Formally Justifying MDL-based Inference of Cause and Effect

Alexander Marx,[1] Jilles Vreeken,[2]

[1] ETH Zurich and ETH AI Center
[2] CISPA Helmholtz Center for Information Security
alexander.marx@ai.ethz.ch, jv@cispa.de

Abstract

The algorithmic independence of conditionals, which postulates that the causal mechanism is algorithmically independent of the cause, has recently inspired many highly successful approaches to distinguish cause from effect given only observational data. Most popular among these is the idea to approximate algorithmic independence via two-part Minimum Description Length (MDL). Although intuitively sensible, the link between the original postulate and practical two-part MDL encodings has so far been left vague. In this work, we close this gap by deriving a two-part formulation of this postulate, in terms of Kolmogorov complexity, which directly links to practical MDL encodings. To close the cycle, we prove that this formulation leads on expectation to the same inference result as the original postulate.

1 Introduction

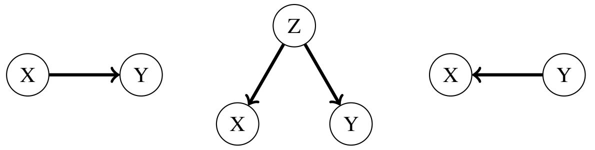
Recovering causal networks from observational data is a challenging, yet important problem in science. Traditional methods are only able to infer the true causal graph up to its Markov equivalence class, i.e., recover the correct undirected network and some of the edge directions (Spirtes et al. 2000). Inferring the network structure beyond such a Markov equivalence class boils down to identifying the causal direction in a bi-variate setting (Peters, Janzing, and Schölkopf 2017). That is, given two dependent random variables X and Y, we need to distinguish between the two Markov equivalent graphs $X \rightarrow Y$ an $X \leftarrow Y$. Recent results show that it is possible solve this task and thus infer all edge directions, if we are willing to make assumptions about the causal mechanism (Peters, Janzing, and Schölkopf 2017; Shimizu et al. 2006). One such assumption is the principle of independent mechanisms, which postulates that the causal mechanism (i.e. the conditional distribution $P_{\text{effect}|\text{cause}}$) is independent of the distribution of the cause P_{cause} (Janzing et al. 2012; Sgouritsa et al. 2015). Vice versa, it is unlikely to find such an independence for the anti-causal direction. In practice, it is difficult to measure the independence between mechanism and cause directly. Thus, many recent approaches build upon an algorithmic version of this principle, the algorithmic independence of conditionals postulate (Janzing and Schölkopf 2010). This postulate states that if $X \rightarrow Y$ is the true graph, the Kolmogorov complexity of the factorization of P_{XY} for the causal model, i.e., $K(P_X) + K(P_{Y|X})$, is shorter than for the anti-causal model $K(P_Y) + K(P_{X|Y})$. If Kolmogorov complexity was computable, we could infer the causal direction between a pair X and Y by comparing those quantities.

Since Kolmogorov complexity is not computable, practical instantiations of the above postulate, build upon two-part Minimum Description Length (MDL) (Rissanen 1978). Methods based on this idea are among the state-of-the-art for cause-effect inference on discrete (Budhathoki and Vreeken 2017a,b), continuous (Kalainathan 2019; Marx and Vreeken 2019; Mitrovic, Sejdinovic, and Teh 2018; Tagasovska, Chavez-Demoulin, and Vatter 2020) and mixed-type data (Marx and Vreeken 2018). Despite the success of two-part MDL approaches in causal inference, the link between the theory of algorithmic independence and practical encodings is crude at best. While the postulate, formulated in terms of Kolmogorov complexity, only compares the complexities of the factorizations of the true distribution, e.g., $K(P_X) + K(P_{Y|X})$, the MDL formulations consider the data with respect to a model under consideration and the model itself. Slightly abusing the notation, we show that those methods rather approximate

$$K(P_X) + K(x \mid P_x) + K(P_{Y|X}) + K(y \mid x, P_{Y|X}) \qquad (1)$$

instead of $K(P_X) + K(P_{Y|X})$; and similarly so for the inverse direction. Simply put, an explicit compression of the data x and y does not appear in the postulate itself, but only in the MDL approximation. Naively including the data breaks the asymmetry between the causal and anti-causal direction since information is symmetric. That is, given the shortest programs x^* and y^* to compute x resp. y, $K(x) + K(y \mid x^*) = K(y) + K(x \mid y^*)$ holds up to an additive constant. Hence, it is crucial to analyze the link between both concepts more thoroughly, which is what we do in this paper.

In particular, starting from the MDL framework, we first derive a two-part formulation of the algorithmic independence of conditionals in terms of Kolmogorov complexity (building upon Eq. 1). Then, we prove that our new formulation leads to an equivalent inference principle as the original postulate, which closes the cycle to connect both ideas. As a corollary, we investigate the implications of our findings for joint encodings, which encode data and model jointly. We



(a) (b) (c)

Figure 1: Illustration of Reichenbach's Common Cause Principle: X and Y are two dependent random variables. Their dependence is either due to an unobserved confounder Z (middle), or one of the two causes the other, i.e., either X causes Y (left) or Y causes X (right).

emphasize that for such encodings is important to encode the model independent of the data, as otherwise the asymmetry between the description length of the causal and anti-causal model might vanish.

This paper is organized as follows. We discuss the principle of independent mechanisms in Sec. 2 and state the algorithmic independence of conditionals postulate in Sec. 3. Then we review two-part MDL approximations of this postulate in Sec. 4 and provide our main result, the link between both inference criteria in Sec. 5. In Sec. 6 we discuss practical implications of our findings for joint descriptions and in Sec. 7, we round up with a conclusion.

2 The Principle of Independent Mechanisms

Before we introduce the algorithmic independence of conditionals postulate, we first discuss its statistical equivalent, the principle of independent mechanisms.

To illustrate the problem, we will first consider the smallest working example. That is, assume we are given two dependent random variables X and Y for which we want to infer the underlying causal graph from observational data. That is, we assume the data is passively collected and represents an i.i.d. sample of joint distribution P_{XY}. According to *Reichenbach's common cause principle* (Reichenbach 1956), the dependence between X and Y can be explained by three possible graphs.

Definition 1 (Reichenbach's Common Cause Principle) *If two random variables X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then there exists a third variable Z that causally influences both, that is $X \leftarrow Z \rightarrow Y$. As a special case, Z may coincide with either X or Y, which results in the causal graphs $X \rightarrow Y$ resp. $X \leftarrow Y$. See Figure 1 for illustration.*

It could also be that the dependence between X and Y is due to a combination of the above graphs, e.g. $X \rightarrow Y$, $X \leftarrow Z \rightarrow Y$. In this paper, however, we assume all causal relations to be acyclic and further assume *causal sufficiency*, i.e. all relevant variables are observed. Further, we assume that there is no selection bias. Thus, we only need to decide between $X \rightarrow Y$ and $X \leftarrow Y$, which is already a difficult problem since these DAGs are Markov equivalent. Hence, it is not possible to tell apart both graphs if we rely on a constraint-based causal discovery approach (Pearl 2009).

This is, however, not the end of the story. We can distinguish $X \rightarrow Y$ from $X \leftarrow Y$ if we additionally make assumptions about the generating mechanism. A very general such assumption, which has gained a lot of attention in recent years, is the *principle of independent mechanisms*, which focuses on the possible factorizations of the joint distribution P_{XY}. In particular, we can write P_{XY} as the product $P_X P_{Y|X}$ or $P_Y P_{X|Y}$; but why does this help? Consider an example inspired by (Peters, Janzing, and Schölkopf 2017, Ch. 2.1), in which the cause X corresponds to the altitude and the effect Y to the temperature as measured for different cities. Assume we consider a set of different cities from the same climate zone. If we observe the altitude for a random city, we will have a mechanism in mind to derive the corresponding temperature value (i.e. $P_{Y|X}$), which is independent of P_X. Further, we can make a thought experiment and think about how the temperature would change, if we were to change the altitude of the city, e.g., by magically lifting it into the air. Vice versa, it is hard to imagine that increasing or decreasing the temperature in a city, e.g., by putting on the heating system in every house, will change the altitude. In other words, the independence of the mechanism does not hold for the anti-causal direction: the mechanism $P_{X|Y}$ would need to take a rather particular form to be independent of P_Y, which only holds in specific settings, such as a linear model with both the cause and additive noise being Gaussian distributed (Peters, Janzing, and Schölkopf 2017). More generally, we can formulate the principle of independent mechanisms as follows (Janzing and Schölkopf 2010; Peters, Janzing, and Schölkopf 2017).

Postulate 1 (Principle of Independent Mechanisms) *The causal generative process of a systems variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the conditional distributions of other variables.*

Projected on our two-variable example, if we assume that the principle of independent mechanisms holds, we get that $P_X \perp\!\!\!\perp P_{Y|X}$, while the same does not necessarily hold for the factorization w.r.t. the anti-causal direction. There exist several approaches that aim at using this asymmetry to infer the causal direction between two random variables from observational data (Janzing et al. 2012; Sgouritsa et al. 2015); in practice, however, it is difficult to precisely estimate this dependence. Therefore, we focus on the information theoretic variant of this principle (Janzing and Schölkopf 2010), which initiated the development of MDL-based estimators, that are among the state-of-the-art in the field (Kaltenpoth and Vreeken 2019; Marx and Vreeken 2019; Mitrovic, Sejdinovic, and Teh 2018; Tagasovska, Chavez-Demoulin, and Vatter 2020).

3 The Algorithmic Model of Causality

In the following, we first introduce the algorithmic model of causality and the algorithmic independence of conditionals. After that, we state the commonly used two-part MDL approximation (Budhathoki and Vreeken 2017b).

Since the algorithmic model of causality is defined through Kolmogorov complexity, we will first provide a brief introduction to that topic. Intuitively, the Kolmogorov complexity of a finite binary string x is the length of the shortest program that can output x and then halt. The idea is that the more structure of a string can be expressed algorithmically as opposed to just printing its verbatim, the smaller is its Kolmogorov complexity.

Formally, we will in this paper refer to prefix Kolmogorov complexity (Kolmogorov 1965; Li and Vitányi 2019).

Definition 2 (Kolmogorov Complexity) *The prefix Kolmogorov complexity of a finite binary string x is the length of the shortest self-delimiting binary program p^* for a universal prefix Turing machine $\mathcal{U}$ that generates x, and then halts, i.e.,*

$$K(x) = \min\{|p| \mid p \in \{0,1\}^*, \mathcal{U}(p) = x\} .$$

To define conditional Kolmogorov complexity of a binary string x given string y, we provide y as input to the program that computes x for free, that is

$$K(x \mid y) = \min\{|q| \mid q \in \{0,1\}^*, \mathcal{U}(y, q) = x\} .$$

Finally, building upon the above definitions, we can define the algorithmic equivalent to mutual information, which we need below. For two binary strings x and y, *algorithmic mutual information* (Chaitin 1975) is defined as

$$I_A(x; y) := K(x) + K(y) - K(x, y) .$$

Simply put, algorithmic mutual information is greater than zero if we can extract more structure by jointly compressing x and y than with two individual programs. Equivalently, we can define I_A as $K(x) + K(y \mid x^*)$, which holds up to an additive constant, which we denote by $\stackrel{+}{=}$. Similarly, we could have inequalities which hold up to an additive constant, which we write as $\stackrel{+}{\leq}$. The x^* in the conditional refers to the shortest program that describes x. Note that if we would instead use x in the conditional, the equality would only hold up to a logarithmic constant dependent on x—i.e. a constant in $O(\log K(x))$, which breaks the symmetry of the formulation (Li and Vitányi 2019).

Now that we discussed the preliminary concepts, we can state the *algorithmic model of causality* (AMC) (Janzing and Schölkopf 2010), which was proposed as an algorithmic equivalent of the statistical model of causality. Simply put, we can compute the value of X with a program of constant complexity that gets as input the data over the parents of X in the corresponding causal graph, and data w.r.t. an independent noise term.

Postulate 2 (Algorithmic Model of Causality) *Let G be a DAG formalizing the causal structure among the strings $x_1, \ldots, x_m$. Then every x_i is computed by a program q_i with length $O(1)$ from its parents pa_i and an additional input n_i. We write*

$$x_i = q_i(pa_i, n_i) ,$$

meaning that the Turing machine computes x_i from the input pa_i, n_i using the additional program q_i and halts. The inputs n_i are jointly independent, i.e.,

$$n_i \perp\!\!\!\perp n_1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_m .$$

Janzing and Schölkopf (2010) justified this model by showing that similar to the statistical model, we can also derive an algorithmic version of the causal Markov condition. That is, the *algorithmic Markov condition* (AMC) states that the joint complexity over all nodes is given by the sum of the complexities of each individual node given the optimal compression of its parents

$$K(x_1, \ldots, x_m) \stackrel{+}{=} \sum_{i=1}^{m} K(x_i \mid pa_i^*) . \tag{2}$$

Due to the symmetry of information, i.e.,

$$K(x) + K(y \mid x^*) \stackrel{+}{=} K(y) + K(x \mid y^*) ,$$

the algorithmic Markov condition only allows for identifying the Markov equivalence class. To be able to distinguish between Markov equivalence classes, Janzing and Schölkopf (2010) further postulated the algorithmic equivalent of the principle of independent mechanisms.

Postulate 3 (Algorithmic Independence of Conditionals) *Let G be a causal DAG over a set of m variables $\mathbf{X} = (X_1, \ldots, X_m)$ with joint distribution $P_{\mathbf{X}}$, which is lower semi-computable, that is, $K(P_{\mathbf{X}}) < \infty$. A causal hypothesis is only acceptable if the shortest description of the joint distribution $P_{\mathbf{X}}$ is given by the concatenation of the shortest descriptions of the Markov kernels, i.e.,*

$$K(P_{X_1,\ldots,X_m}) \stackrel{+}{=} \sum_{i=1}^{m} K(P_{X_i|Pa_i}) ,$$

where Pa_i are the parents of X_i in G. Equivalently,

$$I_A(P_{X_1|Pa_1}; \ldots; P_{X_m|Pa_m}) \stackrel{+}{=} 0 .$$

If we apply the above postulate to the case where the true graph only consists of the edge $X \to Y$, we get that

$$I_A(P_X; P_{Y|X}) \stackrel{+}{=} 0 . \tag{3}$$

Note that it is assumed that this independence only holds for the true causal direction. For additive noise models, for example, it has been shown that for the anti-causal direction we get a significant dependence (Janzing and Steudel 2010), that is, $I_A(P_Y; P_{X|Y}) \gg 0$. If we combine those results, we can derive an inference rule as follows. If $X \to Y$ is the true graph, then

$$K(P_X) + K(P_{Y|X}) \stackrel{+}{\leq} K(P_Y) + K(P_{X|Y}) . \tag{4}$$

In other words, we can infer the true causal direction by selecting the factorization with the smallest Kolmogorov complexity. To use this idea in practice, we first need to solve two problems. First, Kolmogorov complexity is not computable (Li and Vitányi 2019), and second, we are not given the true distribution but just a limited number of data points. A principled way to solve at least the first part of the problem is to approximate Kolmogorov complexity via the Minimum Description Length principle, which we explain below.

4 MDL as a Practical Solution

The Minimum Description Length (MDL) principle (Grünwald 2007; Rissanen 1978) is a practical variant of Kolmogorov Complexity, which approximates K from above. Instead of considering all programs, we restrict ourselves to a certain model class $\mathcal{M}$, for example, a certain class of parametric probability distributions. Given data D, which may represent a sample from a distribution P_X, our goal is to find that model $M^* \in \mathcal{M}$, such that

$$M^* = \operatorname*{argmin}_{M \in \mathcal{M}} L(D \mid M) + L(M) \ , \tag{5}$$

where $L(M)$ is the length in bits needed to describe the model M or identify M within the model class $\mathcal{M}$, and $L(D \mid M)$ is the length in bits of the description of data D given M. This specific version of MDL is also referred to as two-part or crude MDL. As an example, consider that $\mathcal{M}$ is the model class which refers to a multinomial distribution with k categories. In this case M refers to a specific k-dimensional parameter vector $\hat{\theta}$. Accordingly, $L(M)$ measures the complexity of identifying $M \in \mathcal{M}$ or describing the parameter vector $\hat{\theta}$, and $L(D \mid M)$ refers to the negative log-likelihood of the data D, under the assumption that D follows a multinomial distribution with parameter vector $\hat{\theta}$. That is, we encode a single data point x as $-\log P(x \mid \hat{\theta})$. Note that by taking the negative logarithm with log base 2, we arrive at a code length in bits. For more details to MDL, we refer to Grünwald (2007).

The first idea on how the algorithmic independence of conditionals could lead to an MDL-based inference rule was sketched out by Janzing and Schölkopf (2010), however, they do neither instantiate nor evaluate this idea in practice. They suggest that, the probabilistic models $\hat{P}_X$ and $\hat{P}_{Y|X}$, which are learned from a finite number of observations, together define a joint distribution $\hat{P}_{X\to Y}$, which is not necessarily equal to the description of $\hat{P}_{Y\to X}$ in the inverse direction. As common in MDL, they first encode the complexity of the model, i.e., $\hat{P}_X$ and $\hat{P}_{Y|X}$, and then encode the data given the model as the negative log-likelihood w.r.t. $\hat{P}_{X\to Y}$ resp. $\hat{P}_{Y\to X}$ and select the direction with the smaller complexity as the causal one.

Budhathoki and Vreeken (2017b) suggested a more practical approximation of Eq. 4 via two-part MDL as follows. For the causal direction, we define a model as $M_{X\to Y} = (M_X, M_{Y|X})$ from the class $\mathcal{M}_{X\to Y} = \mathcal{M}_X \times \mathcal{M}_{Y|X}$ that best describes the data over Y by exploiting as much structure of X as possible to save bits. By MDL, we identify the optimal model $M_{X\to Y} \in \mathcal{M}_{X\to Y}$ for data (x^n, y^n) over X and Y as the one minimizing

$$L_{X\to Y} := L(M_X) + L(x^n \mid M_X) + L(M_{Y|X}) + L(y^n | x^n, M_{Y|X}) \ . \tag{6}$$

We can define $L_{Y\to X}$ analogously and infer $X \to Y$ if $L_{X\to Y} < L_{Y\to X}$, $X \leftarrow Y$ if $L_{X\to Y} > L_{Y\to X}$, and do not decide if both terms are equal. Consequently, to use this idea in practice, we need to define the model class. Budhathoki and Vreeken (2017b) implemented their idea for multivariate binary data and used binary trees as their models. Following this example, two-part MDL approaches have been developed for univariate discrete pairs (Budhathoki and Vreeken 2017a; SY and Nagaraj 2020), univariate continuous random variables (Kalainathan 2019; Marx and Vreeken 2017, 2019; Mitrovic, Sejdinovic, and Teh 2018; Tagasovska, Chavez-Demoulin, and Vatter 2020) and multivariate mixed-type data (Marx and Vreeken 2018). Further, Kaltenpoth and Vreeken (2019) adapted this idea to tell whether two random variables are causally related or whether they are likely to be confounded and Mian et al. (2021) build upon a two-part MDL score to discover the complete causal DAG G between a set of random variables.[1]

Although these approaches perform well in practice, Eq. 4 only considers the true distribution, while Eq. 6 is formulated via a two-part description of a model and the data given this model. In the following section, we will present our main result and formally analyze the connection between both inference rules. We bridge the gap between both variants by deriving a two-part variant of Eq. 4, in terms of Kolmogorov complexity, and show that on expectation both versions lead to the same inference.

5 Linking Algorithmic Independence and Two-Part Descriptions

Given an i.i.d. sample x^n drawn from a computable distribution P, the shortest encoding of the data that is theoretically possible converges to the Shannon entropy

$$H(P) = -\sum_x P(x) \log P(x) \ ,$$

as proven by Shannon's source coding theorem (Shannon 1948). Hence, if P is a computable distribution with parameter vector θ, the sample estimate $\hat{\theta}$ will in the limit converge to the true parameter. Therefore, we could in the limit encode the data x^n conditional on P to arrive at the shortest code-length of the data given the model that describes P. Thus, the shortest encoding for our causal setup can be achieved if the model class $\mathcal{M}_X$ contains P_X and similarly, $\mathcal{M}_{Y|X}$ contains $P_{Y|X}$. Slightly abusing the notation, we define

$$L^*_{X\to Y} := L(P_X) + L(x^n \mid P_X) + L(P_{Y|X}) + L(y^n \mid x^n, P_{Y|X}) \ . \tag{7}$$

The above equation already comes close to an MDL version of the algorithmic independence of conditionals, however, we still need to explain how the data encoded by the model fits into the equation. To this end, we will show that the equivalent formulation of $L^*_{X\to Y}$ in terms of Kolmogorov complexity, i.e.,

$$K_{X\to Y} := K(P_X) + K(x \mid P_X) + K(P_{Y|X}) + K(y \mid x, P_{Y|X}) \ ,$$

is on expectation equal to $K(P_X) + K(P_{Y|X}) + H(P_{XY})$, where $H(P_{XY})$ relates to the Shannon entropy of the joint distribution P_{XY}. The analogoue holds for the anti-causal direction, that is, on expectation $K_{Y\to X}$ is equal to $K(P_Y) + K(P_{X|Y}) + H(P_{XY})$. Thus, assuming that the algorithmic independence

[1] There also exists approaches for continuous i.i.d. data (Mooij et al. 2010) and time series data (Hlaváčková-Schindler and Plant 2020) based on the Minimum Message Length (Wallace and Boulton 1968), however, these do not directly build upon Eq. 4.

of conditionals holds, we get that on expectation the inequality between cause and effect holds similar to Eq. 4. That is,

$$K_{X \to Y} \stackrel{+}{\leq} K_{Y \to X} \ ,$$

if $X \to Y$ is the true causal direction.

Before we prove this statement, we need to introduce a more general Lemma that links Kolmogorov complexity to Shannon entropy (Li and Vitányi 2019, Ch. 8.1).

Lemma 1 *Let $H(P)$ be the entropy of a computable probability distribution P and $H(P) < \infty$. Then,*

$$\left| \sum_x P(x) K(x \mid P) - H(P) \right| \leq O(1), \tag{8}$$

with a constant precision that is independent of x and P.

Note that it is important to sum over $P(x)K(x \mid P)$ and not over $P(x)K(x)$, as otherwise the inequality becomes less precise and only holds up to a logarithmic constant $c_P = K(P) + O(1)$, which is dependent on P (Li and Vitányi 2019, Ch. 8.1). For conditional codes such as $K(y \mid x, P_{Y|X})$ assume that given input x there exists an $O(1)$ program that selects the correct probability table $P_{Y|X=x}$ from the auxiliary conditional probability table that is given as input. Based on these insights, we can derive of our main theorem.

Theorem 1 *Given distribution P_{XY} with finite support and rational values, for which all factorizations are lower semi-computable, i.e., $K(P_X) + K(P_{Y|X}) + K(P_Y) + K(P_{X|Y}) < \infty$, it holds that*

$$\sum_x \sum_y P_{XY}(x, y) \left(K(P_X) + K(x \mid P_X) + K(P_{Y|X}) + K(y \mid x, P_{Y|X}) \right)$$

is equal to $K(P_X) + K(P_{Y|X}) + H(P_{XY})$ up to an additive constant that is independent of P_X and $P_{Y|X}$. Equivalently, $K(P_Y) + K(P_{X|Y}) + H(P_{XY})$ is equal to the expectation over $K(P_Y) + K(y \mid P_Y) + K(P_{X|Y}) + K(x \mid y, P_{X|Y})$ up to an additive constant independent of P_Y and $P_{X|Y}$.

Proof: *In the following, we prove the statement for the factorization $P_X P_{Y|X}$; the proof for the factorization $P_Y P_{X|Y}$ follows analogously. First, note that we can compute $P_X(x)$ as $P_X(x) = \sum_y P_{XY}(x, y)$. Thus, we can rewrite the first part as*

$$(\star_1) = \sum_x \sum_y P_{XY}(x, y) \left(K(P_X) + K(x \mid P_X) \right) \tag{9}$$

$$= \sum_x P_X(x) \left(K(P_X) + K(x \mid P_X) \right) \tag{10}$$

$$\stackrel{+}{=} K(P_X) + H(P_X) \ . \tag{11}$$

To get from line 10 to 11, we apply Lemma 1. Similarly, we can proceed with the second part

$$(\star_2) = \sum_x \sum_y P_{XY}(x, y) \left(K(P_{Y|X}) + K(y \mid x, P_{Y|X}) \right) \tag{12}$$

$$= \sum_x P_X(x) \sum_y \frac{P_{XY}(x, y)}{P_X(x)} \left(K(P_{Y|X}) + K(y \mid x, P_{Y|X}) \right) \tag{13}$$

$$= K(P_{Y|X}) + \sum_x P_X(x) \sum_y \frac{P_{XY}(x, y)}{P_X(x)} K(y \mid x, P_{Y|X}) \tag{14}$$

$$\stackrel{+}{=} K(P_{Y|X}) + \sum_x P_X(x) \sum_y P_{Y|X=x}(y) K(y \mid P_{Y|X=x}) \tag{15}$$

$$\stackrel{+}{=} K(P_{Y|X}) + \sum_x P_X(x) H(P_{Y|X=x}) \tag{16}$$

$$= K(P_{Y|X}) + H(P_{Y|X}) \ . \tag{17}$$

Importantly, in the step from line 14 to 15, we assume that we can select the correct probability table $P_{Y|X=x}$ form inputs $P_{Y|X}$ and x with an $O(1)$ program.

If we combine $(\star_1)$ and $(\star_2)$, we get $K(P_X) + K(P_{Y|X}) + H(P_{XY})$ and obtain an equivalent result for the inverse direction due to the symmetry of the joint entropy. □

Although the theorem is only stated for two random variables, it is straight forward to extend it to the general formulation of the algorithmic independence of conditionals. In particular, we have

$$\sum_{i=1}^{m} \sum_{x_i} \sum_{pa_i} P_{X_i \mathrm{Pa}_i}(x_i, pa_i) \left(K(P_{X_i|\mathrm{Pa}_i}) + K(x_i \mid pa_i, P_{X_i|\mathrm{Pa}_i}) \right)$$

is equal to $\sum_{i=1}^{m} K(P_{X_i|\mathrm{Pa}_i}) + H(P_{X_1,\ldots,X_m})$ up to a constant.

This concludes the main contribution of this paper. In the following section, we point out that the results of Theorem 1 do not necessarily hold for joint descriptions and discuss implications of these observations for MDL encodings.

6 Connection to Joint Descriptions

The optimization goal of a two-part encoding, e.g., two-part MDL, is often formalized as finding that model $M^* \in \mathcal{M}$, which mimimizes the *joint* costs of data and model, that is,

$$M^* = \operatorname*{argmin}_{M \in \mathcal{M}} L(D, M) \ .$$

Hence, intuitively we could rewrite $L_{X \to Y}$ as $L(x^n, M_X) + L(y^n, M_{Y|X} \mid x^n)$. The problem is, as follows. If we rigorously expand the second term, we need to encode the model $M_{Y|X}$ given the data x^n, i.e.,

$$L(y^n, M_{Y|X} \mid x^n) = L(M_{Y|X} \mid x^n) + L(y^n \mid x^n, M_{Y|X}) \ .$$

In terms of MDL, we can argue that the model is independent of the data x^n. In general, while technically possible, it is not common to encode a model conditioned on the data. Thus, we do not elaborate further on this ambiguity and jump into Kolmogorov land.

In particular, assume that $X \to Y$ is the true causal model. If we were to prove that $K(x, P_X) + K(y, P_{Y|X} \mid x)$ is on average equal to $K(P_X) + K(P_{Y|X}) + H(P_{XY})$, the proof would become slightly more involved. It is inevitable, however, that to split off $P_{Y|X}$ from $K(y, P_{Y|X} \mid x)$ we need to keep x in the conditional term. That is, we arrive at the term $K(P_{Y|X} \mid x)$ and would need to argue that it is equal to $K(P_{Y|X})$. Since $P_X \perp\!\!\!\perp P_{Y|X}$ and x is sampled from P_X, we can indeed conclude that $K(P_{Y|X} \mid x) = K(P_{Y|X}) + O(1)$. For the anti-causal direction, this independence does not hold, i.e., $P_Y \not\perp\!\!\!\perp P_{X|Y}$. Hence, $K(P_{X|Y} \mid y) \overset{+}{\leq} K(P_{X|Y})$. Due to this asymmetry, we get that

$$\sum_x \sum_y P_{XY}(x, y)(K(y, P_Y) + K(x, P_{X|Y} \mid y)) \tag{18}$$

$$\overset{+}{\leq} K(P_Y) + K(P_{X|Y}) + H(P_{XY}) \tag{19}$$

$$\overset{+}{\geq} K(P_X) + K(P_{Y|X}) + H(P_{XY}) \; . \tag{20}$$

In other words, an approximation of this formulation does not allow us to distinguish $X \to Y$ from $Y \to X$, because we cannot guarantee that the inequality between the causal and anti-causal direction still holds. Thus, encodings that approximate the joint description with the goal to do causal inference should be designed with caution, as the description length of the conditional model should be independent of the data of the conditioning variable.

7 Conclusion

In this paper, we focused on causal discovery from observational data. We revisited the idea of using two-part MDL encodings to approximate the postulate of algorithmic independence of conditionals, which allows us to infer causal DAGs beyond Markov equivalence classes. Our main contribution is that we link this postulate to an alternative formulation, which considers both the complexity of the data conditioned on the distribution as well as the complexity of the distribution. Thus we arrive at a two-part description in terms of Kolmogorov complexity $K_{X \to Y}$, which has a one-to-one mapping to the proposed two-part MDL approximation $L_{X \to Y}$ (Budhathoki and Vreeken 2017b). In addition, we prove that approximating $K_{X \to Y}$ is equivalent to approximating the algorithmic independence of conditionals, which closes the cycle. We further analyze the implications of our new formulation for joint descriptions of data and model and conclude that it is important that the model for the conditional distributions, e.g., $M_{Y|X}$ is independent of the data x^n, resp. that $M_{X|Y}$ is independent of y^n.

In a nutshell, drawing this connection is crucial to understand how two-part MDL approaches can be used to approximate the algorithmic independence of conditionals. We expect that these insights make MDL-based methods for causal inference more accessible to a broader audience.

Acknowledgment

The authors would like to thank Bruno Bauwens and David Kaltenpoth for insightful discussions.

References

Budhathoki, K.; and Vreeken, J. 2017a. MDL for Causal Inference on Discrete Data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 751–756. IEEE.

Budhathoki, K.; and Vreeken, J. 2017b. Origo: causal inference by compression. *Knowledge and Information Systems*, 1–23.

Chaitin, G. J. 1975. A Theory of Program Size Formally Identical to Information Theory. *Journal of the ACM*, 22(3): 329–340.

Grünwald, P. 2007. *The Minimum Description Length Principle*. MIT Press.

Hlaváčková-Schindler, K.; and Plant, C. 2020. Heterogeneous Graphical Granger Causality by Minimum Message Length. *Entropy*, 22(12): 1400.

Janzing, D.; Mooij, J.; Zhang, K.; Lemeire, J.; Zscheischler, J.; Daniušis, P.; Steudel, B.; and Schölkopf, B. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183: 1–31.

Janzing, D.; and Schölkopf, B. 2010. Causal Inference Using the Algorithmic Markov Condition. *IEEE Transactions on Information Technology*, 56(10): 5168–5194.

Janzing, D.; and Steudel, B. 2010. Justifying Additive Noise Model-Based Causal Discovery via Algorithmic Information Theory. *Open Systems and Information Dynamics*, 17(2): 189–212.

Kalainathan, D. 2019. *Generative Neural Networks to infer Causal Mechanisms: algorithms and applications*. Ph.D. thesis, Université Paris-Saclay.

Kaltenpoth, D.; and Vreeken, J. 2019. We are not your real parents: Telling causal from confounded using mdl. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 199–207. SIAM.

Kolmogorov, A. 1965. Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii*, 1(1): 3–11.

Li, M.; and Vitányi, P. 2019. *An Introduction to Kolmogorov Complexity and its Applications*, volume 4. Springer.

Marx, A.; and Vreeken, J. 2017. Telling Cause from Effect using MDL-based Local and Global Regression. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 307–316. IEEE.

Marx, A.; and Vreeken, J. 2018. Causal Inference on Multivariate and Mixed-Type Data. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 655–671. IEEE, Springer.

Marx, A.; and Vreeken, J. 2019. Identifiability of Cause and Effect using Regularized Regression. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM.

Mian, O.; Marx, A.; and Vreeken, J. 2021. Discovering Fully Oriented Causal Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

Mitrovic, J.; Sejdinovic, D.; and Teh, Y. W. 2018. Causal inference via kernel deviance measures. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 6986–6994.

Mooij, J.; Stegle, O.; Janzing, D.; Zhang, K.; and Schölkopf, B. 2010. Probabilistic latent variable models for distinguishing between cause and effect. *Advances in Neural Information Processing Systems*, 1687–1695.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. New York, NY, USA: Cambridge University Press, 2nd edition.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Reichenbach, H. 1956. *The direction of time*, volume 65. Univ of California Press.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica*, 14(1): 465–471.

Sgouritsa, E.; Janzing, D.; Hennig, P.; and Schölkopf, B. 2015. Inference of Cause and Effect with Unsupervised Inverse Regression. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 38: 847–855.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7: 2003–2030.

Spirtes, P.; Glymour, C. N.; Scheines, R.; Heckerman, D.; Meek, C.; Cooper, G.; and Richardson, T. 2000. *Causation, prediction, and search*. MIT press.

SY, P.; and Nagaraj, N. 2020. Causal Discovery using Compression-Complexity Measures. *arXiv preprint arXiv:2010.09336*.

Tagasovska, N.; Chavez-Demoulin, V.; and Vatter, T. 2020. Distinguishing Cause from Effect Using Quantiles: Bivariate Quantile Causal Discovery. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, 9311–9323. PMLR.

Wallace, C. S.; and Boulton, D. M. 1968. An information measure for classification. *The Computer Journal*, 11(1): 185–194.