Fundamental Properties of Causal Entropy and Information Gain

Francisco N. F. Q. Simoes F.SIMOES@UU.NL
Mehdi Dastani M.M.DASTANI@UU.NL
Thijs van Ommen T.VANOMMEN@UU.NL
Department of Information and Computing Sciences, Utrecht University

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Recent developments enable the quantification of *causal control* given a structural causal model (SCM). This has been accomplished by introducing quantities which encode changes in the entropy of one variable when intervening on another. These measures, named causal entropy and causal information gain, aim to address limitations in existing information theoretical approaches for machine learning tasks where causality plays a crucial role. They have not yet been properly mathematically studied. Our research contributes to the formal understanding of the notions of causal entropy and causal information gain by establishing and analyzing fundamental properties of these concepts, including bounds and chain rules. Furthermore, we elucidate the relationship between causal entropy and stochastic interventions. We also propose definitions for causal conditional entropy and causal conditional information gain. Overall, this exploration paves the way for enhancing causal machine learning tasks through the study of recently-proposed information theoretic quantities grounded in considerations about causality.

Keywords: Structural Causal Models, Information Theory, Causal Inference

1. Introduction

Information theoretical quantities are ubiquitous in machine learning. Cross-entropy losses are commonly used in deep learning (Goodfellow et al., 2016). Decision tree learning algorithms commonly use mutual information (often called information gain in that context) to decide what variable to split at each step (James et al., 2013) — see for example the ID3 (Quinlan, 1986) and CR4.5 (Quinlan, 1993) algorithms. In Haarnoja et al. (2018), reinforcement learning tasks see their stability and robustness improved when the agents learn by maximizing not only their expected rewards but also the entropy of their policies. Also in Seitzer et al. (2021), mutual information is used for reinforcement learning tasks to measure whether or not an agent's action has "causal influence" on a state, and they use this quantity to decide if said action has "control" over that state. In Achille and Soatto (2018), an optimal representation $\mathbf{Z}$ of a random vector $\mathbf{X}$ for a given task $\mathbf{Y}$ is learned by minimizing the information bottleneck lagrangian, which is a difference between the mutual informations $I(\mathbf{X};\mathbf{Z})$ and $I(\mathbf{Y};\mathbf{Z})$ (Tishby et al., 2000), the idea being that $\mathbf{Z}$ should keep as little information about $\mathbf{X}$ as possible while retaining as much information about $\mathbf{Y}$ as possible. In Höltgen (2021), the aforementioned information bottleneck lagrangian is used in the loss function of an autoencoder in order to learn a causally-relevant representation for a given task. The goal of this approach is to enable us to intervene on the representation instead of on the original variables without losing much control over the task variable.

In the cases just described the existence of confounders or selection bias could lead to misleading results if one ascribes a causal interpretation. This is a common issue when using standard information theoretical quantities in situations that require consideration of the underlying causal relationships. A version of mutual information which takes into account the causal structure of the system would solve this problem. Preliminary work in this direction was recently done by Simoes et al. (2023) in the context of interpretable machine learning and inspired by the earlier philosophical work of Griffiths et al. (2015). The former define "causal entropy" and "causal information gain" and study their relationship with total causal effect. They also argue that causal information gain provides an adequate measure of causal control, so that it can be used when deciding which variables in an SCM provide more control over a chosen target variable. However, a proper mathematical study of the properties of causal entropy and causal information gain is missing.

In this paper, we relate causal entropy with other quantities such as conditional entropy and post-stochastic intervention entropy, revealing potentially more convenient approaches to its computation as well as new interpretations of this quantity. We also define conditional causal entropy and conditional causal information gain. Additionally, we derive fundamental properties of both causal entropy and causal information gain, drawing upon analogous results from information theory.

The novelty of our work consists of deriving key properties of causal entropy and causal information gain for the first time, and defining conditional causal entropy and information gain. Concretely:

- We show that, perhaps unexpectedly, causal entropy is not the same as the entropy after a stochastic intervention, and establish formal relations between causal entropy and stochastic interventions. Furthermore, we show that causal entropy can be seen as a post-stochastic-intervention conditional entropy.

- We check that causal entropy is non-negative, but that, surprisingly, causal information gain can be negative. We also find upper bounds on the causal entropy, including an independence bound mirroring the independence bound on standard entropy. We check that, unexpectedly, the causal version of the data processing inequality does not hold.

- We define conditional causal entropy and conditional causal information gain for the first time. We use these to derive chain rules for both causal entropy and causal information gain. Finally, we discuss how alternative causal versions of conditional information gain are possible, and study one in particular.

This paper is organized as follows. Section 2 introduces the assumptions and the definitions of information theoretical quantities that will be used throughout the paper, including causal entropy and causal information gain. In Section 3 we discuss how causal entropy differs from the entropy of the post-stochastic-intervention distribution. We study how causal entropy can be linked to post-stochastic interventions, resulting in measures of causal entropy which can be both elucidating and easier to compute in practice. We also establish lower and upper bounds on the causal entropy, define conditional causal entropy, and present some crucial findings that culminate in a chain rule for causal entropy. Section 4 lays down fundamental properties of causal information gain, including a chain rule. It also introduces conditional causal information gain and discusses its interpretation, along with the consideration of an alternative causal extension of conditional mutual information termed post-intervention mutual information. Section 5 compares the definitions and results in this text with that of other work that has been done before. In Section 6 we discuss the results obtained in this work and propose future avenues of research. The proofs of all the results presented in this paper can be found in the appendix.

2. Formal Setting

Many basic concepts from causal inference and information theory are used throughout this paper. For completeness, we include the necessary definitions from causal inference in Appendix A. All random variables are henceforth assumed to be discrete and have finite range. In this paper, a "random variable" can be a random vector. In cases where we want to restrict ourselves to random variables which are in fact random vectors, boldface is used. *E.g.* X can be both a single-value random variable or a random vector, while $\mathbf{X}$ must be a random vector. Furthermore, the symbols of the form X, Y, Z, X_i, Y_i or Z_i are taken to be endogenous variables of some SCM $\mathcal{C}$.

In this section we present the definitions from information theory which are necessary for the rest of this paper. We also include the definitions of causal entropy and causal information gain.

2.1. Entropy and Mutual Information

In this subsection we will start by stating the definitions of entropy, conditional entropy and mutual entropy. In the interest of space, we will not try to motivate these definitions. For more information, see Cover and Thomas (2006).

Definition 1 (Entropy and Cond. Entropy (Cover and Thomas, 2006)) *Let X be a discrete random variable with range R_X and p be a probability distribution for X. The* entropy of X w.r.t. the distribution p *is*[1]

$$H_{X\sim p}(X) := -\sum_{x\in R_X} p(x)\log p(x). \tag{1}$$

Entropy is measured in bit. *If the context suggests a canonical probability distribution for X, one can write $H(X)$ and refers to it simply as the* entropy of X.
The conditional entropy *$H(Y \mid X)$ of Y conditioned on X is the expected value w.r.t. p_X of the entropy $H(Y \mid X = x) := H_{Y\sim p_{Y\mid X=x}}(Y)$:*

$$H(Y \mid X) := \mathbb{E}_{x\sim p_X}\left[H(Y \mid X = x)\right]. \tag{2}$$

This means that the conditional entropy $H(Y \mid X)$ is the entropy of $H(Y)$ that remains on average if one conditions on X.

Remark 2 *Notice that $H(Y \mid X = x)$ is seen as a function of x and the expected value in Equation (2) is taken over the random variable x with distribution p_X. This disrespects the convention that random variables are represented by capital letters, but preserves the convention that the specific value conditioned upon is represented by a lower case letter. We will follow the common practice and opt to use lower case letters for random variables in these cases.*

There are two common equivalent ways to define mutual information (often called information gain).

Definition 3 (Mutual Information and Cond. Mutual Information (Cover and Thomas, 2006)) *Let X and Y be discrete random variables with ranges R_X and R_Y and distributions p_X and p_Y, respectively. The* mutual information *between X and Y is*

$$I(X;Y) := \sum_{x,y\in R_X\times R_Y} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}. \tag{3}$$

[1] In this article, log denotes the logarithm to the base 2.

Or equivalently:

$$\begin{aligned} I(X;Y) &:= H(Y) - H(Y \mid X) \\ &= H(X) - H(X \mid Y). \end{aligned} \tag{4}$$

Let Z be another discrete random variable. The conditional mutual information *between X and Y conditioned on Z is:*

$$\begin{aligned} I(X;Y \mid Z) &:= H(Y \mid Z) - H(Y \mid X, Z) \\ &= H(X \mid Z) - H(X \mid Y, Z). \end{aligned} \tag{5}$$

The view of mutual information as entropy reduction from Equation (4) is the starting point for the definition of causal information gain.

2.2. Causal Entropy and Causal Information Gain

We will now define causal entropy and causal information gain. See Simoes et al. (2023) for a thorough discussion about these concepts. The causal entropy of Y for X is the entropy of Y that is left, on average, after one atomically intervenes on X. It is defined in a manner analogous to conditional entropy (see Definition 1). Concretely, causal entropy is the average uncertainty one has about Y if one sets X to x with probability $p_{X'}(x)$, where X' is a new auxiliary variable with the same range as X but independent of all other variables, including X.

Definition 4 (Causal Entropy, H_c (Simoes et al., 2023)) *Let Y, X and X' be random variables such that X and X' have the same range and X' is independent of all variables in $\mathcal{C}$. We say that X' is an* intervention protocol *for X. The* causal entropy *$H_c(Y \mid do(X \sim X'))$ of Y for X given the intervention protocol X' is the expected value w.r.t. $p_{X'}$ of the entropy $H(Y \mid do(X = x)) := H_{Y \sim p_Y^{do(X=x)}}(Y)$ of the interventional distribution $p_Y^{do(X=x)}$. That is:*

$$H_c(Y \mid do(X \sim X')) := \mathbb{E}_{x \sim p_{X'}} \left[H(Y \mid do(X = x)) \right]. \tag{6}$$

Causal information gain extends mutual information/information gain to the causal context. While mutual information between two variables X and Y is the average reduction in uncertainty about Y if one observes the value of X (see Equation (4)), the causal information gain of Y for X is the average decrease in the entropy of Y after one atomically intervenes on X (folowing an intervention protocol X').

Definition 5 (Causal Information Gain, I_c (Simoes et al., 2023)) *Let Y, X and X' be random variables such that X' is an intervention protocol for X. The* causal information gain *$I_c(Y \mid do(X \sim X'))$ of Y for X given the intervention protocol X' is the difference between the entropy of Y w.r.t. its prior and the causal entropy of Y for X given the intervention protocol X'. That is:*

$$I_c(Y \mid do(X \sim X')) := H(Y) - H_c(Y \mid do(X \sim X')). \tag{7}$$

The causal information gain of Y for X was proposed in Simoes et al. (2023) as a measure of the "(causal) control that variable X has over the variable Y". This is a qualitative concept used in the philosophy of science literature (Pocheville et al., 2015) and defined in Simoes et al. (2023) as the reduction of uncertainty about Y that results from intervening on X. This is precisely what causal information gain measures, by construction. It is important to note that a common measure of causal strength such as average causal effect (ACE) would not be a suitable measure of causal control. Indeed, the uncertainty about Y can be reduced by intervening on X while maintaining the average of $p_Y^{do(X=1)}$ the same as that of $p_Y^{do(X=0)}$, yielding an ACE of zero even though uncertainty is reduced by intervening on X.

3. Properties of Causal Entropy

We start this section by showing that causal entropy is distinct from the entropy after a stochastic intervention $do(X = X')$. We study the relation of causal entropy with other quantities, providing new insights about causal entropy and establishing some first basic properties. This section ends with the definition of conditional causal entropy and the derivation of a chain rule for causal entropy.

3.1. Comparison with Entropy After a Stochastic Intervention

One could think that $H_c(Y \mid do(X \sim X')) = H(Y \mid do(X = X'))$: both are entropies of Y resulting from making X follow the distribution $p_{X'}$, albeit through two distinct procedures. While these may appear identical, this is, in reality, not accurate. In other words, the average uncertainty about Y after *atomically* intervening on X by setting $X = x$ with probability $p_{X'}$ is not the same as the average uncertainty after *stochastically* intervening on X by setting X to X'. Example 1 will illustrate this. The underlying reason for this difference will be made clear by Proposition 8.

Example 1 *We will look at an example where* $H_c(Y \mid do(X \sim X')) \neq H(Y \mid do(X = X'))$. *Consider the SCM over the variables* X, Y *with ranges* $R_X = \{0, 1\}$ *and* $R_Y = \{0, 1, 2\}$, *characterized by the following structural assignments and noise distributions:*

$$\begin{cases} f_X(N_X) = N_X \\ f_Y(X, N_Y) = X + N_Y \\ N_X, N_Y \sim \text{Bern}(\frac{1}{2}) \end{cases} \tag{8}$$

Notice that the causal graph is then simply $X \to Y$. *Further, let* X' *be an intervention protocol for* X *with* $p_{X'} = \text{Bern}(\frac{1}{3})$. *The atomic and stochastic interventions on* Y *can then be written as in Table 1. Hence*[2] $H_c(Y \mid do(X \sim X')) = 1(\text{bit})$ *and:*

Y	$p_Y^{do(X=0)}$	$p_Y^{do(X=1)}$	$p_Y^{do(X=X')}$
0	1/2	0	1/3
1	1/2	1/2	1/2
2	0	1/2	1/6

Table 1: Post-intervention distributions for Y in Example 1. (Computed in Appendix C).

$$H(Y \mid do(X = X')) = \frac{1}{3}\underbrace{\log 3}_{>1} + \frac{1}{2} + \frac{1}{6}\underbrace{\log 6}_{>1} > 1(\text{bit}). \tag{9}$$

Thus in particular $H(Y \mid do(X = X')) \neq H_c(Y \mid do(X \sim X'))$ *in this example.*

[2] The details of the computations of these entropies can be found in Appendix C.

Remark 6 (Intuition for Example 1) *The difference between causal and post-intervention entropies observed in Example 1 stems from the fact that, since $Y = X + N_Y$, fixing X renders the distribution of Y equal in shape to that of N_Y. Hence the post-atomic intervention entropies $H_{Y \sim p_Y^{do(X=x)}}(Y)$ are all the same (1 bit), so that averaging just gives 1 bit. Notice that this does not depend on the choice of $p_{X'}$. Now, $H(Y \mid do(X = X'))$ is the entropy of the distribution of the sum of random variables $X' + N_Y$, meaning that in this case the shape of the distribution of Y is changed, not simply shifted.*

3.2. Alternative Views of Causal Entropy

We will now see that causal entropy can be written as an expected value over a post-stochastic intervention joint distribution. It is often useful to have an expression for a statistic as an average w.r.t. the joint distribution of the variables involved. This enables, for instance, the straightforward construction of an estimator for the statistic, known as the "plug-in estimator", achieved by substituting the joint distribution figuring in the expected value by the empirical joint distribution (Wasserman, 2004).

Proposition 7 (Causal Entropy as a Single Expected Value) *The causal entropy of Y given the intervention protocol X' for X can be written as an expected value w.r.t. the post-stochastic-intervention distribution $p_{X,Y}^{do(X=X')}$ as follows:*

$$H_c(Y \mid do(X \sim X')) = -\mathbb{E}_{x,y \sim p_{X,Y}^{do(X=X')}} \Big[\log p(y \mid do(X = x)) \Big]. \tag{10}$$

We will now see how causal entropy relates with post-stochastic intervention entropies. We use $H(Y \mid X = x, do(X = X'))$ as notation for the entropy $H_{Y \sim p_{Y|X=x}^{do(X=X')}}(Y)$ of the covariate-specific effect[3] $p_{Y|X=x}^{do(X=X')}$. That is, $H(Y \mid X = x, do(X = X'))$ is the entropy resulting from conditioning on $X = x$ *after* having performed the stochastic intervention $do(X = X')$.

Proposition 8 (Causal Entropy as Average Entropy of Covariate-Specific Effects) *The causal entropy of Y given the intervention protocol X' for X can be seen as the expected value w.r.t. $x \sim p_{X'}$ of the entropies $H(Y \mid X = x, do(X = X'))$ of the covariate-specific effects $p_{Y|X=x}^{do(X=X')}$. That is:*

$$\begin{aligned} H_c(Y \mid do(X \sim X')) &= \mathbb{E}_{x \sim p_{X'}} \big[H(Y \mid X = x, do(X = X')) \big] \\ &= H(Y \mid X, do(X = X')) \end{aligned} \tag{11}$$

where the notation used in the second equality is analogous to the notation for conditional entropy.

This result further elucidates the origin of the difference discussed in Example 1. Concretely, Proposition 8 shows us that $H_c(Y \mid do(X \sim X'))$ is the average of the entropies of Y obtained from stochastically intervening (according to $p_{X'}$) AND knowing the value that X was set to, while $H(Y \mid do(X = X'))$ is simply the entropy of Y after performing the stochastic intervention – which we can interpret as having performed an atomic intervention on X, but not knowing exactly which value X was set to.

[3] The term "covariate-specific effect" is commonly used when conditioning on a variable distinct from the intervened variable. In this paper we use the term also when the conditioned variable coincides with the intervened variable.

Notice that Proposition 8 implies that the causal entropy is a *bona fide* conditional entropy, where the conditioning is performed after a stochastic intervention setting $X = X'$. Indeed, this is precisely the meaning of $H(Y \mid X, do(X = X'))$. Consequently, we can harness known properties of conditional entropy to prove properties of causal entropy.

3.3. Bounds on Causal Entropy

Just like conditional entropy, causal entropy is a non-negative quantity. This follows simply from the fact that causal entropy is an average of entropies, which are themselves non-negative.

Proposition 9 *Causal entropy is non-negative.*

The following result is not so much a property of causal entropy as it is the absence of one – namely, causal entropy is not necessarily smaller than the initial entropy. This can be surprising, since it is in stark contrast with conditional entropy, which is always less that the initial entropy. Meaning that, on average, information about the conditioning variable X can never increase the uncertainty about Y (Cover and Thomas, 2006). In contrast, Proposition 10 tells us that, on average, intervening on X can in fact increase the uncertainty about Y.

Proposition 10 *For some SCMs and intervention protocols X', the causal entropy of Y for X given an intervention protocol X' is greater than the initial entropy of Y,* i.e. $H_c(Y \mid do(X \sim X')) > H(Y)$.

There is however an upper bound on causal entropy. It follows immediately from Proposition 8 that causal entropy, being related to the entropy after the stochastic intervention $do(X = X')$ by conditioning, cannot be greater than the latter.

Corollary 11 *The causal entropy of Y for X given an intervention protocol X' cannot be greater than the post-stochastic intervention entropy,* i.e. $H_c(Y \mid do(X \sim X')) \leq H(Y \mid do(X = X'))$.

One can make use of the connection between causal entropy and post-stochastic intervention conditional entropy together with the independence bound on entropy (Cover and Thomas, 2006, Theorem 2.6.6) to derive an independence bound on causal entropy.

Proposition 12 (Independence Bound on Causal Entropy) *Let $\mathbf{Y}$ be a random vector of length n_Y. Then:*

$$H_c(\mathbf{Y} \mid do(X \sim X')) \leq \sum_{i=1}^{n_Y} H_c(Y_i \mid do(X \sim X')). \tag{12}$$

and equality holds if and only if the Y_i are independent.

3.4. Conditional Causal Entropy and the Chain Rule

We can of course mix intervening and conditioning. In this section we will start by defining conditional causal entropy $H_c(Y \mid Z, do(X \sim X'))$, which will capture the uncertainty that we have about Y given that we intervened on X according to the intervention protocol X' and then conditioned on Z. We will then see that, unsurprisingly, conditioning reduces causal entropy on average, and that, just like causal entropy, conditional causal entropy can also be seen as a conditional entropy after a stochastic intervention. We will conclude this section with a chain rule for causal entropy.

Definition 13 (Conditional Causal Entropy) *Let X, Y and Z be endogenous variables of an SCM $\mathcal{C}$ and $x \in R_X$. The* atomic conditional causal entropy $H_c(Y \mid Z, do(X = x))$ *of Y conditioned on Z for the atomic intervention $do(X = x)$ is defined as the post-atomic intervention conditional entropy of Y conditioned on Z,* i.e.*:*

$$\begin{aligned} H_c(Y \mid Z, do(X = x)) &:= \mathbb{E}_{z \sim p_Z^{do(X=x)}}[H(Y \mid Z = z, do(X = x))] \\ &= \mathbb{E}_{z \sim p_Z^{do(X=x)}}[H_{Y \sim p_{Y \mid Z=z}^{do(X=x)}}(Y)]. \end{aligned} \tag{13}$$

Moreover, the conditional causal entropy $H_c(Y \mid Z, do(X \sim X'))$ *of Y conditioned on Z given the intervention protocol X' for X is defined as the expected value given the intervention protocol X' of the atomic conditional causal entropies,* i.e.*:*

$$H_c(Y \mid Z, do(X \sim X')) := \mathbb{E}_{x \sim p_{X'}}[H_c(Y \mid Z, do(X = x))]. \tag{14}$$

Notice that our definition of conditional causal entropy assumes that the intervention precedes the conditioning operation. This assumption will hold throughout the paper. A definition of a “condition-first conditional causal entropy” where one intervenes after conditioning would be more involved, and demands the use of counterfactuals. This quantity should coincide with the conditional causal entropy here defined whenever X does not have a causal effect on Z (see *e.g.* Figure 1(b))— in those cases, conditioning before or after intervening will result in the same distribution. In the interest of space, we leave further discussion about this topic for future work.

Similarly to causal entropy, conditional causal entropy can also be regarded as the entropy of the distribution resulting from conditioning on X following the stochastic intervention $do(X = X')$. They differ only in the conditioning set.

Proposition 14 (Conditional Causal Entropy as Conditional Entropy) *Let X, Y, Z and X' be as in Definition 13. Then:*

$$H_c(Y \mid Z, do(X \sim X')) = H(Y \mid Z, X, do(X = X')). \tag{15}$$

Since this conditioning set is a superset of the conditioning set in $H(Y \mid X, do(X = X'))$, we can use the fact that conditioning cannot increase entropy to conclude that the conditional causal entropy cannot be larger than the causal entropy.

Proposition 15 (Conditioning Reduces Causal Entropy) *The conditional causal entropy is never larger than the causal entropy.* I.e. *for X, Y, Z and X' as in Definition 13, we have:*

$$H_c(Y \mid Z, do(X \sim X')) \leq H_c(Y \mid do(X \sim X')). \tag{16}$$

Utilizing Proposition 8 and Proposition 14 to express causal entropy and conditional causal entropy as conditional entropies enables us to use the standard chain rule for conditional entropy to derive a chain rule for causal entropy.

Proposition 16 (Two-variable Chain Rule for Causal Entropy) *Let X, Y, Z and X' be as in Definition 13. Then:*

$$H_c(Y, Z \mid do(X \sim X')) = H_c(Y \mid do(X \sim X')) + H_c(Z \mid Y, do(X \sim X')). \tag{17}$$

Similarly, we can obtain a chain rule specifically for random vectors by leveraging the general chain rule for the conditional entropy.

Proposition 17 (Chain Rule for Causal Entropy) *Let* $\mathbf{Y}$ *be a random vector of length* n_Y *and* $\mathbf{Y}_{<i} = Y_1, \ldots, Y_{i-1}$*. Then:*

$$H_c(\mathbf{Y} \mid do(X \sim X')) = \sum_{i=1}^{n_Y} H_c(Y_i \mid \mathbf{Y}_{<i}, do(X \sim X')). \tag{18}$$

4. Properties of Causal Information Gain

We start this section by noting basic properties of causal information gain. We continue by defining conditional causal information gain and deriving a chain rule for causal information gain. This section ends with a discussion about the interpretation of conditional causal information gain and post-intervention mutual information. The latter is a distinct quantity from conditional causal information gain which serves as an alternative reasonable extension of conditional information gain in the context of causality.

4.1. Immediate Properties of Causal Information Gain

A few properties of causal information gain can be immediately gleaned from its definition. In contrast with mutual information, causal information gain is *not* symmetric. Similarly to causal entropy, one needs to specify an intervention protocol X' which specifies the probability of each atomic intervention on X. As shown in Simoes et al. (2023), if X has no total effect on Y, then $I_c(Y \mid do(X \sim X')) = 0$ for any protocol X'. In contrast with mutual information, causal information gain can be negative. This is an immediate corollary of Proposition 10.

Corollary 18 *For some SCMs and intervention protocols* X'*, the causal information gain of* Y *given the intervention protocol* X' *for* X *is negative,* i.e. $I_c(Y \mid do(X \sim X')) < 0$.

Since KL divergences are non-negative, this means in particular that causal information gain cannot be written as a KL divergence of two distributions, again contrary to its non-causal counterpart.

One of the most important results in information theory is the data processing inequality. It tells us that, for a Markov chain $X \to Y \to Z$, Z can never have more information about X than Y has (Cover and Thomas, 2006). It is natural to wonder whether a similar result holds for causal information gain. Specifically, one might ask if, for a causal chain $X \to Y \to Z$, the causal information gain of Z for X can never be larger than the causal information gain of Y for X. Such a proposition aligns with the intuitive notion that we have less control over variables "farther away" from the intervened variable. Surprisingly, this is false. No such causal data processing inequality holds for causal information gain. To see this, we just need to devise an SCM whose causal graph is a chain $X \to Y \to Z$ and $I_c(Y \mid do(X \sim X')) < I_c(Z \mid do(X \sim X'))$. This situation can arise, for instance, if performing an atomic intervention on X still results in some uncertainty regarding a subset of values of Y, but every such value leads to the same Z.

Example 2 *Consider a causal chain* $X \to Y \to Z$ *with ranges* $R_X = \{x_1, x_2\}$, $R_Y = \{y_1, y_2, y_3\}$ *and* $R_Z = \{z_1, z_2\}$*, and whose structural assignments and noise distributions are given by:*

$$\begin{aligned} &N_X \sim \mathcal{U}[R_X] \\ &N_Y \sim \mathcal{U}\{y_1, y_2\} \\ &X := N_X \end{aligned} \qquad Y := \begin{cases} N_Y, & X = x_1 \\ y_3, & X = x_2 \end{cases} \qquad Z := \begin{cases} z_1, & Y = y_1 \text{ or } Y = y_2 \\ z_2, & Y = y_3 \end{cases}$$

Furthermore, we choose an intervention protocol X' *with a point mass at* x_1, i.e. $X' \sim \delta(x_1)$. *Then* $H(X) = H(Z) = 1$ *and* $H(Y) = 2 \times \frac{1}{4}\log(4) + \frac{1}{2}\log(2) = \frac{3}{2}$. *The relevant causal entropies are* $H_c(Y \mid do(X \sim X')) = H_c(Y \mid do(X = x_1)) = 1$ *and* $H_c(Z \mid do(X \sim X')) = H_c(Z \mid do(X = x_1)) = 0$. *Hence* $I_c(Y \mid do(X \sim X')) = \frac{1}{2} < I_c(Z \mid do(X \sim X')) = 1$.

4.2. Conditional Causal Information Gain and the Chain Rule

Recall the definition of conditional mutual information in Equation (5). It captures how much the uncertainty about Y is reduced on average after one *observes* X, *if one knows* Z. Similarly, the "conditional causal information gain" will be defined such that it captures how much the uncertainty about Y is reduced on average if one *sets* X to x with probability $p_{X'}(x)$, *and*[4]*one knows* Z. Accordingly, its definition will be the gap between the conditional entropy of Y conditioned on Z and the conditional causal entropy of Y conditioned on Z given an intervention protocol X' for X.

Definition 19 (Conditional Causal Information Gain) *Let* X, Y *and* Z *be endogenous variables of an SCM* $\mathcal{C}$. *The* conditional causal information gain $I_c(Y \mid Z, do(X \sim X'))$ *of* Y *conditioned on* Z *given the intervention protocol* X' *for* X *is defined as follows:*

$$I_c(Y \mid Z, do(X \sim X')) := H(Y \mid Z) - H_c(Y \mid Z, do(X \sim X')). \tag{19}$$

We can now leverage the chain rule for the causal entropy in Proposition 17 to obtain a chain rule for the causal information gain.

Proposition 20 (Chain Rule for Causal Information Gain) *Let* $\mathbf{Y}$ *be a random vector of length* n_Y *and* $\mathbf{Y}_{<i} = Y_1, \ldots, Y_{i-1}$. *Then:*

$$I_c(\mathbf{Y} \mid do(X \sim X)) = \sum_{i=1}^{n_Y} I_c(Y_i \mid \mathbf{Y}_{<i}, do(X \sim X)). \tag{20}$$

4.3. Choices and Interpretations

In the causal inference literature, if both the do operator and a random variable appear after the conditioning bar, it is to be understood that the intervention precedes conditioning (Peters et al., 2017). We chose to respect this convention, so that here too interventions precede conditioning. Indeed, the second term in Equation (19) relies on conditional causal entropy, which itself respects this convention. Another choice was made by using $H(Y \mid Z)$ as the first term in Equation (19). Notice that this is the average entropy of Y due to conditioning on Z, *with respect to the pre-intervention joint distribution* $p_{Y,Z}$. We will now look at the interpretation of conditional causal information gain and subsequenty introduce and interpret the quantity resulting from making a different choice for this first term, originating another reasonable causal generalization of conditional mutual information.

Definition 19 tells us that $I_c(Y \mid Z, do(X \sim X'))$ is the information that is gained about Y if one intervenes on X before observing Z as opposed to only observing Z. In other words, it measures how much intervening on X improves the information that is gained about Y by observing Z.

[4] The word "and" does not establish the order between intervening and conditioning. As explained in Section 3.4, we assume that interventions take precedence. See also the discussion in Section 4.3.

Example 3 *Radiologists often use substances called contrast agents before performing MRI (magnetic resonance imaging) scans to enhance image quality. Consult the causal graph in Figure 1(a). We want to assess the impact of using a contrast agent ($X = 1$) on the information that can be extracted about the disease Y from the MRI image Z. This is represented by the conditional causal information gain $I_c(Y \mid Z, do(X \sim X')) = H(Y \mid Z) - H(Y \mid Z, do(X = 1))$, where X' was chosen to have a point mass at 1. This precisely measures the information gained about disease Y by employing the contrast agent before viewing the image Z, as opposed to solely observing Z without the use of the contrast agent.*

In a case where X has no total causal effect on Z such as the one depicted in Figure 1(b), $I_c(Y \mid Z, do(X \sim X'))$ can also be interpreted as the average information that is gained by intervening in strata of the population with the same value of Z. This comes about because in such situations the order between conditioning and intervening is irrelevant. In the case depicted this would be the information that is gained about the probability of patients having a stroke given that we intervene on their blood pressure, averaged over the age groups $Z = z$.

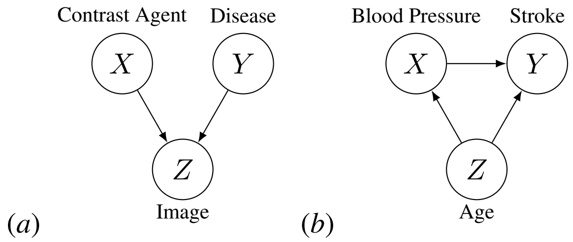

Figure 1: Causal graphs illustrating the interpretations of I_c and MI_c.

We will now see that replacing $H(Y \mid Z)$ with the causal entropy $H_c(Y \mid do(X \sim X'))$ in the first term of Equation (19) also results in a sensible quantity.

Definition 21 *Let X, Y and Z be endogenous variables of an SCM $\mathcal{C}$. The* post-intervention mutual information $\mathrm{MI}_c(Y \mid Z, do(X \sim X'))$ *between Y and Z given the intervention protocol X' for X is defined as follows:*

$$\mathrm{MI}_c(Y \mid Z, do(X \sim X')) := H_c(Y \mid do(X \sim X')) - H_c(Y \mid Z, do(X \sim X')). \tag{21}$$

This quantity measures the average information that is gained about Y due to observing Z, given that X was intervened on following the protocol X'. There should therefore be a close connection between $\mathrm{MI}_c(Y \mid Z, do(X \sim X'))$ and the mutual information between Y and Z after one performs an atomic intervention on X[5]. Indeed, the post-intervention mutual intervention turns out to be the average mutual information between Y and Z after an atomic intervention on X:

Proposition 22 *Let X' be an intervention protocol for X. Then $\mathrm{MI}_c(Y \mid Z, do(X \sim X')) = \mathbb{E}_{x \sim p_{X'}}[I(Y; Z \mid do(X = x))]$, where the notation $I(Y; Z \mid do(X = x))$ indicates that the mutual information is computed with respect to the joint post-atomic-intervention distribution.*

Example 4 *We now revisit the scenario outlined in Example 3 and juxtapose the meanings of conditional causal information gain and post-intervention mutual information within this context. Recall*

[5] This is the reason for the name of this quantity.

that the intervention protocol had been chosen to be $X' \sim \delta(1)$. While $I_c(Y \mid Z, do(X \sim X'))$ quantifies the impact that the contrast agent X has on the information that is gained about Y by observing Z, $\mathrm{MI}_c(Y \mid Z, do(X \sim X')) = I(Y; Z \mid do(X = 1))$ is the information that is gained about the disease Y by observing the image Z, given that the contrast agent has been injected.

For the case of Figure 1(b), $\mathrm{MI}_c(Y \mid Z, do(X \sim X'))$ is simply the average information that is gained about the probability of having a stroke by knowing the age, given that one intervened on blood pressure following a chosen intervention protocol.

5. Related Work

We build upon the work developed in Simoes et al. (2023), which was itself inspired in work from the philophy of science literature (Griffiths et al., 2015). Simoes et al. (2023) define causal counterparts of entropy and mutual information, referred to as causal entropy and causal information gain. These metrics are designed to evaluate the extent to which a feature has control over a chosen outcome variable. They do this by capturing changes in the entropy of a variable resulting from intervening on other variables. The authors also study the relationship between these quantities and the existence of total causal effect.

Another causal generalization of mutual information had been proposed before (Ay and Polani, 2008). This quantity, termed “information flow”, is both conceptually and numerically distinct from causal information gain: Information flow was introduced both as a measure of “causal strength” and of “causal independence”, similarly to how standard mutual information is a measure of statistical independence. This is accomplished by starting from the definition of mutual information as a KL divergence and proceeding to “make it causal” by replacing conditioning with interventions. In contrast, Simoes et al. (2023) treat entropy as the main quantity of interest. They start from the definition of mutual entropy as the change in entropy due to conditioning, and define causal entropy as the change in entropy due to intervening. This then results in a quantity that is appropriate for evaluating the control that a variable has over another, where control is taken to be how much one can reduce the uncertainty about the second by intervening on the first. As confirmation that these two causal generalizations of mutual information are indeed distinct, one can simply notice that the former can be written as a KL divergence (Ay and Polani, 2008), while the latter cannot (Corollary 18). It should be noted that there are metrics other than the ACE and the information flow which purport to measure strength. See Janzing et al. (2013) for a compilation of such metrics. From those, the only one conceptually close to ours is information flow, given that it also serves as a causal version of mutual information.

6. Discussion and Conclusion

The motivation for extending traditional entropy and mutual information to interventional settings stems from the desire to develop algorithms that utilize information theoretical quantities in the presence of non-causal statistical dependencies (*e.g.* due to unobserved confounding). Extending these quantities to handle interventions allows them to capture the effects of manipulating one variable on another. The causal entropy, conditional causal entropy, causal information gain, and conditional causal information gain, together with their basic properties proved herein provide the foundation for developing new algorithms in areas where one has or can obtain knowledge of the causal relationships involved.

Establishing practical methods for computing these quantities remains a topic for future investigation. This includes designing appropriate estimators, evaluating their performance characteristics (such as consistency, bias and rate of convergence), and determining whether these estimators are identifiable based on the available information about the underlying causal structure. Furthermore, one can leverage the foundational results established in this paper (such as the chain rules) to aid the computation of causal control between variables in a complex structural causal model. More generally, the properties presented herein can aid in examining how intervening on a variable within a structural causal model impacts the uncertainty associated with other variables. As alluded to in the Introduction, potential applications of these concepts include guiding action selection in reinforcement learning, devising causal adaptations of entropy-based decision tree algorithms, and developing causal versions of representation learning algorithms which rely on information theoretical quantities such as the mutual information between the representation and a target variable. On the theoretical front, one may try to generalize other important results from information theory to this causal information theoretical framework. Additionally, other causal generalizations of conditional causal entropy and information gain can be studied, in particular those for which the assumption that interventions are performed before conditioning is dropped. Lastly, the precise connection between causal information gain and putative measures of causal strength, such as information flow, could be studied in detail.

Acknowledgments

This publication is part of the CAUSES project (KIVI.2019.004) of the research programme Responsible Use of Artificial Intelligence which is financed by the Dutch Research Council (NWO) and ProRail.

References

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40 (12):2897–2905, 2018.

Nihat Ay and Daniel Polani. Information flows in causal networks. *Advances in complex systems*, 11(01):17–41, 2008.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, second edition, 2006.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Paul E Griffiths, Arnaud Pocheville, Brett Calcott, Karola Stotz, Hyunju Kim, and Rob Knight. Measuring causal specificity. *Philosophy of science*, 82(4):529–555, 2015.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

Benedikt Höltgen. Encoding causal macrovariables. *arXiv preprint arXiv:2111.14724*, 2021.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kevin B Korb, Lucas R Hope, Ann E Nicholson, and Karl Axnick. Varieties of causal intervention. In *PRICAI 2004: Trends in Artificial Intelligence: 8th Pacific Rim International Conference on Artificial Intelligence, Auckland, New Zealand, August 9-13, 2004. Proceedings 8*, pages 322–331. Springer, 2004.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Arnaud Pocheville, Paul Griffiths, and Karola Stotz. Comparing causes – an information-theoretic approach to specificity, proportionality and stability. *15th Congress of Logic, Methodology, and Philosophy of Science*, 08 2015.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 1993.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 22905–22918, 2021.

Francisco Nunes Ferreira Quialheiro Simoes, Mehdi Dastani, and Thijs van Ommen. Causal entropy and information gain for measuring causal control. In *European Conference on Artificial Intelligence*, pages 216–231. Springer, 2023.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.

Appendix A. Structural Causal Models

One can model the causal structure of a system by means of a "structural causal model", which can be seen as a Bayesian network (Koller and Friedman, 2009) whose graph G has a causal interpretation and each conditional probability distribution (CPD) $P(X_i \mid \mathrm{PA}_{X_i})$ of the Bayesian network stems from a deterministic function f_{X_i} (called "structural assignment") of the parents of X_i. In this context, it is common to separate the parent-less random variables (which are called "exogenous" or "noise" variables) from the rest (called "endogenous" variables). Only the endogenous variables are represented in the structural causal model graph. As is commonly done (Peters et al., 2017), we assume that the noise variables are jointly independent and that exactly one noise variable N_{X_i} appears as an argument in the structural assignment f_{X_i} of X_i. In full rigor[6](Peters et al., 2017):

Definition 23 (Structural Causal Model) *Let X be a random variable with range R_X and $\mathbf{W}$ a random vector with range $R_{\mathbf{W}}$. A* structural assignment for X from $\mathbf{W}$ *is a function $f_X \colon R_{\mathbf{W}} \to R_X$. A* structural causal model (SCM) $\mathcal{C} = (\mathbf{X}, \mathbf{N}, S, p_{\mathbf{N}})$ *consists of:*

1. *A random vector $\mathbf{X} = (X_1, \dots, X_n)$ whose variables we call* endogenous.

2. *A random vector $\mathbf{N} = (N_{X_1}, \dots, N_{X_n})$ whose variables we call* exogenous *or* noise.

3. *A set S of n structural assignments f_{X_i} for X_i from $(\mathrm{PA}_{X_i}, N_{X_i})$, where $\mathrm{PA}_{X_i} \subseteq \mathbf{X}$ are called* parents *of X_i. The* causal graph $G^{\mathcal{C}} := (\mathbf{X}, E)$ *of $\mathcal{C}$ has as its edge set $E = \{(P, X_i) : X_i \in \mathbf{X},\ P \in \mathrm{PA}_{X_i}\}$. The PA_{X_i} must be such that the $G^{\mathcal{C}}$ is a directed acyclic graph (DAG).*

4. *A jointly independent probability distribution $p_{\mathbf{N}}$ over the noise variables. We call it simply the* noise distribution.

We denote by $\mathcal{C}(\mathbf{X})$ the set of SCMs with vector of endogenous variables $\mathbf{X}$. Furthermore, we write $X := f_X(X, N_X)$ to mean that $f_X(X, N_X)$ is a structural assignment for X.

Notice that for a given SCM the noise variables have a known distribution $p_{\mathbf{N}}$ and the endogenous variables can be written as functions of the noise variables. Therefore the distributions of the endogenous variables are themselves determined if one fixes the SCM. This brings us to the notion of the entailed distribution[6] (Peters et al., 2017):

Definition 24 (Entailed distribution) *Let $\mathcal{C} = (\mathbf{X}, \mathbf{N}, S, p_{\mathbf{N}})$ be an SCM. Its* entailed distribution $p^{\mathcal{C}}_{\mathbf{X}}$ *is the unique joint distribution over $\mathbf{X}$ such that $\forall X_i \in \mathbf{X},\ X_i = f_{X_i}(\mathrm{PA}_{X_i}, N_{X_i})$. It is often simply denoted by $p^{\mathcal{C}}$. Let $\mathbf{x}_{-i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$. For a given $X_i \in \mathbf{X}$, the marginalized distribution $p^{\mathcal{C}}_{X_i}$ given by $p^{\mathcal{C}}_{X_i}(x_i) = \sum_{\mathbf{x}_{-i}} p^{\mathcal{C}}_{\mathbf{X}}(\mathbf{x})$ is also referred to as* entailed distribution (of *X_i).*

Having an SCM in hand allows us to model interventions on the system. The idea is that an SCM represents how the values of the random variables are generated, and by intervening on a variable system we are effectively changing its generating process – read: its structural assignment. Thus intervening on a variable can be modeled by modifying the structural assignment of said variable, resulting in a new SCM differing from the original only in the structural assignment of the intervened variable, and possibly introducing a new noise variable for it, in place of the old one. Naturally, the new SCM will have an entailed distribution which is in general different from the distribution entailed by the original SCM.

Definition 25 (General intervention) *Let $\mathcal{C} = (\mathbf{X}, \mathbf{N}, S, p_{\mathbf{N}})$ be an SCM, $X_i \in \mathbf{X}$, $\widetilde{\mathrm{PA}}_{X_i} \subseteq \mathbf{X}$, $\tilde{N}_i$ be a random variable and $\tilde{f}_{X_i}$ be a structural assignment for X_i from $\widetilde{\mathrm{PA}}_{X_i}, \tilde{N}_i$.*

The intervention $do(X_i = \tilde{f}_{X_i}(\widetilde{\mathrm{PA}}_{X_i}, \tilde{N}_i))$ *is the function $\mathcal{C}(\mathbf{X}) \to \mathcal{C}(\mathbf{X})$ given by $\mathcal{C} \mapsto \mathcal{C}^{do(X_i = \tilde{f}_{X_i}(\widetilde{\mathrm{PA}}_{X_i}, \tilde{N}_i))}$, where $\mathcal{C}^{do(X_i = \tilde{f}_{X_i}(\widetilde{\mathrm{PA}}_{X_i}, \tilde{N}_i))}$ is the ordered pair $(\mathbf{X}, \tilde{\mathbf{N}}, \tilde{S}, p_{\tilde{N}})$, with $\tilde{\mathbf{N}} = (\mathbf{N} \setminus \{N_i\}) \cup \{\tilde{N}_i\}$ and $\tilde{S} = (S \setminus \{f_{X_i}\}) \cup \{\tilde{f}_{X_i}\}$. We call it the* post-intervention SCM (w.r.t. the intervention $do(X_i = \tilde{f}_{X_i}(\widetilde{\mathrm{PA}}_{X_i}, \tilde{N}_i))$). *It is also denoted $\tilde{\mathcal{C}} := \mathcal{C}^{do(X_i = \tilde{f}_{X_i}(\widetilde{\mathrm{PA}}_{X_i}, \tilde{N}_i))}$.*

[6] We slightly rephrase the definition provided in Peters et al. (2017) for our purposes.

Note that in order for $\tilde{\mathcal{C}}$ to be an SCM, $\widetilde{\mathrm{PA}}_{X_i}$ *must be such that the causal graph* $G^{\tilde{\mathcal{C}}}$ *is a DAG. We then say that the variable X_i was* intervened on.

The distribution $p^{do(X_i=\tilde{f}_{X_i}(\widetilde{\mathrm{PA}}_{X_i},\tilde{N}_i))} := p^{\tilde{\mathcal{C}}}$ *entailed by $\tilde{\mathcal{C}}$ is called the* post-intervention distribution (w.r.t. the intervention $do(X_i = \tilde{f}_{X_i}(\widetilde{\mathrm{PA}}_{X_i}, \tilde{N}_i))$ on $\mathcal{C}$).

The most common type of interventions are the so-called "atomic interventions", where one sets a variable to a chosen value, effectively replacing the distribution of the intervened variable with a point mass distribution. In particular, this means that the intervened variable has no parents after the intervention.

Definition 26 (Atomic intervention) *Let $\mathcal{C} = (\mathbf{X}, \mathbf{N}, S, p_{\mathbf{N}})$ be an SCM and $X_i \in \mathbf{X}$. An* atomic intervention on *X_i is an intervention of the type $do(X_i = \tilde{N}_i)$, where $\tilde{N}_i$ is a random variable with range R_{X_i} and $p_{\tilde{N}_i}(x_i) = \delta_{x,x_i}$ for some $x \in R_{X_i}$. Such an intervention is usually denoted simply by $do(X_i = x)$.*

Another special type of intervention is the "stochastic intervention" (Korb et al., 2004), where again the intervened variable has no parents, but its distribution can be any distribution. Thus, an atomic intervention is a particular type of stochastic intervention.

Definition 27 (Stochastic intervention) *Let $\mathcal{C} = (\mathbf{X}, \mathbf{N}, S, p_{\mathbf{N}})$ be an SCM and $X_i \in \mathbf{X}$. A* stochastic intervention on *X_i is an intervention of the type $do(X_i = \tilde{N}_i)$, where $\tilde{N}_i$ is a random variable with range R_{X_i} and $p_{\tilde{N}_i}$ can be any probability distribution.*

We can also define what we mean by "X having a total causal effect on Y". Following Peters et al. (2017); Pearl (2009), there is such a total causal effect if there is an atomic intervention on X which modifies the initial distribution of Y.

Definition 28 (Total Causal Effect) *Let X, Y be random variables of an SCM $\mathcal{C}$. X has a* total causal effect on *Y if there is $x \in R_X$ such that $p_Y^{do(X=x)} \neq p_Y$. We then write $X \rightarrow Y$.*

Appendix B. Relating Stochastic and Atomic Post-intervention Distributions

We will here prove a lemma that is used to prove many of the results in this paper.

Lemma 29 (Atomic Intervention Equals Conditioning After Stochastic Intervention) *Let X be an endogenous variable of an SCM $\mathcal{C}$ and $\mathbf{Y}$ be a vector of endogenous variables of $\mathcal{C}$ distinct from X. Furthermore, let X' be an intervention protocol for X and x be an X-value in the support of $p_{X'}$. The post-intervention distribution of $\mathbf{Y}$ resulting from an atomic intervention $do(X = x)$ equals the conditional post-intervention distribution of $\mathbf{Y}$ resulting from the stochastic intervention $do(X = X')$ and conditioning on $X = x$. That is:*

$$p_{\mathbf{Y}}^{do(X=x)} = p_{\mathbf{Y}|X=x}^{do(X=X')}. \tag{22}$$

Proof This proof will be easier if we introduce operators corresponding to intervening and conditioning. That will allow us to not overburden our notation with subscripts and superscripts. Denote by $\mathbf{X}$ the set of endogenous variables of $\mathcal{C}$, and by $\mathcal{M}_{\mathcal{P}(\mathbf{X})}$ the set of all probability mass functions for any of the variable subsets in the powerset $\mathcal{P}(\mathbf{X})$ of $\mathbf{X}$. Let x be an X-value. Define operators $\mathrm{Do}[X = x]\colon \mathcal{M}_{\mathcal{P}(\mathbf{X})} \to \mathcal{M}_{\mathcal{P}(\mathbf{X})}$ and $\mathrm{Cond}[X = x]\colon \mathcal{M}_{\mathcal{P}(\mathbf{X})} \to \mathcal{M}_{\mathcal{P}(\mathbf{X})}$ such that

$\forall \mathbf{Y} \subseteq \mathbf{X} \backslash X,\ \mathrm{Do}[X=x](p_{\mathbf{Y}}) = p_{\mathbf{Y}}^{do(X=x)}$ and $\forall \mathbf{Y} \subseteq \mathbf{X} \backslash X,\ \mathrm{Cond}[X=x](p_{\mathbf{Y}}) = p_{\mathbf{Y}|X=x}$. Let $\mathbf{Y} \subseteq \mathbf{X} \backslash X$. Then what we want to prove can be written

$$\mathrm{Cond}[X=x]\left(\mathrm{Do}[X=X']\left(p_{\mathbf{Y}}\right)\right) = \mathrm{Do}[X=x]\left(p_{\mathbf{Y}}\right). \tag{23}$$

Denote by f_X the structural assignment for X in $\mathcal{C}$. The only difference between $\mathcal{C}^{do(X=X')}$ and $\mathcal{C}$ is that $f_X(\mathrm{PA}_X, N_X)$ is replaced by another structural assignment $\tilde{f}_X(\tilde{N}_X = X') = X'$ and there is a new variable X' in $\mathcal{C}^{do(X=X')}$ which did not figure in $\mathcal{C}$. If one proceeds by performing the intervention $do(X=x)$, one obtains the SCM $(\mathcal{C}^{do(X=X')})^{do(X=x)}$ which differs from $\mathcal{C}^{do(X=X')}$ only in that $\tilde{f}_X(X')$ is replaced by $\tilde{\tilde{f}}_X(\tilde{\tilde{N}}_X \sim \delta_x) = \tilde{\tilde{N}}_X$. Notice that the marginal distribution $p_{\mathbf{Y}}^{(\mathcal{C}^{do(X=X')})^{do(X=x)}}$ entailed by this SCM is precisely the result of $\mathrm{Do}[X=x]\left(\mathrm{Do}[X=X'](p_{\mathbf{Y}})\right)$. Furthermore, the SCM entailing the RHS of (23) differs from $(\mathcal{C}^{do(X=X')})^{do(X=x)}$ only in that the latter contains a childless exogenous variable X'. Therefore they entail the same marginal distribution of $\mathbf{Y}$. Hence:

$$\mathrm{Do}[X=x]\left(\mathrm{Do}[X=X'](p_{\mathbf{Y}})\right) = \mathrm{Do}[X=x]\left(p_{\mathbf{Y}}\right).$$

On the other hand, starting again with $\mathcal{C}$ and setting $X = X'$ sets $\mathrm{PA}_X = \emptyset$, which means that intervening on X after that will be equivalent to conditioning:

$$\mathrm{Do}[X=x]\left(\mathrm{Do}[X=X'](p_{\mathbf{Y}})\right) = \mathrm{Cond}[X=x]\left(\mathrm{Do}[X=X'](p_{\mathbf{Y}})\right).$$

Hence (23) holds, which proves the lemma. ■

Appendix C. Proofs for Results on Causal Entropy

Computations for Example 1

$$\begin{cases} p_Y^{do(X=0)}(0) = p_{N_Y}(0) = {}^1\!/_2 \\ p_Y^{do(X=0)}(1) = p_{N_Y}(1) = {}^1\!/_2 \\ p_Y^{do(X=0)}(2) = 0 \end{cases} \tag{24}$$

$$\begin{cases} p_Y^{do(X=1)}(0) = 0 \\ p_Y^{do(X=1)}(1) = p_{N_Y}(0) = {}^1\!/_2 \\ p_Y^{do(X=1)}(2) = p_{N_Y}(1) = {}^1\!/_2 \end{cases} \tag{25}$$

$$\begin{cases} p_Y^{do(X=X')}(0) = p_{X'}(0)p_{N_Y}(0) = \frac{2}{3} \times \frac{1}{2} = {}^1\!/_3 \\ p_Y^{do(X=X')}(1) = p_{X'}(0)p_{N_Y}(1) + p_{X'}(1)p_{N_Y}(0) = {}^1\!/_2 \\ p_Y^{do(X=X')}(2) = p_{X'}(1)p_{N_Y}(1) = \frac{1}{3} \times \frac{1}{2} = {}^1\!/_6 \end{cases} \tag{26}$$

$$\begin{aligned}H_c(Y \mid do(X \sim X')) = &- p_{X'}(0)\Big(p_Y^{do(X=0)}(0) \log p_Y^{do(X=0)}(0)\\ &\qquad + p_Y^{do(X=0)}(1) \log p_Y^{do(X=0)}(1)\\ &\qquad + p_Y^{do(X=0)}(2) \log p_Y^{do(X=0)}(2)\Big)\\ &- p_{X'}(1)\Big(p_Y^{do(X=1)}(0) \log p_Y^{do(X=1)}(0)\\ &\qquad + p_Y^{do(X=1)}(1) \log p_Y^{do(X=1)}(1)\\ &\qquad + p_Y^{do(X=1)}(2) \log p_Y^{do(X=1)}(2)\Big)\\ = &- 2/3 \times (-1/2 - 1/2) - 1/3 \times (0 - 1/2 - 1/2) = 1(\text{bit})\end{aligned} \tag{27}$$

$$\begin{aligned}H(Y \mid do(X = X')) = &p_Y^{do(X=X')}(0) \log(1/p_Y^{do(X=X')}(0))\\ &+ p_Y^{do(X=X')}(1) \log(1/p_Y^{do(X=X')}(1))\\ &+ p_Y^{do(X=X')}(2) \log(1/p_Y^{do(X=X')}(2))\\ = &\frac{1}{3}\log 3 + \frac{1}{2} + \frac{1}{6}\log 6.\end{aligned} \tag{28}$$

Proof [Proof of Proposition 7]

$$H_c(Y \mid do(X \sim X')) \equiv - \mathbb{E}_{x' \sim p_{X'}}\left[\mathbb{E}_{y \sim p_Y^{do(X=x')}}\left[\log p_Y^{do(X=x')}(y)\right]\right] \tag{29}$$

$$\equiv - \sum_{x',y} p_{X'}(x') p_Y^{do(X=x')}(y) \log p_Y^{do(X=x')}(y) \tag{30}$$

$$= - \sum_{x',y} p_X^{do(X=X')}(x') p_{Y|X=x'}^{do(X=X')}(y) \log p_Y^{do(X=x')}(y) \tag{31}$$

$$= - \sum_{x,y} p_{X,Y}^{do(X=X')}(x,y) \log p_Y^{do(X=x')}(y) \tag{32}$$

$$= - \mathbb{E}_{x,y \sim p_{X,Y}^{do(X=X')}}\left[\log p(y \mid do(X = x))\right] \tag{33}$$

where to get Equation (31) we used that $p_{X'}(x) = p_X^{do(X=X')}(x)$ and Lemma 29. ■

Proof [Proof of Proposition 8] This comes directly from the definition of H_c and Lemma 29:

$$\begin{aligned}H_c(Y \mid do(X \sim X')) &= \mathbb{E}_{x \sim p_{X'}}\left[H_{Y \sim p_Y^{do(X=x)}}(Y)\right]\\ &= \mathbb{E}_{x \sim p_{X'}}\left[H_{Y \sim p_{Y|X=x}^{do(X=X')}}(Y)\right].\end{aligned} \tag{34}$$

■

Proof [Proof of Proposition 9] The result follows directly from the definition of causal entropy. We can write:

$$H_c(Y \mid do(X \sim X')) = \mathbb{E}_{x' \sim p_{X'}} \left[H_{Y \sim p_Y^{do(X=x')}}(Y) \right]. \tag{35}$$

Since entropies are non-negative quantities, it follows that the causal entropy, being an average of entropies, is also non-negative. ■

Proof [Proof of Proposition 10] It suffices to provide an SCM $\mathcal{C}$ and intervention protocol X' such that the causal entropy of Y given the intervention protocol X' for X is greater than the initial entropy of Y. Consider an SCM $\mathcal{C}$ with exactly two binary endogenous variables X, Y, causal graph $X \rightarrow Y$ and structural assignments given by:

$$\begin{cases} X = N_X \\ Y = \begin{cases} y_0, X = x_0 \\ N_Y, X = x_1 \end{cases} \end{cases}. \tag{36}$$

Furthermore, assume that $N_Y \sim \text{Bern}(1/2)$, $N_X \sim \text{Bern}(1/10)$ and $X' \sim \text{Bern}(1/2)$. Notice that this implies:

$$\begin{cases} p_Y^{do(X=x_0)} = p_{Y|X=x_0} = \delta(x_0) \\ p_Y^{do(X=x_1)} = p_{Y|X=x_1} = \text{Bern}(1/2) \end{cases}. \tag{37}$$

Then:

$$\begin{aligned} H_c(Y \mid do(X \sim X')) &= \mathbb{E}_{x \sim p_{X'}} \left[H_{Y \sim p_Y^{do(X=x)}}(Y) \right] \\ &= \mathbb{E}_{x \sim p_{X'}} \left[H_{Y \sim p_{Y|X=x}}(Y) \right] \\ &= p_{X'}(x_0) \underbrace{H_{Y \sim p_{Y|X=x_0}}(Y)}_{0} + p_{X'}(x_1) \underbrace{H_{Y \sim p_{Y|X=x_1}}(Y)}_{1} \\ &= p_{X'}(x_1) = 1/2 \end{aligned} \tag{38}$$

where in the second equality we used that the causal effect from X to Y is not confounded.

$$\begin{aligned} H(Y) &= -\sum_y p_Y(y) \log\left(p_Y(y)\right) \\ &= -\sum_{x,y} p_{Y|X=x}(y) p_X(x) \log\left(\sum_{\dot{x}} p_{Y|X=\dot{x}}(y) p_X(\dot{x}) \right) \\ &= -\sum_x p_X(x) \sum_y p_{Y|X=x}(y) \log\Big(p(y \mid x_0) p_X(x_0) + p(y \mid x_1) p_X(x_1) \Big) \\ &= -p_X(x_0) \log\Big(p_X(x_0) + \frac{1}{2} p_X(x_1) \Big) \\ &\quad - \frac{1}{2} p_X(x_1) \Big[\log\Big(p_X(x_0) + \frac{1}{2} p_X(x_1) \Big) + \log\Big(\frac{1}{2} p_X(x_1) \Big) \Big] \approx 2.04. \end{aligned} \tag{39}$$

Hence $H_c(Y \mid X \sim X') > H(Y)$ for this SCM and intervention protocol. ■

Proof [Proof of Corollary 11] Conditional entropy cannot be greater than the entropy before conditioning. By Proposition 8 causal entropy is a conditional entropy obtained from the post-stochastic intervention entropy by conditioning. The result follows. ■

Proof [Proof of Proposition 12] We make use of Proposition 8 and the independence bound on (standard) entropy:

$$\begin{aligned} H_c(\mathbf{Y} \mid do(X \sim X')) &= H(\mathbf{Y} \mid X, do(X = X')) \\ &\leq \sum_{i=1}^{n_Y} H(Y_i \mid X, do(X = X')) \\ &= \sum_{i=1}^{n_Y} H_c(Y_i \mid do(X \sim X')). \end{aligned} \tag{40}$$

■

Proof [Proof of Proposition 14]

$$\begin{aligned} H_c(Y \mid Z, do(X \sim X')) &= \mathbb{E}_{x \sim p_{X'}} \left[\mathbb{E}_{z \sim p_Z^{do(X=x)}} \left[H_{Y \sim p_{Y|Z=z}^{do(X=x)}}(Y) \right] \right] \\ &= \sum_{x,z} p_Z^{do(X=x)}(z) p_{X'}(x) H_{Y \sim p_{Y|Z=z}^{do(X=x)}}(Y) \\ &= \sum_{x,z} p_{Z|X=x}^{do(X=X')}(z) p_X^{do(X=X')}(x) H_{Y \sim p_{Y|X=x,Z=z}^{do(X=X')}}(Y) \\ &= \sum_{x,z} p_{X,Z}^{do(X=X')}(x,z) H(Y \mid X = x, Z = z, do(X = X')) \\ &= H(Y \mid X, Z, do(X = X')) \end{aligned} \tag{41}$$

where in the third step we used Lemma 29 twice, and in the last one we used the definition of conditional entropy. ■

Proof [Proof of Proposition 15] We will use Proposition 14 and Proposition 8 together with the fact that conditional entropy can never be larger than the initial entropy (Cover and Thomas, 2006) to prove the result:

$$\begin{aligned} H_c(Y \mid Z, do(X \sim X')) &= H(Y \mid Z, X, do(X = X')) \\ &\leq H(Y \mid X, do(X = X')) \\ &= H_c(Y \mid do(X \sim X')). \end{aligned} \tag{42}$$

■

Proof [Proof of Proposition 16] Due to Proposition 8 and Proposition 14 we can leverage the chain rule for entropy to obtain a chain rule for the causal entropy:

$$\begin{aligned} H_c(Y, Z \mid do(X \sim X')) &= H(Y, Z \mid X, do(X = X')) \\ &= H(Y \mid X, do(X = X')) + H(Z \mid Y, X, do(X = X')) \\ &= H_c(Y \mid do(X \sim X')) + H_c(Z \mid Y, do(X \sim X')). \end{aligned} \tag{43}$$

■

Proof [Proof of Proposition 17]

$$\begin{aligned} H_c(\mathbf{Y} \mid do(X \sim X')) &= H(\mathbf{Y} \mid X, do(X = X')) \\ &= \sum_{i=1}^{n_Y} H(Y_i \mid \mathbf{Y}_{<i}, X, do(X = X')) \\ &= \sum_{i=1}^{n_Y} H_c(Y_i \mid \mathbf{Y}_{<i}, do(X \sim X')) \end{aligned} \tag{44}$$

where in the first step and third steps we used Proposition 14, while the second equality follows from the chain rule for entropy. ■

Appendix D. Proofs for Results on Causal Information Gain

Proof [Proof of Proposition 20]

$$\begin{aligned} I_c(\mathbf{Y} \mid do(X \sim X')) &= H(\mathbf{Y}) - H_c(\mathbf{Y} \mid do(X \sim X')) \\ &= \sum_{i=1}^{n_Y} \Big(H(Y_i \mid \mathbf{Y}_{<i}) - H_c(Y_i \mid \mathbf{Y}_{<i}, do(X \sim X')) \Big) \\ &= \sum_{i=1}^{n_Y} I_c(Y_i \mid \mathbf{Y}_{<i}, do(X \sim X')) \end{aligned} \tag{45}$$

where in the second step we used the chain rules for the entropy (Cover and Thomas, 2006) and for the causal entropy (Proposition 17). ■

Proof [Proof of Proposition 22]

$$\begin{aligned} \mathrm{MI}_c(Y \mid Z, do(X \sim X')) &= \mathbb{E}_{x \sim p_{X'}} \left[H(Y \mid do(X = x)) \right] \\ &\quad - \mathbb{E}_{x \sim p_{X'}} \left[H(Y \mid Z, do(X = x)) \right] \\ &= \mathbb{E}_{x \sim p_{X'}} \left[H(Y \mid do(X = x)) - H(Y \mid Z, do(X = x)) \right] \\ &= \mathbb{E}_{x \sim p_{X'}} \left[I(Y ; Z \mid do(X = x)) \right] \end{aligned} \tag{46}$$

■