Quantifying Context-Dependent Causal Influences

Gabriel Schamberg
Department of Electrical and Computer Engineering
University of California, San Diego
San Diego, CA 92093
gschambe@eng.ucsd.edu

Todd P. Coleman
Department of Bioengineering
University of California, San Diego
San Diego, CA 92093
tpcoleman@ucsd.edu

Abstract

We present an information theoretic framework for quantifying causal influences between random variables in a known causal graph. The proposed measures are "context-dependent" in that their values are determined by a realization of certain nodes in the graph, in particular, the node representing the cause. By contrast, they *do not* depend on the value taken by the effect. This perspective is motivated by the idea that different values of a cause will have different levels of influence on the *distribution* of the effect; however, once those influences have been administered, the exact outcome will occur by chance, according to some conditional distribution. We demonstrate how various approaches to measuring causal influences, such as conditional mutual information and causal strength, can be modified to be context-dependent. The dependence on context gives rise to notions of causal influence that are not captured by non-context-dependent measures, namely *chain reactions*, *caused uncertainty*, and *shared responsibility*. We demonstrate the value of the context-dependent perspective through an analysis of properties and multiple examples and analogies.

1 Introduction

Consider a directed acyclic graph (DAG) $\mathcal{G} = (X, E)$ with nodes $X_i \in X$ for $i = 1, \ldots, n$ and directed edges $(j \to k) \in E$ for j and k such that X_j directly influences X_k. Given such a graph, we seek to provide a *quantification* of the influence that X_j has on X_k. Such a measure has numerous uses, including model selection and exploratory data analysis. Given that human thought is largely driven by causal reasoning, the AI community in particular stands to benefit from a measure of causal influences that is consistent with human intuitions in a variety of settings.

To develop this measure, Janzing et al. [6] propose a set of postulates that, if satisfied, should ensure that the causal measure is consistent with human intuition. They further introduce causal strength (CS), which is shown to satisfy the postulates. While the postulates (and therefore CS) are logical, they are concerned only with the *average* causal influence of one variable on another. To see why this is an incomplete representation of human intuition on causal influences, consider the following simple example: let X represent winning the lottery and Y represent average money spent per month. Clearly, we will have that $X \to Y$; the question is, how much? Any average measure will say that, because X occurs with virtually zero probability, it has a negligible effect on Y. Our intuitions, on the other hand, would say that the level of influence depends on whether or not $X = 1$, i.e. whether or not the lottery was won. While most people would say the lottery has very little influence on their decision making (because most people have not won), a lottery winner would certainly attribute their spending habits to their unlikely victory. Thus, a quantification of $X \to Y$ ought to depend on the value of X. We call this perspective on causal influences "context-dependent."

This perspective is not in itself novel; in fact, it is central to the notion of causality. Pearl and Mackenzie [12] recently introduced the Ladder of Causation, which is topped by the questions "Was

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

it X that caused Y?" and "What if X had not occurred?", both of which are fundamentally concerned with the *values* taken by X. In an analysis of causal interpretations of ANOVA, Northcott [9] argues that ANOVA is insufficient because (among other reasons) a measure of causal strength should be "context-specific." Existing methods that are dependent upon context, however, do not satisfy the postulates proposed by Janzing et al. [6], and therefore are not well suited to address the problem of quantifying causal influences in full generality. For example, the average causal effect considers the average difference in effect that results from switching the value of the cause [5, 11], and as a result, only captures influences on the first-order moment of the effect. The same applies to the path-specific measure presented by Pearl [10].

As such, we propose a framework that bridges the information-theoretic approach of Janzing et al. [6] with the context-dependent perspective. The paper proceeds as follows: Section 2 defines relevant notation. Section 3 provides a brief summary of non-context dependent measures from which this work is built upon. Section 4 defines the context-dependent measures and discusses some key properties. Section 5 provides examples that illustrate new notions of causal influence enabled by the context-dependent perspective. Finally, Section 6 summarizes our contribution and suggests directions for future work.

2 Notation

Define a DAG $\mathcal{G} = (X, E)$ with nodes $X = \{X_1, \ldots, X_n\}$ and edges $E \subset [n] \times [n]$, where $X_i \in \mathcal{X}_i$, $[n] \triangleq \{1, \ldots, n\}$, and the possible edges in E are constrained to be such that $\mathcal{G}$ contains no cycles. At times, the set of nodes will instead be given by $\{X, Y, Z\}$. For the sake of clarity, assume that all random variables are discrete, i.e. that each $\mathcal{X}_i$ is finite with cardinality $|\mathcal{X}_i|$, though all results can be trivially extended to settings with continuous or mixed random variables. Let the joint pmf of the nodes be p (i.e. $X \sim p$), with p satisfying the causal Markov condition with respect to $\mathcal{G}$ [11]. For any two disjoint sets of indices $S_1, S_2 \subseteq [n]$ we have the pmf and conditional pmf given by $p(X_{S_1})$ and $p(X_{S_1} \mid X_{S_2})$, respectively. For any $S \subseteq [n]$, define the set of indices of parent nodes of that set to be $\mathcal{P}_S \triangleq \{k : \exists s \in S \; s.t. \; (X_k \to X_s) \in E\}$. Using this notation, we denote the set of parent nodes by $X_{\mathcal{P}_S}$. Note that, with slight abuse of notation, curly braces may be omitted when the meaning is clear (i.e. $X_{i,j} = X_{\{i,j\}} = \{X_i, X_j\}$). For nodes given by $S \subseteq [n]$, the set of values taken by those nodes to be $\mathcal{X}_S \triangleq \prod_{s \in S} \mathcal{X}_s$. In general, capital letters will be used to represent random variables, while lower case letters are used to represent values (or realizations) of those variables. For example, $p(x_i)$ is the probability of the event $X_i = x_i$. All logarithms will be assumed to be base two. Finally, we define two information theoretic quantities used frequently throughout. First, the entropy and conditional entropy of random variables $(X, Y) \sim p$ are given by $H(X) = -\sum_x p(x) \log p(x)$ and $H(X \mid Y) = -\sum_{x,y} p(x, y) \log p(x \mid y)$, respectively. Second, for two distributions p and p' over $\mathcal{X}$, the KL-divergence from p to p' is given by $D\left(p \mid\mid p'\right) = \sum_x p(x) \log {}^{p(x)}/_{p'(x)}$. In the special case where X is a Bernoulli random variable with $p = Bern(p_1)$ and $p' = Bern(p_2)$, we will equivalently write the KL-divergence as $D\left(p_1 \mid\mid p_2\right)$.

3 Related Work

We will now introduce two information theoretic quantifications of causal influence which can be adapted to be context-dependent.

3.1 Conditional Mutual Information (CMI)

While mutual information is a symmetric measure, it may be interpreted as a representation of causal influence in scenarios where there is a known causal model. Suppose, for example, we wish to measure the influence of a random variable X on another random variable Y in a DAG containing the edge $X \to Y$. If we let Z represent some subset of other nodes (i.e. the other parents or predecessors of Y), then the CMI is given by [3]:

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

These equivalent definitions of CMI give rise to two interpretations. First, we can interpret CMI as the reduction in uncertainty of Y that is obtained by additionally conditioning on X. Intuitively, we

only attribute a causal influence of X on Y if it provides information about Y that is not provided by other non-descendants of Y. Second, we can think of the CMI as measuring the extent to which X aids in the prediction of Y. This interpretation is obtained by viewing the CMI as the expected value of a log-likelihood ratio, where the numerator is given by the true generative distribution of Y and the denominator is given by a distribution of Y that is not conditioned on X. While these interpretations are barely distinguishable from one another (because they are equivalent), we will see that they are no longer equivalent when introducing a context-dependency. A critique of CMI provided by Janzing et al. [6] is that there are cases where X is the only cause of Y, but when taking a limit, Y can be perfectly inferred from Z, yielding a CMI of zero. We will show that these missed causal influences can be avoided by introducing a dependence on context.

A popular notion of causality between time series studied in the information theory literature is directed information (DI) [7, 8], which can be thought of as a generalization of Granger causality [4] for non-linear and/or non-Gaussian settings [1]. The time series causality problem addressed by DI can be seen as a special case of using CMI for the problem considered in this paper, where nodes represent a single sample of a time series and the edges are restricted to point forward in time.

3.2 Causal Strength

The causal strength (CS) was introduced by Janzing et al. [6] as an example of a causal measure that satisfies a set of postulates proposed in the same paper. The CS bears a resemblance to the CMI in that it may be represented as an expected log-likelihood ratio between two distributions:

$$\mathfrak{C}_{X\to Y} \triangleq \mathbb{E}\left[\log \frac{p(Y \mid X, Z)}{\tilde{p}(Y \mid Z)}\right] \tag{2}$$

where $\tilde{p}(Y \mid Z)$ is known as the "post-cutting" distribution and differs from the standard conditional distribution as follows:

$$\tilde{p}(y \mid z) \triangleq \sum_x p(y \mid z, x)p(x) \neq \sum_x p(y \mid z, x)p(x \mid z) = p(y \mid z) \tag{3}$$

As we can see, the post-cutting distribution weights different conditional distributions of Y given X and Z based on the marginal distribution of X. Thus, if Z has a significant influence on the distribution of Y *through* X, then this influence is removed from the post-cutting distribution. The difference between CMI and CS is subtle — in both cases we are comparing the predictability of Y given all of its parents with the predictability of Y given all parents excluding X. The difference between CMI and CS is *how much is known about the hidden* X. In particular, CMI measures the difference in predictability when we can infer something about X from other nodes in the graph while CS enforces that we know nothing about the hidden X, aside from its marginal distribution.

4 Context Dependent Causal Measures

Let X_j be a node whose causal influence upon X_i we wish to measure. We begin by defining a *context* of node X_i to be $x_{\mathcal{P}_i} \in \mathcal{X}_{\mathcal{P}_i}$. It is important to make the distinction that a context of X_i differs from the parent set $X_{\mathcal{P}_i}$ in that the parent set is a set of nodes, whereas a context is a set of values that those nodes have taken on. In other words, a context is a *realization of the parent set*. We further define a context of X_i with the value x_j hidden to be:

$$x_{\mathcal{P}_{i\setminus j}} \triangleq \{x_k : k \in \mathcal{P}_i, k \neq j\} \tag{4}$$

Using the notion of contexts, we will present two context dependent measures of causal influence.

4.1 Context-Dependent Conditional Mutual Information (CDMI)

Begin by defining two conditional pmfs of X_i:

$$p_i(x_i) \triangleq p(x_i \mid x_{\mathcal{P}_i}) \tag{5}$$

$$p_{i\setminus j}(x_i) \triangleq p(x_i \mid x_{\mathcal{P}_{i\setminus j}}, x_{\mathcal{P}_j}) \tag{6}$$

For ease of (and with slight abuse of) notation, it is implied that p_i and $p_{i\setminus j}$ are functions of the appropriate contexts — i.e. $p_i(x_i)$ will be different for two contexts $x_{\mathcal{P}_i} \neq x'_{\mathcal{P}_i}$ despite their absence

from the term $p_i(x_i)$. We can interpret p_i as describing a generative model for our graph, noting that the likelihood of a collection of observations $x = \{x_1, \ldots, x_n\}$ can be factorized as:

$$p(x) = \prod_{i=1}^{n} p_i(x_i) \tag{7}$$

To interpret $p_{i\setminus j}$, note that it may be decomposed as:

$$p_{i\setminus j}(x_i) = \sum_{x'_j \in \mathcal{X}_j} p(x_i \mid x_{\mathcal{P}_{i\setminus j}}, x'_j) p(x'_j \mid x_{\mathcal{P}_j}) \triangleq \sum_{x'_j \in \mathcal{X}_j} p'_i(x_i) p_j(x'_j) \tag{8}$$

where we define p'_i to be the p_i that results from substituting x'_j for x_j in the context $x_{\mathcal{P}_i}$. Using this decomposition, $p_{i\setminus j}$ can be interpreted as an estimate of p_i that is acquired by weighting possible distributions p'_i by the conditional probability of each possible value of the hidden x_j given its context $x_{\mathcal{P}_j}$. Using these two conditional distributions, we define the CDMI from X_j to X_i as:

$$C_{j\to i}(x_{\mathcal{P}_i}, x_{\mathcal{P}_j}) \triangleq D(p_i \mid\mid p_{i\setminus j}) \tag{9}$$

where the causal measure is now a function of the contexts of the cause and the effect. Importantly, we note that the causal measure is not dependent on the value taken by the effect. This aspect of the proposed measure formalizes the perspective that different values of a cause will have different effects on the distribution of an effect, but the particular value of the effect will be randomly sampled from that distribution once the influence has been administered.

As a result of the context dependency, the CDMI is itself a random variable. Taking the expectation over all contexts recovers the non-context-dependent CMI:

$$\mathbb{E}[C_{j\to i}(X_{\mathcal{P}_i}, X_{\mathcal{P}_j})] = I(X_j; X_i \mid X_{\mathcal{P}_{i\setminus j}}, X_{\mathcal{P}_j}) \tag{10}$$

This provides the clearest intuition regarding the insufficiency of a non-context dependent measure of causal influence, namely that if a particular cause is unlikely, then its impact on the causal influence is small, regardless of the impact it may have.

Finally, we note that, as a direct result of the properties of the KL-divergence, the CDMI is always non-negative and is zero for a given context if and only if $p_i(x_i) = p_{i\setminus j}(x_i)$ for all $x_i \in \mathcal{X}_i$. This property is not a given for a context dependent version of CMI. In particular, we note that the equivalent definitions of CMI given by (1) are no longer equivalent when conditioning upon contexts:

$$C_{X\to Y}(x, z) = \mathbb{E}\left[\log \frac{p(Y \mid X, Z)}{p(Y \mid Z)} \Big| X = x, Z = z\right] \neq H(Y \mid Z = z) - H(Y \mid X = x, Z = z) \tag{11}$$

It is clear that the two definitions of CMI yield vastly different interpretations when adapting to the context-dependent setting. The most notable difference is that, for certain contexts, the difference in conditional entropies on the RHS of (11) may be negative, while the proposed CDMI on the LHS is always non-negative. This is because it is possible that there is a greater level of uncertainty in Y when conditioning on a particular value of x, despite that fact that this cannot happen *on average* because conditioning reduces entropy [3]. An example of this scenario is presented in Section 5.2 where we discuss a notion of *caused uncertainty*.

4.2 Context Dependent Causal Strength (CDCS)

Next we demonstrate how the causal strength introduced by [6] can be made context-dependent. First, using notation from the decomposition for $p_{i\setminus j}$ given by (8), define the "post cutting" distribution [6, Definition 1] as:

$$\tilde{p}_{i\setminus j}(x_i) \triangleq \sum_{x'_j \in \mathcal{X}_j} p'_i(x_i) p(x'_j) \tag{12}$$

where $p(x'_j)$ is the marginal distribution of X_j evaluated at x'_j, as opposed to the conditional distribution of X_j given a context as in (8). Using this distribution, we can define the CDCS as:

$$\tilde{C}_{j\to i}(x_{\mathcal{P}_i}) \triangleq D\left(p_i \mid\mid \tilde{p}_{i\setminus j}\right) \tag{13}$$

As with the CDMI, taking the expectation of the CDCS with respect to possible contexts recovers the standard CS. It is important to note that the CDCS is only a function of $x_{\mathcal{P}_i}$, the context of the effect. This enables CDCS to satisfy the locality postulate [6, P2], which states that the influence of X_j on X_i should only depend on how X_i depends on its parents, $X_{\mathcal{P}_i}$, and the joint distribution of $X_{\mathcal{P}_i}$. While CDMI does not satisfy this postulate (as is made clear by (6)), a counterargument to the need for locality is that it removes any ability to detect *chain reactions* (see Section 5.1). Thus, locality may not be desired in all causal inference settings, and the decision between use of CDMI and CDCS should be problem-specific.

4.3 Direct vs. Total Effects, Interventions, and Counterfactuals in the Context Dependent Framework

The proposed context-dependent measures of causal influence have interesting interpretations with regard to common discussion topics in causal inference. In particular, we note that simple adjustments to what constitutes a context can enable a distinction between total and direct effects or the ability to utilize the do-operator of Pearl [11]. While a complete discussion of these extensions is postponed for future work, they are briefly discussed here. We will show in Section 5 that, even in simple observational settings where only direct effects are considered, the context-dependent measures of causal influence give rise to interesting results.

Both the CDMI and CDCS are measurements of the direct effect of one variable on another. If $X_j \notin \mathcal{P}_i$, then $C_{j\to i}(x_{\mathcal{P}_i}, x_{\mathcal{P}_j}) = \tilde{C}_{j\to i}(x_{\mathcal{P}_i}) = 0$ for all $(x_{\mathcal{P}_i}, x_{\mathcal{P}_j}) \in \mathcal{X}_{\mathcal{P}_i} \times \mathcal{X}_{\mathcal{P}_j}$, i.e. X_j will have a causal influence of zero upon X_i for any context if it is not a parent of X_i. Furthermore, consider a three-node DAG given by $X_3 \leftarrow X_1 \rightarrow X_2 \rightarrow X_3$. Then, when measuring the influence of X_1 on X_3 for a given context (x_1, x_2), the distributions that take an average over possible values of X_1 given by (6) and (12) neglect to consider the effects those values have on the value taken by X_2, which remains fixed at x_2. As such, $C_{1\to 3}$ and $\tilde{C}_{1\to 3}$ do not capture any effects that X_1 has on X_3 *through* X_2. In order to extend context-dependent perspective to measure the *total effect* of X_j on X_i, we can modify the contexts to exclude any intermediaries of the two variables in question (i.e. X_2 in the above example) and instead include the parents of those intermediaries. The total context dependent causal influence would not depend on the values of the intermediate causes. This can be seen as an extension of the previously noted perspective (that a cause will only impact the *distribution* of an effect) to intermediate effects.

Building on the measurement of total effects, we can incorporate the notion of interventions using the do-operator [11]. In particular, we note that if the contexts are set up as described above to account for the total effect of X_j on X_i, then we can equivalently think of the value x_j as being replaced by the intervention $do(x_j)$. This is not in itself a novel contribution, as it follows from Pearl's causal calculus, which determines the scenarios in which we can substitute an intervention for an observation [11]. A more interesting note is the relationship to counterfactual reasoning provided by this approach. In particular, we note that even when we use the interventional lens of $do(x_j)$, the context $x_{\mathcal{P}_j}$ still factors into equation (6). While it is, at first, counterintuitive to include the value taken by the parents of X_j when considering the influence of an intervention on X_j, it makes sense when we frame the causal measure as asking the counterfactual question: "How different would I expect the distribution of X_i to be if I *had not* intervened on X_j?" Thus, rather than considering the effect of an intervention (or observation) relative to a particular alternative intervention, the proposed measures compare the influence of a particular cause with the expected course taken by nature in a particular context.

5 Context-Dependent Notions of Causal Influence

We now introduce three notions of causal influence that rely upon the context-dependent perspective. In each case, the concept is illustrated by example. A formal characterization of these concepts will be the subject of future work. For all of the following examples our focus will be on the CDMI.

5.1 Chain Reactions

For the first example we will consider a simplified version of the example proposed by Ay and Polani [2] and modified to include noise by Janzing et al. [6]. Specifically, consider the scenario where a

message is being passed through a chain of messengers (i.e. random variables). In such a scenario, the corresponding DAG is given by a straight line, i.e. $X_1 \to X_2 \to X_3 \to \ldots \to X_n$. We will consider the simplest case, where a binary message is being passed between nodes, with the message being flipped with probability ϵ and interpret X_i as "the message passed from messenger i to messenger $i+1$". Formally, let $X_i \in \{0,1\}$ with:

$$X_i = \begin{cases} X_{i-1} & w.p.\ 1-\epsilon \\ X_{i-1} \oplus 1 & w.p.\ \epsilon \end{cases} \tag{14}$$

where $\oplus$ is the XOR operation. In determining $C_{j\to i}$, we need only to consider the case $j = i-1$, for otherwise it will be zero (for all contexts). The distributions that determine the causal measure are given by $p_i(x_i) = p(x_i \mid x_{i-1})$ and $p_{i\setminus j}(x_i) = p(x_i \mid x_{i-2})$ and the probabilities of $X_i = 1$ are given by:

$$p_i(1) = \begin{cases} \epsilon & x_{i-1} = 0 \\ 1-\epsilon & x_{i-1} = 1 \end{cases} \qquad p_{i\setminus j}(1) = \begin{cases} 2\epsilon(1-\epsilon) & x_{i-2} = 0 \\ \epsilon^2 + (1-\epsilon)^2 & x_{i-2} = 1 \end{cases} \tag{15}$$

with $p_i(0) = 1 - p_i(1)$ and $p_{i\setminus j}(0) = 1 - p_{i\setminus j}(1)$. Finally, noting that the relevant contexts are given by $\{x_{\mathcal{P}_i}, x_{\mathcal{P}_j}\} = \{x_{i-1}, x_{i-2}\}$, the causal measure is given by:

$$C_{j\to i}(x_{i-1}, x_{i-2}) = \begin{cases} D\left(\epsilon \mid\mid 2\epsilon(1-\epsilon)\right) & x_{i-1} = x_{i-2} \\ D\left(\epsilon \mid\mid \epsilon^2 + (1-\epsilon)^2\right) & x_{i-1} \neq x_{i-2} \end{cases} \tag{16}$$

We can see that when $\epsilon = 1/2$, the causal measure is zero for any context. This is consistent with intuition, as each messenger will pass on the message zero or one with equal probability, regardless of the message it receives. Then as ϵ approaches zero, we get:

$$C_{j\to i}(x_{i-1}, x_{i-2}) \to \begin{cases} 0 & x_{i-1} = x_{i-2} \\ \infty & x_{i-1} \neq x_{i-2} \end{cases} \tag{17}$$

To understand this result, fix ϵ to be an arbitrarily small number and we can say with very high confidence that each messenger will pass on its received message accurately. Thus, when $x_{i-1} = x_{i-2}$, it is, in a sense, unreasonable to endow X_{i-1} with responsibility for causing the value taken by X_i when it is propagating the message in a nearly deterministic manner (note that for any fixed $\epsilon > 0$ the causal measure will not be *exactly* zero). In such a case, it is not so much x_{i-1} that is causing X_i, but rather an earlier x_{i-k} for some $k \in \{1, \ldots, i-1\}$ that initiated a *chain reaction*. On the other hand, in the unlikely occurrence that $x_{i-1} \neq x_{i-2}$, intuition would say that X_{i-1} absolutely has a causal effect on X_i. This scenario can be thought of as X_{i-1} acting of its own volition in selecting a message to pass to X_i.

We acknowledge that the notion of an unbounded causal influence is initially unsettling. When looking closer, however, this property of the causal measure is intuitive. First, we note that for any fixed $\epsilon > 0$, the CDMI will be finite. It is only for $\epsilon = 0$ that the CDMI could be infinite, but in that case, the context that results in infinite CDMI 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.

5.1.1 Analogy – Telephone Wire

Consider a signal passing along a telephone wire, and suppose we segment the wire into arbitrarily small intervals. Let X_1 give the signal sent on one end and X_n be the signal received on the other end, with each segment X_i passing on the signal it received with very high probability. Recalling that the CDMI is itself a random variable, we expect that, with very high probability, X_1 has a large causal influence on X_2, but the influence of X_2 on X_3, X_3 on X_4, and so on, would be negligible. In this sense, we can think of the signal initiation as starting a *chain reaction*. On the other hand, in the unlikely event that the message was corrupted by a malfunction in the wire, we would see a spike in causal influence at the point where the wire malfunctioned. This is consistent with our intuition – when our phones function properly, we attribute what we hear to the person speaking on the other side, but when a call is dropped, we attribute it to a malfunction somewhere in the chain of communication.

5.2 Causing Uncertainty

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

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

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

Moving on to the CDMI, we consider $C_{X\to Z}(x,y)$ and $C_{Y\to Z}(x,y)$ for $(x,y) \in \{0,1\}^2$. Given the symmetry of the problem with respect to X, we only need to consider two of the four possible contexts, namely $(x_0,y_0) \triangleq (0,0)$ and $(x_0,y_1) \triangleq (0,1)$. In order to compute the CDMI for each X and Y to Z for both contexts, we need the following distributions:

$$\begin{aligned} &p(Z \mid x_0,y_0) = Bern(0.1) && p(Z \mid x_0,y_1) = Bern(0.5) \\ &p(Z \mid y_0) = Bern(0.5) && p(Z \mid y_1) = Bern(0.5) \\ &p(Z \mid x_0) = Bern(0.14) \end{aligned}$$

For a given context, the CDMI is given by $C_{X\to Z}(x,y) = D\left(p(Z \mid x,y) \,||\, p(Z \mid y)\right)$ and $C_{Y\to Z}(x,y) = D\left(p(Z \mid x,y) \,||\, p(Z \mid x)\right)$:

$$C_{X\to Z}(x,y) \approx \begin{cases} 0.53 & y = 0 \\ 0.00 & y = 1 \end{cases} \qquad C_{Y\to Z}(x,y) \approx \begin{cases} 0.01 & y = 0 \\ 0.52 & y = 1 \end{cases} \tag{18}$$

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

5.3 Shared Responsibility

Consider a scenario where a collection of n iid variables $X_i \sim Bern(\epsilon)$ 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 and CMI are equivalent for this problem setting. While there is no simple computation of the CS or CMI 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 $\epsilon \to 0$, 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 or CS.

Now consider a context $\{x_i\}_{i=1}^n$ and the corresponding CDMI $C_{X_1 \to Y}(\{x_i\}_{i=1}^n)$. While the influence of each x_i on Y will *not* be the same for a given context, 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:

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

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

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

$$C_{X_1 \to Y}(x_1, k_1) = D\left(p(Y \mid k) \,||\, p(Y \mid k_1)\right) = \begin{cases} D\left(\frac{1}{2^{k_1}} \,||\, \frac{\epsilon}{2^{k_1+1}} + \frac{1-\epsilon}{2^{k_1}}\right) & x_1 = 0 \\ D\left(\frac{1}{2^{k_1+1}} \,||\, \frac{\epsilon}{2^{k_1+1}} + \frac{1-\epsilon}{2^{k_1}}\right) & x_1 = 1 \end{cases}$$

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 $C_{X_1 \to Y}(0, k_1) < C_{X_1 \to Y}(1, k_1)$, i.e. an inhibitor has a greater influence when it is activated. More interestingly, we note that $C_{X_1 \to Y}(x_1, k_1)$ 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 $\epsilon \to 0$, its influence will be infinite.

5.3.1 Analogy – Faulty Parts

Consider a scenario as described above, where X_i represents whether or not a particular airplane component is constructed with defects and Y represents whether or not the airplane crashes as a result of malfunctioning parts. In real life scenarios, it is reasonable to expect that an airplane component has a very small probability of being constructed with defects. As such, both CS and DI would say that the manufacturing of a plane part has very little causal influence on the functionality of the part. This is not a complete picture – intuitively, we do not think of a properly manufactured plane component as having a causal influence on the outcome of a flight, at least not to the extent that a defective component does. Moreover, if by some strange event there were numerous faulty parts, then we would expect each component to have a lesser responsibility than the case where a single component was faulty. In this sense, the context-dependent measures are uniquely able to capture *shared responsibility*.

6 Discussion

This work is motivated by the observation that different values of a cause will have different levels of influence upon an effect. The particular value taken by the effect, however, will in general be random, according to some distribution determined in part by the cause. In order to incorporate this observation into a method for measuring causal influences, we have introduced context-dependent extensions to conditional mutual information and causal strength. These methods, in their non-context-dependent form, are determined entirely by the underlying joint distribution of the variables in question, and thus are incomplete characterizations of causal influence. We have shown that by introducing a context-dependence, three new notions of causal influence — *chain reactions*, *causing uncertainty*, and *shared responsibility* — are easily identified by context-dependent measures. Moreover, these context-dependent measures are random variables whose expectations recover their non-context-dependent counterparts.

There are numerous directions for future explorations within the proposed framework. First, there are a number of ideas discussed here in need of formalization, including measurement of total effects, a modification of causal strength postulates [6] to accommodate a measure whose value is random, and formal definitions of the notions of influence that were discussed in Section 5. Beyond the ideas discussed above, natural next steps include measuring group effects or accounting for unobserved variables in the graph. A final important direction for future work is development of estimators and associated bounds on accuracy. In particular, estimating the causal measures for a given context requires estimation of two distributions (i.e. (5) and (6)), so it may be possible to relate performance guarantees of density estimators to the performance of estimators of the causal measures.

References

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

[2] N. Ay and D. Polani. Information flows in causal networks. *Advances in complex systems*, 11 (01):17–41, 2008.

[3] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[4] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[5] P. W. Holland. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):i–50, 1988.

[6] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, B. Schölkopf, et al. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

[7] H. Marko. The bidirectional communication theory–a generalization of information theory. *IEEE Transactions on communications*, 21(12):1345–1351, 1973.

[8] J. Massey. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pages 303–305. Citeseer, 1990.

[9] R. Northcott. Can ANOVA measure causal strength? *The Quarterly review of biology*, 83(1): 47–55, 2008.

[10] 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.

[11] J. Pearl. *Causality*. Cambridge university press, 2009.

[12] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.