Justifying additive-noise-model based causal discovery via algorithmic information theory

Dominik Janzing[1] and Bastian Steudel[2]

1) Max Planck Institute for Biological Cybernetics
Tübingen, Germany

2) Max Planck Institute for Mathematics in the Sciences
Leipzig, Germany

April 29, 2010

Abstract

A recent method for causal discovery is in many cases able to infer whether X causes Y or Y causes X for just two observed variables X and Y. It is based on the observation that there exist (non-Gaussian) joint distributions $P(X, Y)$ for which Y may be written as a function of X up to an additive noise term that is independent of X and no such model exists from Y to X. Whenever this is the case, one prefers the causal model $X \to Y$.

Here we justify this method by showing that the causal hypothesis $Y \to X$ is unlikely because it requires a specific tuning between $P(Y)$ and $P(X|Y)$ to generate a distribution that admits an additive noise model from X to Y. To quantify the amount of tuning needed we derive lower bounds on the *algorithmic* information shared by $P(Y)$ and $P(X|Y)$. This way, our justification is consistent with recent approaches for using algorithmic information theory for causal reasoning. We extend this principle to the case where $P(X, Y)$ *almost* admits an additive noise model.

Our results suggest that the above conclusion is more reliable if the complexity of $P(Y)$ is high.

1 Additive noise models in causal discovery

Causal inference from statistical data is a field of research that obtained increasing interest in recent years. To infer causal relations among several random variables by purely observing their joint distribution is unsolvable from the point of view of traditional statistics. During the 90s, however, it was more and more believed that also puerly observational data contain at least *hints* on the causal directions. The most important postulate that links the observed statistical dependencies on the one hand to the causal structure (which is here assumed to be a DAG, i.e., a directed acyclic graph) on the other hand is the causal Markov condition [16]. It states that every variable is conditionally independent of its non-effects, given its causes. If the joint distribution $P(X_1, \ldots, X_n)$ has a density $p(x_1, \ldots, x_n)$ with respect to some product measure, then the density factorizes [9] into

$$p(x_1, \ldots, x_n) = \prod_{j=1}^{n} p(x_j|pa_j)\,,$$

where $p(x_j|pa_j)$ denotes the conditional probability density of X_j, given the values pa_j of its parents PA_j.

The Markov condition already rules out some DAGs as being incompatible with the observed conditional dependencies. However, often a large set of DAGs still is compatible. In particular, for n variables, there are $n!$ DAGs that are consistent with every joint distribution because they do not impose any conditional independence. They are given by defining an order $X_1, \ldots, X_n$ and drawing an error from $X_i \to X_j$ for every $i < j$. For this reason, additional inference rules are required to choose the most plausible ones among the compatible DAGs. Spirtes at al. [20] and Pearl [16] use the causal faithfulness principle that prefers those DAGs for which the causal Markov condition imposes all the observed *in*dependencies. In other words, it is considered unlikely that independencies are due to particular (non-generic) choices of the conditionals $p(x_j|pa_j)$. The underlying idea is, so to speak, that "nature chooses" the conditionals independently from each other, while the generation of additional independencies (that are not imposed by the structure of the DAG) would require to mutually adjust these conditionals. A more general perspective on such an independence assumption has been provided by Lemeire and Dirkx [10]. Following their idea we postulate:

Postulate 1 (Algorithmic independence of conditionals)**.**
If the true causal structure is given by the directed acyclic graph G with random variables $X_1, \ldots, X_n$ as nodes, the shortest description of the joint density $p(x_1, \ldots, x_n)$ is given by separate descriptions of the conditionals[1] *$p(x_j|pa_j)$.*

In [10] the description length has been defined in terms of algorithmic information, also called "Kolmogorov complexity" (the details will be explained in Section 2). In [10] the postulate is mainly used to justify the causal faithfulness assumption [20], since it rules out mutual adjustments among conditionals like those required for unfaithful distributions. However, in [7] it has been argued that the complete determination of the joint distribution is never feasible which makes it hard to give empirical content to it. Moreover, [7] shows that Lemeire and Dirkx's principle can be seen as an implication of a general framework for causal inference via algorithmic information. There, the postulate is rephrased in a way that avoids the complexity of conditionals and uses only empirical observations. Furthermore, the general framework may impose many causal inference rules yet to be discovered. Here we focus on a method [6, 14, 13] that yielded quite encouraging results on real data sets and show that it also can be justified via algorithmic information theory. We briefly rephrase the idea of [6] for the special case of two real-valued variables X and Y. To this end we introduce the following terminology:

Definition 1 (Additive noise model)**.**
The joint density $p(x, y)$ of two real-valued random variables X and Y is said to admit an additive noise model from X to Y if there is a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that

$$Y = f(X) + E\,, \tag{1}$$

where E is some unobserved noise variable that is statistically independent of X. The joint density thus is of the form

$$p(x, y) = p_X(x) p_E(y - f(x))\,,$$

where $p_X(x)$ is the density of X and $p_E(e)$ the density of E.

Whenever this causes no confusion, we will drop the indices and write $p(x)$ instead of $p_X(x)$ and, similarly, write $p(y - f(x))$. We will write p_X if we want to emphasize that we refer to the entire density and not one specific value $p(x)$.

[1] For sake of simple terminology, we also consider the density $p(x_j)$ of parentless nodes as a "conditional", given an empty set of variables.

It can be shown [6] that for generic choices of f, distribution of the noise, and distribution of X, there is no additive noise model from Y to X. In other words, if causality in nature would always be of the form of additive noise models (which is certainly not the case[2]), we could almost always identify causal directions because a joint distribution that admits an additive noise model in the true direction usually does not admit one in the wrong direction. This paper addresses the question whether a causal structure $Y \to X$ that is *not* of the form of an additive noise model could induce a joint distribution that admits an additive noise model in the wrong direction (i.e., from X to Y). The basic observation of this paper is that this would be a rare coincidence because it requires that p_Y (which would be the distribution of the cause) and the transition probabilities $p_{X|Y}$ (which would describe the mechanism generating the relation between cause and effect) satisfy an untypical relation that makes this scenario unlikely. However, instead of deriving probability values for such a coincidence (which required to assign priors to probability distributions) we will take a non-Bayesian view and follow the algorithmic information theory approach developed in [7] and [10]. The following lemma makes explicit what kind of coincidence is meant:

Lemma 1 (Relation between p_Y and $p_{X|Y}$)**.**
Let $p(x,y)$ be positive definite and let f as well as all logarithms of marginal and conditional densities be two times differentiable. If $p(x,y)$ admits an additive noise model from X to Y, then the marginal $p(y)$ and the conditional $p(x|y)$ are related via the differential equation

$$\frac{\partial^2}{\partial y^2}\log p(y) = -\frac{\partial^2}{\partial y^2}\log p(x|y) - \frac{1}{f'(x)}\frac{\partial^2}{\partial x \partial y}\log p(x|y)\,. \tag{2}$$

Proof: Applying $\partial^2/\partial y^2$ and $\partial^2/(\partial x \partial y)$ to the identity

$$\log p(y) = -\log p(x|y) + \log p(y|x) + \log p(x)$$

yields

$$\begin{aligned} \frac{\partial^2}{\partial y^2}\log p(y) &= -\frac{\partial^2}{\partial y^2}\log p(x|y) + \frac{\partial^2}{\partial y^2}\log p(y|x) &(3)\\ 0 &= -\frac{\partial^2}{\partial x \partial y}\log p(x|y) + \frac{\partial^2}{\partial x \partial y}\log p(y|x)\,, &(4)\end{aligned}$$

respectively. Using

$$\begin{aligned} \frac{\partial^2}{\partial y^2}\log p(y|x) &= \frac{\partial^2}{\partial y^2}\log p_E(y - f(x))\\ &= -\frac{1}{f'(x)}\frac{\partial^2}{\partial x \partial y}\log p_E(y - f(x))\\ &= -\frac{1}{f'(x)}\frac{\partial^2}{\partial x \partial y}\log p(y|x)\end{aligned}$$

and eq. (4) we can replace the second term in eq. (3) and convert it into the desired eq. (2). □

Eq. (2) implies

$$\log p(y) = -\int_0^y \int_0^{y''} \frac{\partial^2}{\partial y^2}\log p(x|y') - \frac{1}{f'(x)}\frac{\partial^2}{\partial x \partial y}\log p(x|y')dy'dy'' + ay + b\,,$$

[2]For instance, [21] discusses an interesting generalization.

where b is determined by $\int p(y)dy = 1$. Since the equation has to be valid for all x, we can choose an arbitrary x_0 with $f'(x_0) \neq 0$. Then p_Y can already be determined from $f'(x_0)$, the function $y \mapsto p(x_0|y)$ and a. Given the conditional $p_{X|Y}$, the tupel $(x_0, f'(x_0))$ and a are sufficient to describe the marginal p_Y. In general, these are much fewer parameters than those required for describing p_Y without knowing $p_{X|Y}$. This already suggests that p_Y and $p_{X|Y}$ have algorithmic information in common because knowing $p_{X|Y}$ shortens the description of p_Y.

The paper is structured as follows. In Section 2 we briefly rephrase algorithmic information theory based causal inference as developed in [7]. In Section 3 we show that additive noise models from X to Y induce densities p_Y and $p_{X|Y}$ that have algorithmic information in common. In Section 4 we consider additive noise models over finite fields and show that p_Y and $p_{X|Y}$ also share algorithmic information if the distribution is only *close to* an additive noise model from X to Y. Since our bounds on the information shared by these objects depend on the Kolmogorov complexity of p_Y (which cannot be determined) we discuss a method to estimate the latter in Section 5. Section 6 and Section 7 discuss how to apply the insights gained from the discrete case to empirical and to continuous distributions respectively.

2 Algorithmic information theory and the causal principle

Reichenbach's Principle of Common Cause [17] is meanwhile the cornerstone of causal reasoning from statistical data: Every statistical dependence between two random variables X and Y indicates at least one of the three causal relations (1) "X causes Y", (2) "Y causes X", or (3) there is a common cause Z influencing both X and Y. As an extension of this principle, we have argued [7] that causal inference is not always based on *statistical* dependencies. Instead, similarities between single objects also indicate causal links: if, for instance, two T-shirts produced by different companies have the same sophisticated pattern we would not believe that the designer came up with the patterns independently. Since this conclusion only requires one copy of each T-shirt, the statistical sample size is one.

We have therefore postulated the "causal principle" stating that there is a causal link between two objects (without referring to any statistical sampling) whenever the joint description of them is shorter than the concatenation of their separate descriptions To formalize this, we first introduce some concepts of algorithmic information theory [8, 19, 2, 3, 12]. Let s, t be two binary strings that describe the observed objects and let $K(s)$ denote the algorithmic information (or "Kolmogorov complexity"), i.e., the length of the shortest program that generates s on a prefix free universal Turing machine [8, 18, 3, 2]. Let $K(s|t)$ denote the length of the shortest program that generates s from the input t. Then we define [5]:

Definition 2 (algorithmic mutual information)**.**
Let s, t be two binary strings. Then the algorithmic mutual information between s and t reads

$$I(s:t) := K(t) - K(t|s^*) \stackrel{+}{=} K(s) + K(t) - K(s,t)\,, \tag{5}$$

where s^* denotes the shortest program that computes s and $K(s,t)$ is the length of the shortest program generating the concatenation of s and t.

As usual in algorithmic information theory, all (in)equalities are only understood up to an additive constant that depends on the Turing machine [12]. For this reason, we write $\stackrel{+}{=}$ instead of $=$. Since s can be computed from s^*, but usually not vice versa, we

have

$$K(t|s^*) \stackrel{+}{\leq} K(t|s)\,. \tag{6}$$

We will later also need the conditional version of (5), see [5]:

Definition 3 (conditional algorithmic mutual information)**.**
Let s, t, v be binary strings. Then the conditional algorithmic mutual information reads

$$I(s:t|v) := K(t|v) - K(t|s, K(s|v), v) \stackrel{+}{=} K(s|v) + K(t|v) - K(s,t|v)\,. \tag{7}$$

Eq. (5) is formally similar to the statistical mutual information [4]

$$I(X;Y) := H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)\,,$$

phrased in terms of the Shannon entropy $H(\cdot)$. Reichenbach's principle can then be rephrased as:

> "$I(X;Y) > 0$ indicates that there is at least one of the three possible causal links between X and Y."

In analogy to this principle, we have postulated in [7]:

Postulate 2 (Causal Principle)**.**
Let s and t be binary strings that formalize the descriptions of two objects in nature. Whenever

$$I(s:t) \gg 0\,,$$

there is a causal link between the two objects s and t in the sense that $s \to t$ or $t \to s$ or there is a third object u with $s \leftarrow u \to t$.

Here, it is up to the researcher's decision how to set the threshold above which a dependence is considered significant. This is similar to setting the significance value in a statistical test.

Note that the condition $K(t) - K(t|s) \gg 0$ implies $I(s:t) \gg 0$ due to ineq. (6). We will use the former condition since it is easier to work with. To interpret Postulate 1 as a special case of Postulate 2, we consider the following model [7] of a causal structure $X \to Y$ for two random variables X and Y. Let the two objects be (1) a source S that generates x-values according to $p(x)$ and (2) a machine M that takes x-values as input and generates y-values according to $p(y|x)$ (see Figure 1).

If S and M have been designed independently, their optimal joint description should be given by separate descriptions of S and M. However, the only feature of S that is relevant for our observations is given by the distribution of x-values, i.e., p_X. Similarly, $p_{Y|X}$ is the only relevant feature of M. These features are directly given by observing the x and y-values infinitely often. We therefore consider the algorithmic dependencies between p_X and $p_{Y|X}$. Since the objects of our descriptions will be probability distributions, we introduce the following concept:

Definition 4 (computable functions and distributions)**.**
Let $\mathcal{S}$ denote some subset of $\mathbb{R}^k$. A function $f : \mathcal{S} \to \mathbb{R}$ is computable if there is a program that computes $f(x)$ up to a precision $\epsilon > 0$ for every input (x, ϵ) having a finite description. Then $K(f)$ denotes the length of the shortest program of this kind. A probability distribution on a finite probability space $\mathcal{S}$ is called computable if its density is a computable function.

In the following section we apply the concepts introduced above to the case of strictly positive continuous densities $p(x, y)$.

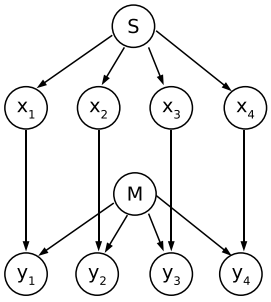

Figure 1: Causal structure obtained by resolving the causal structure $X \to Y$ between the random variables X and Y into causal relations among single events

3 Algorithmic dependencies induced by additive noise models

We have already argued that an additive noise model from X to Y makes the causal structure $Y \to X$ unlikely because p_Y and $p_{X|Y}$ then satisfy the non-generic relation of eq. (2). We now express this fact in terms of algorithmic information theory:

Theorem 1 (algorithmic dependence induced by an additive noise model)**.**
Let $p(x, y)$ be a two-times differentiable computable strictly positive probability density over $\mathbb{R}^2$. If $p(x, y)$ admits an additive noise model from X to Y with a computable differentiable non-constant function f, then

$$I(p_Y : p_{X|Y}) \stackrel{+}{\geq} K(p_Y) - K(y_0, (\log p)'(y_0)) - K(x_0, f'(x_0)) \tag{8}$$

where x_0 and y_0 are computable x- and y-values, respectively, where y_0 is arbitrary and x_0 satisfies $f'(x_0) \neq 0$.

Proof: Eq. (2) expresses the second derivative $(\log p_Y)''$ in terms of $p_{X|Y}$ and $f'(x_0)$. Hence,

$$K((\log p_Y)''|p_{X|Y}) \stackrel{+}{\leq} K(x_0, f'(x_0))\,. \tag{9}$$

We have by definition and due to ineq. (6)

$$I(p_Y : p_{X|Y}) \stackrel{+}{=} K(p_Y) - K(p_Y|p^*_{X|Y}) \stackrel{+}{\geq} K(p_Y) - K(p_Y|p_{X|Y})\,. \tag{10}$$

The density p_Y is already determined by $(\log p_Y)''$ and $((\log p)'(y_0), y_0)$ for some y_0 because $\log p_Y(y_0)$ then follows from normalization. Therefore,

$$K(p_Y|z) \stackrel{+}{\leq} K((\log p_Y)''|z) + K(y_0, (\log p)'(y_0)|z)\,,$$

where z is some arbitrary prior information. Using $z = p_{X|Y}$, the right hand term of ineq. (10) yields

$$\begin{aligned} I(p_Y : p_{X|Y}) &\stackrel{+}{\geq} K(p_Y) - K((\log p_Y)''|p_{X|Y}) - K(y_0, (\log p)'(y_0)|p_{X|Y}) \\ &\stackrel{+}{\geq} K(p_Y) - K(x_0, f'(x_0)|p_{X|Y}) - K(y_0, (\log p)'(y_0)|p_{X|Y}) \\ &\stackrel{+}{\geq} K(p_Y) - K(x_0, f'(x_0)) - K(y_0, (\log p)'(y_0))\,, \end{aligned}$$

where the second inequality is due to ineq. (9). □

Roughly speaking, Theorem 1 states that $p_{X|Y}$ helps in describing p_Y unless the additional information given by the second and third term on the rhs. of ineq. (8) exceeds the complexity of p_Y. We now want to discuss one of the rare examples where the complexity of p_Y indeed is low enough to make ineq. (8) void. Assume we are given the information that p_{XY} belongs to the family of bivariate Gaussians. Then its description reduces to the following five parameters: μ_X and μ_Y denoting the expectations of X and Y, respectivley; σ_X, σ_Y and c_{XY} denoting standard deviation of X and Y and the correlation between them, respectively. One can also choose an asymmetric parametrization of the form $\mu_X, \sigma_X, c, \mu_E, \sigma_E$, where the first two parameters correspond to p_X and the last three describe $p_{Y|X}$ via the linear addtive noise model

$$Y = cX + E\,,$$

where μ_E and σ_E denote the expectation and width of the independent additive noise E. Similarly, one can parameterize p_{XY} via describing p_Y and $p_{X|Y}$, because p_{XY} admits an additive noise model in both directions and both causal directions are possible. This is consistent with the fact that our argument above fails for the bivariate Gaussian case because a and $f'(x_0)$ then coincides with the information that also would be required to describe p_Y *without* knowing $p_{X|Y}$. To see this, set

$$\log p(x) \stackrel{+}{=} \frac{(x-\mu_X)^2}{2\sigma_X^2}\,,$$

where $\stackrel{+}{=}$ denotes here equality up to a term that neither depends on x nor on y (by slightly overloading notation). Furthermore, let

$$\log p(y|x) \stackrel{+}{=} \frac{(y-cx-\mu_E)^2}{2\sigma_E^2}\,,$$

where $c = f'(x_0)$. We then get

$$\log p(y) \stackrel{+}{=} \frac{(y-\mu_X-\mu_E)^2}{2(c^2\sigma_X^2+\sigma_E^2)}.$$

Hence,

$$\log p(x|y) \stackrel{+}{=} \frac{(x-\mu)^2}{2\sigma_X^2} + \frac{(y-cx)^2}{2\sigma_E^2} - \frac{(y-\mu_X-\mu_E)^2}{2(c^2\sigma_X^2+\sigma_E^2)}\,,$$

which implies

$$\frac{\partial^2}{\partial x \partial y}\log p(x|y) \stackrel{+}{=} -\frac{c}{\sigma_E^2} =: \alpha\,,$$

and

$$\frac{\partial^2}{\partial y^2}\log p(x|y) \stackrel{+}{=} -\frac{1}{\sigma_E^2} =: \beta\,.$$

The constants α and β can be derived from observing $p_{X|Y}$, but to determine the second derivative of $\log p_Y$ one needs to know c since eq. (2) imposes

$$\frac{\partial^2}{\partial y^2}\log p(y) = \beta - \frac{1}{c}\alpha\,. \tag{11}$$

To determine p_Y completely, we also need to know the first derivative

$$a := \frac{\partial}{\partial y}\log p(y=0) = -\frac{\mu_Y}{\sigma_Y^2}\,.$$

Moreover, we observe that c specifies the standard deviation σ_Y of Y because the left hand side of eq. (11) is given by $-1/\sigma_Y^2$. This shows, that given $p_{X|Y}$, we still need to describe the two parameters μ_Y and σ_Y. These are exactly the two parameters that describe the Gaussian p_Y also *without* knowing $p_{X|Y}$. Hence, knowing $p_{X|Y}$ is worthless for the description of p_Y.

The arguments above show that knowing $p_{X|Y}$ makes the description of p_Y shorter except for some rare cases where p_Y already has a short description. The interpretation of Theorem 1 raises two problems: First, we cannot determine the exact "true" probabilities[3] from the observations, and second, we do not expect these probabilities to be computable, and hence it may require an infinite amount of information to describe p_Y and $p_{X|Y}$ if we could. As already pointed out in [7], algorithmic dependencies among the *empirical* distributions q_Y and $q_{X|Y}$ after finite sampling do not show algorithmic dependencies between S and M. For continuous variables, this is already obvious from the fact that the conditional distribution of X, given Y, is only defined for the support of q_Y. If the true distribution is a density, the empirical distribution contains every y-value only once and knowing the support of q_Y thus already implies knowing q_Y.

To circumvent this problem, we will in the following section consider additive noise models over a finite probability space. Within this setting, we derive statements on distributions that are *close to* additive noise models. Since the finite case has the advantage that empirical frequencies converge pointwise to the true probabilities, this result also implies statements for the corresponding empirical distribution.

4 Stronger statements in finite probability spaces

The following theorem is a modification of Theorem 1 for additive noise models over the finite field $\mathbb{Z}_m$ for some prime number m.

Theorem 2 (Algorithmic information between p_Y and $p_{X|Y}$ for the discrete model)**.**
Let $p_{X,Y}$ be a computable strictly positive distribution on $\mathbb{Z}_m^2$ for some prime number m that admits an additive noise model, i.e., there is a function $f : \mathbb{Z}_m \to \mathbb{Z}_m$ such that $E := Y - f(X)$ and X are statistically independent. Here, subtraction is understood with respect to $\mathbb{Z}_m$. Then, if f is non-constant, we have

$$I(p_Y : p_{X|Y}) \stackrel{+}{\geq} K(p_Y) - 2\log m\,. \tag{12}$$

Proof: The idea is, again, to derive an equation that shows that p_Y is essentially determined by $p_{X|Y}$ up to some small amount of additional information. We have

$$\log p(x,y) = \log p_X(x) + \log p_E(y - f(x))\,. \tag{13}$$

Defining $\delta := f(x_0) - f(x_0 - 1)$, for some x_0 for which $\delta \neq 0$, we calculate

$$\begin{aligned}
\frac{p(y+\delta)}{p(y)} &\stackrel{(a)}{=} \frac{p(x_0|y)}{p(x_0|y+\delta)}\,\frac{p(x_0, y+\delta)}{p(x_0,y)} \\
&\stackrel{(b)}{=} \frac{p(x_0|y)}{p(x_0|y+\delta)}\,\frac{p(x_0-1, y)}{p(x_0-1, y-\delta)} \\
&\stackrel{(c)}{=} \frac{p(x_0|y)}{p(x_0|y+\delta)}\,\frac{p(x_0-1|y)}{p(x_0-1|y-\delta)}\,\frac{p(y)}{p(y-\delta)},
\end{aligned}$$

where (a) and (c) are just applications of Bayes' rule and (b) follows from relation (13). Taking logarithms and introducing

$$k_{(x|y)} := \log p(x|y) - \log p(x|y+\delta) + \log p(x-1|y) - \log p(x-1|y-\delta)\,, \tag{14}$$

[3]It is, anyway, a philosophical problem to what extent they are well-defined.

yields the equation

$$\log p(y+\delta) - \log p(y) = k_{(x_0|y)} + \log p(y) - \log p(y-\delta). \tag{15}$$

We interpret eq. (15) as a discrete version of eq. (2) because it relates differences between the values $\log p(y)$ at different points y to the quantity $k_{(x|y)}$, which is a property of the conditional $p_{X|Y}$ alone. Eq. (15) implies for arbitrary y_0

$$\log p(y_0+(j+1)\delta) - \log p(y_0+j\delta) = \log p(y_0+j\delta) - \log p(y_0+(j-1)\delta) + k_{(x_0|y+j\delta)}\,,$$

for all $j = 1, \ldots, m$. Writing $\log p_Y$ for the vector with coefficients $\log p(y_0+(j+1)\delta)$ and k for the vector with coefficients $k_{(x_0|y+j\delta)}$ for $j = 0, \ldots, m-1$, we rewrite eq. (15) as

$$(S-I)^2 \log p_Y = k\,,$$

where S denotes the cyclic shift in dimension m. Using the fact that $(S-I)$ is invertible on the space of vectors with zero sum of coefficients, we thus obtain

$$\log p_Y = (S-I)^{-2} k + \alpha\, \mathbf{e}\,, \tag{16}$$

where α is given by normalization and $\mathbf{e}$ is the vector with only ones as entries. This shows that x_0, δ, and $p_{X|Y}$ determine p_Y. Denoting $i := (x_0, \delta)$ we can summarize the above into $K(p_Y|p_{X|Y}, i) \stackrel{+}{=} 0$. This implies

$$K(p_Y|p_{X|Y}) \stackrel{+}{\leq} K(i)\,,$$

because

$$\begin{aligned} K(p_Y|p_{X|Y}) - K(p_Y|p_{X|Y}, i) &\stackrel{+}{=} K(p_Y|p_{X|Y}) - K(p_Y|p_{X|Y}, K(i|p_{X|Y}), i) \\ &\stackrel{+}{=} I(p_Y : i|p_{X|Y}) \stackrel{+}{\leq} K(i)\,, \end{aligned}$$

where the second equality is due to the definition of conditional algorithmic mutual information (7). Then the statement follows because the algorithmic information of i is bounded from above by two times the number of bits of m. □

We want to derive a similar lower bound for the case where p_{XY} *almost* admits an additive noise model. To this end, we first define a precision dependent Kolmogorov complexity of a probability distribution:

Definition 5 (Precision dependent algorithmic information)**.**
Let p be a density on finite probability space. Let r be a computable probability density and $K(r)$ be the length of the shortest program on a universal Turing machine that computes $r(x)$ from x. Then

$$K_\epsilon(p) := \min_{r \ \textit{with}\ D(p||r)<\epsilon} K(r|\epsilon)\,,$$

where $D(\cdot||\cdot)$ denotes the relative entropy distance. Similarly, we define the conditional complexity $K_\epsilon(p|i)$ given some prior information i.

If q is an arbitrary approximation of a distribution p in the sense that $|\log p(x) - \log q(x)| \leq \epsilon$ holds for all x, then $D(p||q) \leq \epsilon$ and thus the precision dependent algorithmic information can be bounded from above by the complexity of the approximation: $K_\epsilon(p) \leq K(q)$. For computable p, we obviously have

$$\lim_{\epsilon \to 0} K_\epsilon(p) = K(p)\,,$$

but for uncomputable p, the complexity tends to infinity. The following lemma shows the empirical content of precision-dependent complexity:

Lemma 2 (precision-dependent complexity of empirical distributions)**.**
Let p be a positive definite distribution on a finite probability space and $q^{(n)}$ be the empirical distribution after n-fold sampling from p. Then

$$\lim_{n\to\infty} K_\epsilon(q^{(n)}) = K_\epsilon(p)\,,$$

with probability one.

Proof: Let r be a distribution for which $K_\epsilon(p) = K(r)$ and $D(p||r) < \epsilon$. Due to $D(q^{(n)}||r) \to D(p||r)$ with probability one and because of the continuity of relative entropy for positive definite distributions we also have $D(q^{(n)}||r) < \epsilon$ for all sufficiently large n. Hence $K_\epsilon(q^{(n)}) \leq K_\epsilon(p)$.

To prove that $K_\epsilon(q^{(n)}) \geq K_\epsilon(p)$, let $r^{(n)}$ be a sequence of distributions such that $K_\epsilon(q^{(n)}) = K(r^{(n)})$ and $D(q^{(n)}||r^{(n)}) < \epsilon$. Hence, $D(p||r^{(n)}) < \epsilon$ for sufficiently large n which completes the proof.□

The following lemma will later be used to derive a lower bound on $I(p_Y : p_{X|Y})$ in terms of $K_\epsilon(p_Y)$:

Lemma 3 (mutual information and approximative descriptions)**.**
Let p be a computable distribution on a finite probability space, z an arbitrary string and $\epsilon > 0$ computable. Let q be a distribution that is ϵ-close to p, i.e.,

$$D(p||q) < \epsilon\,. \tag{17}$$

If q can be derived from z and from p in the sense that

$$K(q|p, i_p) \stackrel{+}{=} K(q|z, i_z) \stackrel{+}{=} 0\,, \tag{18}$$

for additional strings i_p and i_z, then

$$I(p : z) \stackrel{+}{\geq} K_\epsilon(p) - K(i_p) - K(i_z).$$

Proof: Using the definition of conditional mutual information (7) we get

$$I(q : i_p|p^*) \quad \stackrel{+}{=} \quad K(q|p^*) - K(q|i_p, K(i_p|p^*), p^*) \stackrel{+}{=} K(q|p^*),$$

because Eq. (18) implies $K(q|i_p, K(i_p|p^*), p^*) \stackrel{+}{=} 0$. On the other hand $I(q : i_p|p^*) \stackrel{+}{\leq} K(i_p)$ and therefore

$$K(q|p^*) \stackrel{+}{\leq} K(i_p).$$

In the same way, eq. (18) implies $K(q|z^*) \stackrel{+}{\leq} K(i_z)$. A data processing inequality (Theorem II.7 in [5]) then implies

$$I(p : z) \stackrel{+}{\geq} I(q : q) - K(i_p) - K(i_z)\,.$$

We conclude with $I(q : q) = K(q) \stackrel{+}{\geq} K_\epsilon(p)$ due to ineq. (17). □

We will moreover need the following Lemma:

Lemma 4 (bound on the differences of logarithms)**.**
Given a vector $v \in \mathbb{R}^m$, we define a probability distribution by

$$p_j := \frac{1}{z_v} e^{-v_j}\,,$$

where z_v is the partition function. Let $\tilde{p}$ be defined by $\tilde{v}$ in the same way. Then

$$|\log p_j - \log \tilde{p}_j| \leq 2\|v - \tilde{v}\|_\infty\,.$$

Proof: Due to

$$\log p_j - \log \tilde{p}_j = v_j - \tilde{v}_j - \log z_v + \log z_{\tilde{v}}$$

we only have to show

$$|\log z_v - \log z_{\tilde{v}}| \leq \|v - \tilde{v}\|_\infty \,.$$

To this end, we define

$$\log z(\epsilon) := \log z_{v+\epsilon(\tilde{v}-v)} \,.$$

Using the mean value theorem we have for an appropriate value $\eta \in (0,1)$

$$\begin{aligned} \log z_{\tilde{v}} - \log z_v &= \log z(1) - \log z(0) \\ &= (\log z)'(\eta) \\ &= \sum_j (v_j - \tilde{v}_j) \frac{1}{z(\eta)} e^{-v_j + \eta(v_j - \tilde{v}_j)} \,. \end{aligned}$$

The last expression is the expected value of $v_j - \tilde{v}_j$ with respect to the probability distribution corresponding to $v + \eta(\tilde{v} - v)$, which cannot be greater than $\|v - \tilde{v}\|_\infty$. □

We now have introduced the technical requirements to formulate a theorem for approximate additive noise models:

Theorem 3 (approximate additive noise model)**.**
Let $p_{X,Y}$ be as in Theorem 2, but only admitting an approximative additive noise model in the sense that the statistical mutual information between X and the residual $E := Y - f(X)$ satisfies

$$I(X;E) \leq \frac{\beta}{2} \left(\frac{\epsilon\beta}{4m^3} \right)^2 \,, \tag{19}$$

where β is a lower bound on $p(x,y)$. Here, subtraction is understood with respect to $\mathbb{Z}_m$. Then, if f is non-constant, we have

$$I(p_Y : p_{X|Y}) \stackrel{+}{\geq} K_\epsilon(p_Y) - 2\log m - m - 2K(\epsilon) \,. \tag{20}$$

Proof: The idea is to define a distribution $\tilde{p}_{X,Y}$ that is close to $p_{X,Y}$ and admits an *exact* additive noise model: Define a joint distribution on X and E by the product

$$\tilde{p}_{X,E} := p_X \otimes p_E \,.$$

By variable transformation, $\tilde{p}_{X,E}$ defines a distribution $\tilde{p}_{X,Y}$ that admits an additive noise model from X to Y. Eq. (15) now holds for $\tilde{p}_{X|Y}$ and $\tilde{p}_Y$ with $\tilde{k}_{(x_0|y)}$ instead of $k_{(x_0|y)}$, which is defined similar to eq. (14). Denote the corresponding vector by $\tilde{k} = (\tilde{k}_{(x_0|y)})_y$. In analogy to eq. (16) and the proof of Theorem 2, we now have

$$\log \tilde{p}_Y = (S - I)^{-2} \tilde{k} + \tilde{\alpha}\, \mathbf{e} \,,$$

where $\tilde{\alpha}$ is the appropriate normalization constant and $\mathbf{e}$ the all-one vector. To show that $p_{X|Y}$ allows an approximative description of p_Y we have to replace $\tilde{k}$ and $\tilde{p}_Y$ with k and p_Y, respectively. We define

$$\log r_Y := (S - I)^{-2} k + \alpha\, \mathbf{e} \,,$$

and using Lemma 4 we obtain

$$\begin{aligned} \|\log p_Y - \log r_Y\|_\infty &\leq \|\log p_Y - \log \tilde{p}_Y\|_\infty + \|\log \tilde{p}_Y - \log r_Y\|_\infty \\ &\leq \|\log p_Y - \log \tilde{p}_Y\|_\infty + 2\|(S - I)^{-2}(k - \tilde{k})\|_\infty . \end{aligned} \tag{21}$$

The modulus of the eigenvalues of $(S-I)^{-1}$ on this subspace are all smaller than $m/4$ (for $m \geq 2$) since they read

$$\frac{1}{e^{2\pi\, i/m}-1}, \frac{1}{e^{2\pi i\, 2/m}-1}, \ldots, \frac{1}{e^{2\pi i\,(m-1)/m}-1}\,.$$

We thus have

$$\|(S-I)^{-2}(\tilde{k}-k)\|_2 \leq \frac{m^2}{16}\|\tilde{k}-k\|_2 \leq \frac{m^3}{16}\|\tilde{k}-k\|_\infty\,,$$

where the last inequality used $\|\cdot\|_2 \leq \sqrt{m}\|\cdot\|_\infty$. Together with $\|\cdot\|_\infty \leq \|\cdot\|_2$, ineq. (21) then yields

$$\|\log p_Y - \log r_Y\|_\infty \leq \|\log p_Y - \log \tilde{p}_Y\|_\infty + \frac{m^3}{8}\|\tilde{k}-k\|_\infty. \tag{22}$$

Now we derive an upper bound on the two summands of the rhs. using our assumption on the limited statistical mutual information between X and E. To this end, we observe that

$$D(p_{X,Y}||\tilde{p}_{X,Y}) = D(p_{X,E}||\tilde{p}_{X,E}) = I(X:E)\,, \tag{23}$$

where the first equality is due to the invariance of relative entropy under variable transformation and the second uses a well-known reformulation of mutual information [4]. Moreover, we have

$$D(p_{X|Y}||\tilde{p}_{X|Y}) = \sum_y D(p_{X|y}||\tilde{p}_{X|y})p(y) \leq \frac{\beta}{2}\left(\frac{\epsilon\beta}{4m^3}\right)^2\,,$$

where $p_{X|y}$ denotes the conditional distribution for one specific value y of Y. Using the lower bound on $p(y)$ we obtain

$$D(p_{X|y}||\tilde{p}_{X|y}) \leq \frac{1}{2}\left(\frac{\epsilon\beta}{4m^3}\right)^2 \quad \forall y\,.$$

Due to the well-known relation $D(p||q) \geq (2\ln 2)^{-1}\|p-q\|_1^2$ between relative entropy and ℓ_1-distance for two distributions [4], we obtain

$$|p(x|y) - \tilde{p}(x|y)| \leq \frac{\epsilon\beta}{4m^3}\,.$$

This implies

$$|\log p(x|y) - \log \tilde{p}(x|y)| \leq \frac{\epsilon}{4m^3}\,, \tag{24}$$

by applying the mean value theorem to the function $a \mapsto \log a$. From the definition of $\tilde{k}_{(x|y)}$ and $k_{(x|y)}$ in eq. (14) we conclude

$$\|\tilde{k}-k\|_\infty \leq \frac{\epsilon}{m^3}\,. \tag{25}$$

On the other hand, (23) implies

$$D(p_Y||\tilde{p}_Y) \leq \frac{\beta}{2}\left(\frac{\epsilon\beta}{4m^3}\right)^2 \leq \frac{1}{2}\left(\frac{\epsilon\beta}{4m^3}\right)^2\,,$$

and hence

$$\|\log p(y) - \log \tilde{p}(y)\|_\infty \leq \frac{\epsilon\beta}{4m^3} < \frac{\epsilon}{8}. \tag{26}$$

Using ineqs. (25) and (26), ineq. (22) yields for all y

$$|\log p(y) - \log r(y)| < \frac{\epsilon}{4}\,. \tag{27}$$

Let $\log q_p(y)$ be given by discretizing all values $\log p(y)$ up to an accuracy of $\epsilon/4$. Then

$$K(q_p|p_Y, \epsilon) \stackrel{+}{=} 0.$$

On the other hand, let $\log q_r(y)$ be given by discretizing all values $\log r(y)$ up to an accuracy of $\epsilon/4$. Then $K(q_r|r, \epsilon) \stackrel{+}{=} 0$ and thus

$$K(q_r|p_{X|Y}, \delta, x_0, \epsilon) \stackrel{+}{=} 0\,.$$

Due to (27), both discretizations coincide up to one bit for each value y, say $b_m(y)$. To illustrate this, consider the binary strings $0.111\ldots$ and $1.000\ldots$ which can be arbitrarily close despite their truncation being different. We conclude that

$$K(q_p|p_{X|Y}, \delta, x_0, \epsilon, b_m) \stackrel{+}{=} 0.$$

Let q be the distribution generated by $\log q_p$ through normalization

$$\log q(y) := \log q_p - \log \sum_y q_p(y).$$

Due to the upper bound (27), Lemma 4 gives

$$D(p||q) \leq 2\|\log p(y) - \log q_p(y)\|_\infty < \epsilon.$$

The theorem now follows from Lemma 3 applied to $z = p_{X|Y}$, $i_z = (\delta, x_0, \epsilon, b_m)$, $p = p_Y$ and $i_p = \epsilon$. □

The complexity of p_Y in the bound (20) will typically exceed the terms with m because we will need several bits for every bin to describe the corresponding probability (this will be discussed in Section 5 in more detail). Moreover, $K(\epsilon)$ can be quite low, in particular if we choose $\epsilon = 2^{-k}$ for some k. Therefore, the mutual information between p_Y and $p_{X|Y}$ is almost as large as the complexity of p_Y. This shows that the amount of adjustments required to mimic an additive noise model in the wrong direction depends essentially on the complexity of p_Y. In the following section we consider the complexity in the case in which p_Y is typical with respect to some known parametric family of distributions.

5 Kolmogorov complexity of distributions from a parametric family

The problem with applying Theorems 2 and 3 to real data is that the term $K_\epsilon(p_Y)$ cannot be known due to the uncomputability of Kolmogorov complexity in general. Fortunately, we can estimate how the complexity of typical elements in a *family* of distributions increases with decreasing ϵ. This is shown by the following lemma:

Lemma 5 (typical distributions in parametric families)**.**
Let $(p_\theta)_{\theta\in\Lambda}$ be a parametric family of probability densities over some finite probability space, where θ runs over a d-dimensional compact manifold $\Lambda \subset \mathbb{R}^d$. Let (p_θ) satisfy the following technical requirements: p_θ is strictly positive, the third derivative of the function $\theta \mapsto \log p_\theta$ is continuous, and the Fisher matrices F_θ have full rank for all θ. Moreover, let

the family p_θ be computable in the following sense: there exists a program that computes $p_\theta(y)$ for any computable input θ.

If Pr *denotes the uniform probability distribution on Λ, then there is an ϵ_0 such that for every $\delta > 0$ there exists an $\ell \in \mathbb{N}$ with*

$$\Pr\left(K_\epsilon(p_\theta) \in \left[-\frac{d}{2}\log\epsilon - \ell, -\frac{d}{2}\log\epsilon + \ell\right] \quad \forall \epsilon < \epsilon_0\right) \geq 1 - \delta\,.$$

In other words: for most of the distributions p_θ, the complexity $K_\epsilon(p_\theta)$ grows logarithmically with decreasing ϵ.

Proof: Let $S_\epsilon(\ell)$ be the set of distributions p_θ with

$$K_\epsilon(p_\theta) \leq -\frac{d}{2}\log\epsilon - \ell\,. \tag{28}$$

Every such p_θ is ϵ-close to some computable distribution $p_{\theta'}$ with

$$K(p_{\theta'}) \leq -\frac{d}{2}\log\epsilon - \ell\,. \tag{29}$$

Note that the number $N_\epsilon(\ell)$ of such distributions $p_{\theta'}$ is bounded from above by usual counting arguments:

$$N_\epsilon(\ell) \leq 2^{-\frac{d}{2}\log\epsilon - \ell} = \frac{2^{-\ell}}{\epsilon^{d/2}}\,.$$

To estimate the total volume of $S_\epsilon(\ell)$ we first estimate the volume of the set of p_θ that are ϵ-close to *one specific* $p_{\theta'}$ satisfying (29). The second order Taylor expansion of the relative entropy yields [1],

$$D(p_\theta||p_{\theta'}) = \frac{1}{2}(\theta-\theta')^T F_{\theta'}(\theta-\theta') + E(\theta-\theta')\,, \tag{30}$$

where $F_{\theta'}$ is the Fisher matrix and the error $E(\theta-\theta')$ is in $O(\|\theta-\theta'\|^3)$. Since the third derivative of $\log p_\theta$ is conituous (and thus bounded on a compact manifold), Taylor's theorem provides an error bound of the form constant times $\|\theta-\theta'\|^3$ that holds for all θ uniformly. For sufficiently small ϵ, the set of all θ with $D(p_\theta||p_{\theta'}) \leq \epsilon$ is thus contained in the ellipsoid

$$(\theta-\theta')^T F_{\theta'}(\theta-\theta') \leq 4\epsilon\,.$$

Hence, we can choose ϵ_0 such that this statement holds all $\epsilon < \epsilon_0$ and all θ'. The volume $V(\epsilon)$ of such an ellipsoid with respect to the Lebesgue measure is given by

$$V(\epsilon) = (\det F_{\theta'})^{-1/2}\frac{\pi^{d/2}}{\Gamma(d/2+1)}(4\epsilon)^{d/2}\,.$$

This can be seen by transforming the ellipsoid into a sphere of radius $\sqrt{4\epsilon}$ via the linear map $(F_{\theta'})^{-1/2}$.

If c denotes an upper bound on $(\det F_\theta)^{-1/2}$ over all parameter vectors θ (the eigenvalues of F_θ are bounded away from zero since Λ is compact), the total volume of $S_\epsilon(\ell)$ is bounded from above by

$$N_\epsilon(\ell)\frac{c\,\pi^{d/2}}{\Gamma(d/2+1)}(4\epsilon)^{d/2} \leq \frac{2^{-\ell}c}{\epsilon^{d/2}}\frac{\pi^{d/2}}{\Gamma(d/2+1)}(4\epsilon)^{d/2} = 2^{-\ell}\,\tilde{c}\,,$$

where $\tilde{c}$ is an appropriate constant. Choosing ℓ such that $2^{-\ell}\tilde{c}$ is smaller than δ times the volume of Λ shows that

$$K_\epsilon(p_\theta) \geq -\frac{d}{2}\log\epsilon - \ell\,,$$

holds at least with probability $1-\delta$.

Now we show that we can find an $\ell' \in \mathbb{N}$ such that

$$K_\epsilon(p_\theta) \leq -\frac{d}{2}\log\epsilon + \ell' \,, \tag{31}$$

for all θ and all $\epsilon < \epsilon_0$. To this end, we assume without loss of generality that Λ is contained in the cube $[0,1]^d$. We divide the cube into N equally sized cubes of side length Δ with middle points $\theta_1, \ldots, \theta_N$ such that

$$(\theta - \theta_j)^T F_\theta (\theta - \theta_j) \leq \epsilon$$

for any point θ in the same cube. By (30), this ensures for all $\theta, \theta_j \in \Lambda$ and sufficiently small ϵ that $D(p_\theta || p_{\theta_j}) \leq \epsilon$. If μ is an upper bound for all eigenvalues of all F_θ (uniformly in θ) it is sufficient to guarantee

$$\|\theta - \theta_j\|^2 \leq \frac{\epsilon}{\mu} \,.$$

This can be achieved by choosing

$$\Delta \leq \sqrt{\frac{\epsilon}{\mu d}} \,.$$

Hence it is sufficient to choose the smallest N that satisfies

$$N \geq \left(\frac{\epsilon}{\mu d}\right)^{d/2} \,,$$

and whose dth root is integer. The grid and thus every vector θ_j can be computed from j and p_{θ_j} can be computed from θ_j by assumption. The total complexity of describing p_{θ_j} is at most given by $\log N$ plus some overhead for the program that computes p_{θ_j} from j. Absorbing $\log(\mu d)$ and the overhead into ℓ' proves eq. (31). $\square$

We will now consider the family of all distributions p_Y on the finite probability space $\{0, \ldots, m-1\}$ for which $p(y)$ is bounded from below by some $\beta > 0$. We use the straightforward parameterization

$$p_\theta(j) := \theta_j \quad \text{for } j = 0, \ldots, m-2 \,,$$

and

$$p_\theta(m-1) = 1 - \sum_{j=1}^{m-2} \theta_j \,.$$

Then we define Λ as the set of $\theta \in \mathbb{R}^{m-1}$ for which $\theta_j \geq \beta$ and $\sum_j \theta_j \leq 1-\beta$. One can check that the Fisher information matrix has full rank and that the eigenvalues are bounded away from zero, i.e., Lemma 5 applies. Hence, we can find an ℓ such that the majority of distributions p_Y satisfy

$$K_\epsilon(p_Y) \stackrel{+}{=} -\frac{d}{2}\log\epsilon \,. \tag{32}$$

For such a "typical" distribution we obtain:

Corollary 1 (algorithmic mutual information for typical distributions)**.**
Let $p_{X,Y}$ be as in Theorem 3. Let p_Y satisfy eq. (32). If $I(X;E)$ satisfies the bound (19) with $\epsilon = 2^{-N}$ for sufficiently large N, then

$$I(p_Y : p_{X|Y}) \stackrel{+}{\geq} \frac{m-1}{2} N - 2\log m - m - 2\log N \,.$$

The proof is given by plugging eq. (32) into the lower bound of Theorem 3 together with $\epsilon = 2^{-N}$. Hence, for typical p_Y, the lower bound is positive if m and n are large enough. This asymptotic statement still holds true if p_Y looks on a coarse-grained scale like some simple distribution q_Y, i.e., a Gaussian, but shows irregular deviations from q_Y if the probabilities are described more accurately.

To give an impression on the amount of information between $p_{Y|X}$ and p_Y that can be inferred after n-fold sampling, we recall that the mutual information between E and X can be estimated up to an accuracy of $O(1/n)$ [15]. The lowest upper bound on ϵ in ineq. (19) that can be guaranteed by the observations thus is proportional to $1/\sqrt{n}$. Hence, for constant m, the best lower bound on the amount of algorithmic information shared by p_Y and $p_{X|Y}$ increases logarithmically in n as long as the sample is not sufficient to reject independence between $Y - f(X)$ and X.

6 Applying the results to empirical distributions

In applying Theorems 2 and 3 to realistic situations, we still have the problem that we have no reason to believe that the true distribution is computable. On the other hand, applying the argument to the empirical distribution (which is, for large sampling close to an additive noise model) is still problematic because algorithmic dependencies between the empirical distribution q_Y and the empirical conditional $q_{X|Y}$ do not prove algorithmic dependencies between the true distributions p_Y and $p_{X|Y}$. One reason is that, for sample size n, every conditional probability $q_{Y|X}(y|x)$ can always be written as a fraction with denominator $q_X(x)n$, which already is an algorithmic dependence.

We now describe how to use Postulate 1 if only a finite list of (x, y)-pairs is observed and the underlying distribution is not known. Given samples $\mathcal{S}_n = \big[(x_1, y_1), \ldots, (x_n, y_n)\big]$, we can generate a non-empty subsample $\mathcal{S}_{\ell(n)} = \big[(x_1, y_1), \ldots, (x_{\ell(n)}, y_{\ell(n)})\big]$ with high probability such that every x-value occurs exactly $\ell(n)/m$-times. The samples $S_{\ell(n)}$ can then be used for the estimation of $p_{Y|X}$. Hereby, $\ell(n)$ is chosen independently of the samples in a way that for $n \to \infty$ we have $\ell(n) \to \infty$ and the probability of obtaining $\mathcal{S}_{\ell(n)}$ from $\mathcal{S}_n$ converges to one.

Now by construction, if M contains no information about S, the empirical distribution

$$q_{Y|X}^{(\ell(n))}$$

of the *subsample* must not contain any information about the empirical distribution

$$q_X^{(n)}$$

of x-values in the entire sample, i.e.,

$$M_{X\to Y} := I(q_X^{(n)} : q_{Y|X}^{(\ell(n))}) \approx 0\,. \tag{33}$$

In the spirit of [10], we postulate that the violation of eq. (33) indicates that the causal hypothesis $X \to Y$ is wrong or the mechanisms generating x-values and the mechanisms generating y-values from x-values have not been generated independently. For a discussion of this case see [11]. Using this terminology, our goal is to derive a lower bound on $M_{Y\to X}$ for the case where $p_{X,Y}$ admits an additive noise model from X to Y. We can apply Theorem 3 to a distribution that is defined by the empirical results via

$$p'(x, y) := q^{(n)}(y)q^{(\ell(n))}(x|y)\,,$$

which is necessarily computable because it only contains rationale values.

We have already argued that the causal hypothesis $Y \to X$ would only be acceptable if

$$I(q^{(n)}(y) : q^{(\ell(n))}(x|y)) \approx 0\,.$$

If the true distribution p almost admits an additive noise model from X to Y in the sense of ineq. (19), the same inequality will also be satisfied by p' if n is sufficiently high and thus

$$I(q_Y^{(n)} : q_{X|Y}^{(\ell(n))}) \gg 0$$

provided that $K_\epsilon(q_Y^{(n)})$, which coincides with $K_\epsilon(p_Y)$ due to Lemma 2 for large n, is high.

7 Approximating continuous variables with discrete ones

Causal inference via additive noise models has been described and tested for continuous variables [6]. We have discussed the discrete case mainly for technical reasons because we were able to prove statements for distributions that are only close to additive noise models. Our results can easily be applied to the continuous case by discretization with increasing number of bins. As already mentioned, the discretized version of the empirical distribution becomes computable, which circumvents the problem that the true distribution may be uncomputable.

Before we discuss the discretization in detail, we emphasize that there is a problem with applying Postulate 1 to the conditionals obtained after discretizing the variables: if we define a discrete variables $X^{(m)}$ and $Y^{(m)}$ by putting X and Y into m bins each, the discretized conditional $p_{Y^{(m)}|X^{(m)}}$ does not only depend on $p_{Y|X}$. Instead, it also contains information about the distribution of X. For this reason, algorithmic dependencies between $p_{Y^{(m)}|X^{(m)}}$ and $p_{X^{(m)}}$ only disprove the causal hypothesis $X \to Y$ if the binning is fine enough to guarantee that the discrete value $x^{(m)}$ is sufficient to determine the conditional probability for $y^{(m)}$, i.e., the relevance of the exact value x is negligible if the discrete value is given. It is therefore essential that the argument below refers to the asymptotic case of infinitely small binning.

To approximate a continuous density $p(x, y)$ on $\mathbb{R}^2$ by $\mathbb{Z}_m^2$ with increasing $m := 2k+1$ we consider the square

$$Q_m := \left[-\frac{1}{2}\sqrt{m}, \frac{1}{2}\sqrt{m}\right]^2$$

for all odd m and replace $p(x, y)$ with $p(x, y|Q_m)$. We discretize Q_m into $m \times m$ bins of equal size $\Delta := 1/\sqrt{m}$, which defines a probability distribution over $\mathbb{Z}_m$-valued variables X_m and Y_m, respectively. We define the function $f_m : \mathbb{Z}_m \to \mathbb{Z}_m$ by putting the values $f(j\Delta)$ with $j = -k, \ldots, k$ to the bin whose middle point has the least distance.

Moreover, appropriate smoothness asumptions on $p(x, y)$ can guarantee that the mutual information between $Y_m - f_m(X_m)$ and X_m converges to $I(X : (Y - f(X)))$ for $m \to \infty$. It is known [15] that there are estimators for mutual information that converge if the binning m is increased proportionally to $\sqrt{n}$ for sample size $n \to \infty$. If $p(x, y)$ admits an additive noise model, i.e., $I(X : (Y - f(X)) = 0$, then $I(X_m : (Y_m - f(X_m)) \to 0$. Hence, the discrete distributions on X_m and Y_m get arbitrarily close to discrete additive noise models. Applying Theorem 2 to these discrete distributions then yields algorithmic dependence between the discretized marginal and the discretized conditional.

8 Conclusions

We have discussed a causal inference method that prefers the causal hypothesis $X \to Y$ to $Y \to X$ whenever the joint distribution $p_{X,Y}$ admits an additive noise model from X to Y and not vice versa. It seems that this way of reasoning assumes that all causal mechanisms in nature can be described by additive noise models, which is certainly not the case. Here we argue that the method is nevertheless justified because it is unlikely that a causal

mechanism that is not of the form of an additive noise model generates a distribution that looks like an additive noise model in the *wrong* direction. This is because such a coincidence would require mutual adjustments between $P(\text{cause})$ and $P(\text{effect}|\text{cause})$. To measure the amount of tuning needed for this situation we have derived a lower bound on the algorithmic information shared by $P(\text{cause})$ and $P(\text{effect}|\text{cause})$. If we assume that "nature chooses" $P(\text{cause})$ and $P(\text{effect}|\text{cause})$ independently, a significant amount of algorithmic information is not acceptable. Our justification of additive-noise-model based causal discovery thus is an application of two recent proposals for using algorithmic information theory in causal inference: [10] postulated that the shortest description of $P(\text{cause}, \text{effect})$ is given by separate descriptions of $P(\text{cause})$ and $P(\text{effect}|\text{cause})$, which would be violated then. [7] argued that algorithmic dependencies between any two objects require a causal explanation. They consider the two mechanisms that determine $P(\text{cause})$ and $P(\text{effect}|\text{cause})$, respectively, as two objects and conclude that the absence of causal links on the level of the two mechanisms imply their algorithmic independence, in agreement with [10].

References

[1] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

[2] G. Chaitin. On the length of programs for computing finite binary sequences. *J. Assoc. Comput. Mach.*, 13:547–569, 1966.

[3] G. Chaitin. A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.*, 22:329–340, 1975.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. Wileys Series in Telecommunications, New York, 1991.

[5] P. Gacs, J. Tromp, and P. Vitányi. Algorithmic statistics. *IEEE Trans. Inf. Theory*, 47(6):2443–2463, 2001.

[6] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Proceedings of the conference Neural Information Processing Systems (NIPS) 2008*, Vancouver, Canada, 2009. MIT Press. `http://books.nips.cc/papers/files/nips21/NIPS2008_0266.pdf`.

[7] D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. to appear in *IEEE Transactions on Information Theory. See also* `http://arxiv.org/abs/0804.3678`, 2008.

[8] A. Kolmogorov. Three approaches to the quantitative definition of information. *Problems Inform. Transmission*, 1(1):1–7, 1965.

[9] S. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, New York, Oxford Statistical Science Series edition, 1996.

[10] J. Lemeire and E. Dirkx. Causal models as minimal descriptions of multivariate systems. `http://parallel.vub.ac.be/~jan/`, 2006.

[11] J. Lemeire and K. Steenhaut. Inference of graphical causal models: Representing the meaningful information of probability distribution. To appear in Proceedings of the NIPS 2008 workshop "Causality - Objectives and Assessment", edited by I. Guyon, D. Janzing, B. Schölkopf., 2009.

[12] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, New York, 1997 (3rd edition: 2008).

[13] J. Mooij and D. Janzing. Distinguishing between cause and effect. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 6:147–146, 2010.

[14] J. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th International Conference on Machine Learning, Montreal, ACM International Conference Proceeding Series*, pages 745–752, New York, NY, USA, 2009. `http://www.cs.mcgill.ca/~icml2009/papers/279.pdf`.

[15] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1254, 2003.

[16] J. Pearl. *Causality*. Cambridge University Press, 2000.

[17] H. Reichenbach. *The direction of time*. Dover, 1999.

[18] R. Solomonoff. A preliminary report on a general theory of inductive inference. *Technical report V-131*, Report ZTB-138 Zator Co., 1960.

[19] R. Solomonoff. A formal theory of inductive inference. *Information and Control, Part II*, 7(2):224–254, 1964.

[20] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search (Lecture notes in statistics)*. Springer-Verlag, New York, NY, 1993.

[21] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.