Vietnam Journal of Mathematics (2021) 49:547–576
https://doi.org/10.1007/s10013-021-00511-w

ORIGINAL ARTICLE

Confounding Ghost Channels and Causality: A New Approach to Causal Information Flows

Nihat Ay[1,2,3,4]

Received: 29 June 2020 / Accepted: 22 April 2021 / Published online: 8 July 2021
© The Author(s) 2021

Abstract
Information theory provides a fundamental framework for the quantification of information flows through channels, formally Markov kernels. However, quantities such as mutual information and conditional mutual information do not necessarily reflect the causal nature of such flows. We argue that this is often the result of conditioning based on σ-algebras that are not associated with the given channels. We propose a version of the (conditional) mutual information based on families of σ-algebras that are coupled with the underlying channel. This leads to filtrations which allow us to prove a corresponding causal chain rule as a basic requirement within the presented approach.

Keywords Information flow · Causality · Mutual information · Conditional mutual information · Filtration

Mathematics Subject Classification (2010) 62A01 · 62B10 · 94A15 · 94A17 · 94A40

1 Introduction: Information Theory and Causality

Classical information theory [20] is based on the definition of *Shannon entropy*, a measure of uncertainty about the outcome of a variable Z:

$$H(Z) = -\sum_z p(z) \log p(z),$$

where $p(z) = \mathbb{P}(Z = z)$ denotes the distribution of Z. (Throughout this introduction, we consider only variables X, Y, and Z with finite state sets X, Y, and Z, respectively.)

Dedicated to Jürgen Jost on the occasion of his 65th birthday.

✉ Nihat Ay
nihat.ay@tuhh.de

1 Hamburg University of Technology, Hamburg, Germany
2 Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
3 Leipzig University, Leipzig, Germany
4 Santa Fe Institute, Santa Fe, NM, USA

Shannon entropy serves as a building block of further important quantities. The flow of information from a sender X to a receiver Z, for instance, can be quantified as the reduction of uncertainty about the outcome of Z based on the outcome of X. More precisely, we compare two uncertainties here, the uncertainty about the outcome of Z, that is $H(Z)$, with the uncertainty about the outcome of Z *after* knowing the outcome of X, that is

$$H(Z|X) = -\sum_x p(x) \sum_z p(z|x) \log p(z|x), \tag{1}$$

where $p(z|x) = \mathbb{P}(Z = z|X = x)$ denotes the conditional distribution of Z given X. (Note that for x with $p(x) = 0$, conditioning is not well-defined. This ambiguity, however, has no effect on (1), due to the multiplication with $p(x)$.) Naturally, the latter uncertainty, $H(Z|X)$, is smaller than or equal to the first one, $H(Z)$, leading to another fundamental quantity of information theory, the *mutual information*:

$$I(X; Z) = H(Z) - H(Z|X). \tag{2}$$

This difference can also be expressed in geometric terms as the *KL-divergence*

$$I(X; Z) = \sum_x p(x) \sum_z p(z|x) \log \frac{p(z|x)}{p(z)}. \tag{3}$$

The KL-divergence plays an important role in information geometry as a canonical divergence [1–4]. Such a divergence is characterised in terms of natural geometric properties. It is remarkable that this purely geometric approach yields the fundamental information-theoretic quantities which were previously derived from a set of axioms that are formulated in non-geometric terms.

Typically, the conditional distribution $p(z|x)$ is interpreted mechanistically as a channel which receives x as an input and generates z as an output. In this interpretation, the stochasticity of a channel is considered to be the effect of external or hidden disturbances of a deterministic map. This is formalised in terms of a so-called *structural equation*

$$Z = f(X, U), \tag{4}$$

with a deterministic map f and a noise variable U that is independent of X [17]. Integrating out the noise variable U, we obtain a Markov kernel, the formal model of a channel. What do we gain by this construction? Formally, we do not gain much, as a Markov kernel is basically a conditional probability distribution, defined for all x. However, with the representation (4) and the associated Markov kernel, the conditional probability distribution $p(z|x)$ can be interpreted as the result of a (probabilitic) causal effect of X on Z. This interpretation provides the basis for Pearl's influential proposal of a general theory of causality [17]. The mutual information (2) then becomes a measure of the causal information flow from X to Z [6], which is consistent with Shannon's original idea of the amount of information transmitted through a channel [20]. This consistency, however, is apparently violated when dealing with variations or extensions of the sender-receiver setting. In Section 2, we are now going to highlight instances of such inconsistency that will play an important role in this article.

Later in this article, a channel will be formally given in terms of a Markov kernel, with a more explicit notation. In what follows, however, we keep the notation of a conditional probability distribution and state explicitly when we interpret it mechanistically as a channel.

2 Confounding Ghost Channels

The mutual information is symmetric, that is $I(X;Z) = I(Z;X)$. Interpreting it as a measure of causal information flow, this symmetry suggests that we have the same amount of causal information flow in both directions, even though the channel goes from X to Z so that there cannot be any flow of information in the opposite direction. What is wrong here? This apparent problem, let us call it "the symmetry puzzle", can be resolved quite easily. We *can* revert the direction and compute the conditional distribution $p(x|z) = \frac{p(x)}{p(z)}\, p(z|x)$, based on elementary rules of probability theory and without reference to any mechanisms. Furthermore, this conditional distribution *can* be mechanistically interpreted and represented in terms of a structural equation (4). (This is always possible for a given conditional distribution.) Such a representation introduces a hypothetical channel for generating the reverted conditional distribution $p(x|z)$, a kind of "ghost" channel that is actually not there. The mutual information then quantifies the causal information flow of this hypothetical channel. The symmetry of the mutual information simply means that the actual causal information flow in forward direction will be equal to the causal information flow of any hypothetical channel in backward direction that is capable of generating the conditional distribution $p(x|z)$. The symmetry puzzle, however, is not the only apparent inconsistency between the (conditional) mutual information and causality. We are now going to highlight another problem, which is closely related to the symmetry puzzle but requires a deeper analysis for its solution.

We now assume that the channel receives x and y as inputs and generates z as an output. With the corresponding conditional distribution $p(z|x,y) = \mathbb{P}(Z = z|X = x, Y = y)$ we have the *conditional mutual information* of Y and Z given X:

$$I(Y;Z|X) = H(Z|X) - H(Z|X,Y) \tag{5}$$

$$= \sum_{x,y} p(x,y) \sum_{z} p(z|x,y) \log \frac{p(z|x,y)}{p(z|x)}. \tag{6}$$

According to (5), the conditional mutual information compares the uncertainty about z given x, before and after observing the outcome y, reflected by the conditional probabilities $p(z|x)$ and $p(z|x,y)$, respectively. The representation (6) makes this comparison more explicit as a deviation of $p(z|x,y)$ from $p(z|x)$. Together with (3), we obtain the chain rule

$$I(X,Y;Z) = \sum_{x,y} p(x,y) \sum_{z} p(z|x,y) \log \frac{p(z|x,y)}{p(z)} \tag{7}$$

$$= \sum_{x,y} p(x,y) \sum_{z} p(z|x,y) \left[\log \frac{p(z|x)}{p(z)} + \log \frac{p(z|x,y)}{p(z|x)} \right] \tag{8}$$

$$= I(X;Z) + I(Y;Z|X). \tag{9}$$

For the computation of both terms, $I(X;Z)$ and $I(Y;Z|X)$, we have to evaluate the "reduced" conditional distribution $p(z|x)$. It is obtained from the original one in the following way:

$$p(z|x) = \sum_{y} p(y|x)\, p(z|x,y). \tag{10}$$

This conditional distribution represents a second kind of hypothetical channel, a "ghost channel", which screens off the actual flow of information. It can be sensitive to information about x that is not necessarily employed by the original channel $p(z|x,y)$. More precisely, given two states x, x' that satisfy $p(z|x,y) = p(z|x',y)$ for all z and all y, we cannot expect

$p(z|x) = p(z|x')$ for all z. This is a consequence of the coupling through $p(y|x)$, on the right-hand side (RHS, for short) of (10). In the most extreme case, y is simply a deterministic map of x, so that the knowledge of y does not provide any additional information about z, that is $p(z|x, y) = p(z|x)$. In the following example we study this case more explicitly and thereby highlight the inconsistency of the terms $I(X; Z)$ and $I(Y; Z|X)$ in (9) with the underlying causal structure. We will argue that the conditional distribution (10) has to be modified in order to allow for a causal interpretation.

Example 1 Consider three variables X, Y, Z with values -1 and $+1$, and assume that Z is obtained as a copy of Y, that is

$$p(z|x, y) = \begin{cases} 1, & \text{if } z = y \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

We interpret this conditional distribution, which is well-defined for *all* arguments x and y, as a mechanism. This means that all information required for the output Z is contained in Y. Intuitively, we would expect from a measure of information flow to assign zero for the flow from X to Z and a positive value to the flow of information from Y to Z given X. This is however not what we get with the usual definitions of mutual information and conditional mutual information. The reason for that is the stochastic dependence of the inputs X and Y. To be more precise, let us assume that the input distribution is given as

$$p(x, y) = \frac{e^{\beta xy}}{\sum_{x', y' \in \{\pm 1\}} e^{\beta x' y'}},$$

where the parameter β controls the coupling of the inputs. This implies $p(x) = \mathbb{P}(X = x) = 1/2$ and $p(y) = \mathbb{P}(Y = y) = 1/2$ for all $x, y \in \{\pm 1\}$. We can decompose the full mutual information, as a measure of information flow from X and Y together to Z, in the following way:

$$I_\beta(X, Y; Z) = \underbrace{I_\beta(Y; Z)}_{=\log 2} + \underbrace{I_\beta(X; Z|Y)}_{=0}. \tag{12}$$

(The subscript β indicates the dependence of the respective information-theoretic quantities on this parameter.) This is consistent with the intuition that Z is receiving all information from Y and no information from X. However, we observe an inconsistency if we decompose the full mutual information in a different way:

$$I_\beta(X, Y; Z) = I_\beta(X; Z) + I_\beta(Y; Z|X). \tag{13}$$

For the two terms on the RHS of (13) we obtain

$$\begin{aligned} I_\beta(X; Z) &= \log(2) - \frac{\log(1 + e^{2\beta})}{1 + e^{2\beta}} - \frac{\log(1 + e^{-2\beta})}{1 + e^{-2\beta}}, \\ I_\beta(Y; Z|X) &= \frac{\log(1 + e^{2\beta})}{1 + e^{2\beta}} + \frac{\log(1 + e^{-2\beta})}{1 + e^{-2\beta}}. \end{aligned}$$

These functions are shown in Fig. 1. In the limit $\beta \to -\infty$ the two inputs become completely anti-correlated with support $(-1, +1)$ and $(+1, -1)$. Correspondingly, for

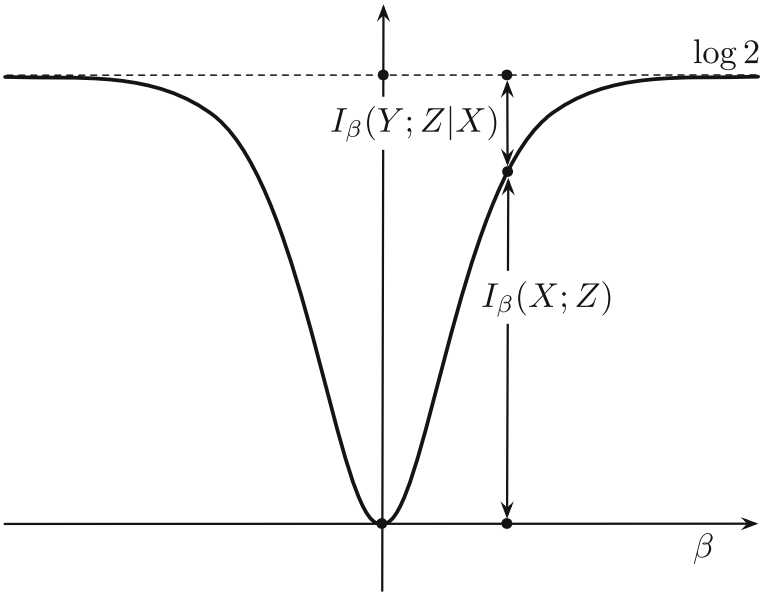

Fig. 1 The mutual information $I_\beta(X;Z)$ and the conditional mutual information $I_\beta(Y;Z|X)$ as functions of β. Even though the channel does not employ any information from X, the mutual information $I_\beta(X;Z)$ converges to the maximal value for $\beta \to \infty$

$\beta \to +\infty$ we have complete correlation, and the support is $(-1,-1)$ and $(+1,+1)$. With (13), we obtain the following decomposition:

$$\begin{aligned} I(X,Y;Z) &= \lim_{\beta\to\infty} I_\beta(X,Y;Z) \\ &= \lim_{\beta\to\infty} I_\beta(X;Z) + \lim_{\beta\to\infty} I_\beta(Y;Z|X) \\ &= \underbrace{I(X;Z)}_{=\log 2} + \underbrace{I(Y;Z|X)}_{=0}. \end{aligned} \tag{14}$$

The decomposition (14) gives the impression that Z is receiving all information from X and no information from Y. However, we know, by the definition of the mechanism (11), that this is not the case. The actual situation is better reflected by the decomposition (12).

The problem highlighted in Example 1 can be resolved by an appropriate modification of the conditional probability (10). We are now going to outline this modification, which will provide the main idea of this article. In a first step, let us assume that $\bar{y}$ is fixed as an input to the channel. Which information about x does the channel then use for generating z? In order to qualitatively describe that information, we lump any two states x and x' together whenever the channel cannot distinguish them, that is

$$p(z|x,\bar{y}) = p(z|x',\bar{y})$$

for all z. This defines a partition $\alpha_{X,\bar{y}}$ of the state set of X that depends on $\bar{y}$. In a second step, we consider the join of all these partitions, that is their coarsest refinement. More

precisely, we define

$$\alpha_X := \bigvee_{\bar{y}\in\mathsf{Y}} \alpha_{X,\bar{y}} := \left\{ \bigcap_{\bar{y}\in\mathsf{Y}} A_{\bar{y}} \;:\; A_{\bar{y}} \in \alpha_{X,\bar{y}},\ \bar{y} \in \mathsf{Y} \right\}. \tag{15}$$

The partition α_X represents a qualitative description of the information in X that is used by the channel $p(z|x, y)$. Denote by A_x the set in α_X that contains x. When the channel receives x, in addition to y, then it does not "see" the full x but only the class A_x, and it is easy to verify $p(z|x, y) = p(z|A_x, y)$. Therefore, we replace the conditioning $p(z|x)$ in the above formula (10) by

$$\hat{p}(z|x) := p(z|A_x) = \sum_y p(y|A_x)\, p(z|A_x, y) = \sum_{x'\in A_x} p(x'|A_x)\, p(z|x'). \tag{16}$$

This shows that the new conditional distribution $\hat{p}(z|x)$ is obtained by averaging the previous one, $p(z|x)$, according to the information that is actually used by the channel $p(z|x, y)$. Now, replacing in (8) the conditional distribution $p(z|x)$ by $\hat{p}(z|x)$ leads to a corresponding modification of the mutual information and the conditional mutual information:

$$I(X \to Z) := \sum_x p(x) \sum_z \hat{p}(z|x) \log \frac{\hat{p}(z|x)}{p(z)},$$

$$I(Y \to Z|X) := \sum_{x,y} p(x, y) \sum_z p(z|x, y) \log \frac{p(z|x, y)}{\hat{p}(z|x)}.$$

It is easy to see that

$$I(X \to Z) \le I(X; Z), \qquad I(Y \to Z|X) \ge I(Y; Z|X).$$

However, the sum does not change and we have the chain rule

$$I(X, Y; Z) = I(X \to Z) + I(Y \to Z|X).$$

With this new definition, we come back to Example 1. The channel defined by (11) does not use any information from X. Therefore, $\alpha_{X,\bar{y}} = \{\mathsf{X}\}$ for all $\bar{y} \in \mathsf{Y}$, which implies $\alpha_X = \{\mathsf{X}\}$. With formula (16) we obtain $\hat{p}(z|x) = p(z|\mathsf{X}) = p(z)$, and therefore

$$I(X \to Z) = 0 \qquad \text{and} \qquad I(Y \to Z|X) = \log 2.$$

If we compare this with (14) we see that the information is shifted from the first to the second term which corresponds to the variable that has the actual causal effect on Z. On the other hand, in both cases the sum of the two contributions equals $\log 2$, the full mutual information $I(X, Y; Z)$.

Causality plays an important role in time series analysis. In this context, *Granger causality* [12, 13] has been the subject of extensive debates which tend to highlight its non-causal nature. Schreiber proposed an information-theoretic quantification of Granger causality, referred to as *transfer entropy*, which is based on conditional mutual information [9, 19]. Even though transfer entropy is an extremely useful and widely applied quantity, it is generally accepted that it has shortcomings as a measure of causal information flow. In particular, it can vanish in cases where the causal effect is the strongest possible. We argue that this is again a result of a ghost channel that is involved in the computation of the classical conditional mutual information and screens off the actual causal information flow. This is demonstrated in the following example which is taken from [6]. Essentially, this example is a reformulation of Example 1, thereby adjusted to the context of time series and stochastic processes.

Example 2 (Transfer entropy) Consider a stochastic process (X_m, Y_m), $m = 1, 2, \ldots$, with state space $\{\pm 1\}^2$ and define $X^m := (X_1, \ldots, X_m)$ and $Y^m := (Y_1, \ldots, Y_m)$. The transfer entropy at time m is defined as

$$T(Y^{m-1} \to X_m) := I(Y^{m-1}; X_m | X^{m-1}).$$

Thus, the transfer entropy quantifies how much information the variables $Y_1, \ldots, Y_{m-1}$ contribute to the evaluation of X_m, in addition to the information in $X_1, \ldots, X_{m-1}$. We assume that the process is a Markov chain, given by a transition matrix of the form

$$p(x', y'|x, y) = p(x'|x, y)\, p(y'|x, y),$$

where

$$p(x'|x, y) = \frac{1}{1 + e^{2\beta x' y}}, \qquad p(y'|x, y) = \frac{1}{1 + e^{2\beta y' y}}.$$

The causal structure of the dynamics is represented by the following diagram:

$$\begin{tikzcd}
Y_1 \arrow[r] \arrow[dr] & Y_2 \arrow[r] \arrow[dr] & Y_3 \arrow[r] \arrow[dr] & \cdots \arrow[r] \arrow[dr] & Y_{m-1} \arrow[r] \arrow[dr] & Y_m \\
X_1 & X_2 & X_3 & \cdots & X_{m-1} & X_m
\end{tikzcd}$$

As a stationary distribution we have

$$p(+1, +1) = p(-1, -1) = \frac{1}{2} - ab, \qquad p(+1, -1) = p(-1, +1) = ab,$$

where

$$a = \frac{1}{1 + e^{2\beta}}, \qquad b = \frac{1}{1 + e^{-2\beta}}.$$

The transfer entropy can be upper bounded as follows (the subscript β indicates the dependence on the coupling parameter β):

$$\begin{aligned}
T_\beta(Y^{m-1} \to X_m) &= I_\beta(Y^{m-1}; X_m | X^{m-1}) \\
&= I_\beta(Y_{m-1}; X_m | X^{m-1}) \\
&= H_\beta(X_m | X^{m-1}) - H_\beta(X_m | X^{m-1}, Y_{m-1}) \\
&= H_\beta(X_m | X^{m-1}) - H_\beta(X_m | X_{m-1}, Y_{m-1}) \\
&\leq H_\beta(X_m | X_{m-1}) - H_\beta(X_m | X_{m-1}, Y_{m-1}) \\
&= I_\beta(Y_{m-1}; X_m | X_{m-1}).
\end{aligned}$$

For $\beta = 0$, we have an i.i.d. process with uniform distribution over the states $(+1, +1)$, $(-1, +1)$, $(+1, -1)$, and $(-1, -1)$. For $\beta \to \infty$, we obtain the deterministic transition

$$(x, y) \mapsto (-y, -y).$$

In this limit, the variables (X_m, Y_m) are completely correlated with $p(+1, +1) = p(-1, -1) = \frac{1}{2}$. In both cases, $\beta = 0$ and $\beta \to \infty$, the conditional mutual information $I_\beta(Y_{m-1}; X_m | X_{m-1})$, and therefore the transfer entropy $T_\beta(Y^{m-1} \to X_m)$, vanishes. For $\beta = 0$, this does not represent a problem because any measure of causal information flow should vanish in the i.i.d. case. However, for $\beta \to \infty$, the variable X_m is causally determined by Y_{m-1}. Therefore, a measure of casual information flow should be maximal in this case. This is not reflected by the transfer entropy. Let us compare this with the information flow measure proposed in this article. Given that X_m only depends on Y_{m-1}, the partition (15) is trivial, that is $\alpha = \{\mathsf{X}\}$. Therefore,

$$I(Y^{m-1} \to X_m | X^{m-1}) = I(Y_{m-1}; X_m).$$

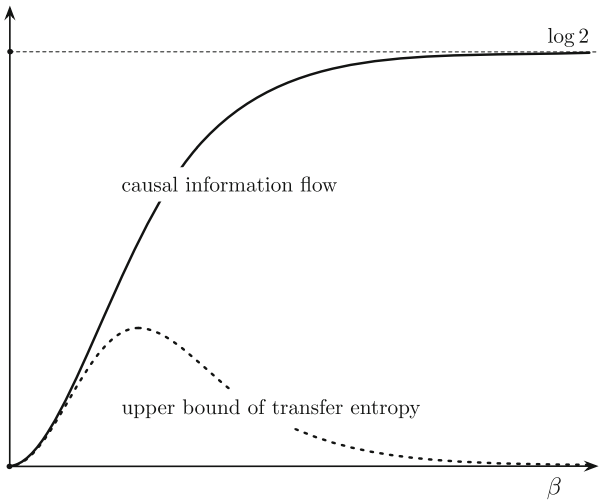

Fig. 2 Dashed line: the conditional mutual information $I_\beta(Y_{m-1}; X_m|X_{m-1})$ as an upper bound of the transfer entropy $T_\beta(Y^{m-1} \to X_m)$; solid line: the causal information flow $I_\beta(Y^{m-1} \to X_m|X^{m-1})$ which coincides with the mutual information $I_\beta(Y_{m-1}; X_m)$ in this example

This quantity is converging to the maximal value $\log 2$ for $\beta \to \infty$. For comparison, both functions are plotted in Fig. 2.

In what follows, we will extend the idea outlined in this section to a more general context of measurable spaces, probability measures, and Markov kernels. In further steps, we will also consider more input nodes.

3 General Information-Theoretic Quantities

In the previous sections, we reviewed fundamental information-theoretic quantities as they are introduced in standard textbooks such as [10]. In this section, we offer an alternative review from a measure-theoretic perspective (see, for instance, [15]). This more abstract setting will allow us to identify natural operations and definitions which are not always visible when dealing with finite state spaces.

Shannon Entropy For a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a finite measurable partition $\gamma = \{C_1, \ldots, C_m\}$, that is $C_i \in \mathscr{F}$, $C_i \cap C_j = \emptyset$ for all $i \neq j$, and $\bigcup_{i=1}^m C_i = \Omega$, the *Shannon entropy* of γ is given by

$$H(\gamma) := -\sum_{C \in \gamma} \mathbb{P}(C) \log \mathbb{P}(C). \tag{17}$$

As a local version of the Shannon entropy, we define

$$h(\gamma) := -\sum_{C \in \gamma} \mathbb{1}_C \log \mathbb{P}(C),$$

where $\mathbb{1}_C$ is the indicator function of C. Denoting by C_ω the set in γ that contains $\omega \in \Omega$, we have $h(\gamma)(\omega) = -\log \mathbb{P}(C_\omega)$. If we integrate the function $h(\gamma)$, we recover the entropy

(17) of the partition γ:

$$\begin{aligned}\int_\Omega h(\gamma)\mathrm{d}\mathbb{P} &= -\sum_{C\in\gamma}\left\{\int_\Omega \mathbb{1}_C\,\mathrm{d}\mathbb{P}\right\}\log\mathbb{P}(C)\\ &= -\sum_{C\in\gamma}\mathbb{P}(C)\log\mathbb{P}(C)\\ &= H(\gamma).\end{aligned}$$

Conditional Entropy Consider two finite measurable partitions α and γ of Ω, where we assume $\mathbb{P}(A) > 0$ for all $A \in \alpha$. The *conditional entropy* of γ given α is then defined by

$$H(\gamma|\alpha) := -\sum_{A\in\alpha}\mathbb{P}(A)\sum_{C\in\gamma}\mathbb{P}(C|A)\log\mathbb{P}(C|A). \tag{18}$$

As a local version $h(\gamma|\alpha)$ of $H(\gamma|\alpha)$, we define

$$h(\gamma|\alpha) := -\sum_{C\in\gamma}\mathbb{1}_C\log\left(\sum_{A\in\alpha}\mathbb{P}(C|A)\mathbb{1}_A\right). \tag{19}$$

If we evaluate this function for $\omega \in \Omega$ we obtain $h(\gamma|\alpha)(\omega) = -\log\mathbb{P}(C_\omega|A_\omega)$, where A_ω and C_ω are the atoms in α and γ, respectively, that contain ω. Integrating $h(\gamma|\alpha)$, we recover (18):

$$\begin{aligned}\int_\Omega h(\gamma|\alpha)\,\mathrm{d}\mathbb{P} &= -\int_\Omega \log\mathbb{P}(C_\omega|A_\omega)\mathbb{P}(\mathrm{d}\omega)\\ &= -\sum_{A\in\alpha}\sum_{C\in\gamma}\int_{A\cap C}\log\mathbb{P}(C_\omega|A_\omega)\,\mathbb{P}(\mathrm{d}\omega)\\ &= -\sum_{A\in\alpha}\sum_{C\in\gamma}\mathbb{P}(A\cap C)\log\mathbb{P}(C|A)\\ &= H(\gamma|\alpha).\end{aligned}$$

The function $h(\gamma|\alpha)$ can be generalised by replacing the partition α by an arbitrary σ-subalgebra $\mathscr{A}$ of $\mathscr{F}$:

$$h(\gamma|\mathscr{A}) := -\sum_{C\in\gamma}\mathbb{1}_C\log\mathbb{P}(C|\mathscr{A}), \tag{20}$$

where $\mathbb{P}(C|\mathscr{A}) = \mathbb{E}(\mathbb{1}_C|\mathscr{A})$. Note that this function is only $\mathbb{P}$-almost everywhere defined (abbreviated as $\mathbb{P}$-a.e.). In the case where the σ-algebra $\mathscr{A}$ is given by a finite partition α with $\mathbb{P}(A) > 0$ for all $A \in \alpha$, we have

$$\mathbb{P}(C|\mathscr{A}) = \sum_{A\in\alpha}\mathbb{P}(C|A)\,\mathbb{1}_A, \qquad \mathbb{P}\text{-a.e.}$$

This shows that the definition (20) is indeed an extension of (19). Correspondingly, integrating (20) yields a generalistaion of (18):

$$\begin{aligned} H(\gamma|\mathscr{A}) &:= \int_\Omega h(\gamma|\mathscr{A})\,\mathrm{d}\mathbb{P} \\ &= -\sum_{C\in\gamma}\int_\Omega \mathbb{1}_C \log \mathbb{P}(C|\mathscr{A})\,\mathrm{d}\mathbb{P} \\ &= -\sum_{C\in\gamma}\int_\Omega \mathbb{P}(C|\mathscr{A}) \log \mathbb{P}(C|\mathscr{A})\,\mathrm{d}\mathbb{P}. \end{aligned}$$

Mutual Information If we subtract from the entropy of a partition γ the conditional entropy of γ given a partition α, we obtain the *mutual information*:

$$I(\alpha;\gamma) := \sum_{A\in\alpha}\mathbb{P}(A)\sum_{C\in\gamma}\mathbb{P}(C|A)\log\frac{\mathbb{P}(C|A)}{\mathbb{P}(C)}. \tag{21}$$

Let us relate this function to the corresponding local functions $h(\gamma)$ and $h(\gamma|\alpha)$. Taking the difference, we obtain

$$\begin{aligned} i(\alpha;\gamma) &:= h(\gamma) - h(\gamma|\alpha) \\ &= -\sum_{C\in\gamma}\mathbb{1}_C\log\mathbb{P}(C) + \sum_{C\in\gamma}\mathbb{1}_C\log\left(\sum_{A\in\alpha}\mathbb{P}(C|A)\,\mathbb{1}_A\right) \\ &= \sum_{C\in\gamma}\mathbb{1}_C\log\left(\sum_{A\in\alpha}\frac{\mathbb{P}(C|A)}{\mathbb{P}(C)}\,\mathbb{1}_A\right). \end{aligned}$$

If we evaluate this for $\omega\in\Omega$ we obtain $i(\alpha;\gamma)(\omega) = \log\frac{\mathbb{P}(C_\omega|A_\omega)}{\mathbb{P}(C_\omega)}$, and thus we have

$$I(\alpha;\gamma) = \int_\Omega i(\alpha;\gamma)\,\mathrm{d}\mathbb{P}.$$

For the general case where the partition α is replaced by a σ-subalgebra $\mathscr{A}$ of $\mathscr{F}$, we obtain

$$\begin{aligned} i(\mathscr{A};\gamma) &:= h(\gamma) - h(\gamma|\mathscr{A}) \\ &= -\sum_{C\in\gamma}\mathbb{1}_C\log\mathbb{P}(C) + \sum_{C\in\gamma}\mathbb{1}_C\log\mathbb{P}(C|\mathscr{A}) \\ &= \sum_{C\in\gamma}\mathbb{1}_C\log\frac{\mathbb{P}(C|\mathscr{A})}{\mathbb{P}(C)}. \end{aligned}$$

This leads to a corresponding generalisation of (21):

$$\begin{aligned} I(\mathscr{A};\gamma) &:= \int_\Omega i(\mathscr{A};\gamma)\,\mathrm{d}\mathbb{P} \\ &= \sum_{C\in\gamma}\int_\Omega \mathbb{1}_C\log\frac{\mathbb{P}(C|\mathscr{A})}{\mathbb{P}(C)}\,\mathrm{d}\mathbb{P} \\ &= \sum_{C\in\gamma}\int_\Omega \mathbb{P}(C|\mathscr{A})\log\frac{\mathbb{P}(C|\mathscr{A})}{\mathbb{P}(C)}\,\mathrm{d}\mathbb{P}. \end{aligned} \tag{22}$$

Conditional Mutual Information Finally, we define the *conditional mutual information*. With two σ-subalgebras $\mathscr{A}$ and $\mathscr{B}$ of $\mathscr{F}$, we define

$$i(\mathscr{B};\gamma|\mathscr{A}) := h(\gamma|\mathscr{A}) - h(\gamma|\mathscr{A}\vee\mathscr{B}) = \sum_{C\in\gamma} \mathbb{1}_C \log \frac{\mathbb{P}(C|\mathscr{A}\vee\mathscr{B})}{\mathbb{P}(C|\mathscr{A})}.$$

(Here, $\mathscr{A}\vee\mathscr{B}$ denotes the smallest σ-algebra that contains $\mathscr{A}$ and $\mathscr{B}$.) Integration of this function leads to

$$\begin{aligned} I(\mathscr{B};\gamma|\mathscr{A}) &= \int_\Omega i(\mathscr{B};\gamma|\mathscr{A})\,\mathrm{d}\mathbb{P} \\ &= \sum_{C\in\gamma}\int_\Omega \mathbb{1}_C \log\frac{\mathbb{P}(C|\mathscr{A}\vee\mathscr{B})}{\mathbb{P}(C|\mathscr{A})}\,\mathrm{d}\mathbb{P} \\ &= \sum_{C\in\gamma}\int_\Omega \mathbb{P}(C|\mathscr{A}\vee\mathscr{B})\log\frac{\mathbb{P}(C|\mathscr{A}\vee\mathscr{B})}{\mathbb{P}(C|\mathscr{A})}\,\mathrm{d}\mathbb{P}. \end{aligned} \tag{23}$$

In a final step, we could further extend $i(\mathscr{B};\gamma|\mathscr{A})$ and $I(\mathscr{B};\gamma|\mathscr{A})$ to the case where γ is replaced by a σ-algebra $\mathscr{C}$, by taking the supremum over all finite partitions γ in $\mathscr{C}$. However, in this article we restrict attention to a fixed finite partition γ.

4 The Chain Rule as a Guiding Scheme

4.1 Two Inputs

In the introduction, Section 1, we have used the two-input case for discrete random variables in order to highlight the main issue with the classical definitions of the mutual information and the conditional mutual information and to outline the core idea of this article. After having introduced the required information-theoretic quantities for more general variables in Section 3, we now revisit the instructive two-input case and demonstrate how measure-theoretic concepts come into play here very naturally.

Consider measurable spaces $(\mathsf{X},\mathscr{X})$, $(\mathsf{Y},\mathscr{Y})$, $(\mathsf{Z},\mathscr{Z})$, and their product

$$(\Omega,\mathscr{F}) := (\mathsf{X}\times\mathsf{Y}\times\mathsf{Z}, \mathscr{X}\otimes\mathscr{Y}\otimes\mathscr{Z}).$$

In order to ensure the existence of various (regular versions of) conditional distributions, we need to assume that these measurable spaces carry a further structure. Typically, it is sufficient to require that $(\mathsf{X},\mathscr{X})$, $(\mathsf{Y},\mathscr{Y})$, and $(\mathsf{Z},\mathscr{Z})$ are Polish spaces (see [11], Theorem 13.1.1), which will be implicitly assumed hereinafter for all measurable spaces.

Now, consider a probability measure μ on $(\mathsf{X}\times\mathsf{Y},\mathscr{X}\otimes\mathscr{Y})$ and a Markov kernel

$$\nu:\ \mathsf{X}\times\mathsf{Y}\times\mathscr{Z}\to[0,1],$$

which models a channel that takes two inputs, $x\in\mathsf{X}$ and $y\in\mathsf{Y}$, and generates a possibly random output $z\in\mathsf{Z}$. This allows us to define a probability measure on the joint space $(\Omega,\mathscr{F})$, given by

$$\mathbb{P}(A\times B\times C) := \int_{A\times B} \nu(x,y;C)\,\mu(\mathrm{d}x,\mathrm{d}y).$$

With the natural projections

$$X:\Omega\to\mathsf{X},\ (x,y,z)\mapsto x,\quad Y:\Omega\to\mathsf{Y},\ (x,y,z)\mapsto y,\quad Z:\Omega\to\mathsf{Z},\ (x,y,z)\mapsto z,$$

we have

$$\mu(A \times B) = \mathbb{P}(X \in A, Y \in B), \tag{24}$$

$$\nu(x, y; C) = \mathbb{P}(Z \in C | X = x, Y = y), \qquad \mu\text{-almost everywhere}, \tag{25}$$

where the set of probability one to which "μ-almost everywhere" refers in (25) is independent of C. Furthermore, we have the marginals

$$\mu_X(A) := \mu(A \times \mathsf{Y}) = \mathbb{P}(X \in A),$$
$$\mu_Y(B) := \mu(\mathsf{X} \times B) = \mathbb{P}(Y \in B),$$

and, finally, the ν-push-forward measure of μ,

$$\nu_*(\mu)(C) := \mathbb{P}(Z \in C).$$

Note that the definition of the conditional distribution $\mathbb{P}(Z \in C | X = x, Y = y)$ on the RHS of (25) is quite general and does not exclude cases where $\mathbb{P}(X = x, Y = y) = 0$. It requires a formalism that goes beyond the context of variables with finitely many state sets X, Y, and Z. It is important to outline this formalism in some detail here, which will provide the basis for an appropriate definition of marginal channels. The definition of the conditional distribution

$$\mathbb{P}(Z \in C | X = x, Y = y) \tag{26}$$

involves two steps:

Step 1. We interpret the indicator function $\mathbb{1}_{\{Z \in C\}}$ as an element of the Hilbert space $L^2(\Omega, \mathscr{F}, \mathbb{P})$ and project it onto the (closed) linear subspace of (X, Y)-measurable functions $\Omega \to \mathbb{R}$. Its projection is referred to as *conditional expectation* and denoted by

$$\mathbb{E}(\mathbb{1}_{\{Z \in C\}} | X, Y). \tag{27}$$

Note that the elements of the Hilbert space $L^2(\Omega, \mathscr{F}, \mathbb{P})$ are equivalence classes of functions where two functions are identified if they coincide on a measurable set of probability one. Therefore, the conditional expectation (27) is only almost surely well defined, where the set of probability one to which "almost surely" refers is dependent on C.

Step 2. Formally, $\mathbb{E}(\mathbb{1}_{\{Z \in C\}} | X, Y)$ is a real-valued function defined on Ω. On the other hand, it is (X, Y)-measurable so that we should be able to interpret it as a function of x and y. Indeed, it follows from the factorisation lemma that there is a unique measurable function $\varphi_C : (\mathsf{X} \times \mathsf{Y}, \mathscr{X} \otimes \mathscr{Y}) \to \mathbb{R}$ satisfying $\mathbb{E}(\mathbb{1}_{\{Z \in C\}} | X, Y) = \varphi_C \circ (X, Y)$. The conditional distribution (26) is then simply defined to be the function φ_C, which has x and y as arguments. In the special situation where we start with a Markov kernel ν, we recover it, μ-almost everywhere, in terms of equation (25). It turns out that this equation already describes a quite general situation. Under mild conditions, assuming, for instance, that all measurable spaces are Polish spaces, the conditional distribution (26) can be considered to be a Markov kernel, as a function of x, y, and C.

For the definition of mutual information and conditional mutual information, as generalisations of (3) and (6), respectively, we have to find an appropriate notion of a marginal kernel. We begin with the conditional distribution $\mathbb{P}(Z \in C | X = x)$, as generalisation of $p(z|x)$. For its evaluation we repeat the arguments of the above two steps and consider the conditional expectation

$$\mathbb{E}(\mathbb{1}_{\{Z \in C\}} | X). \tag{28}$$

This is an X-measurable random variable $\Omega \to \mathbb{R}$. By the factorisation lemma we have a unique measurable function $\nu_X(\cdot; C) : (\mathsf{X}, \mathscr{X}) \to \mathbb{R}$ satisfying $\mathbb{E}(\mathbb{1}_{\{Z\in C\}}|X) = \nu_X(X; C)$, and we set

$$\mathbb{P}(Z \in C|X = x) := \nu_X(x; C).$$

Under mild conditions we can assume that $\nu_X(x; C)$ defines a Markov kernel when considered as a function $\nu_X : \mathsf{X} \times \mathscr{Z} \to [0, 1]$ in x and C.

We can now easily extend the classical definitions of mutual information and conditional mutual information to the context of this section. For a finite measurable partition ξ of Z, the sets $Z^{-1}(C)$, $C \in \xi$, form a corresponding finite measurable partition $Z^{-1}(\xi)$ of Ω, and we can use (22) to define the mutual informations

$$\begin{aligned} I_\xi(X, Y; Z) &:= I(\sigma(X) \vee \sigma(Y); Z^{-1}(\xi)) \\ &= \sum_{C\in\xi} \int_{\mathsf{X}\times Y} \nu(x, y; C) \log \frac{\nu(x, y; C)}{\nu_*(\mu)(C)}\, \mu(\mathrm{d}x, \mathrm{d}y), \end{aligned}$$

and

$$\begin{aligned} I_\xi(X; Z) &:= I(\sigma(X); Z^{-1}(\xi)) \\ &= \sum_{C\in\xi} \int_{\mathsf{X}} \nu_X(x; C) \log \frac{\nu_X(x; C)}{\nu_*(\mu)(C)}\, \mu_X(\mathrm{d}x). \end{aligned} \tag{29}$$

Furthermore, with (23) we define the conditional mutual information

$$\begin{aligned} I_\xi(Y; Z|X) &:= I(\sigma(Y); Z^{-1}(\xi)|\sigma(X)) \\ &= \sum_{C\in\xi} \int_{\mathsf{X}\times\mathsf{Y}} \nu(x, y; C) \log \frac{\nu(x, y; C)}{\nu_X(x; C)}\, \mu(\mathrm{d}x, \mathrm{d}y). \end{aligned} \tag{30}$$

We repeat the computation (8) and decompose the mutual information $I_\xi(X, Y; Z)$ as follows:

$$\begin{aligned} I_\xi(X, Y; Z) &= \sum_{C\in\xi} \int_{\mathsf{X}\times Y} \nu(x, y; C) \log \frac{\nu(x, y; C)}{\nu_*(\mu)(C)}\, \mu(\mathrm{d}x, \mathrm{d}y) \\ &= \sum_{C\in\xi} \int_{\mathsf{X}\times Y} \nu(x, y; C) \left[\log \frac{\nu_X(x; C)}{\nu_*(\mu)(C)} + \log \frac{\nu(x, y; C)}{\nu_X(x; C)}\right] \mu(\mathrm{d}x, \mathrm{d}y) \quad (31) \\ &= I_\xi(X; Z) + I_\xi(Y; Z|X). \quad (32) \end{aligned}$$

We argue that, in order to have a causal decomposition of the full mutual information $I_\xi(X, Y; Z)$ into two terms similar to $I_\xi(X; Z)$ and $I_\xi(Y; Z|X)$, we have to modify the marginal channel

$$\nu_X(x; C) = \mathbb{P}(Z \in C|X = x)$$

in (31). In this modification, the conditioning with respect to X has to be adjusted to the actual information used by the kernel $\nu(x, y; C)$. To this end, we consider the smallest σ-subalgebra $\mathscr{A}_{X,\bar{y}}$ of $\mathscr{X}$ for which all functions $\nu_{X,\bar{y}}(\cdot; C) := \nu(\cdot, \bar{y}; C)$, $C \in \mathscr{Z}$, are measurable. It corresponds to the partition $\alpha_{X,\bar{y}}$ that appears in (15). Now we generalise the definition (15) of the partition α_X by combining all the σ-algebras $\mathscr{A}_{X,\bar{y}}$:

$$\mathscr{A}_X := \bigvee_{\bar{y}\in\mathsf{Y}} \mathscr{A}_{X,\bar{y}} \subseteq \mathscr{X}. \tag{33}$$

Note that, typically, $\mathscr{A}_{X,\bar{y}}$, $\bar{y} \in \mathsf{Y}$, as well as $\mathscr{A}_X$ cannot be naturally identified with corresponding σ-subalgebras of the σ-algebra generated by the channel ν, that is the smallest

σ-subalgebra of $\mathscr{X} \otimes \mathscr{Y}$ for which $(x, y) \mapsto \nu(x, y; C)$ is measurable for all $C \in \mathscr{Z}$. This is illustrated by the following example.

Example 3 We consider

$$(\mathsf{X}, \mathscr{X}) = (\mathsf{Y}, \mathscr{Y}) = (\mathsf{Z}, \mathscr{Z}) = (\mathbb{R}, \mathscr{B}(\mathbb{R})),$$

where $\mathscr{B}(\mathbb{R})$ denotes the Borel σ-algebra of $\mathbb{R}$. We assume that the channel ν is simply given by the addition $(x, y) \mapsto x + y$:

$$\nu(x, y; C) = \mathbb{1}_C(x + y).$$

As $\mathscr{B}(\mathbb{R})$ is generated by the intervals $[r - \varepsilon, r + \varepsilon] \subseteq \mathbb{R}$, the smallest σ-algebra $\mathscr{A} \subseteq \mathscr{X} \otimes \mathscr{Y}$ for which all functions $\nu(\cdot, \cdot; C)$ are measurable is generated by the following sets

$$A(r, \varepsilon) := \left\{(x, y) \in \mathbb{R}^2 : r - \varepsilon \leq x + y \leq r + \varepsilon\right\}, \qquad r \in \mathbb{R},\ \varepsilon \in \mathbb{R}_+.$$

Now let us consider $\mathscr{A}_{X,\bar{y}}$, the σ-algebra generated by the kernel

$$\nu_{X,\bar{y}} : \mathbb{R} \times \mathscr{B}(\mathbb{R}) \to [0, 1], \qquad (x, C) \mapsto \nu_{X,\bar{y}}(x; C) := \nu(x, \bar{y}; C).$$

It is easy to see that $\mathscr{A}_{X,\bar{y}}$ is the smallest σ-subalgebra of $\mathscr{X}$ containing the $\bar{y}$-sections

$$A_{X,\bar{y}}(r, \varepsilon) := \{x \in \mathbb{R} : (x, \bar{y}) \in A(r, \varepsilon)\} = \{x \in \mathbb{R} : r - \bar{y} - \varepsilon \leq x \leq r - \bar{y} + \varepsilon\},$$

that is, $\mathscr{A}_{X,\bar{y}} = \mathscr{B}(\mathbb{R})$. This example shows that the cylinder sets $A \times \mathbb{R}$, $A \in \mathscr{A}_{X,\bar{y}}$, are not necessarily contained in $\mathscr{A}$ (see illustration in Fig. 3).

With the σ-subalgebra $\mathscr{A}_X$ of $\mathscr{X}$, we can now modify the random variable $X : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathsf{X}, \mathscr{X})$ by simply reducing the image σ-algebra to $\mathscr{A}_X$:

$$\widehat{X} : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathsf{X}, \mathscr{A}_X).$$

We will see that this step is crucial here, even though it might appear like a minor technical step at first sight. It allows us to modify (28) by replacing the full σ-algebra of X, $\mathscr{X}$, by the σ-algebra of $\widehat{X}$:

$$\mathbb{E}(\mathbb{1}_{\{Z \in C\}} | \widehat{X}).$$

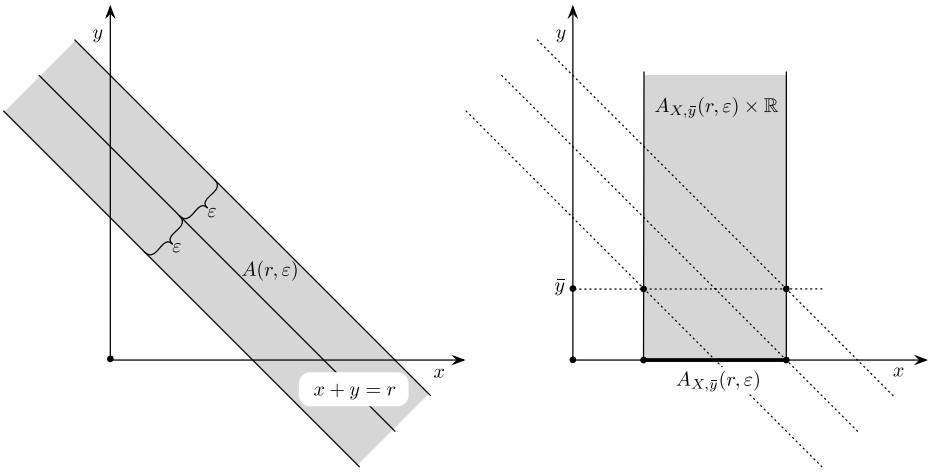

Fig. 3 Illustration of the ν-measurable sets in $\mathbb{R}^2$ and their sections in $\mathbb{R}$

This is, by definition, an $\widehat{X}$-measurable random variable $\Omega \to \mathbb{R}$. By the factorisation lemma, we can find a unique measurable function $\hat{\nu}_X(\cdot; C) : (\mathsf{X}, \mathscr{A}_X) \to \mathbb{R}$ satisfying $\mathbb{E}(\mathbb{1}_{\{Z\in C\}}|\widehat{X}) = \hat{\nu}_X(\widehat{X}; C)$. This yields a new marginal channel,

$$\mathbb{P}(Z \in C|\widehat{X} = x) := \hat{\nu}_X(x; C),$$

as a modification of $\nu_X(x; C)$ which appears twice in (31). Note that the kernel $\hat{\nu}_X(x; C)$ is defined almost surely. However, the definition of a conditional mutual information will be independent of the version of that kernel.

Now we come to the definition of a causal version of the mutual information (29) and the conditional mutual information (30). We simply replace in these definitions $\nu_X(x; C)$ by $\hat{\nu}_X(x; C)$:

$$\begin{aligned} I_\xi(X \to Z) &:= I(\sigma(\widehat{X}); Z^{-1}(\xi)) \\ &= \sum_{C\in\xi} \int_{\mathsf{X}} \hat{\nu}_X(x; C) \log \frac{\hat{\nu}_X(x; C)}{\nu_*(\mu)(C)} \, \mu_X(\mathrm{d}x), \end{aligned} \tag{34}$$

$$\begin{aligned} I_\xi(Y \to Z|X) &:= I(\sigma(Y); Z^{-1}(\xi)|\sigma(\widehat{X})) \\ &= \sum_{C\in\xi} \int_{\mathsf{X}\times\mathsf{Y}} \nu(x, y; C) \log \frac{\nu(x, y; C)}{\hat{\nu}_X(x; C)} \, \mu(\mathrm{d}x, \mathrm{d}y). \end{aligned} \tag{35}$$

The following proposition relates the causal quantities (34) and (35) to the corresponding non-causal ones, (29) and (30).

Proposition 4 *We have the chain rule*

$$I_\xi(X, Y; Z) = I_\xi(X \to Z) + I_\xi(Y \to Z|X). \tag{36}$$

Furthermore,

$$I_\xi(X \to Z) \le I_\xi(X; Z), \qquad I_\xi(Y \to Z|X) \ge I_\xi(Y; Z|X). \tag{37}$$

Proof With C_z denoting the set in ξ that contains z, we have

$$\log \frac{\nu(x, y; C_z)}{\nu_*(\mu)(C_z)} = \log \frac{\nu(x, y; C_z)}{\hat{\nu}_X(x; C_z)} + \log \frac{\hat{\nu}_X(x; C_z)}{\nu_*(\mu)(C_z)}.$$

Integrating this with respect to $\nu(x, y; \mathrm{d}z)$ we get

$$\begin{aligned} &\int_{\mathsf{Z}} \log \frac{\nu(x, y; C_z)}{\nu_*(\mu)(C_z)} \, \nu(x, y; \mathrm{d}z) \\ &\quad = \int_{\mathsf{Z}} \log \frac{\nu(x, y; C_z)}{\hat{\nu}_X(x; C_z)} \, \nu(x, y; \mathrm{d}z) + \int_{\mathsf{Z}} \log \frac{\hat{\nu}_X(x; C_z)}{\nu_*(\mu)(C_z)} \, \nu(x, y; \mathrm{d}z) \\ &\quad = \sum_{C\in\xi} \nu(x, y; C) \log \frac{\nu(x, y; C)}{\hat{\nu}_X(x; C)} + \sum_{C\in\xi} \nu(x, y; C) \log \frac{\hat{\nu}_X(x; C)}{\nu_*(\mu)(C)}. \end{aligned} \tag{38}$$

Further integrating the first term of (38) with respect to μ gives us $I_\xi(Y \to Z|X)$ (see (35)). For the corresponding integration of the second term, we obtain

$$\sum_{C\in\xi}\int_{\mathsf{X}\times\mathsf{Y}} \nu(x,y;C)\log\frac{\hat{\nu}_X(x;C)}{\nu_*(\mu)(C)}\,\mu(\mathrm{d}x,\mathrm{d}y)$$

$$= \sum_{C\in\xi}\int_\Omega \mathbb{P}(Z\in C|X,Y)\log\frac{\mathbb{P}(Z\in C|\widehat{X})}{\mathbb{P}(Z\in C)}\,\mathrm{d}\mathbb{P} \tag{39}$$

$$= \sum_{C\in\xi}\int_\Omega \mathbb{P}(Z\in C|\widehat{X})\log\frac{\mathbb{P}(Z\in C|\widehat{X})}{\mathbb{P}(Z\in C)}\,\mathrm{d}\mathbb{P} \tag{40}$$

$$= \sum_{C\in\xi}\int_{\mathsf{X}} \hat{\nu}_X(x;C)\log\frac{\hat{\nu}_X(x;C)}{\nu_*(\mu)(C)}\,\mu_X(\mathrm{d}x). \tag{41}$$

The crucial step (40) follows from the general property of the conditional expectation of a function f with respect to a σ-subalgebra $\mathscr{A}$:

$$\int_\Omega fg\,\mathrm{d}\mathbb{P} = \int_\Omega \mathbb{E}(f|\mathscr{A})g\,\mathrm{d}\mathbb{P}, \qquad \text{for all } \mathscr{A}\text{-measurable functions } g.$$

Here, f is given by $\mathbb{P}(Z\in C|X,Y)$, $\mathscr{A}$ is the σ-algebra generated by $\widehat{X}$, and g is given by $\log\frac{\mathbb{P}(Z\in C|\widehat{X})}{\mathbb{P}(Z\in C)}$ which is $\widehat{X}$-measurable. The steps (39) and (41) follow directly from the definitions of the Markov kernels, and we finally obtain $I_\xi(X\to Z)$ (see (34)). This concludes the proof of the chain rule (36).

We now prove the inequalities (37) where we can restrict attention to the first one. We consider the convex function $\phi(r) := r\log\frac{r}{\nu_*(\mu)(C)}$, for $r>0$, and $\phi(0):=0$, and apply Jensen's inequality for conditional expectations:

$$\mathbb{E}\left(\phi\big(\mathbb{P}(Z\in C|X)\big)\middle|\widehat{X}\right) \geq \phi\left(\mathbb{E}\left(\mathbb{P}(Z\in C|X)\middle|\widehat{X}\right)\right) = \phi\left(\mathbb{P}(Z\in C|\widehat{X})\right). \tag{42}$$

This implies

$$\begin{aligned}
I_\xi(X;Z) &= \sum_{C\in\xi}\int_\Omega \mathbb{P}(Z\in C|X)\log\frac{\mathbb{P}(Z\in C|X)}{\mathbb{P}(Z\in C)}\,\mathrm{d}\mathbb{P}\\
&= \sum_{C\in\xi}\mathbb{E}\left(\phi\big(\mathbb{P}(Z\in C|X)\big)\right)\\
&= \sum_{C\in\xi}\mathbb{E}\left(\mathbb{E}\left(\phi\big(\mathbb{P}(Z\in C|X)\big)\middle|\widehat{X}\right)\right)\\
&\geq \sum_{C\in\xi}\mathbb{E}\left(\phi\left(\mathbb{P}(Z\in C|\widehat{X})\right)\right) \qquad \text{(by (42))}\\
&= \sum_{C\in\xi}\int_\Omega \mathbb{P}(Z\in C|\widehat{X})\log\frac{\mathbb{P}(Z\in C|\widehat{X})}{\mathbb{P}(Z\in C)}\,\mathrm{d}\mathbb{P}\\
&= I_\xi(X\to Z).
\end{aligned}$$

The second inequality in (37) follows from the first one and the chain rule (36). □

Let us interpret this result. The first inequality in (37) highlights the fact that the stochastic dependence between X and Z, here quantified by the usual mutual information $I_\xi(X;Z)$, cannot be fully attributed to the causal effect of X on Z. Some part of $I_\xi(X;Z)$ is purely

associational, and $I_\xi(X \to Z)$ constitutes the causal part of it. The second inequality in (37) highlights a different fact. Conditioning on the variable X may "screen off" some part of the causal effect of Y on Z. More precisely, the uncertainty reduction about the outcome of Z through X can be so strong that a further reduction through Y becomes "invisible". Therefore, the classical conditional mutual information, $I_\xi(Y; Z|X)$, tends to reflect only part of the causal influence of Y on Z given X, $I_\xi(Y \to Z|X)$. Even though the classical information-theoretic quantities are replaced by their causal versions, the full mutual information can still be decomposed according to the chain rule (36). However, in comparison to the decomposition (32), some amount of it is shifted from one term to the other so that both terms can be interpreted causally.

It turns out that the definitions (34) and (35) require a careful extension if we want to have a general chain rule for more than two input variables. We are now going to highlight this for three input variables.

4.2 Three Inputs

We now consider three input variables. This will reveal that the previous case with two input variables is quite special. An extension to more than two variables requires an adjustment of our definition of causal information flow.

We consider a third input variable (denoted below by W) with values in a measurable space $(\mathsf{W}, \mathscr{W})$, a probability measure

$$\mu \qquad \text{on} \qquad (\mathsf{W} \times \mathsf{X} \times \mathsf{Y}, \mathscr{W} \otimes \mathscr{X} \otimes \mathscr{Y}),$$

and an input-output channel, given by a Markov kernel

$$\nu : \mathsf{W} \times \mathsf{X} \times \mathsf{Y} \times \mathscr{Z} \to [0, 1].$$

This gives rise to a probability space, consisting of the measurable space

$$(\Omega, \mathscr{F}) := (\mathsf{W} \times \mathsf{X} \times \mathsf{Y} \times \mathsf{Z}, \mathscr{W} \otimes \mathscr{X} \otimes \mathscr{Y} \otimes \mathscr{Z}),$$

and the probability measure $\mathbb{P}$ defined by

$$\mathbb{P}(A \times B \times C \times D) := \int_{A \times B \times C} \nu(w, x, y; D)\, \mu(\mathrm{d}w, \mathrm{d}x, \mathrm{d}y).$$

Finally, we have the natural projections $W : \Omega \to \mathsf{W}$, $X : \Omega \to \mathsf{X}$, $Y : \Omega \to \mathsf{Y}$, and $Z : \Omega \to \mathsf{Z}$.

The definition of the marginal kernel $\hat{\nu}_X(x; C)$, as introduced in Section 4.1, is directly applicable to the situation of three input variables. It allows us to define marginal channels by an appropriate grouping of two input variables into one input variable, which formally reduces the three-input case to a two-input case. In particular, we can define the channels $\hat{\nu}_{W,X}(w, x; C)$ and $\hat{\nu}_W(w; C)$, by grouping W, X and X, Y, respectively, into one variable. Denoting the set in ξ that contains z by C_z, we then have

$$\log \frac{\nu(w, x, y; C_z)}{\nu_*(\mu)(C_z)} = \log \frac{\nu(w, x, y; C_z)}{\hat{\nu}_{W,X}(w, x; C_z)} + \log \frac{\hat{\nu}_{W,X}(w, x; C_z)}{\hat{\nu}_W(w; C_z)} + \log \frac{\hat{\nu}_W(w; C_z)}{\nu_*(\mu)(C_z)}.$$

By integration we obtain

$$\begin{aligned}
\int_{\mathsf{Z}} \log \frac{\nu(w,x,y;C_z)}{\nu_*(\mu)(C_z)}\,\nu(w,x,y;\mathrm{d}z) &= \int_{\mathsf{Z}} \log \frac{\nu(w,x,y;C_z)}{\hat{\nu}_{W,X}(w,x;C_z)}\,\nu(w,x,y;\mathrm{d}z) \\
&\quad + \int_{\mathsf{Z}} \log \frac{\hat{\nu}_{W,X}(w,x;C_z)}{\hat{\nu}_W(w;C_z)}\,\nu(w,x,y;\mathrm{d}z) \\
&\quad + \int_{\mathsf{Z}} \log \frac{\hat{\nu}_W(w;C_z)}{\nu_*(\mu)(C_z)}\,\nu(w,x,y;\mathrm{d}z) \\
&= \sum_{C\in\xi} \nu(w,x,y;C) \log \frac{\nu(w,x,y;C)}{\hat{\nu}_{W,X}(w,x;C)} \qquad (43) \\
&\quad + \sum_{C\in\xi} \nu(w,x,y;C) \log \frac{\hat{\nu}_{W,X}(w,x;C)}{\hat{\nu}_W(w;C)} \qquad (44) \\
&\quad + \sum_{C\in\xi} \nu(w,x,y;C) \log \frac{\hat{\nu}_W(w;C)}{\nu_*(\mu)(C)}. \qquad (45)
\end{aligned}$$

An integration of the last term (45) with respect to μ yields, by the same reasoning as in the steps (39), (40), and (41),

$$I_\xi(W \to Z) = \sum_{C\in\xi} \int_{\mathsf{W}} \hat{\nu}_W(w;C) \log \frac{\hat{\nu}_W(w;C)}{\nu_*(\mu)(C)}\,\mu_W(\mathrm{d}w).$$

A corresponding integration of the first term (43) with respect to μ yields a non-negative quantity that can be interpreted as $I_\xi(Y \to Z|W,X)$ (see definition (35)). Even though we will have to slightly adjust this first term, the problem we are facing here is most clearly highlighted by the second term, (44). In order to naturally generalise the chain rule (36) we have to interpret the integral of the second term as $I_\xi(X \to Z|W)$. However, it turns out that, in general,

$$\begin{aligned}
I_\xi(X \to Z|W) &= \sum_{C\in\xi} \int_{\mathsf{W}\times\mathsf{X}} \hat{\nu}_{W,X}(w,x;C) \log \frac{\hat{\nu}_{W,X}(w,x;C)}{\hat{\nu}_W(w;C)}\,\mu_{W,X}(\mathrm{d}w,\mathrm{d}x) \\
&\neq \sum_{C\in\xi} \int_{\mathsf{W}\times\mathsf{X}\times\mathsf{Y}} \nu(w,x,y;C) \log \frac{\hat{\nu}_{W,X}(w,x;C)}{\hat{\nu}_W(w;C)}\,\mu_{W,X}(\mathrm{d}w,\mathrm{d}x,\mathrm{d}y),
\end{aligned} \qquad (46)$$

where (46) is the integral of the term (44) with respect to μ. We cannot even ensure that this integral is non-negative. The reason is that the σ-algebra used for the definition of $\hat{\nu}_W(w;C)$ is not necessarily a σ-subalgebra of the one used for the definition of the kernel $\hat{\nu}_{W,X}(w,x;C)$ (the situation is similar to the one of Example 3). Therefore, the reasoning of the steps (39), (40), and (41), cannot be applied here.

The problem highlighted in this section will now be resolved. This will be done by a modification of the involved σ-algebras, which should define a filtration in order to imply a general causal version of the chain rule. In the next section, this modification will be presented for the general case of n input variables.

5 The General Definition of Causal Information Flow

5.1 Filtrations and Information

Let $(\mathsf{X}_i, \mathscr{X}_i)$, $i \in N := [n] = \{1, \ldots, n\}$, be a family of measurable spaces, the state spaces of the input variables. For each subset M of N, we have the corresponding product space $(\mathsf{X}_M, \mathscr{X}_M)$ consisting of $\mathsf{X}_M := \times_{i \in M} \mathsf{X}_i$ and $\mathscr{X}_M := \otimes_{i \in M} \mathscr{X}_i$. Note that for $M = \emptyset$, the set $\mathsf{X}_\emptyset$ consists of one element, the empty sequence ϵ, and $\mathscr{X}_\emptyset = \{\emptyset, \{\epsilon\}\}$ is the trivial σ-algebra with two elements. In addition to the input variables, we consider an output variable with state space $(\mathsf{Z}, \mathscr{Z})$. The input-output channel is given by a Markov kernel

$$\nu : \mathsf{X}_N \times \mathscr{Z} \to [0, 1].$$

Together with a probability measure μ on $(\mathsf{X}_N, \mathscr{X}_N)$ this defines the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ where

$$\Omega := \mathsf{X}_N \times \mathsf{Z}, \qquad \mathscr{F} := \mathscr{X}_N \otimes \mathscr{Z},$$

and

$$\mathbb{P}(A \times C) := \int_A \nu(x; C)\, \mu(\mathrm{d}x), \qquad A \in \mathscr{X}_N,\ C \in \mathscr{Z}.$$

Finally, we have the canonical projections

$$X_M : \Omega \to \mathsf{X}_M, \quad M \subseteq N, \qquad \text{and} \qquad Z : \Omega \to \mathsf{Z}.$$

We are now going to define the M-marginal of the channel based on a general σ-subalgebra $\mathscr{B}_M$ of $\mathscr{X}_M$. Below, in Section 5.2, this will allow us to incorporate causal aspects of ν by an appropriate adaptation of $\mathscr{B}_M$ to ν. In order to highlight the flexibility that we have here, let us begin with the usual definition where $\mathscr{B}_M$ equals the largest σ-subalgebra of $\mathscr{X}_M$, that is $\mathscr{X}_M$ itself. Given a measurable set $C \subseteq \mathsf{Z}$, we have the conditional expectation

$$\mathbb{E}\left(\mathbb{1}_{\{Z \in C\}} | X_M\right). \tag{47}$$

This is by definition an X_M-measurable function $\Omega \to \mathbb{R}$. By the factorisation lemma we can represent it as a composition $\mathbb{E}(\mathbb{1}_{\{Z \in C\}} | X_M) = \nu_M(X_M; C)$ with a measurable function $\nu_M(\cdot; C) : (\mathsf{X}_M, \mathscr{X}_M) \to \mathbb{R}$. This allows us to define the conditional distribution

$$\mathbb{P}(Z \in C | X_M = x_M) := \nu_M(x_M; C) \tag{48}$$

which can be interpreted as a channel

$$\nu_M : \mathsf{X}_M \times \mathscr{Z} \to [0, 1], \qquad (x_M, C) \mapsto \nu_M(x_M; C).$$

We now modify the outlined marginalisation of ν by reducing the maximal σ-algebra $\mathscr{X}_M$ to the σ-subalgebra $\mathscr{B}_M$. More precisely, we replace X_M in (47) by

$$\widehat{X}_M : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathsf{X}_M, \mathscr{B}_M)$$

and consider the conditional expectation

$$\mathbb{E}\left(\mathbb{1}_{\{Z \in C\}} | \widehat{X}_M\right).$$

This is now an $\widehat{X}_M$-measurable function $\Omega \to \mathbb{R}$, and, by the factorisation lemma, we can represent it as a composition $\mathbb{E}(\mathbb{1}_{\{Z \in C\}} | \widehat{X}_M) = \hat{\nu}(\widehat{X}_M; C)$ with a measurable function $\hat{\nu}(\cdot; C) : (\mathsf{X}_M, \mathscr{B}_M) \to \mathbb{R}$. Finally, we have the modification

$$\mathbb{P}(Z \in C | \widehat{X}_M = x_M) := \hat{\nu}(x_M; C)$$

of the conditional distribution (48), which corresponds to a modified channel

$$\hat{\nu}_M : \mathsf{X}_M \times \mathscr{Z} \to [0,1], \qquad (x_M, C) \mapsto \hat{\nu}_M(x_M; C).$$

By construction, $\hat{\nu}_M$ is $\mathscr{B}_M$-measurable, that is, it uses only information that is contained in $\mathscr{B}_M$. For the maximal σ-algebra we recover ν_M. We can also consider the other extreme where $\mathscr{B}_M$ equals the smallest σ-algebra, $\{\emptyset, \mathsf{X}_M\}$. In that case, we obtain $\hat{\nu}(x_M; C) = \nu_*(\mu)(C)$. An adjustment of $\mathscr{B}_M$ to the information actually used by ν will allow us to interpret $\hat{\nu}$ causally. In contrast, if we do not have such an adjustment, $\hat{\nu}_M$ will represent a hypothetical channel, a "ghost channel", based on the σ-algebra of an external observer rather than the σ-algebra of the actual mechanisms of the channel. As we will see in Section 5.2, there are various natural ways to adjust $\mathscr{B}_M$ to ν. However, for the largest subset M of inputs, N, ν will always be measurable with respect to $\mathscr{B}_N$. This ensures that we recover ν when we condition with respect to $\mathscr{B}_N$, that is, $\hat{\nu}_N = \nu$.

We now consider a family $\mathscr{B} = (\mathscr{B}_M)_{M \subseteq N}$ of σ-algebras. It gives rise to a corresponding family

$$\mathscr{F}_M := X_M^{-1}(\mathscr{B}_M) \subseteq \mathscr{F}, \qquad M \subseteq N,$$

of σ-algebras on Ω. We call the family $\mathscr{B}$ *projective*, if the maps

$$\pi_L^M : \mathsf{X}_M \to \mathsf{X}_L, \qquad x_M = (x_i)_{i \in M} \mapsto x_L = (x_i)_{i \in L}, \qquad L \subseteq M \subseteq N,$$

are $\mathscr{B}_M$-$\mathscr{B}_L$-measurable. For projective families, we have the following monotonicity:

$$L \subseteq M \qquad \Rightarrow \qquad \mathscr{F}_L \subseteq \mathscr{F}_M. \tag{49}$$

Given a projective family $\mathscr{B}$, we now define a corresponding family of information-theoretic quantities which generalise (conditional) mutual information. We begin with a local version, applied to a measurable partition ξ of Z. For $z \in \mathsf{Z}$, we denote the set in ξ that contains z by C_z. For $L \subseteq M \subseteq N$, we consider $x_M = (x_L, x_{M \setminus L}) \in \mathsf{X}_M$ and define

$$i_\xi^{\mathscr{B}}(x_{M \setminus L} \to z | x_L) := \log \frac{\hat{\nu}_M(x_M; C_z)}{\hat{\nu}_L(x_L; C_z)}.$$

This is a local version of the conditional mutual information. Integration over z yields

$$\int_{\mathsf{Z}} \log \frac{\hat{\nu}_M(x_M; C_z)}{\hat{\nu}_L(x_L; C_z)} \, \nu(x; \mathrm{d}z) = \sum_{C \in \xi} \nu(x; C) \log \frac{\hat{\nu}_M(x_M; C)}{\hat{\nu}_L(x_L; C)}.$$

With a second integration, with respect to μ, we obtain

$$\begin{aligned}
&\sum_{C \in \xi} \int_{\mathsf{X}} \nu(x; C) \log \frac{\hat{\nu}_M(x_M; C)}{\hat{\nu}_L(x_L; C)} \, \mu(\mathrm{d}x) \\
&= \sum_{C \in \xi} \int_\Omega \mathbb{E}\left(\mathbb{1}_{\{Z \in C\}} | X\right) \log \frac{\mathbb{E}\left(\mathbb{1}_{\{Z \in C\}} | \widehat{X}_M\right)}{\mathbb{E}\left(\mathbb{1}_{\{Z \in C\}} | \widehat{X}_L\right)} \, \mathrm{d}\mathbb{P} \\
&= \sum_{C \in \xi} \int_\Omega \mathbb{E}\left(\mathbb{1}_{\{Z \in C\}} | \widehat{X}_M\right) \log \frac{\mathbb{E}\left(\mathbb{1}_{\{Z \in C\}} | \widehat{X}_M\right)}{\mathbb{E}\left(\mathbb{1}_{\{Z \in C\}} | \widehat{X}_L\right)} \, \mathrm{d}\mathbb{P} \qquad \text{(by the monotonicity (49))} \\
&= \sum_{C \in \xi} \int_{\mathsf{X}_M} \hat{\nu}_M(x_M; C) \log \frac{\hat{\nu}_M(x_M; C)}{\hat{\nu}_L(x_L; C)} \, \mu_M(\mathrm{d}x_M),
\end{aligned}$$

where μ_M denotes the M-marginal of μ. This suggests the following version of the conditional mutual information which we refer to as information flow.

Definition 5 Let $\mathscr{B} = (\mathscr{B}_M)_{M\subseteq N}$ be a projective family of σ-algebras, let ξ be a finite measurable partition of Z, and let $L \subseteq M \subseteq N$. Then we define the *information flow from $X_{M\setminus L}$ to Z given X_L* as

$$I_\xi^{\mathscr{B}}(X_{M\setminus L} \to Z|X_L) := \sum_{C\in\xi} \int_{\mathsf{X}_M} \hat{\nu}_M(x_M; C) \log \frac{\hat{\nu}_M(x_M; C)}{\hat{\nu}_L(x_L; C)}\, \mu_M(\mathrm{d}x_M).$$

For $L = \emptyset$ we simplify the notation by $I_\xi^{\mathscr{B}}(X_M \to Z)$ and refer to the *information flow from X_M to Z*.

Given disjoint subsets $M_1, M_2, \ldots, M_k$ of N, we use a filtration of σ-algebras for proving a general chain rule for information flows, thereby extending the chain rule in Proposition 4 and resolving the problem highlighted in Section 4.2 for three inputs.

Theorem 6 (General chain rule) *Let $\mathscr{B} = (\mathscr{B}_M)_{M\subseteq N}$ be a projective family of σ-algebras, let ξ be a finite measurable partition of Z, and let $M_1, M_2, \ldots, M_k$ be disjoint subsets of N. Then*

$$\begin{aligned} I_\xi^{\mathscr{B}}(X_{M_1}, \ldots, X_{M_k} \to Z) \;=\; & I_\xi^{\mathscr{B}}(X_{M_1} \to Z) + I_\xi^{\mathscr{B}}(X_{M_2} \to Z|X_{M_1}) \\ & + \cdots + I_\xi^{\mathscr{B}}(X_{M_k} \to Z|X_{M_1}, X_{M_2}, \ldots, X_{M_{k-1}}). \end{aligned} \tag{50}$$

Proof Let $M^j := \cup_{i=1}^j M_i$, $j = 0, 1, \ldots, k$. The monotonicity (49) implies that the sequence

$$\mathscr{F}_j := \mathscr{F}_{M^j} := X_{M^j}^{-1}(\mathscr{B}_{M^j}), \qquad j = 0, \ldots, k,$$

is increasing and therefore represents a filtration of σ-algebras. This implies

$$\begin{aligned} & I_\xi^{\mathscr{B}}(X_{M_1}, \ldots, X_{M_k} \to Z) \\ &= \sum_{C\in\xi} \int_{\mathsf{X}_{M^k}} \hat{\nu}(x_{M^k}; C) \log \frac{\hat{\nu}(x_{M^k}; C)}{\nu_*(\mu)(C)}\, \mu(\mathrm{d}x_{M^k}) \\ &= \sum_{C\in\xi} \int_{\mathsf{X}_{M^k}} \hat{\nu}(x_{M^k}; C) \log \left(\prod_{j=1}^k \frac{\hat{\nu}_{M^j}(x_{M^j}; C)}{\hat{\nu}_{M^{j-1}}(x_{M^{j-1}}; C)} \right) \mu(\mathrm{d}x_{M^k}) \\ &= \sum_{j=1}^k \sum_{C\in\xi} \int_{\mathsf{X}_{M^k}} \hat{\nu}(x_{M^k}; C) \log \frac{\hat{\nu}_{M^j}(x_{M^j}; C)}{\hat{\nu}_{M^{j-1}}(x_{M^{j-1}}; C)}\, \mu(\mathrm{d}x_{M^k}) \\ &= \sum_{j=1}^k \sum_{C\in\xi} \int_\Omega \mathbb{P}(Z \in C|\widehat{X}_{M^k}) \log \frac{\mathbb{P}(Z \in C|\widehat{X}_{M^j})}{\mathbb{P}(Z \in C|\widehat{X}_{M^{j-1}})}\, \mathrm{d}\mathbb{P} \\ &= \sum_{j=1}^k \sum_{C\in\xi} \int_\Omega \mathbb{P}(Z \in C|\widehat{X}_{M^j}) \log \frac{\mathbb{P}(Z \in C|\widehat{X}_{M^j})}{\mathbb{P}(Z \in C|\widehat{X}_{M^{j-1}})}\, \mathrm{d}\mathbb{P} \\ &= \sum_{j=1}^k \sum_{C\in\xi} \int_{\mathsf{X}_{M^j}} \hat{\nu}_{M^j}(x_{M^j}; C) \log \frac{\hat{\nu}_{M^j}(x_{M^j}; C)}{\hat{\nu}_{M^{j-1}}(x_{M^{j-1}}; C)}\, \mu(\mathrm{d}x_{M^j}) \\ &= \sum_{j=1}^k I_\xi^{\mathscr{B}}(X_{M_j} \to Z|X_{M_1}, \ldots, X_{M_{j-1}}). \end{aligned}$$

□

We now state basic properties of the information flow. Some of these properties are listed in [14] as natural postulates (P0–P4) for a measure of causal strength.

Proposition 7 (Natural properties) *Let $\mathscr{B} = (\mathscr{B}_M)_{M \subseteq N}$ be a projective family of σ-algebras such that ν is measurable with respect to $\mathscr{B}_N$, and let ξ be a finite measurable partition of Z. Then the following properties hold:*

(a) *The information flow from all input variables to the output variable coincides with the mutual information: $I_\xi^{\mathscr{B}}(X_N \to Z) = I_\xi(X_N; Z)$.*
(b) *For a subset M of N, the set of all input variables, the information flow $I_\xi^{\mathscr{B}}(X_M \to Z)$ is smaller than or equal to the mutual information $I_\xi(X_M; Z)$.*
(c) *For a subset M of N, the information flow $I_\xi^{\mathscr{B}}(X_M \to Z|X_{N\setminus M})$ is greater than or equal to the conditional mutual information $I_\xi(X_M; Z|X_{N\setminus M})$.*
(d) *If the information flow $I_\xi^{\mathscr{B}}(X_M \to Z|X_{N\setminus M})$ vanishes then Z is independent of X_M given $X_{N\setminus M}$.*
(e) *For $L \subseteq M \subseteq N$, we have $I_\xi^{\mathscr{B}}(X_L \to Z) \leq I_\xi^{\mathscr{B}}(X_M \to Z)$. In particular, if $I_\xi^{\mathscr{B}}(X_M \to Z) = 0$ then $I_\xi^{\mathscr{B}}(X_L \to Z) = 0$.*

Proof Statement (a) follows from $\hat{\nu}_N(x_N; C) = \nu(x_N; C)$ and $\hat{\nu}_\emptyset(x; C) = \nu_*(\mu)(C)$. The statements (b) and (c) can be proven in the same way as the corresponding inequalities (37) of Proposition 4, thereby using the chain rule

$$I_\xi^{\mathscr{B}}(X_N \to Z) = I_\xi^{\mathscr{B}}(X_M \to Z) + I_\xi^{\mathscr{B}}(X_{N\setminus M} \to Z|X_M)$$

for $M \subseteq N$ (this follows from the general chain rule (50), with $M_1 = M$ and $M_2 = N \setminus M$). In order to prove (d), note that with (c) we have

$$I_\xi^{\mathscr{B}}(X_M \to Z|X_{N\setminus M}) = 0 \quad \Rightarrow \quad I_\xi(X_M; Z|X_{N\setminus M}) = 0.$$

This implies that X_M is independent of Z given $X_{N\setminus M}$. Finally, (e) follows from the chain rule

$$I_\xi^{\mathscr{B}}(X_M \to Z) = I_\xi^{\mathscr{B}}(X_L \to Z) + I_\xi^{\mathscr{B}}(X_{M\setminus L} \to Z|X_L),$$

by the general chain rule (50), with $M_1 = L$ and $M_2 = M \setminus L$. □

Now let us come back to the two- and three-input cases which we studied in Sections 4.1 and 4.2, respectively. These cases guided our search for the setting that allows us to prove the general chain rule. In the two-input case, we have considered the σ-algebra $\mathscr{A}_X$, defined by (33). Similarly, we can consider the correspondingly defined σ-algebras $\mathscr{A}_Y$ and $\mathscr{A}_{X,Y}$. The latter is simply the σ-algebra generated by the channel ν. Example 3 then demonstrates that the projections $(x, y) \mapsto x$ and $(x, y) \mapsto y$ do not have to be $\mathscr{A}_{X,Y}$-$\mathscr{A}_X$- and $\mathscr{A}_{X,Y}$-$\mathscr{A}_Y$-measurable, respectively. This simply means that the family, extended by $\mathscr{A}_\emptyset = \{\emptyset, \{\epsilon\}\}$, is not necessarily projective. Why does the chain rule still hold this case? The reason is that we can simply extend $\mathscr{A}_{X,Y}$ to the full σ-algebra $\mathscr{X} \otimes \mathscr{Y}$, thereby making the family projective. This extension is trivial in the sense that it has no "visible" effect, simply because conditioning ν with respect to $\mathscr{X} \otimes \mathscr{Y}$ is the same as conditioning it with respect to $\mathscr{A}_{X,Y}$. As a result, we obtain the chain rule in Proposition 4 from Theorem 6. Going from the two-input case to the three-input case, we can define the corresponding family of σ-algebras, which is again not necessarily projective. However, in this case it is not possible to make it projective without visible effect. This demonstrates that the two-input case is quite special. If we want to consider the chain rule as a fundamental property of causal information flows, we have to accept the construction of a projective family of σ-algebras that are associated

with ν as a natural step. The individual causal information flows will then crucially depend on this family. In the following section, we are going to propose two natural ways of such a construction.

5.2 Adaptation of the Filtration to the Channel

We are now going to couple the family $(\mathscr{B}_M)_{M\subseteq N}$ to the channel ν so that we can interpret the corresponding marginals $(\hat{\nu}_M)_{M\subseteq N}$ causally. In order to simplify the presentation, we first consider an arbitrary σ-subalgebra $\mathscr{A}$ of $\mathscr{X}_N$. (Below, $\mathscr{A}$ will be chosen to be the σ-algebra generated by ν.) We begin with information in M in the context of a configuration $\bar{x}$ outside of M, that is $\bar{x} \in \mathsf{X}_{N\setminus M}$. Given such an $\bar{x}$, we define the $(M, \bar{x})$-*trace* of $\mathscr{A}$ as follows: For each $A \in \mathscr{A}$, we consider the $(M, \bar{x})$-*section* of A,

$$\mathrm{sec}_{M,\bar{x}}(A) := A_{M,\bar{x}} := \{x \in \mathsf{X}_M : (x, \bar{x}) \in A\}.$$

These sections then form the $(M, \bar{x})$-trace of $\mathscr{A}$, that is

$$\mathrm{tr}_{M,\bar{x}}(\mathscr{A}) := \mathscr{A}_{M,\bar{x}} := \left\{A_{M,\bar{x}} : A \in \mathscr{A}\right\}.$$

Considering all possible contexts $\bar{x} \in \mathsf{X}_{N\setminus M}$, we finally define the M-*trace* of $\mathscr{A}$ as

$$\mathrm{tr}_M(\mathscr{A}) := \mathscr{A}_M := \bigvee_{\bar{x}\in\mathsf{X}_{N\setminus M}} \mathscr{A}_{M,\bar{x}}.$$

The $(M, \bar{x})$-trace as well as the M-trace of $\mathscr{A}$ are σ-subalgebras of $\mathscr{X}_M$. Note that in the extreme cases $M = \emptyset$ and $M = N$, we recover $\mathscr{A}_\emptyset = \{\emptyset, \{\epsilon\} = \mathsf{X}_\emptyset\}$ (where ϵ denotes the empty sequence), and $\mathscr{A}_N = \mathscr{A}$, respectively.

The family of all M-traces of $\mathscr{A}$ describes how $\mathscr{A}$ is "distributed" over the subsets M of N. However, there is a problem here: The canonical projections π_L^M are not necessarily $\mathscr{A}_M$-$\mathscr{A}_L$-measurable. This projectivity property is required for the definition of a measure of causal information flow that satisfies the general chain rule of Theorem 6. We highlighted this problem for the three-input case in Section 4.2. There are two ways to recover the projectivity, first by extending and second by reducing $\mathscr{A}_M$ appropriately. Let us begin with the extension:

$$\overline{\mathscr{A}}_M := \bigvee_{L\subseteq M} \left(\pi_L^M\right)^{-1}\left(\mathscr{A}_L\right). \tag{51}$$

We have the following characterisation of the family $\overline{\mathscr{A}}_M$, $M \subseteq N$, as the smallest projective extension of the family $\mathscr{A}_M$, $M \subseteq N$.

Proposition 8 (Extension of $\mathscr{A}_M$, $M \subseteq N$) *The family $\overline{\mathscr{A}}_M$, $M \subseteq N$, satisfies the following two conditions:*

1. *For all $M \subseteq N$, $\mathscr{A}_M$ is contained in $\overline{\mathscr{A}}_M$.*
2. *For all $L \subseteq M \subseteq N$, the canonical projection π_L^M is $\overline{\mathscr{A}}_M$-$\overline{\mathscr{A}}_L$-measurable.*

Furthermore, for every family $\mathscr{A}'_M$, $M \subseteq N$, that satisfies these two conditions (where $\overline{\mathscr{A}}_M$ is replaced by $\mathscr{A}'_M$), we have

$$\overline{\mathscr{A}}_M \subseteq \mathscr{A}'_M \qquad \textit{for all } M \subseteq N. \tag{52}$$

Proof The first statement is clear (simply choose on the RHS of (51) $L = M$). For the second statement, we have to show

$$\left(\pi_L^M\right)^{-1}\left(\overline{\mathscr{A}}_L\right) = \overline{\mathscr{A}}_M.$$

Given that

$$\overline{\mathscr{A}}_L = \bigvee_{K \subseteq L} \left(\pi_K^L\right)^{-1}\left(\mathscr{A}_K\right)$$

it is sufficient to verify

$$\left(\pi_L^M\right)^{-1}\left(\left(\pi_K^L\right)^{-1}\left(\mathscr{A}_K\right)\right) \subseteq \overline{\mathscr{A}}_M \qquad \text{for all } K \subseteq L. \tag{53}$$

The LHS of (53) reduces to $(\pi_K^M)^{-1}(\mathscr{A}_K)$ which is by definition contained in $\overline{\mathscr{A}}_M$.

Finally, we prove the minimality. It is easy to see that any family $\mathscr{A}'_M$, $M \subseteq N$, that satisfies the two conditions has to contain $(\pi_L^M)^{-1}(\mathscr{A}_L)$, $L \subseteq M$. By definition, $\overline{\mathscr{A}}_M$ is the smallest σ-algebra that contains these σ-subalgebras (see (51)). This implies (52). □

After having defined the smallest extension of the family $\mathscr{A}_M$, $M \subseteq N$, as one way to recover projectivity, we now come to the alternative way, which is by reduction of that family. More precisely, we define

$$\underline{\mathscr{A}}_M := \left\{A \in \mathscr{X}_M : \left(\pi_M^N\right)^{-1}(A) \in \mathscr{A}\right\}. \tag{54}$$

We have the following characterisation of this family as the largest projective reduction of $\mathscr{A}_M$, $M \subseteq N$.

Proposition 9 (Reduction of $\mathscr{A}_M$, $M \subseteq N$) *The family $\underline{\mathscr{A}}_M$, $M \subseteq N$, satisfies the following two conditions:*

1. *For all $M \subseteq N$, $\underline{\mathscr{A}}_M$ is contained in $\mathscr{A}_M$.*
2. *For all $L \subseteq M \subseteq N$, the canonical projection π_L^M is $\underline{\mathscr{A}}_M$-$\underline{\mathscr{A}}_L$-measurable.*

Furthermore, for every family $\mathscr{A}'_M$, $M \subseteq N$, that satisfies these two conditions (where $\underline{\mathscr{A}}_M$ is replaced by $\mathscr{A}'_M$), we have

$$\mathscr{A}'_M \subseteq \underline{\mathscr{A}}_M \qquad \textit{for all } M \subseteq N.$$

Proof In order to prove the first statement, let $A \in \underline{\mathscr{A}}_M$. This means that

$$\widetilde{A} := \left(\pi_M^N\right)^{-1}(A) = A \times \mathsf{X}_{N \setminus M} \in \mathscr{A}.$$

For all $\bar{x} \in \mathsf{X}_{N \setminus M}$, we have

$$\sec_{M,\bar{x}}(\widetilde{A}) = \left\{x \in \mathsf{X}_M : (x, \bar{x}) \in \widetilde{A}\right\} = A.$$

This means that $A \in \mathrm{tr}_M(\mathscr{A}) = \mathscr{A}_M$, which concludes the proof of the first statement. Now we come to the measurability of the canonical projection π_L^M. For this, we choose $A \in \underline{\mathscr{A}}_L$ and have to show $(\pi_L^M)^{-1}(A) \in \underline{\mathscr{A}}_M$:

$$\left(\pi_M^N\right)^{-1}\left(\left(\pi_L^M\right)^{-1}(A)\right) = \left(\pi_L^N\right)^{-1}(A) \in \mathscr{A} \qquad \text{(by definition (54))}.$$

Finally, we have to prove the maximality. Let $\mathscr{A}'_M$, $M \subseteq N$, be a family that satisfies the two conditions. Then

$$\left(\pi_M^N\right)^{-1}\left(\mathscr{A}'_M\right) \subseteq \mathscr{A}'_N \subseteq \mathscr{A}_N = \mathscr{A}.$$

This means that $\mathscr{A}'_M \subseteq \underline{\mathscr{A}}_M$. □

By the definitions, we have

$$\underline{\mathscr{A}}_M \subseteq \mathscr{A} \subseteq \overline{\mathscr{A}}_M, \qquad M \subseteq N,$$

where equalities hold for $M = N$ and $M = \emptyset$. More precisely,

$$\underline{\mathscr{A}}_N = \mathscr{A} = \overline{\mathscr{A}}_N, \qquad \underline{\mathscr{A}}_\emptyset = \{\emptyset, \{\epsilon\}\} = \overline{\mathscr{A}}_\emptyset.$$

This concludes the constructions for a given σ-algebra $\mathscr{A}$, without explicit reference to the channel $\nu : \mathsf{X} \times \mathscr{Z} \to [0, 1]$. We now couple the studied σ-algebras with the channel ν and therefore choose $\mathscr{A}$ to be the σ-algebra generated by the channel ν, that is $\sigma(\nu)$. We highlight this coupling by writing $\mathscr{A}^\nu$, as a particular choice of $\mathscr{A}$, and consider the family $(\mathscr{A}_M^\nu)_{M\subseteq N}$ of its traces, together with the corresponding smallest projective extension $(\overline{\mathscr{A}}_M^\nu)_{M\subseteq N}$ and the largest projective reduction $(\underline{\mathscr{A}}_M^\nu)_{M\subseteq N}$. In the context of a channel, the traces of $\mathscr{A}^\nu$ have a natural interpretation. In order to see this, we first consider a configuration $\bar{x} \in \mathsf{X}_{N\setminus M}$ and define the "constrained" Markov kernel

$$\nu_{M,\bar{x}} : \mathsf{X}_M \times \mathscr{Z} \to [0, 1], \qquad \nu_{M,\bar{x}}(x; C) := \nu(x, \bar{x}; C).$$

We denote the σ-algebra generated by $\nu_{M,\bar{x}}$ by $\sigma_{M,\bar{x}}(\nu)$. Taking all "constraints" $\bar{x}$ into account, we then define

$$\sigma_M(\nu) := \bigvee_{\bar{x}\in\mathsf{X}_{N\setminus M}} \sigma_{M,\bar{x}}(\nu).$$

Proposition 10 *Let $\mathscr{A}^\nu \subseteq \mathscr{X}_N$ be the σ-algebra generated by the Markov kernel $\nu : \mathsf{X}_N \times \mathscr{Z} \to [0, 1]$. Then for all $M \subseteq N$ and all $\bar{x} \in \mathsf{X}_{N\setminus M}$,*

$$\sigma_{M,\bar{x}}(\nu) = \mathrm{tr}_{M,\bar{x}}(\mathscr{A}^\nu) \quad \textit{and} \quad \sigma_M(\nu) = \mathrm{tr}_M(\mathscr{A}^\nu).$$

Proof The σ-algebra $\mathscr{A}^\nu$ is the smallest σ-algebra that contains all measurable sets of the form

$$A = \{x \in \mathsf{X}_N : \nu(x; C) \in B\}, \tag{55}$$

with some $C \in \mathscr{Z}$ and a Borel set B in $\mathscr{B}([0, 1])$. Now consider the $(M, \bar{x})$-section of such a set A:

$$\begin{aligned}\mathrm{sec}_{M,\bar{x}}(A) &= \{x \in \mathsf{X}_{N\setminus M} : (x, \bar{x}) \in A\}\\ &= \{x \in \mathsf{X}_{N\setminus M} : \nu(x, \bar{x}; C) \in B\}\\ &= \{x \in \mathsf{X}_{N\setminus M} : \nu_{M,\bar{x}}(x; C) \in B\}.\end{aligned}$$

This shows that the sections $\mathrm{sec}_{M,\bar{x}}(A)$ of measurable sets A of the form (55) generate $\sigma_{M,\bar{x}}(\nu)$, which proves the first equality. The second equality is a direct implication of the first one. □

The results of the previous section, Theorem 6 and Proposition 7, apply to the information flows, defined for the projective families $(\overline{\mathscr{A}}_M^\nu)_{M\subseteq N}$ and $(\underline{\mathscr{A}}_M^\nu)_{M\subseteq N}$. These families take into account the information that is actually used by the channel ν. Therefore, we can interpret the corresponding marginal channels $\hat{\nu}_M$ causally, where we have to distinguish

two kinds of causality. For the projective family $(\overline{\mathscr{A}}^{\nu}_M)_{M\subseteq N}$, the channel $\hat{\nu}_M$ incorporates the information in any input configuration x_K, $K \subseteq M$, that is used by ν in conjunction with a context configuration $\bar{x}_{N\setminus K} = \bar{x}_{N\setminus M}\bar{x}_{M\setminus K}$ outside of K. For the projective family $(\underline{\mathscr{A}}^{\nu}_M)_{M\subseteq N}$, on the other hand, the channel $\hat{\nu}_M$ incorporates the information used by ν that is solely contained in x_M, independent of any context. When comparing a marginal channel $\hat{\nu}_M$ with another marginal channel $\hat{\nu}_L$, where $L \subseteq M$, the corresponding information flows $\overline{I}_{\xi}(X_{M\setminus L} \to Z|X_L)$ and $\underline{I}_{\xi}(X_{M\setminus L} \to Z|X_L)$, respectively, quantify the causal effects in $\hat{\nu}_M$ that exceed those in $\hat{\nu}_L$. These measures will capture different causal aspects, where the difference can be large. This is illustrated by the following extension of Example 3.

Example 11 Let

$$(\mathsf{X}_i, \mathscr{X}_i) = (\mathbb{R}, \mathscr{B}(\mathbb{R})), \qquad i \in \{1, \ldots, n\} = N,$$

where $\mathscr{B}(\mathbb{R})$ denotes the Borel σ-algebra of $\mathbb{R}$. We define the channel simply by the sum of the input states, interpreted as a Markov kernel,

$$\nu(x_1, \ldots, x_n; C) := \mathbb{1}_C(x_1 + \cdots + x_n).$$

As $\mathscr{B}(\mathbb{R})$ is generated by the intervals $[r - \varepsilon, r + \varepsilon] \subseteq \mathbb{R}$, the smallest σ-algebra $\mathscr{A}^{\nu}$ for which all functions $\nu(\cdot; C)$ are measurable is generated by the following sets

$$A(r, \varepsilon) := \{(x_1, \ldots, x_n) \in \mathbb{R}^n : r - \varepsilon \le x_1 + \cdots + x_n \le r + \varepsilon\}, \quad r \in \mathbb{R},\ \varepsilon \in \mathbb{R}_+.$$

For a set $M \subseteq N$ and a context configuration $\bar{x} = (\bar{x}_i)_{i\in N\setminus M} \in \mathbb{R}^{N\setminus M}$, the $(M, \bar{x})$-section of $A_{r,\varepsilon}$ is given by

$$\mathrm{sec}_{M,\bar{x}}(A(r, \varepsilon)) = \left\{ x = (x_i)_{i\in M} \in \mathbb{R}^M : r - \sum_{i\in N\setminus M} \bar{x}_i - \varepsilon \le \sum_{i\in M} x_i \le r - \sum_{i\in N\setminus M} \bar{x}_i + \varepsilon \right\}.$$

Therefore, the M-trace of $\mathscr{A}^{\nu}$, $\mathscr{A}^{\nu}_M$, is generated by the halfspaces

$$H_{\vartheta} := \left\{ x = (x_i)_{i\in M} \in \mathbb{R}^M : \sum_{i\in M} x_i \le \vartheta \right\}, \qquad \vartheta \in \mathbb{R}.$$

For $|M| = 1$, we recover the half lines, so that $\mathscr{A}^{\nu}_{\{i\}} = \mathscr{B}(\mathbb{R})$. The projective extension then leads to the largest σ-algebra, the Borel algebra of $\mathbb{R}^M$:

$$\overline{\mathscr{A}}^{\nu}_M = \mathscr{B}(\mathbb{R}^M).$$

Therefore, the marginal channel $\hat{\nu}_M(x; C)$ equals the usual marginal $\nu_M(x; C)$ for the projective extension. For the projective reduction, on the other hand, we obtain the trivial σ-algebra except for $M = N$:

$$\underline{\mathscr{A}}^{\nu}_M = \begin{cases} \{\emptyset, \mathbb{R}^M\} & \text{if } M \subsetneq N, \\ \mathscr{A}^{\nu} & \text{if } M = N. \end{cases}$$

In this case we have $\hat{\nu}_M(x; C) = \nu_*(\mu)(C)$ for $M \neq N$ and $\hat{\nu}_N(x; C) = \nu(x; C)$, where μ is the joint distribution of the input variables.

We now consider the information flows associated with $L \subsetneq M \subseteq N$, for the projective extension as well as for the projective reduction. In both cases these flows coincide with usual (conditional) mutual informations, in an instructive way. More precisely, for the extension we have

$$\overline{I}_{\xi}(X_{M\setminus L} \to Z|X_L) = I_{\xi}(X_{M\setminus L}; Z|X_L).$$

For the reduction, we obtain

$$\underline{I}_\xi(X_{M\setminus L} \to Z|X_L) = \begin{cases} 0 & \text{if } M \subsetneq N, \\ I_\xi(X_N; Z) & \text{if } M = N. \end{cases} \tag{56}$$

Interestingly, (56) does not depend on L. The vanishing of the information flow for $M \neq N$ is due to the fact that the output of the channel, the sum $x_1 + \cdots + x_n$, cannot be computed from a proper subset of the inputs. The flow of information only takes place if all inputs are given.

The following example, which is closely related to Example 1, highlights continuity issues of the introduced measures of information flow. They result from the fact that small changes in the mechanisms can lead to large differences of the involved σ-algebras.

Example 12 Consider two input variables, X and Y, and one output variable Z, with corresponding state spaces

$$(\mathsf{X}, \mathscr{X}) = (\mathsf{Y}, \mathscr{Y}) = (\mathsf{Z}, \mathscr{Z}) = (\mathbb{R}, \mathscr{B}(\mathbb{R})).$$

Furthermore, consider two channels, ν and ν', where ν simply copies the second input, y, and ν' the first one, x. More precisely,

$$\begin{aligned} \nu(x, y; C) &:= \mathbb{1}_C(y), \\ \nu'(x, y; C) &:= \mathbb{1}_C(x). \end{aligned}$$

With $0 \leq \varepsilon \leq 1$, we define the convex combination

$$\nu^{(\varepsilon)} := (1 - \varepsilon)\nu + \varepsilon\nu'.$$

For $\varepsilon = 0$, only the channel ν is acting. Obviously, we have $\underline{I}_\xi(X \to Z) = \overline{I}_\xi(X \to Z) = 0$, as expected in this case, because ν simply copies y and is not sensitive to x at all. One might intuitively expect that the causal information flow from X to Z stays close to 0 if ε is small but greater than 0. However, this intuition is not reflected by the actual quantities, as defined in this article. It is easy to see that for $\varepsilon \neq 0$, $\underline{I}_\xi(X \to Z) = \overline{I}_\xi(X \to Z) = I_\xi(X; Z)$. Thus, these quantities are not at all sensitive to the parameter ε and behave discontinuously in the limit $\varepsilon \to 0$.

6 Conclusions

Conditioning is an important operation within the study of causality. The theory of causal networks, pioneered by Pearl [17], introduces interventional conditioning as an operation, the so-called *do-operation*, that is fundamentally different from the classical conditioning based on the general rule $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$. It models more appropriately experimental setups and avoids confusion with purely associational dependencies. Information theory has been classically used for the quantification of such dependencies, in terms of mutual information and conditional mutual information [20]. Within the original setting of information theory, the mutual information between the input and the output of a channel can be interpreted causally. In the more general context of causal networks, however, confounding effects make a distinction between associations and causal effects more difficult. In such cases, information-theoretic quantities can be misleading as measures of causal effects. In order to overcome this problem, information theory has been coupled with the

interventional calculus of causal networks, and corresponding measures of causal information flow have been proposed [5, 6]. Given that such measures are based on the notion of an experimental intervention, which represents a perturbation of the system, it remains unclear to what extent they quantify causal information flows in the unperturbed system. As another consequence of the interventional conditioning, one cannot expect that causal information flow, as defined in [6], decomposes according to a chain rule. The current article is based on an idea of the author from 2003 which precedes the above-mentioned works on combining the theory of causal networks with information theory. It proposes a way to quantify causal information flows without perturbing the system through intervention. Instead, it is based on classical conditioning in terms of the conditional distribution $\mathbb{P}(B|\mathscr{A})$, where the σ-algebra $\mathscr{A}$ is adjusted to the intrinsic mechanisms of the system. The derived information flow measure satisfies the chain rule and the natural properties of a general measure of causal strength postulated in [14]. The chain rule, together with the generalised Pythagoras relation from information geometry, provide powerful tools within the study of the problem of partial information decomposition [7, 8, 16].

Even though the introduced information flows satisfy natural properties, the aim of the present article is relatively moderate. For instance, the analysis is focussed on a simple network consisting of a number of inputs and one output, which is a strong restriction compared to the setting of [6]. The extension of the present work to more general casual networks remains to be worked out. Furthermore, this article does not address the important problem of causal inference [18]. In addition to these general directions of research, there are various ways to modify and extend the constructions of the present work and thereby potentially highlight further causal aspects of a given channel. The following perspectives are particularly important:

1. In the present article, the information flow has been defined for a fixed finite measurable partition ξ of the state space $(\mathsf{Z}, \mathscr{Z})$ of the output variable Z. A natural further step would be to consider the limit of information flows with respect to an increasing sequence ξ_n, $n = 1, 2, \ldots$, so that
$$\bigvee_{n=1}^{\infty} \sigma(\xi_n) = \mathscr{Z}.$$
This limit will be an information flow measure that is independent of a particular partition.
2. Throughout this article, the partition ξ has not been coupled with the σ-algebra of the channel ν. This is the smallest σ-algebra for which all functions $\nu(x; C)$, $C \in \mathscr{Z}$, are measurable. Given that the channel is analysed with respect to the partition ξ, one can restrict attention to the smallest σ-algebra for which the functions $\nu(x; C)$, $C \in \xi$, are measurable. This will be a potentially small σ-subalgebra of the one generated by the channel. We would then have a natural coupling of the partition ξ with the information used by the channel.
3. We started with the family $\mathscr{A}_M^\nu$ of M-traces of $\mathscr{A}^\nu$, the σ-algebra generated by ν, as the natural family associated with the channel. However, these traces do not form a projective family of σ-algebras. Such a projectivity is required for the chain rule for corresponding information flows. One can recover projectivity by extension and by reduction, leading to $\overline{\mathscr{A}}_M^\nu$ and $\underline{\mathscr{A}}_M^\nu$, respectively. Example 11 shows that the extension can lead to the largest σ-algebra and the reduction to the trivial one. Given this fact, one might ask whether the extension is too large and the reduction is too small to capture the causal aspects of ν. Even though we argued above that these two projective

families associated with ν capture two different kinds of causal aspects, this question remains to be further pursued. One possible direction would be the analysis of the context-dependent traces of $\mathscr{A}^\nu$, that is the family of $\mathrm{tr}_{M,\bar{x}}(\mathscr{A}^\nu)$, $\bar{x} \in \mathsf{X}_{N\setminus M}$. Instead of conditioning with respect to the join

$$\mathrm{tr}_M(\mathscr{A}^\nu) = \bigvee_{\bar{x}\in\mathsf{X}_{N\setminus M}} \mathrm{tr}_{M,\bar{x}}(\mathscr{A}^\nu),$$

one could adjust the conditioning to the individual σ-algebras $\mathrm{tr}_{M,\bar{x}}(\mathscr{A}^\nu)$. This would represent an important refinement of the presented theory.

Acknowledgements The author is grateful for valuable comments of two anonymous reviewers. He acknowledges the support of the Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134).

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Amari, S.-i.: Information Geometry and its Applications. Applied Mathematical Sciences, vol. 194. Springer, Tokyo (2016)
2. Amari, S.-i., Nagaoka, H.: Methods of Information Geometry. Translations of Mathematical Monographs, vol. 191. Oxford University Press, Oxford (2000)
3. Ay, N., Amari, S.-i.: A novel approach to canonical divergences within information geometry. Entropy **17**, 8111–8129 (2015)
4. Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information Geometry. A Series of Modern Surveys in Mathematics, vol. 64. Springer International Publishing, New York (2017)
5. Ay, N., Krakauer, D.C.: Geometric robustness theory and biological networks. Theory Biosci. **125**, 93–121 (2007)
6. Ay, N., Polani, D.: Information flows in causal networks. Adv. Complex Syst. **11**, 17–41 (2008)
7. Ay, N., Polani, D., Virgo, N.: Information decomposition based on cooperative game theory. Kybernetika **56**, 979–1014 (2020)
8. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., Ay, N.: Quantifying unique information. Entropy **16**, 2161–2183 (2014)
9. Bossomaier, T., Barnett, L., Harré, M., Lizier, J.T.: An Introduction to Transfer Entropy. Springer International Publishing (2016)
10. Cover, T.M., Thomas, J.A. Elements of Information Theory, 2nd edn. Wiley-Interscience, Hoboken (2006)
11. Dudley, R.M.: Real Analysis and Probability, 2nd edn. Cambridge Studies in Advanced Mathematics, vol. 74. Cambridge University Press, Cambridge (2002)
12. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**, 424–438 (1969)
13. Granger, C.W.J.: Testing for causality: A personal viewpoint. J. Econ. Dyn. Control **2**, 329–352 (1980)
14. Janzing, D., Balduzzi, D., Grosse-Wentrup, M., Schölkopf, B.: Quantifying causal influences. Ann. Stat. **41**, 2324–2358 (2013)
15. Kakihara, Y.: Abstract Methods in Information Theory. Multivariate Analysis, vol. 4. World Scientific, New Jersey (1999)

16. Lizier, J., Bertschinger, N., Jost, J., Wibral, M.: Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work. Entropy **20**, 307 (2018)
17. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
18. Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge (2017)
19. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. **85**, 461–464 (2000)
20. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 623–656 (1948)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.