aligned}
L_{R_n}(r_n \mid x_n) &= -\log f_{R_n|X_n}(r_n \mid x_n) \\
&= -\log \frac{1}{\sqrt{2\pi\sigma_R^2}} \exp\left(-\frac{(r_n-\psi-\omega x_n)^2}{2\sigma_R^2}\right) \\
&\propto \frac{(r_n-\psi-\omega x_n)^2}{2\sigma_R^2}
\end{aligned}$$

$$\begin{aligned}
L_{\mathbf{S}_n}(\mathbf{s}_n \mid x_n) &= -\log p_{\mathbf{S}_n|X_n}(\mathbf{s}_n \mid x_n) \\
&= -\log \exp\left(\sum_{j=1}^{J}\left[\log(\Lambda_{n,j})s_{n,j} - \Lambda_{n,j}\Delta t\right]\right) \\
&= -\sum_{j=1}^{J}\left(\xi + a x_n + \sum_{m=1}^{M} c_m s_{n,j-m}\right) n_{n,j} + \sum_{j=1}^{J} \exp\left(\xi + a x_n + \sum_{m=1}^{M} c_m s_{n,j-m}\right)\Delta t \\
&\propto -a x_n \sum_{j=1}^{J} n_{n,j} + \sum_{j=1}^{J} \exp\left(\xi + a x_n + \sum_{m=1}^{M} c_m s_{n,j-m}\right)\Delta t \\
&= \Delta t \exp(\xi + a x_n) \sum_{j=1}^{J} \exp\left(\sum_{m=1}^{M} c_m s_{n,j-m}\right) - a x_n \sum_{j=1}^{J} s_{n,j}
\end{aligned}$$

These expressions can be plugged into equation (2.25) to obtain the measurement model update equation, which can in turn be solved using Newton's method.

Next, the system model update can be solved in closed form:

$$\begin{aligned}\mathbf{w}^{(i+1)} &= \underset{\mathbf{w}}{\operatorname{argmin}}\ \phi(\mathbf{w}) + \frac{\rho}{2}\left|\left|\mathbf{w} - \tilde{\mathbf{w}}^{(i)}\right|\right|_2^2 \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{n=1}^{N}\left(\frac{(w_n-\gamma)^2}{2\sigma_V^2} + \frac{\rho}{2}(w_n - \tilde{w}_n^{(i)})^2\right)\end{aligned}$$

where $\tilde{w}_n^{(i)} := z_n^{(i)} - \kappa z_{n-1}^{(i)} - \alpha_n^{(i)}/\rho$. Thus, we can solve for each w_n separately:

$$\begin{aligned}w_n^{(i+1)} &= \underset{w_n}{\operatorname{argmin}} \frac{(w_n-\gamma)^2}{2\sigma_V^2} + \frac{\rho}{2}(w_n - \tilde{w}_n^{(i)})^2 \\ &= \underset{w_n}{\operatorname{argmin}} \left(\frac{1}{2\sigma_V^2} + \frac{\rho}{2}\right) w_n^2 - \left(\frac{\gamma}{\sigma_V^2} + \rho\tilde{w}_n^{(i)}\right) w_n \\ &= \underset{w_n}{\operatorname{argmin}} \left(w_n - \frac{\frac{\gamma}{\sigma_V^2} + \rho\tilde{w}_n^{(i)}}{\frac{1}{\sigma_V^2} + \rho}\right)^2 \\ &= \frac{\frac{\gamma}{\sigma_V^2} + \rho\tilde{w}_n^{(i)}}{\frac{1}{\sigma_V^2} + \rho}.\end{aligned}$$

Finally, given its relatively low dimensionality, we can efficiently solve the consensus update in closed form by posing it as a least squares problem. First, we note that $\mathcal{A}(\mathbf{z}) = \mathbf{G}\mathbf{z}$ when we

define:

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -\kappa & 1 & 0 & \dots & 0 & 0 \\ 0 & -\kappa & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -\kappa & 1 \end{bmatrix} \tag{A.15}$$

with $\mathbf{G} \in \mathbb{R}^{N \times N}$. Thus, we have:

$$\mathbf{z}^{(i+1)} = \underset{\mathbf{z}}{\operatorname{argmin}} \, ||\mathbf{z} - \tilde{\mathbf{z}}_{\mathbf{x}}^{(i)}||_F^2 + ||\mathbf{G}\mathbf{z} - \tilde{\mathbf{z}}_{\mathbf{w}}^{(i)}||_F^2. \tag{A.16}$$

Taking the gradient of the RHS and setting to zero yields:

$$\mathbf{z}^{(i+1)} = (\mathbf{I} + \mathbf{G}^T\mathbf{G})^{-1}(\tilde{\mathbf{z}}_{\mathbf{x}}^{(i)} + \mathbf{G}^T\tilde{\mathbf{z}}_{\mathbf{w}}^{(i)}). \tag{A.17}$$

Given that $\mathbf{G}$ is known a-priori, we can find $(\mathbf{I} + \mathbf{G}^T\mathbf{G})^{-1}$ once and each consensus update becomes a matrix multiplication problem.

A.4 Convexity State-Space Model of Learning Negative Log-Likelihood

Given that L is the sum of the negative log-likelihoods for each of the observation modalities as in (2.25), it is sufficient to show that they are each convex in x_n, which is made easier by use of the simplifications derived in Appendix A.3. Noting that addition of a constant does not affect convexity, we can assess the final simplification provided in each case. As such, we see that $L_{B_n}(b_n \mid x_n)$ is the sum of a term that is linear in x_n and a special case of the log

sum exponential (LSE) function with an added auxiliary variable constrained to equal zero (giving $e^0 = 1$). Given the convexity of LSE, its sum with a linear term is also convex, and thus $L_{B_n}(b_n \mid x_n)$ is convex. Next, $L_{R_n}(r_n \mid x_n)$ is quadratic in x_n and thus convex. Finally, $L_{\mathbf{S}_n}(\mathbf{s}_n \mid x_n)$ is the sum of a term that is linear in x_n and a term that is exponential in x_n, both of which are convex. As a result, $L_{B_n}(b_n \mid x_n)$, $L_{R_n}(r_n \mid x_n)$, and $L_{\mathbf{S}_n}(\mathbf{s}_n \mid x_n)$ are all convex in x_n for any $(b_n, r_n, \mathbf{s}_n) \in \{0,1\} \times \mathbb{R} \times \{0,1\}^J$, and thus so is their sum L.

Appendix B

Appendix to Chapter 3

B.1 Proof of Theorem 2

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

$$
\begin{aligned}
& p(X_i \mid X^{i-1}, Z^{i-1}) \\
&= \sum_{y_{i-d}^{i-1}} p(X_i \mid X^{i-1}, y_{i-d}^{i-1}, Z^{i-1}) \prod_{j=i-d}^{i-1} p(y_j \mid X^{i-1}, Z^{i-1}) && \text{(B.1)} \\
&= \sum_{y_{i-d}^{i-1}} p(X_i \mid X_{i-d}^{i-1}, y_{i-d}^{i-1}, Z_{i-d}^{i-1}) \prod_{j=i-d}^{i-1} p(y_j \mid X_{j-d}^{i-1}, Z_{j-d}^{i-1}) && \text{(B.2)} \\
&= \sum_{y_{i-d}^{i-1}} p(X_i \mid X_{i-2d}^{i-1}, y_{i-d}^{i-1}, Z_{i-2d}^{i-1}) \prod_{j=i-d}^{i-1} p(y_j \mid X_{i-2d}^{i-1}, Z_{i-2d}^{i-1}) && \text{(B.3)} \\
&= p(X_i \mid X_{i-2d}^{i-1}, Z_{i-2d}^{i-1})
\end{aligned}
$$

where (B.1) follows from the chain rule and that y_{i-d}^{i-1} are conditionally independent given (X^{i-1}, Z^{i-1}), (B.2) follows from the joint Markovicity of X and Y and the conditional indepen-

dence of y_{i-d}^{i-1}, and (B.3) follows from the conditional independence of the past and the future given the present for Markov processes.

Next we will show that if there is some $j < k \leq i$ such that $I(Y_j; Y_k \mid X^i, Z^i) > 0$, 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 $I(Y^n \to X^n \mid Z^n) > 0$ implies that for some $q \leq m$, there is a directed edge from Y_q to X_m. Letting $\tau_2 = m - q$, we know from the joint stationarity of (X, Y, Z) that for every X_i, there is an incoming directed edge from $Y_{i-\tau_2}$. As such, at the end of step 4, every X_i will be part of an undirected path connecting $Y_{i-\tau_2}$, $Y_{i-\tau_2-\tau_1}$, $Y_{i-\tau_2-2\tau_1}$, ... 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. $\square$

B.2 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:

$$p(x_i \mid x_{i-l-1}^{i-1}, z_{i-l-1}^{i-1}) = p(x_i \mid \tilde{x}_{i-l-1}^{i-1}, \tilde{z}_{i-l-1}^{i-1}) \tag{B.4}$$

where we define $\tilde{x}_{i-l-d}^{i-1} \triangleq \{x_{i-l}^{i-1}, x'_{i-l-1}\}$ and $\tilde{z}_{i-l-1}^{i-1} \triangleq \{z_{i-l}^{i-1}, z'_{i-l-1}\}$. We will demonstrate that the equation given by (B.4) amounts to solving a polynomial function of the parameters θ. It is shown in [87] 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 (B.4), we see that:

$$
\begin{aligned}
& p(x_i \mid x_{i-l-1}^{i-1}, z_{i-l-1}^{i-1}) \\
&= \sum_{y_{i-l-1}^{i-1}} \theta_{x,y,z}^{x_i} p(y_{i-l-1}^{i-1} \mid x_{i-l-1}^{i-1}, z_{i-l-1}^{i-1}) \\
&= \sum_{y_{i-l-1}^{i-1}} \theta_{x,y,z}^{x_i} \frac{p(x_{i-l-1}^{i-1}, y_{i-l-1}^{i-1}, z_{i-l-1}^{i-1})}{p(x_{i-l-1}^{i-1}, z_{i-l-1}^{i-1})} \\
&= \frac{\sum_{y_{i-l-1}^{i-1}} \theta_{x,y,z}^{x_i} \pi(x_{i-l-1}, y_{i-l-1}, z_{i-l-1}) \prod_{j=1}^{l} \theta_{x,y,z}^{(x,y,z)_{i-j}}}{\sum_{\tilde{y}_{i-l-1}^{i-1}} \pi(x_{i-l-1}, \tilde{y}_{i-l-1}, z_{i-l-1}) \prod_{j=1}^{l} \theta_{x,\tilde{y},z}^{(x,\tilde{y},z)_{i-j}}}
\end{aligned}
\tag{B.5}
$$

where $\pi : |\mathcal{X}| \times |\mathcal{Y}| \times |\mathcal{Z}| \to [0,1]$ 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 $\tilde{\pi} : |\mathcal{X}| \times |\mathcal{Y}| \times |\mathcal{Z}| \to \mathbb{R}$ in (B.5) we see that C cancels in the numerator and denominator and thus each side of (B.4) can be written entirely as fractions of polynomial functions of θ. Next, repeat the process on the right hand side of (B.4) 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 (B.4) by the denominator and repeat until (B.4) is a polynomial function of θ. Finally, we note that the polynomial given by (B.4) is trivial only if *every* process is a solution. Though

omitted here for brevity, one can easily show that the polynomial is non-trivial by constructing a counterexample. □

B.3 Proof of Theorem 4

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

$$
\begin{aligned}
& H(X_i \mid X^{i-1}, Y^{i-k_1}, Z^i) - H(X_i \mid X^{i-1}, Y^i, Z^i) && \text{(B.6)} \\
\leq & H(X_i \mid X^{i-1}, Z^i) - H(X_i \mid X^{i-1}, Y^i, Z^i) && \text{(B.7)} \\
\leq & H(X_i \mid X_{i-k_2}^{i-1}, Z_{i-k_2}^i) - H(X_i \mid X^{i-1}, Y^i, Z^i) && \text{(B.8)} \\
= & H(X_i \mid X_{i-k_2}^{i-1}, Z_{i-k_2}^i) - H(X_i \mid X_{i-d}^{i-1}, Y_{i-d}^i, Z_{i-d}^i) && \text{(B.9)} \\
\leq & H(X_i \mid X_{i-k_2}^{i-1}, Z_{i-k_2}^i) - H(X_i \mid X_{i-k_2}^{i-1}, Y_{i-k_2}^i, Z_{i-k_2}^i) && \text{(B.10)}
\end{aligned}
$$

where (B.7), (B.8), and (B.10) follow from conditioning reduces entropy and (B.9) follows from joint d-Markovicity of (X,Y,Z). Taking the sum over $i = 1, \ldots, n$ and the normalized limit as $n \to \infty$ gives the desired result, noting that (B.6), (B.7), and (B.10) become the PDI, DI, and TDI rates, respectively. □

Appendix C

Appendix to Chapter 4

C.1 Equivalence of the Interventional and Non-Interventional Measures in Section 4.3.1

First, we introduce the *causal model* defined by Pearl [92, 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 4.3.1, we have that the causal structure is given by Figure 4.1. For the first two examples, the functions are given by equations (4.19) and (4.20), 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 [92, Definition 2.2.2]. Once these causal models are established, showing the equivalence between the interventional and non-interventional measures discussed in Remark 3 can easily be shown using the second rule of the so-called *do*-calculus [90, Theorem 3].

Specifically, showing that $p(x_i \mid x^{i-1}, do(y_{i-1})) = p(x_i \mid x^{i-1}, y_{i-1})$ 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 4.1 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.

C.2 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 4.4.1, the restricted distribution is, in general, non-Markov. As such, it is desirable to have an efficient method for computing the true restricted distribution $p(y_i \mid y^{i-1})$. Here we derive update equations for recursively computing $p(y_i \mid y^{i-1})$. 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 $p(y_i \mid x_i)$ and the one-step prediction distribution is given by $p(x_i \mid x_{i-1})$ [57, Ch. 9].

First, assume X and Y are jointly first order Markov as in Section 4.5.1 and decompose the restricted distribution as the product of "likelihood" and "prior" terms:

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

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:

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

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:

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

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

C.3 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 $p \in \tilde{\mathcal{P}}_n$:*

$$\sum_{i=1}^{n} D(p_i \,||\, \hat{p}_i) \leq M(n) \tag{C.1}$$

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 $g_i : \mathcal{X} \to \mathbb{R}$ for $i = 1, \ldots, n$:*

$$\sum_{i=1}^{n} \left| \mathbb{E}_{p_i^*}[g_i(X)] - \mathbb{E}_{\hat{p}_i}[g_i(X)] \right| \leq \frac{||\vec{c}_n||_2}{\sqrt{2}} \sqrt{M(n)} \tag{C.2}$$

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

$$c_i = \sum_{x\in\mathcal{X}} |g_i(x)| \tag{C.3}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n}\left|\mathbb{E}_{p_i^*}[g_i(X)]-\mathbb{E}_{\hat{p}_i}[g_i(X)]\right| &= \sum_{i=1}^{n}\left|\sum_{x\in\mathcal{X}}\left[p_i^*(x)-\hat{p}_i(x)\right]g_i(x)\right| \\
&\leq \sum_{i=1}^{n}\sum_{x\in\mathcal{X}}\left|p_i^*(x)-\hat{p}_i(x)\right|\left|g_i(x)\right| && \text{(C.4)} \\
&\leq \sum_{i=1}^{n}\left[\sum_{x\in\mathcal{X}}\left|p_i^*(x)-\hat{p}_i(x)\right|\right]\left[\sum_{x\in\mathcal{X}}\left|g_i(x)\right|\right] && \text{(C.5)} \\
&\leq \sum_{i=1}^{n}\sqrt{\frac{1}{2}D(p_i^*\,||\,\hat{p}_i)}\sum_{x\in\mathcal{X}}\left|g_i(x)\right| && \text{(C.6)} \\
&= \frac{1}{\sqrt{2}}\sum_{i=1}^{n}c_i\sqrt{D(p_i^*\,||\,\hat{p}_i)}
\end{aligned}$$

where (C.4) uses the triangle inequality, (C.5) follows from both terms of the sum being positive, and (C.6) uses Pinsker's inequality. Focusing on the sum, we define a vector $\vec{v}_n = [v_1,\ldots,v_n]$ such that $v_i = \sqrt{D(p_i^*\,||\,\hat{p}_i)}$ for $i = 1,\ldots,n$:

$$\begin{aligned}
\sum_{i=1}^{n}c_i\sqrt{D(p_i^*\,||\,\hat{p}_i)} &= \left|\vec{c}_n\cdot\vec{v}_n\right| && \text{(C.7)} \\
&\leq ||\vec{c}||_2\,||\vec{v}||_2 && \text{(C.8)} \\
&= ||\vec{c}||_2\left(\sum_{i=1}^{n}D(p_i^*\,||\,\hat{p}_i)\right)^{\frac{1}{2}} \\
&\leq ||\vec{c}||_2\sqrt{M(n)} && \text{(C.9)}
\end{aligned}$$

where (C.7) follows from the fact that $c_i \geq 0$ and $v_i \geq 0$ for all i, (C.8) uses the Cauchy–Schwarz inequality and (C.9) uses Lemma 2 and the assumption that $M(n) \geq 1$. □

C.4 Proof of Propositions

C.4.1 Proof of Proposition 1

Using the definition of the causal measure, we get that the left hand side of (4.16) is:

$$\begin{aligned}
\sum_{\mathcal{H}_i^{(c)}} p(\mathcal{H}_i^{(c)}) D(p_{X_i}^{(c)} \,||\, p_{X_i}^{(r)}) &= \sum_{\mathcal{H}_i^{(c)}} p(\mathcal{H}_i^{(c)}) \sum_{x_i} p_{X_i}^{(c)}(x_i) \log \frac{p_{X_i}^{(c)}(x_i)}{p_{X_i}^{(r)}(x_i)} \\
&= \sum_{\mathcal{H}_i^{(c)}} p(\mathcal{H}_i^{(c)}) \sum_{x_i} p(x_i \mid \mathcal{H}_i^{(c)}) \log \frac{p(x_i \mid \mathcal{H}_i^{(c)})}{p(x_i \mid \mathcal{H}_i^{(r)})} \\
&= \sum_{\mathcal{H}_i^{(c)}, x_i} p(\mathcal{H}_i^{(c)}, x_i) \log \frac{p(x_i \mid \mathcal{H}_i^{(c)})}{p(x_i \mid \mathcal{H}_i^{(r)})} \\
&= \mathbb{E}_{X^i, Y^{i-1}, Z^{i-1}} \left[\log \frac{p(X_i \mid X^{i-1}, Y^{i-1}, Z^{i-1})}{p(X_i \mid X^{i-1}, Z^{i-1})} \right] \\
&= I(Y^{n-1} \to X^n \,||\, Z^{n-1})
\end{aligned}$$

□

C.4.2 Proof of Proposition 2

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 [66], otherwise known as the KT-estimator. Then it is known that the worst case regret is given by [119]:

$$\sup_{x_n} \log \frac{p(x^n)}{p_e(x^n)} \leq \frac{|\mathcal{X}| - 1}{2} \log n + |\mathcal{X}| - 1 \tag{C.10}$$

We next define the KT-tree estimator with side information $p_c(x^n \,||\, y^n)$ 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 $p_e^{(l)}(x^n \,||\, y^n) \triangleq p_e^{(l)}(x^{(l)})$ 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:

$$\begin{aligned}
\sup_{x^n,y^n} \log \frac{p(x^n)}{p_c(x^n)} &= \sup_{x^n,y^n} \log \frac{p(x^n \,||\, y^n)}{\prod_{l\in\mathcal{L}} p_e^{(l)}(x^n \,||\, y^n)} \\
&= \sup_{x^n,y^n} \log \prod_{l\in\mathcal{L}} \frac{p(x^{(l)})}{p_e^{(l)}(x^{(l)})} \\
&= \sup_{x^n,y^n} \sum_{l\in\mathcal{L}} \log \frac{p(x^{(l)})}{p_e^{(l)}(x^{(l)})} \\
&\leq \sum_{l\in\mathcal{L}} \left(\frac{|\mathcal{X}-1|}{2} \log n_l + |\mathcal{X}| - 1 \right) \qquad \text{(C.11)} \\
&= \frac{L(|\mathcal{X}|-1)}{2} \sum_{l\in\mathcal{L}} \frac{1}{L} \log n_l + L(|\mathcal{X}|-1) \\
&\leq \frac{L(|\mathcal{X}|-1)}{2} \log \sum_{l\in\mathcal{L}} \frac{n_l}{L} + L(|\mathcal{X}|-1) \qquad \text{(C.12)} \\
&= \frac{L(|\mathcal{X}|-1)}{2} \log \frac{n}{L} + L(|\mathcal{X}|-1) \qquad \text{(C.13)}
\end{aligned}$$

where (C.11) follows from the bound in (C.10), (C.12) follows from Jensen's inequality, and (C.13) follows from the fact that $\sum_l n_l = n$. We now define the set of all nodes to be $\mathcal{S} \triangleq \{(x_{i-k}^{i-1}, y_{i-k}^{i-1}) : (x_{i-k}^{i-1}, y_{i-k}^{i-1}) \in \mathcal{X}^k \times \mathcal{Y}^k, k = 1, \ldots, d\}$, with $|\mathcal{S}| = S$. Then, we can define a context tree by letting defining a probability $p_w^{(s)}(x^n \,||\, y^n)$ for each node $s \in \mathcal{S}$ as follows:

$$p_w^{(s)}(x^n \,||\, y^n) = \begin{cases} \frac{1}{2} p_e^{(s)}(x^n \,||\, y^n) + \frac{1}{2} \prod_{s' \in \mathcal{X}\times\mathcal{Y}} p_w^{(s's)}(x^n \,||\, y^n) & s \notin \mathcal{L} \\ p_e^{(s)}(x^n \,||\, y^n) & s \in \mathcal{L} \end{cases} \qquad \text{(C.14)}$$

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 $s' = (x_{i-k-1}, y_{i-k-1})$. Letting λ be the root node of the tree (i.e. $s\lambda = s$), the CTW probability assignment is given by $p_w(x^n \,||\, y^n) \triangleq f_w^{(\lambda)}(x^n \,||\, y^n)$. This probability assignment may be recursively

lower-bounded as:

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

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:

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

as was to be shown. □

C.5 Proof of Theorems

C.5.1 Proof of Theorem 5

We begin by defining the functions:

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

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

$$\sum_{i=1}^{n} \left|\hat{C}_{Y\to X}(i) - C^*_{Y\to X}(i)\right| - \left|\mathbb{E}_{p^{(c)*}_{X_i}}[\hat{g}_i(X)] - \mathbb{E}_{\hat{p}^{(c)}_{X_i}}[\hat{g}_i(X)]\right| \tag{C.15}$$

$$= \sum_{i=1}^{n} \left|\mathbb{E}_{p^{(c)*}_{X_i}}[g^*_i(X)] - \mathbb{E}_{\hat{p}^{(c)}_{X_i}}[\hat{g}_i(X)]\right| - \left|\mathbb{E}_{p^{(c)*}_{X_i}}[\hat{g}_i(X)] - \mathbb{E}_{\hat{p}^{(c)}_{X_i}}[\hat{g}_i(X)]\right|$$

$$\leq \sum_{i=1}^{n} \left|\left|\mathbb{E}_{p^{(c)*}_{X_i}}[g^*_i(X)] - \mathbb{E}_{\hat{p}^{(c)}_{X_i}}[\hat{g}_i(X)]\right| - \left|\mathbb{E}_{p^{(c)*}_{X_i}}[\hat{g}_i(X)] - \mathbb{E}_{\hat{p}^{(c)}_{X_i}}[\hat{g}_i(X)]\right|\right| \tag{C.16}$$

$$\leq \sum_{i=1}^{n} \left|\mathbb{E}_{p^{(c)*}_{X_i}}[g^*_i(X)] - \mathbb{E}_{\hat{p}^{(c)}_{X_i}}[\hat{g}_i(X)] - \mathbb{E}_{p^{(c)*}_{X_i}}[\hat{g}_i(X)] + \mathbb{E}_{\hat{p}^{(c)}_{X_i}}[\hat{g}_i(X)]\right| \tag{C.17}$$

$$= \sum_{i=1}^{n} \left|\mathbb{E}_{p^{(c)*}_{X_i}}[g^*_i(X) - \hat{g}_i(X)]\right|$$

$$= \sum_{i=1}^{n} \left|\mathbb{E}_{p^{(c)*}_{X_i}}\left[\log \frac{p^{(c)*}_{X_i}(X)}{\hat{p}^{(c)}_{X_i}(X)} - \log \frac{p^{(r)*}_{X_i}(X)}{\hat{p}^{(r)}_{X_i}(X)}\right]\right|$$

$$\leq \sum_{i=1}^{n} \left|D(p^{(c)*}_{X_i} \,||\, \hat{p}^{(c)}_{X_i})\right| + \left|\mathbb{E}_{p^{(c)*}_{X_i}}\left[\log \frac{p^{(r)*}_{X_i}(X)}{\hat{p}^{(r)}_{X_i}(X)}\right]\right| \tag{C.18}$$

$$\leq M^{(c)}(n) + M^{(r)}(n) \tag{C.19}$$

where (C.16) follows from the properties of absolute value, (C.17) follows from the reverse triangle inequality, (C.18) follows from the triangle inequality, and (C.19) follows from non-

negativity of the KL-divergence, Lemma 2, and Assumption 2. Moving the second term of (C.15) to the other side of the inequality yields:

$$\begin{aligned}\sum_{i=1}^{n}\left|\hat{C}_{Y\to X}(i)-C^{*}_{Y\to X}(i)\right| &\leq M^{(c)}(n)+M^{(r)}(n)+\sum_{i=1}^{n}\left|\mathbb{E}_{p^{(c)*}_{X_i}}[\hat{g}_i(X)]-\mathbb{E}_{\hat{p}^{(c)}_{X_i}}[\hat{g}_i(X)]\right| \\ &\leq M^{(c)}(n)+M^{(r)}(n)+\frac{||\vec{c}_n||_2}{\sqrt{2}}\sqrt{M^{(c)}(n)} \qquad \text{(C.20)}\end{aligned}$$

where (C.20) follows from Lemma 3. This concludes the proof. □

C.5.2 Proof of Theorem 6

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

$$\begin{aligned}p(X_i,Y_{i-k+1}\mid X^{i-1},Y^{i-k}) &= \sum_{y^{i-1}_{i-k+2}} p(X_i,Y_{i-k+1},y^{i-1}_{i-k+2}\mid X^{i-1},Y^{i-k}) \\ &\triangleq \sum_{y^{i-1}_{i-k+2}} p(X_i,\tilde{Y}^{i-1}_{i-k+1}\mid X^{i-1},Y^{i-k}) \\ &= \sum_{y^{i-1}_{i-k+2}} p(X_i\mid X^{i-1},\tilde{Y}^{i-1})p(\tilde{Y}^{i-1}_{i-k+1}\mid X^{i-1},Y^{i-k}) \\ &= \sum_{y^{i-1}_{i-k+2}} p(X_i\mid X^{i-1}_{i-k-d},\tilde{Y}^{i-1}_{i-k-d})p(\tilde{Y}^{i-1}_{i-k+1}\mid X^{i-1}_{i-k-d},Y^{i-k}_{i-k-d}) \qquad \text{(C.21)} \\ &= \sum_{y^{i-1}_{i-k+2}} p(X_i,\tilde{Y}^{i-1}_{i-k+1}\mid X^{i-1}_{i-k-d},Y^{i-k}_{i-k-d}) \\ &= p(X_i,Y_{i-k+1}\mid X^{i-1}_{i-k-d},Y^{i-k}_{i-k-d})\end{aligned}$$

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 (C.21) follows from the joint Markovicity of X and Y and:

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

where we define $\tilde{Y}_a^b = \emptyset$ when $b < a$. This proves the Markovicity of $\tilde{X}$. To get the equality given by (4.40), 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 $\tilde{x} \in \tilde{\mathcal{X}}$ 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:

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

implying that $\tilde{X}$ is aperiodic. □

C.5.3 Proof of Theorem 7

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

$$\hat{I}^{(k)}_{P,n}(Y \to X) \triangleq \frac{1}{n}\sum_{i=1}^{n} D(\hat{p}^{(c)}_{X_i} \,||\, \hat{p}^{(k)}_{X_i}). \tag{C.22}$$

Then the theorem states that $\hat{I}^{(k)}_{P,n}$ converges to $\bar{I}^{(k)}_{P}$ almost surely. Following the proof of Theorem 3 in [55], decompose the estimate as:

$$\hat{I}^{(k)}_{P,n}(Y \to X) = \frac{1}{n}\sum_{i=1}^{n}\sum_{x_i}\hat{p}^{(c)}_{X_i}(x_i)\log\frac{1}{\hat{p}^{(k)}_{X_i}(x_i)} - \frac{1}{n}\sum_{i=1}^{n}\sum_{x_i}\hat{p}^{(c)}_{X_i}(x_i)\log\frac{1}{\hat{p}^{(c)}_{X_i}(x_i)}.$$

It was shown in [55] 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:

$$F^{(k)}_{n} \triangleq \frac{1}{n}\sum_{i=1}^{n}\sum_{x_i}\hat{p}^{(c)}_{X_i}(x_i)\log\frac{1}{\hat{p}^{(k)}_{X_i}(x_i)}. \tag{C.23}$$

Then it remains to be shown that $F^{(k)}_{n}$ converges to $\bar{H}^{(k)}(X \,||\, Y)$ almost surely. Next, define $R^{(k)}_{n}$

and $S_n^{(k)}$ as:

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

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 [55] 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 6 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:

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

Then, from Breiman's generalized ergodic theorem [17], it follows that the following equality holds almost surely:

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

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

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

Thus, we conclude that:

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

as was to be shown. □

Appendix D

Appendix to Chapter 5

D.1 Exchanging Interventions and Observations

The *do*-calculus provides a set of rules to aid using the *do*-operator in practice and to enable identifying if and how interventional probabilities can be computed. Of particular interest is computing interventional probabilities (i.e. those using the *do*-operator) from the standard conditional probabilities that represent observing variables. This is particularly important in scenarios such as the one considered in Section 5.5, wherein it is infeasible to actually perform interventions. The *do*-calculus consists of three rules, each of which involves an equivalence statement between probabilities that is implied by a d-separation criterion. We here focus on Rule 2, which provides a condition for which observations can be exchanged for actions. Specifically, this rule says that for a DAG $\mathcal{G}$ and any disjoint sets of variables X,Y,Z, and W:

$$(Y \perp\!\!\!\perp_d Z \mid X, W)_{\mathcal{G}_{\overline{X}\underline{Z}}} \implies p(y \mid \hat{x}, \hat{z}, w) = p(y \mid \hat{x}, z, w) \tag{D.1}$$

where $(\cdot \perp\!\!\!\perp_d \cdot \mid \cdot)_{\mathcal{G}}$ represents d-separation with respect to the DAG $\mathcal{G}$ and $\mathcal{G}_{\overline{X}\underline{Z}}$ represents an augmented DAG with all incoming arrows to X and outgoing arrows from Z removed. The rule is framed in a general form in that it allows other variables to be observed or intervened upon

(i.e. W and X) on both sides of the equality. Roughly speaking, this rule says that if the only way Z relates to Y is via descendants of Z, than knowing whether or not a particular value z was observed or forced will not change the distribution of Y. To see this, first let $X = \emptyset$, and note that the d-separation condition becomes $(Y \perp\!\!\!\perp_d Z \mid W)_{\mathcal{G}_{\underline{Z}}}$, i.e. Y is d-separated from Z by W if we ignore all paths coming *out* of Z. If that condition is not satisfied, then observing a value of Z informs us about the values of Z's parents, which then may provide further information on the distribution of Y. By contrast, if we intervene on Z, then no information is conveyed about Z's parents, and the distribution of Y will not be the same. Next, letting $X \neq \emptyset$, we see that the condition now requires removing all incoming arrows to X. This is because if X is intervened upon, it will contain no information about the values of its parents.

This rule is applied in a straightforward manner in two ways in Section 5.5. First, when measuring the effect of ENSO on temperature, we need to exchange an intervention on the ENSO phase for an observation of an ENSO phase. Focusing on the graph on the right side of Figure 5.4, the augmented graph $\mathcal{G}_{\underline{E}}$ is given by E being an isolated node. Thus, in this augmented graph E is d-separated from T by either $\emptyset$ or S, and we have $p(t \mid \hat{e}) = p(t \mid e)$ and $p(t \mid s, \hat{e}) = p(t \mid s, e)$. Similarly, for measuring the effect of S on T, we need to consider the augmented graph $\mathcal{G}_{\underline{S}}$ given by $S \leftarrow E \rightarrow T$. Using the d-separation algorithm described in Algorithm 1, it is straightforward to see that $(S \perp\!\!\!\perp_d T \mid E)_{\mathcal{G}_{\underline{S}}}$ and thus $p(t \mid e, \hat{s}) = p(t \mid e, s)$.

D.2 Conditional Specific Causal Measures

Definition 8. *The partially observed conditional SCDE of x on Y with mediator z in setting $\tilde{u}$ is defined as:*

$$SCDE(x \rightarrow Y; z \mid \tilde{u}) \triangleq D(p(Y \mid \hat{x}, \hat{z}, \tilde{u}) \mid\mid \sum_{x'} p(x' \mid \tilde{u}) p(Y \mid \hat{x}', \hat{z}, \tilde{u}))$$

In the fully observable setting $\tilde{U} = U$ we have:

$$SCDE(x \to Y; z \mid u) \triangleq D(p(Y \mid \hat{x}, \hat{z}, u_Y) \mid\mid \sum_{x'} p(x' \mid u_X) p(Y \mid \hat{x}', \hat{z}, u_Y))$$

Definition 9. *The partially observed conditional SNDE of x on Y in setting $\tilde{u}$ is defined as:*

$$SNDE(x \to Y \mid \tilde{u}) \triangleq D(p(Y \mid \hat{x}, \tilde{u}) \mid\mid \sum_{x', z'} p(x' \mid \tilde{u}) p(z' \mid \hat{x}, \tilde{u}) p(Y \mid \hat{x}', z', \tilde{u}))$$

In the fully observable setting $\tilde{U} = U$ we have:

$$SNDE(x \to Y \mid u) \triangleq D(p(Y \mid \hat{x}, u_Y, u_Z) \mid\mid \sum_{x', z'} p(x' \mid u_X) p(z' \mid \hat{x}, u_Z) p(Y \mid \hat{x}', z', u_Y))$$

Definition 10. *The partially observed conditional SNIE of x on Y in setting $\tilde{u}$ is defined as:*

$$SNIE(x \to Y \mid \tilde{u}) \triangleq D(p(Y \mid \hat{x}, \tilde{u}) \mid\mid \sum_{x', z'} p(x' \mid \tilde{u}) p(z' \mid \hat{x}', \tilde{u}) p(Y \mid \hat{x}, z', \tilde{u}))$$

In the fully observable setting $\tilde{U} = U$ we have:

$$SNIE(x \to Y \mid u) \triangleq D(p(Y \mid \hat{x}, u_Y, u_Z) \mid\mid \sum_{x', z'} p(x' \mid u_X) p(z' \mid \hat{x}', u_Z) p(Y \mid \hat{x}, z', u_Y))$$

D.3 Proof of Theorems

D.3.1 Proof of Theorem 8

The proposition follows directly from the definitions in (5.1) and (5.7):

$$\mathbb{E}_{p(X)}[STE(X \to Y)] = \sum_x p(x) D(p(Y \mid \hat{x}) \mid\mid \sum_{x'} p(x') p(Y \mid \hat{x}')) \tag{D.2}$$

$$= \sum_x p(x) \sum_y p(y \mid \hat{x}) \log \frac{p(y \mid \hat{x})}{\sum_{x'} p(x') p(y \mid \hat{x}')} \tag{D.3}$$

$$= I(X \to Y) \tag{D.4}$$

□

D.3.2 Proof of Theorem 9

Starting with the conditional IF, see that:

$$\begin{aligned} I(X \to Y \mid \hat{Z}) &= \sum_z p(z) \sum_x p(x \mid \hat{z}) \sum_y p(y \mid \hat{x}, \hat{z}) \log \frac{p(y \mid \hat{x}, \hat{z})}{\sum_{x'} p(x' \mid \hat{z}) p(y \mid \hat{x}', \hat{z})} \\ &= \sum_{x,z} p(z) p(x \mid \hat{z}) D(p(y \mid \hat{x}, \hat{z}) \mid\mid \textstyle\sum_{x'} p(x' \mid \hat{z}) p(y \mid \hat{x}', \hat{z})) \\ &= \sum_{x,z} p(z) p(x) D(p(y \mid \hat{x}, \hat{z}) \mid\mid \textstyle\sum_{x'} p(x') p(y \mid \hat{x}', \hat{z})) \qquad \text{(D.5)} \\ &= \mathbb{E}_{p(X)p(Z)}[SCDE(X \to Y; Z)] \end{aligned}$$

where (D.5) follows from the fact that interventions on Z can be ignored in the distribution of X. Moving onto the CS, we have:

$$\begin{aligned}
\mathfrak{C}_{X\to Y} &= D(p(X,Y,Z) \,||\, p_{X\to Y}(X,Y,Z)) && \text{(D.6)}\\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x)p(z \mid x)p(y \mid x,z)}{p(x)p(z \mid x)\left(\sum_{x'} p(x')p(y \mid x',z)\right)} && \text{(D.7)}\\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(y \mid x,z)}{\sum_{x'} p(x')p(y \mid x',z)} && \text{(D.8)}\\
&= \sum_{x,z} p(x,z) \sum_{y} p(y \mid x,z) \log \frac{p(y \mid x,z)}{\sum_{x'} p(x')p(y \mid x',z)} && \text{(D.9)}\\
&= \sum_{x,z} p(x,z) \sum_{y} p(y \mid \hat{x},\hat{z}) \log \frac{p(y \mid \hat{x},\hat{z})}{\sum_{\hat{x}'} p(\hat{x}')p(y \mid \hat{x}',\hat{z})} && \text{(D.10)}\\
&= \sum_{x,z} p(x,z) D(p(Y \mid \hat{x},\hat{z}) \,||\, \textstyle\sum_{x'} p(x')p(Y \mid \hat{x}',\hat{z})) && \text{(D.11)}\\
&= \mathbb{E}_{p(X,Z)}[SCDE(X \to Y;Z)] && \text{(D.12)}
\end{aligned}$$

□

D.3.3 Proof of Theorem 10

Note that the conditional STE, SNDE, and SNIE only utilize three distributions involving interventions, namely $p(y \mid \hat{x},\tilde{u})$, $p(z \mid \hat{x},\tilde{u})$, and $p(y \mid \hat{x},z,\tilde{u})$. We wish to show that we can estimate these distributions can be estimated from observational data, i.e. that the hats can be removed. Assume that the conditions of the theorem hold. We first claim that $(X \perp\!\!\!\perp Y \mid \tilde{U}_1)_{\mathcal{G}_{\underline{X}}} \implies (X \perp\!\!\!\perp Y \mid \tilde{U})_{\mathcal{G}_{\underline{X}}}$ and $(X \perp\!\!\!\perp Z \mid \tilde{U}_2)_{\mathcal{G}_{\underline{X}}} \implies (X \perp\!\!\!\perp Z \mid \tilde{U})_{\mathcal{G}_{\underline{X}}}$. To see this, note that in the DAG $\mathcal{G}_{\underline{X}}$, X has no children, and thus will not be connected to any other nodes in step two of the d-separation algorithm given by Algorithm 1. Since every edge connected to a node in $\tilde{U}$ is removed in step three in the algorithm, the only way for one of the implications to be violated is if there is an undirected path in $\mathcal{G}_{\underline{X}}$ connecting X and Z or X and Y that does

not pass through $\tilde{U}$; however, such a path would necessarily not pass through $\tilde{U}_1$ or $\tilde{U}_2$, which would violate $(X \perp\!\!\!\perp Y \mid \tilde{U}_1)_{\mathcal{G}_{\underline{X}}}$ or $(X \perp\!\!\!\perp Z \mid \tilde{U}_2)_{\mathcal{G}_{\underline{X}}}$. Thus, the claimed implications hold. Next we can directly apply rule two of the *do*-calculus [92, Theorem 3.4.1] to $(X \perp\!\!\!\perp Y \mid \tilde{U})_{\mathcal{G}_{\underline{X}}}$ and $(X \perp\!\!\!\perp Z \mid \tilde{U})_{\mathcal{G}_{\underline{X}}}$ to see that $p(y \mid \hat{x}, \tilde{u}) = p(y \mid x, \tilde{u})$ and $p(z \mid \hat{x}, \tilde{u}) = p(z \mid x, \tilde{u})$. Finally, we claim that $(X \perp\!\!\!\perp Y \mid \tilde{U})_{\mathcal{G}_{\underline{X}}} \implies (X \perp\!\!\!\perp Y \mid Z, \tilde{U})_{\mathcal{G}_{\underline{X}}}$ using the same argument showing the implications above. Applying rule 2 of the *do*-calculus to $(X \perp\!\!\!\perp Y \mid Z, \tilde{U})_{\mathcal{G}_{\underline{X}}}$ yields that $p(y \mid \hat{x}, z, \tilde{u}) = p(y \mid x, z, \tilde{u})$. As such, all three of the interventional distributions needed by the STE, SNDE, and SNIE can be equated to their observational counterparts under the stated assumptions and the proof is completed. □

Bibliography

[1] G. Agamennoni, J. I. Nieto, and E. M. Nebot. Approximate inference in state-space models with heavy-tailed noise. *IEEE Transactions on Signal Processing*, 60(10):5024–5037, 2012.

[2] M. A. Alexander, I. Bladé, M. Newman, J. R. Lanzante, N.-C. Lau, and J. D. Scott. The atmospheric bridge: The influence of ENSO teleconnections on air–sea interaction over the global oceans. *Journal of Climate*, 15(16):2205–2231, 2002.

[3] P.-O. Amblard and O. J. Michel. On directed information theory and Granger causality graphs. *Journal of computational neuroscience*, 30(1):7–16, 2011.

[4] C. Andrieu, A. Doucet, and R. Holenstein. Particle markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

[5] D. Angelosante, G. B. Giannakis, and E. Grossi. Compressed sensing of time-varying signals. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–8. IEEE, 2009.

[6] D. Angelosante, S. I. Roumeliotis, and G. B. Giannakis. Lasso-Kalman smoother for tracking sparse signals. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pages 181–185. IEEE, 2009.

[7] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on signal processing*, 50(2):174–188, 2002.

[8] M. S. Asif and J. Romberg. Dynamic updating for sparse time varying signals. In *Information Sciences and Systems, 2009. CISS 2009. 43rd Annual Conference on*, pages 3–8. IEEE, 2009.

[9] M. S. Asif, A. Charles, J. Romberg, and C. Rozell. Estimation and dynamic updating of time-varying signals with sparse variations. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 3908–3911. IEEE, 2011.

[10] N. Ay and D. Polani. Information flows in causal networks. *Advances in complex systems*,

11(01):17–41, 2008.

[11] D. Ba, B. Babadi, P. L. Purdon, and E. N. Brown. Robust spectrotemporal decomposition by iteratively reweighted least squares. *Proceedings of the National Academy of Sciences*, 111(50):E5336–E5345, 2014.

[12] D. Ba, B. Babadi, P. L. Purdon, and E. N. Brown. Neural spike train denoising by point process re-weighted iterative smoothing. In *Signals, Systems and Computers, 2014 48th Asilomar Conference on*, pages 763–768. IEEE, 2014.

[13] L. Barnett and A. K. Seth. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *J. of Neuroscience Methods*, 2014.

[14] L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical review letters*, 2009.

[15] S. Boyd and L. Vandenberghe. Convex optimization, 2004.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[17] L. Breiman. The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

[18] 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*, pages 2340–2344. IEEE, 2005.

[19] O. Cappé, S. J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924, 2007.

[20] A. Charles, M. S. Asif, J. Romberg, and C. Rozell. Sparsity penalties in dynamical system estimation. In *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, pages 1–6. IEEE, 2011.

[21] A. S. Charles, A. Balavoine, and C. J. Rozell. Dynamic filtering of time-varying sparse signals via L1 minimization. *IEEE Transactions on Signal Processing*, 64(21):5644–5656, 2015.

[22] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, 2016.

[23] T. P. Coleman and S. S. Sarma. A computationally efficient method for nonparametric modeling of neural spiking activity with point processes. *Neural Computation*, 22(8): 2002–2030, 2010.

[24] T. P. Coleman, M. Yanike, W. A. Suzuki, and E. N. Brown. 'A mixed-filter algorithm for dynamically tracking learning from multiple behavioral and neurophysiological measures. *The dynamic brain: an exploration of neuronal variability and its functional significance*, page 1, 2011.

[25] I. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0. `http://cvxr.com/cvx`, Aug. 2012.

[26] M. H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

[27] P. Del Moral. Non-linear filtering: interacting particle resolution. *Markov processes and related fields*, 2(4):555–581, 1996.

[28] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

[29] M. R. DeWeese and M. Meister. How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, 10(4):325–340, 1999.

[30] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[31] R. Dugad and N. Ahuja. Video denoising by combining Kalman and Wiener estimates. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 4, pages 152–156. IEEE, 1999.

[32] J. Eckstein and W. Yao. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pac. J. Optim. To appear*, 2015.

[33] B. Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

[34] M. Eichler and V. Didelez. On granger causality and the effect of interventions in time series. *Lifetime data analysis*, 16(1):3–32, 2010.

[35] J. Etesami and N. Kiyavash. Directed information graphs: A generalization of linear dynamical graphs. In *American Control Conference (ACC), 2014*, pages 2563–2568. IEEE, 2014.

[36] J. Etesami, N. Kiyavash, and T. Coleman. Learning minimal latent directed information polytrees. *Neural computation*, 28(9):1723–1768, 2016.

[37] J. Etesami, A. Habibnia, and N. Kiyavash. Econometric modeling of systemic risk: going beyond pairwise comparison and allowing for nonlinearity, 2017.

[38] J. Etesami, K. Zhang, and N. Kiyavash. A new measure of conditional dependence. *arXiv*

preprint arXiv:1704.00607, 2017.

[39] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development (Online)*, 9(LLNL-JRNL-736881), 2016.

[40] 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*, pages 1917–1925, 2015.

[41] S. J. Godsill, A. Doucet, and M. West. Monte Carlo smoothing for nonlinear time series. *Journal of the american statistical association*, 99(465):156–168, 2004.

[42] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[43] S. Gorantla and T. Coleman. Information-theoretic viewpoints on optimal causal coding-decoding problems. *arXiv preprint arXiv:1102.0250*, 2011.

[44] 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*, 7(1):64–75, 2012.

[45] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[46] R. M. Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

[47] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity*. CRC press, 2015.

[48] G. Hesslow. Two notes on the probabilistic approach to causality. *Philosophy of science*, 43(2):290–292, 1976.

[49] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441 (1):1–46, 2007.

[50] P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[51] P. W. Holland. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):i–50, 1988.

[52] Y. Huang, Y. Zhang, Z. Wu, N. Li, and J. Chambers. A novel robust student's-t based Kalman filter. *IEEE Transactions on Aerospace and Electronic Systems*, 2017.

[53] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554, 2006.

[54] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

[55] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman. Universal estimation of directed information. *IEEE Transactions on Information Theory*, 2013.

[56] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman. Justification of logarithmic loss via the benefit of side information. *IEEE Transactions on Information Theory*, 61(10): 5357–5365, 2015.

[57] D. Jurafsky and J. H. Martin. *Speech and language processing*, volume 3. Pearson London, 2014.

[58] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[59] D. Y. Kang, Y.-S. Kim, G. Ornelas, M. Sinha, K. Naidu, and T. P. Coleman. Scalable microfabrication procedures for adhesive-integrated flexible and stretchable electronic sensors. *Sensors*, 15(9):23459–23476, 2015.

[60] S. Kim, C. J. Quinn, N. Kiyavash, and T. P. Coleman. Dynamic and succinct statistical analysis of neuroscience data. *Proceedings of the IEEE*, 102(5):683–698, 2014.

[61] R. Kleeman. Measuring dynamical prediction utility using relative entropy. *Journal of the atmospheric sciences*, 59(13):2057–2072, 2002.

[62] A. Kolchinsky and B. Corominas-Murtra. Decomposing information into copying versus transformation. *arXiv preprint arXiv:1903.10693*, 2019.

[63] I. Kontoyiannis and M. Skoularidou. Estimating the directed information and testing for causality. *IEEE Transactions on Information Theory*, 62(11):6053–6067, 2016.

[64] G. Krahmann, M. Visbeck, and G. Reverdin. Formation and propagation of temperature anomalies along the north atlantic current. *Journal of Physical Oceanography*, 31(5): 1287–1303, 2001.

[65] G. Kramer. *Directed information for channels with feedback*. PhD thesis, Swiss Institute of Technology, Zurich, 1998.

[66] R. Krichevsky and V. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

[67] A. A. Kulkarni and T. P. Coleman. An optimizer's approach to stochastic control problems with nonclassical information structures. *IEEE Transactions on Automatic Control*, 60(4):

937–949, 2015.

[68] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 1990.

[69] L. D. Lewis, V. S. Weiner, E. A. Mukamel, J. A. Donoghue, E. N. Eskandar, J. R. Madsen, W. S. Anderson, L. R. Hochberg, S. S. Cash, E. N. Brown, and P. L. Purdon. Rapid fragmentation of neuronal networks at the onset of propofol-induced unconsciousness. *Proceedings of the National Academy of Sciences*, 109(49):E3377–E3386, 2012.

[70] J. Li, S.-P. Xie, E. R. Cook, M. S. Morales, D. A. Christie, N. C. Johnson, F. Chen, R. D'Arrigo, A. M. Fowler, X. Gou, et al. El niño modulations over the past seven centuries. *Nature Climate Change*, 3(9):822, 2013.

[71] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[72] S. Liu, J. Chen, P.-Y. Chen, and A. O. Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. *arXiv preprint arXiv:1710.07804*, 2017.

[73] J. T. Lizier. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *arXiv preprint arXiv:1408.3270*, 2014.

[74] J. T. Lizier. Measuring the dynamics of information processing on a local scale in time and space. In *Directed information measures in neuroscience*, pages 161–193. Springer, 2014.

[75] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1):321–353, 2011.

[76] J. G. MacKinnon. Bootstrap hypothesis testing. *Handbook of computational econometrics*, 183:213, 2009.

[77] S. Magnússon, P. C. Weeraddana, M. G. Rabbat, and C. Fischione. On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. *IEEE Transactions on Control of Network Systems*, 3(3):296–309, 2016.

[78] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

[79] H. Marko. Die theorie der bidirektionalen kommunikation und ihre anwendung auf die nachrichtenübermittlung zwischen menschen (subjektive information). *Kybernetik*, 3(3): 128–136, 1966.

[80] H. Marko. The bidirectional communication theory–a generalization of information theory. *IEEE Transactions on communications*, 1973.

[81] 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*, 53(7):1604–1615, 2008.

[82] J. Massey. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pages 303–305. Citeseer, 1990.

[83] V. Matta and A. H. Sayed. Consistent tomography under partial observations over adaptive networks. *IEEE Transactions on Information Theory*, 2018.

[84] C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Proc. of the Eleventh Conf. on Uncertainty in Artificial Intelligence*, 1995.

[85] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[86] Y. Murin. k-NN estimation of directed information. *arXiv preprint arXiv:1711.08516*, 2017.

[87] M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1973.

[88] 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.

[89] N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1 (3):127–239, 2014.

[90] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[91] J. Pearl. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001.

[92] J. Pearl. *Causality*. Cambridge university press, 2009.

[93] J. Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[94] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[95] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Academic press, 2011.

[96] 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, 57(6):3248–3259, 2011.

[97] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[98] A. Pocheville, P. E. Griffiths, and K. Stotz. Comparing causes—an information-theoretic approach to specificity, proportionality and stability. In *Proceedings of the 15th congress of logic, methodology and philosophy of science. College Publications, London*, pages 261–286, 2017.

[99] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. of Comp. Neuroscience*, 2011.

[100] 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*, pages 293–297. IEEE, 2011.

[101] C. J. Quinn, N. Kiyavash, and T. P. Coleman. Directed information graphs. *IEEE Transactions on information theory*, 61(12):6887–6909, 2015.

[102] M. Reichenbach and P. Morrison. The direction of time. *Physics Today*, 9:24, 1956.

[103] J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

[104] R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015.

[105] M. Sahraee-Ardakan and A. K. Fletcher. Estimation and learning of dynamic nonlinear networks (DyNNets). In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2856–2860. IEEE, 2017.

[106] G. Schamberg and T. P. Coleman. Measuring sample path causal influences with relative entropy. *arXiv preprint arXiv:1810.05250*, 2018.

[107] G. Schamberg and T. P. Coleman. A sample path measure of causal influence. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 2251–2255, June 2018. doi: 10.1109/ISIT.2018.8437627.

[108] G. Schamberg, D. Ba, M. Wagner, and T. Coleman. Efficient low-rank spectrotemporal decomposition using ADMM. In *Statistical Signal Processing Workshop (SSP), 2016 IEEE*, pages 1–5. IEEE, 2016.

[109] G. Schamberg, D. Ba, and T. P. Coleman. A modularized efficient framework for non-markov time series estimation. *IEEE Transactions on Signal Processing*, 66(12):3140–3154, 2018.

[110] T. Schreiber. Measuring information transfer. *Physical review letters*, 2000.

[111] A. K. Seth, A. B. Barrett, and L. Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

[112] 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*, 115(17):E3869–E3878, 2018.

[113] A. C. Smith, L. M. Frank, S. Wirth, M. Yanike, D. Hu, Y. Kubota, A. M. Graybiel, W. A. Suzuki, and E. N. Brown. Dynamic analysis of learning in behavioral experiments. *Journal of Neuroscience*, 24(2):447–461, 2004.

[114] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

[115] P. A. Stokes and P. L. Purdon. A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proc. of the National Academy of Sciences*, 2017.

[116] T. Tanaka, P. M. Esfahani, and S. K. Mitter. LQG control with minimum directed information: Semidefinite programming approach. *IEEE Transactions on Automatic Control*, 63 (1):37–52, 2018.

[117] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[118] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[119] 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.

[120] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

[121] N. Vaswani. Kalman filtered compressed sensing. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 893–896. IEEE, 2008.

[122] J. M. Wallace and D. S. Gutzler. Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly Weather Review*, 109(4):784–812, 1981.

[123] E. A. Wan and R. Van Der Merwe. The unscented Kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium*

2000. AS-SPCC. The IEEE 2000, pages 153–158. Ieee, 2000.

[124] H. Wang and A. Banerjee. Online alternating direction method (longer version). *arXiv preprint arXiv:1306.3721*, 2013.

[125] Y. Wang, W. Yin, and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2015.

[126] N. Wiener. The theory of prediction. *Modern mathematics for engineers*, 1956.

[127] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.

[128] Y. Wu and S. Verdú. Optimal phase transitions in compressed sensing. *IEEE Transactions on Information Theory*, 58(10):6241–6263, 2012.

[129] J. Ziniel, L. C. Potter, and P. Schniter. Tracking and smoothing of time-varying sparse signals via approximate belief propagation. In *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, pages 808–812. IEEE, 2010.