Causal inference using the algorithmic Markov condition

Dominik Janzing and Bernhard Schölkopf

***Abstract*—Inferring the causal structure that links n observables is usually based upon detecting statistical dependences and choosing simple graphs that make the joint measure Markovian. Here we argue why causal inference is also possible when the sample size is one.**

We develop a theory how to generate causal graphs explaining similarities between single objects. To this end, we replace the notion of conditional stochastic independence in the causal Markov condition with the vanishing of conditional *algorithmic* mutual information and describe the corresponding causal inference rules.

We explain why a consistent reformulation of causal inference in terms of algorithmic complexity implies a new inference principle that takes into account also the complexity of conditional probability densities, making it possible to select among Markov equivalent causal graphs. This insight provides a theoretical foundation of a heuristic principle proposed in earlier work.

We also sketch some ideas on how to replace Kolmogorov complexity with *decidable* complexity criteria. This can be seen as an algorithmic analog of replacing the empirically undecidable question of statistical independence with practical independence tests that are based on implicit or explicit assumptions on the underlying distribution.

keywords: *algorithmic information, Church-Turing thesis, data compression, graphical models, probability-free causal inference*

CONTENTS

Both authors are with the Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany. E-mail: dominik.janzing@tuebingen.mpg.de, bernhard.schoelkopf@tuebingen.mpg.de

I. INTRODUCTION TO CAUSAL INFERENCE FROM STATISTICAL DATA

Causal inference from statistical data has attracted increasing interest in the past decade. In contrast to traditional statistics where statistical dependences are only taken to prove that some kind of relation between random variables exists, causal inference methods in machine learning are explicitly designed to generate hypotheses on causal directions automatically based upon statistical observations, e.g., via conditional independence tests [1], [2]. The crucial assumption connecting statistics with causality is the causal Markov condition explained below after we have introduced some notations and terminology.

We denote random variables by capitals and their values by the corresponding lowercase letters. Let $X_1, \ldots, X_n$ be random variables and G be a directed acyclic graph (DAG) representing the causal structure where an arrow from node X_i to node X_j indicates a direct causal effect. Here the term *direct* is understood with respect to the chosen set of variables in the sense that the information flow between the two variables considered is not performed via using one or more of the other variables as intermediate nodes. We will next briefly rephrase the postulates that are required in the statistical theory of inferred causation [2], [1].

A. Causal Markov condition

When we consider the causal structure that links n random variables $\mathcal{V} := \{X_1, \ldots, X_n\}$ we will implicitly assume that $\mathcal{V}$ is causally sufficient in the sense that all common causes of two variables in $\mathcal{V}$ are also in $\mathcal{V}$. Then a causal hypothesis G is only acceptable as potential causal structure if the joint distribution $P(X_1, \ldots, X_n)$ satisfies the Markov condition with respect to G. There are several formulations of the Markov condition that are known to coincide under some

technical conditions (see Lemma 1). We will first introduce the following version which is sometimes referred to as the *parental* or the *local* Markov condition [3].

To this end, we introduce the following notations. PA_j is the set of parents of X_j and ND_j the set of non-descendants of X_j except itself. If S, T, R are sets of random variables, $S \perp\!\!\!\perp T\,|R$ means S is statistically independent of T, given R.

Postulate: statistical causal Markov condition, local version
If a directed acyclic graph G formalizes the causal structure among the random variables $X_1, \ldots, X_n$. Then

$$X_j \perp\!\!\!\perp ND_j\,|PA_j\,, \tag{1}$$

for all $j = 1, \ldots, n$.

The strength of violating statistical dependences is often measured in terms of mutual information. For three sets of variables X, Y, Z one defines the conditional mutual information of X and Y, given Z by [4]

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

where the Shannon entropies read as follows. Assume that the distribution $P(X_1, \ldots, X_k, Z)$ has the density $P(x_1, \ldots, x_k, z)$ and a conditional density $P(x_1, \ldots, x_k|z)$ with respect to some measure μ (which may, for instance, be the Lebesgue measure if all variables are continuous and the counting measure if they are discrete), then we have

$$\begin{aligned} H(X_1, \ldots, X_k|Z) &:= -\int P(x_1, \ldots, x_k, z) \\ &\times \log P(x_1, \ldots, x_k|z) d\mu(x_1, \ldots x_k, z)\,. \end{aligned}$$

We call condition (1) the *statistical* causal Markov condition because we will later introduce an algorithmic version. The fact that conditional irrelevance not only occurs in the context of *statistical* dependences has been emphasized in the literature (e.g. [5], [1]) in the context of describing abstract properties (like semi-graphoid axioms) of the relation $\cdot \perp\!\!\!\perp \cdot |\cdot$. We will therefore state the causal Markov condition also in an abstract form that does not refer to any specific notion of conditional informational irrelevance:

Postulate: abstract causal Markov condition, local
Given all the direct causes of an observable O, its non-effects provide no additional information on O.

Here, observables denote something in the real world that can be observed and the observation of which can be formalized in terms of a mathematical language. In this paper, observables will either be random variables (formalizing statistical quantities) or they will be strings (formalizing the description of objects). Accordingly, information will be *statistical* or *algorithmic* mutual information, respectively.

The importance of the causal Markov condition lies in the fact that it links causal terms like "direct causes" and "non-effects" to informational relevance of observables. The local Markov condition is rather intuitive because it echoes the fact that the information flows from direct causes to their effect and every dependence between a node and its non-descendants involves the direct causes. Conditioning on direct causes "screens off" the relation to variables other than the descendants. However, the independences postulated by the local Markov condition imply additional independences. It is therefore hard to decide whether an independence must hold for a Markovian distribution or not, solely on the basis of the local formulation. In contrast, the global Markov condition makes the complete set of independences obvious. To state it we first have to introduce the following graph-theoretical concept.

Definition 1 (d-separation):
A path p in a DAG is said to be blocked by a set of nodes Z if and only if

1) p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or
2) p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set Z is said to d-separate X from Y if Z blocks every (possibly undirected) path from a node in X to a node in Y.

The following Lemma shows that d-separation is the correct condition for deciding whether an independence is implied by the local Markov condition (see [5], Theorem 3.27):

Lemma 1 (equivalent Markov conditions):
Let $P(X_1, \ldots, X_n)$ have a density $P(x_1, \ldots, x_n)$ with respect to a product measure. Then the following three statements are equivalent:

I. **Recursive form:** P admits the factorization

$$P(x_1, \ldots, x_n) = \prod_{j=1}^{n} P(x_j|pa_j)\,, \tag{2}$$

where $P(.|pa_j)$ is shorthand for the conditional probability density, given the values of all parents of X_j.

II. **Local (or parental) Markov condition:** for every node X_j we have

$$X_j \perp\!\!\!\perp ND_j\,|PA_j\,,$$

i.e., it is conditionally independent of its non-descendants (except itself), given its parents.

III. **Global Markov condition:**

$$S \perp\!\!\!\perp T\,|R$$

for all three sets S, T, R of nodes for which S and T are d-separated by R.

Moreover, the local and the global Markov condition are equivalent even if P does not have a density with respect to a product measure.

The conditional densities $P(x_j|pa_j)$ are also called the *Markov kernels* relative to the hypothetical causal graph G. It is important to note that every choice of Markov kernels define a Markovian density P, i.e., the Markov kernels define exactly the set of free parameters remaining after the causal structure has been specified.

To select graphs among those for which P satisfies the Markov condition, we also need an additional postulate:

Postulate: causal faithfulness
Among all graphs G for which P is Markovian, prefer the ones for which all the observed conditional independences in

the joint measure $P(X_1, \ldots, X_n)$ are imposed by the Markov condition.

The idea is that the set of observed independences is typical for the causal structure under consideration rather than being the result of specific choices of the Markov kernels. This becomes even more intuitive when we restrict our attention to random variables with finite range and observe that the values $P(x_j|pa_j)$ then define a natural parameterization of the set of Markovian distributions in a finite dimensional space. The non-faithful distributions form a submanifold of lower dimension, i.e., a set of Lebesgue measure zero [6]. They therefore almost surely don't occur if we assume that "nature chooses" the Markov kernels for the different nodes independently according to some density on the parameter space. There are several objections against faithfulness, we only want to mention that deterministic relations can generate unfaithful distributions. The fact that deterministic relations are not that uncommon shows that "nature does sometimes choose" from sets of measure zero.

The above "zero Lebesgue measure argument" is close to the spirit of Bayesian approaches [7], where priors on the set of Markov kernels are specified for every possible hypothetical causal DAG and causal inference is performed by maximizing posterior probabilities for hypothetical DAGs, given the observed data. This procedure leads to an *implicit* preference of faithful structures in the infinite sampling limit given appropriate conditions for the priors on the parameter space. The assumption that "nature chooses Markov kernels independently", which is also part of the Bayesian approach, will turn out to be closely related to the algorithmic Markov condition postulated in this paper.

We now discuss the justification of the statistical causal Markov condition because we will later justify the algorithmic Markov condition in a similar way. To this end, we introduce functional models [1]:

Postulate: functional model of causality
If a directed acyclic graph G formalizes the causal relation between the random variables $X_1, \ldots, X_N$ then every X_j can be written as a deterministic function of PA_j and a noise variable N_j ,

$$X_j = f_j(PA_j, N_j)\,, \tag{3}$$

where all N_j are jointly independent.

Note that this model does not put any mathematical restriction on the conditionals[1] $P(X_j|PA_j)$. Given that the joint distribution factorizes as in eq. (2) the model thus does not restrict the set of possible joint distributions any further. However, the functional model can be used to justify the causal Markov condition since we have [1], Theorem 1.4.1:

Lemma 2 (Markov condition in functional models):
Every joint distribution $P(X_1, \ldots, X_n)$ generated according to the functional model in Postulate 3 satisfies the local and the global Markov condition relative to G.

We rephrase the proof in [1] because our proof for the algorithmic version will rely on the same idea.

Proof of Lemma 2: extend G to a graph $\tilde{G}$ with nodes $X_1, \ldots, X_n, N_1, \ldots, N_n$ that additionally contains an arrow from each N_j to X_j. The given joint distribution of noise variables induces a joint distribution

$$\tilde{P}(X_1, \ldots, X_n, N_1, \ldots, N_n)\,,$$

that satisfies the local Markov condition with respect to $\tilde{G}$: first, every X_j is completely determined by its parents making the condition trivial. Second, every N_j is parentless and thus we have to check that it is (unconditionally) independent of its non-descendants. The latter are deterministic functions of $\{N_1, \ldots, N_n\} \setminus \{N_j\}$. Hence the independence follows from the joint independence of all N_i.

By Lemma 1, $\tilde{P}$ is also globally Markovian w.r.t. $\tilde{G}$. Then we observe that ND_j and X_j are d-separated in $\tilde{G}$ (where the parents and non-descendants are defined with respect to G), given PA_j. Hence P satisfies the local Markov condition w.r.t. G and hence also the global Markov condition. □

Functional models formalize the idea that the outcome of an experiment is completely determined by the values of all relevant parameters where the only uncertainty stems from the fact that some of these parameters are hidden. Even though this kind of determinism is in contrast with the commonly accepted interpretation of quantum mechanics [8], we still consider functional models as a helpful framework for discussing causality in real life since quantum mechanical laws refer mainly to phenomena in micro-physics. The deterministic function in functional models nicely represent causal mechanisms that persist also after manipulating the distribution of inputs. The framework thus formalizes the modularity of causal structure: every function represents a causal mechanism that exists independently of the others.

Causal inference using the Markov condition and the faithfulness assumption has been implemented in causal learning algorithms [2]. The following fundamental limitations of these methods deserve our further attention:

1) *Markov equivalence:* There are only few cases where the inference rules provide unique causal graphs. Often one ends up with a *class of Markov equivalent* graphs, i.e., graphs that entail the same set of independences. For this reason, additional inference rules are desirable. In particular, deciding whether X causes Y or Y causes X for just two observed variables is a challenging task for novel inference rules [9] since it is unsolvable for independence-based methods.
2) *Dependence on i.i.d. sampling:* the whole setting of causal inference relies on the ability to sample repeatedly and independently from the same joint distribution $P(X_1, \ldots, X_n)$. As opposed to this assumption, causal inference in real life also deals with probability distributions that change in time. Even though there are techniques in conventional statistics to cope with

[1]To see this, let N_j consist of (possibly uncountably many) real-valued random variables $N_j[pa_j]$, one for each value pa_j of parents PA_j. Let $N_j[pa_j]$ be distributed according to $P(X_j|pa_j)$, and define $f_j(PA_j|N_j) := N_j[pa_j]$. Then $X_j|PA_j$ obviously has distribution $P(X_j|PA_j)$.

this problem, there are no methods for inferring causal relations among single observations, i.e., for the case of sample size one.

The idea of this paper is to develop a theory of probability-free causal inference that helps to construct causal hypotheses based on similarities of *single* objects. Then the nodes of the directed acyclic graph formalizing the causal structure are single objects. Here, similarities between these objects will be defined by comparing the length of the shortest description of single objects to the length of their shortest joint description. Despite the analogy to causal inference from statistical data our theory also implies new *statistical* inference rules. In other words, our approach to address weakness 2 also yields new methods to address 1.

The paper is structured as follows. In the remaining part of this Section, i.e., Subsection I-B, we describe recent approaches from the literature to causal inference from statistical data that address problem 1 above. In Section II we develop the general theory on inferring causal relations among individual objects based on algorithmic information. This framework appears, at first sight, as a straightforward adaption of the statistical framework using well-known correspondences between statistical and algorithmic information theory. However, Section III describes that this implies novel causal inference rules for *statistical* inference because *non-statistical* algorithmic dependences can even occur in data that were obtained from statistical sampling. In Section IV we sketch some ideas on how to replace causal inference rules based on the uncomputable *algorithmic information* with decidable criteria that are still motivated by the uncomputable idealization.

The table in fig. 1 summarizes the analogies between the theory of statistical and the theory of algorithmic causal inference described in this paper. The differences, however, which are the main subject of Sections III to IV, can hardly be represented in the table.

B. Developing new statistical inference rules

In [10], [11] we have proposed causal inference rules that are based on the idea that the factorization of $P(\text{cause}, \text{effect})$ into $P(\text{effect}|\text{cause})$ and $P(\text{cause})$ typically leads to simpler terms than the "artificial" factorization into $P(\text{effect})P(\text{cause}|\text{effect})$. The generalization of this principle reads: Among all graphs G that render P Markovian prefer the one for which the decomposition in eq. (2) yields the simplest Markov kernels. We have called this vague idea the "principle of plausible Markov kernels".

Before we describe several options to define simplicity we describe a simple example to illustrate the idea. Assume we have observed that a binary variable X (with values $x = -1, 1$) and a continuous variable Y with values in $\mathbb{R}$ that are distributed according to a mixture of two Gaussians (see fig. 2). Since this will simplify the further discussion let us assume that the two components are equally weighted, i.e.,

$$P(x,y) = \frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu-x\lambda)^2}{2\sigma^2}} ,$$

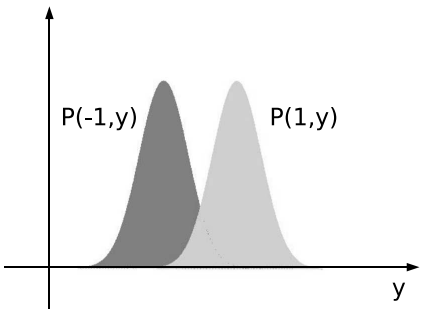

Fig. 2. Observed joint distribution of X and Y consisting of two Gaussians of equal width shifted against each other.

where λ determines the shift of the mean caused by switching between $x = 1$ and $x = -1$.

The marginal $P(Y)$ is given by

$$P(y) = \frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}} \left(e^{-\frac{(y-\mu+\lambda)^2}{2\sigma^2}} + e^{-\frac{(y-\mu-\lambda)^2}{2\sigma^2}} \right) . \quad (4)$$

One will prefer the causal structure $X \to Y$ compared to $Y \to X$ because the former explains in a natural way why $P(Y)$ is bimodal: the effect of X on Y is simply to shift the Gaussian distribution by 2λ. In the latter model the bimodality of $P(Y)$ remains unexplained.[2] To prefer one causal model to another one because the corresponding conditionals are simpler seems to be a natural application of Occam's Razor. However, Section III will show that such an inference rule also follows from the theory developed in the present paper when simplicity is meant in the sense of low Kolmogorov complexity. In the remaining part of this section we will sketch some approaches to implement the "principle of plausible Markov kernels" in practical applications.

In [10] we have defined a family of "plausible Markov kernels" by conditionals $P(X_j|PA_j)$ that are second order exponential models, i.e., $\log P(x_j|pa_j)$ is a polynomial of order two in the variables $\{X_j\} \cup \{PA_j\}$ up to some additive partition function (for normalization) that depends only on the variables PA_j. For every hypothetical causal graph, one thus obtains a family of "plausible joint distributions $P(X_1, \ldots, X_n)$" that are products of the plausible Markov kernels. Then we prefer the causal direction for which the plausible joint distributions provide the best fit for the given observations.

In [11] we have proposed the following principle for causal inference: Given a joint distribution of the random variables $X_1, \ldots, X_n$, prefer a causal structure for which

$$\sum_{j=1}^{n} C(P(X_j|PA_j)) \quad (5)$$

is minimal, where C is some complexity measure on conditional probability densities.

[2] Using Reichenbach's terminology [12] and Salmon's "mark transmission theory" [13], the cause leaves "some kind of mark on its effect": here, the distribution $P(X)$ can be identified from separating the two modes of $P(Y)$.

	statistical	**algorithmic**
observables (vertices of a DAG)	random variables	sequences of strings
observations	i.i.d. sampled data	strings
conditional independence	$X \perp\!\!\!\perp Y \,\vert Z$ $\Updownarrow$ $I(X;Y\vert Z)=0$	$x \perp\!\!\!\perp y \,\vert z$ $\Updownarrow$ $I(x:y\vert z) \stackrel{+}{=} 0$
I. recursion formula	$P(x_1,\ldots,x_n)$ $=$ $\prod_j P(x_j\vert pa_j)$	$K(x_1,\ldots,x_n)$ $=$ $\sum_j K(x_j\vert pa_j^*)$
II. local Markov condition	$X_j \perp\!\!\!\perp ND_j \,\vert PA_j$	$x_j \perp\!\!\!\perp nd_j \,\vert pa_j^*$
III. global Markov condition	d-separation $\Rightarrow$ statistical independence	d-separation $\Rightarrow$ algorithmic independence
equivalence of I-III	Theorem 3.27 in [5]	Theorem 3
functional models	Section 1.4 in [1]	Eq. (21)
functional models imply Markov condition	Theorem 1.4.1 in [1]	Theorem 4
decidable dependence criteria	assumptions on joint distribution	Section IV

Fig. 1. Analogies between statistical and algorithmic causal inference

There is also another recent proposal for new inference rules that refers to a related simplicity assumption, though formally quite different from the ones above. The authors of [14] observe that there are joint distributions of $X_1,\ldots,X_n$ that can be explained by a linear model with additive non-Gaussian noise for one causal direction but require non-linear causal influence for the other causal directions. For real data they prefer the causal graph for which the observations are closer to the linear model.

To justify the belief that conditionals that correspond to the true causal direction tend to be simpler than non-causal conditionals (which is common to all the approaches above) is one of the main goals of this paper.

II. Inferring causal relations among individual objects

It has been emphasized [1] that the application of causal inference principles often benefits from the non-determinism of causal relations between the observed random variables. In contrast, human learning in real-life often is about quite deterministic relations. Apart from that, the most important difference between human causal learning and the inference rules in [2], [1] is that the former is also about causal relations among *single* objects and does not necessarily require sampling. Assume, for instance, that the comparison of two texts show similarities (see e.g. [15]) such that the author of the text that appeared later is blamed to have copied it from the other one or both are blamed to have copied from a third one. The statement that the texts are similar could be based on a statistical analysis of the occurrences of certain words or letter sequences. However, such kind of simple statistical tests can fail in both directions: In Subsection II-B (before Theorem 3) we will discuss an example showing that they can erroneously infer causal relations even though they do not exist. This is because parts that are common to both objects, e.g., the two texts, are only suitable to prove a causal link if they are not "too straightforward" to come up with.

On the other hand, causal relations can generate similarities between texts for which every *efficient* statistical analysis is believed to fail. We will describe an idea from cryptography to show this. A cryptosystem is called ROR-CCA-secure (Real or Random under Chosen Ciphertext Attacks) if there is no

efficient method to decide whether a text is random or the encrypted version of some *known* text without knowing the key [16]. Given that there are ROR-CCA-secure schemes (which is unknown but believed by cryptographers) we have a causal relation leading to similarities that are not detected by any kind of simple counting statistics. However, once an attacker has found the key (maybe by exhaustive search), he/she recognizes similarities between the encrypted text and the plain text and infers a causal relation. The causal relation between plain text and its encrypted version leads to a "similarity" that is not *efficiently* detectable.

This already suggests two things: (1) detecting similarities involves *searching* over potential rules how properties of one object can be algorithmically derived from the structure of the other. (2) It is possible that inferring causal relations therefore relies on *computationally infeasible* decisions (if computable at all) on whether two objects have information in common or not.

A. Algorithmic mutual information

We will now describe how the information one object provides about the other can be measured in terms of Kolmogorov complexity. We start with some notation and terminology. Below, strings will always be binary strings since every description given in terms of a different alphabet can be converted into a binary word. The set of binary strings of arbitrary length will be denoted by $\{0,1\}^*$. Recall that the Kolmogorov complexity $K(s)$ of a string $s \in \{0,1\}^*$ is defined as the length of the shortest program that generates s using a previously defined universal prefix Turing machine [17], [18], [19], [20], [21], [4], [22]. The conditional Kolmogorov complexity $K(t|s)$ [4] of a string t given another string s is the length of the shortest program that can generate t from s. In order to keep our notation simple we use $K(x,y)$ to refer to the complexity of the concatenation $x'y$ where x' is a prefix code of x (equivalently, one can also define a string x,y from the pair (x,y) using a standard bijection between $\mathbb{N}\times\mathbb{N}$ and $\mathbb{N}$).

We will mostly have equations that are valid only up to additive constant terms in the sense that the difference between both sides does not depend on the strings involved in the equation (but it may depend on the Turing machines they refer to). To indicate such constants we denote the corresponding equality by $\stackrel{+}{=}$ and likewise for inequalities. In this context it is important to note that the number n of nodes of the causal graph is considered to be a constant. Moreover, for every string s we define s^* as its shortest description. If the latter is not unique, we consider the first one in a lexicographic order. It is necessary to distinguish between $K(\cdot|s)$ and $K(\cdot|s^*)$. This is because there is a trivial algorithmic method to generate s from s^* (just apply the Turing machine to s^*), but there is no algorithm that computes the shortest description s^* from a general input s. One can show [22] that knowing s^* is equivalent to knowing the pair $(s,K(s))$ since $K(s^*|s,K(s)) \stackrel{+}{=} K(s,K(s)|s^*) \stackrel{+}{=} 0$. The following equation for the joint algorithmic information of two strings x,y will be useful [23]:

$$K(x,y) \stackrel{+}{=} K(x) + K(y|x^*) = K(x) + K(y|x,K(x))\,. \quad (6)$$

The conditional version reads [23]:

$$K(x,y|z) \stackrel{+}{=} K(x|z) + K(y|x,K(x|z),z) \quad (7)$$

The most important notion in this paper will be the algorithmic mutual information measuring the amount of algorithmic information that two objects have in common. Following the literature (e.g. [24], [25]) we define:

Definition 2 (algorithmic mutual information): Let x,y be two strings. Then the algorithmic mutual information of x,y is

$$I(x:y) := K(y) - K(y|x^*)\,.$$

The mutual information is the number of bits that can be saved in the description of y when the shortest description of x is already known. The fact that one uses x^* instead of x ensures that it coincides with the symmetric expression [23]:

Lemma 3 (symmetric version of algorithmic mutual inf.): For two strings x,y we have

$$I(x:y) \stackrel{+}{=} K(x) + K(y) - K(x,y)\,.$$

In the following sections, non-vanishing mutual information will be taken as an indicator for causal relations, but more detailed information on the causal structure will be inferred from *conditional* mutual information. This is in contrast to approaches from the literature to measure similarity versus differences of single objects that we briefly review now. To measure differences between single objects, e.g. pictures [26], [27], one defines the *information distance* $E(x,y)$ between the two corresponding strings as the length of the shortest program that computes x from y *and* y from x. It can be shown [28] that

$$E(x,y) \stackrel{+}{=} \max\{K(x|y),K(y|x)\}\,.$$

However, whether $E(x,y)$ is small or large is not an appropriate condition for the existence and the strength of a causal link. Complex objects can have much information in common even though their distance is large. In order to obtain a measure relating the amount of information that is disjoint for the two strings to the amount they share, Li et al. [27] and Bennett et al. [15] use the "normalized distance measure"

$$d_s(x,y) := \frac{K(x|y^*) - K(y|x^*)}{K(x,y)} \stackrel{+}{=} 1 - \frac{I(x:y)}{K(x,y)}\,,$$

or

$$d(x,y) = \frac{\max\{K(x|y),K(y|x)\}}{\max\{K(x),K(y)\}}\,.$$

The intuitive meaning of $d_s(x,y)$ is obvious from its direct relation to mutual information, and $1-d(x,y)$ measures the fraction of the information of the more complex string that is shared with the other one. Bennett et al. [15] propose to construct evolutionary histories of chain letters using such kinds of information distance measures. The algorithmic mutual information to measure the similarity of two objects has, for instance, been used in [29], [30]. However, like in statistical causal inference, inferring adjacencies on the basis of strongest

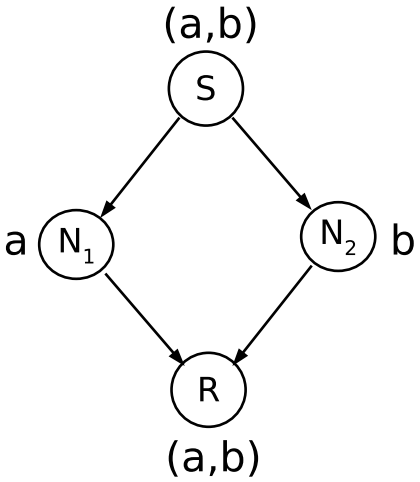

Fig. 3. Even though sender S and receiver R are non-adjacent, they are the pair with maximal mutual information (for the scenario described in the text).

dependences is only possible for simple causal structures like trees. In the general case, non-adjacent nodes can share more information than adjacent ones when information is propagated via more than one path. To see this, consider the "diamond graph" shown in fig. 3: The sender S has two generic strings a, b. He/she copies the information and sends a to the intermediate node N_1 and b to N_2. Then, N_1 and N_2 copy their strings and send them to the receiver R. The pair mazimizing the mutual information is (R, S) (because we have $I(R : S) = K(a, b)$ and the others share at most a or b) even though they are not adjacent.

Instead of constructing causal neighborhood relations by comparing information distances we will therefore use *conditional* mutual information. 'In order to define its algorithmic version, we first observe that Definition 2 can be rewritten into the less concise form

$$I(x : y) \stackrel{+}{=} K(y) - K(y|x, K(x)) \,.$$

This formula generalizes more naturally to the conditional analog [23]:

Definition 3 (conditional algorithmic mutual information):

Let x, y, z be three strings. Then the conditional mutual algorithmic information of x, y, given z is

$$I(x : y|z) := K(y|z) - K(y|x, K(x|z), z) \,.$$

As shown in [23] (Remark II.3), the conditional mutual information also is symmetric up to a constant term:

Lemma 4 (symmetric algorithmic conditional mutual inf.):

For three strings x, y, z one has:

$$I(x : y|z) \stackrel{+}{=} K(x|z) + K(y|z) - K(x, y|z) \,.$$

Definition 4 (algorithmic conditional independence): Given three strings x, y, z, we call x conditionally independent of y, given z (denoted by $x \perp\!\!\!\perp y\,|z$) if

$$I(x : y|z) \approx 0 \,.$$

In words: Given z, the additional knowledge of y does not allow us a stronger compression of x. This remains true if we are given the Kolmogorov complexity of y, given z.

The theory developed below will describe laws where symbols like x, y, z represent arbitrary strings. Then one can always think of *sequences* of strings of increasing complexity and statements like "the equation holds up to constant terms" are well-defined. We will then understand conditional independence in the sense of $I(x : y|z) \stackrel{+}{=} 0$. However, if we are talking about three fixed strings that represent objects in *real-life*, this does not make sense and the threshold for considering two strings dependent will heavily depend on the context. For this reason, we will not specify the symbol $\approx$ any further. This is the same arbitrariness as the cutoff rate for statistical dependence tests.

The definitions and lemmas presented so far were strongly motivated by the statistical analog. Now we want to focus on a theorem in [25] that provides a mathematical relationship between algorithmic and statistical mutual information. First we state the following theorem (see Theorem 7.3.1 of [4] and Brudno's Theorem [31]), showing that the Kolmogorov complexity of a random string is approximatively given by the entropy of the underlying probability distribution:

Theorem 1 (entropy and Kolmogorov complexity): Let $\mathbf{x} = x_1, x_2, \cdots, x_n$ be a string whose symbols $x_j \in \mathcal{A}$ are drawn i.i.d. from a probability distribution $P(X)$ over the finite alphabet $\mathcal{A}$. Slightly overloading notation, set $P(\mathbf{x}) := P(x_1) \cdots P(x_n)$. Let $H(.)$ denote the Shannon entropy of a probability distribution. Then there is a constant c such that

$$H(X) \leq \frac{1}{n}\mathbb{E}(K(\mathbf{x}|n)) \leq H(X) + \frac{|\mathcal{A}| \log n}{n} + \frac{c}{n} \quad \forall n \,,$$

where $\mathbb{E}(.)$ is short hand for the expected value with respect to $P(\mathbf{x})$. Moreover,

$$\lim_{n\to\infty} \frac{1}{n} K(\mathbf{x}) = H(X) \quad \text{ with probability } 1 \,.$$

However, for our purpose, we need to see the relation between algorithmic and statistical *mutual information*. If $\mathbf{x} = x_1, x_2, \cdots, x_n$ and $\mathbf{y} = y_1, y_2, \cdots, y_n$ such that each pair (x_j, y_j) is drawn i.i.d. from a joint distribution $P(X, Y)$, the theorem already shows that

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}(I(\mathbf{x} : \mathbf{y})) = I(X; Y) \,.$$

This can be seen by writing statistical mutual information as

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

The above translations between entropy and algorithmic information refer to a particular setting and to special limits. The focus of this paper is mainly the situation where the above limits are not justified. Before we rephrase Lemma II.4 in [23] which provides insights into the general case, we recall that a function f is called recursive if there is a program on a Turing

machine that computes $f(x)$ from the input x, and halts on all possible inputs.

Theorem 2 (statistical and algorithmic mutual information): Given string-valued random variables X, Y with a recursive probability mass function $P(x, y)$ over pairs (x, y) of strings. We then have

$$I(X;Y) - K(P) \stackrel{+}{\leq} \mathbb{E}(I(x:y)) \stackrel{+}{\leq} I(X;Y) + 2K(P)\,,$$

where $K(P)$ is the length of the shortest prefix-free program that computes $P(x, y)$ from (x, y).

We want to provide an intuition about various aspects of this theorem.

(1) If $I(X;Y)$ is large compared to $K(P)$ the expected algorithmic mutual information is dominated by the statistical mutual information.

(2) If $K(P)$ is no longer assumed to be small, statistical dependences do not necessarily ensure that the knowledge of x allows us to compress y further than without knowing x. It could be that the description of the statistical dependences requires more memory space than its knowledge would save. To see this, consider a distribution $P(X, Y)$ on pairs of binary words that is supported by the ℓ different pairs $(c_1, d_1), \ldots, (c_\ell, d_\ell)$ with

$$(c_j, d_j) \in \{0,1\}^n \times \{0,1\}^n\,.$$

Assume that all these words c_i, d_j are jointly algorithmically independent. If we draw m pairs $(x_1, y_1), \ldots, (x_m, y_m)$ from this distribution with $m \ll \ell$, then all pairs will be different with high probability. Hence, the string $\mathbf{x} := x_1, \ldots, x_m$ will be algorithmically independent of $\mathbf{y} := y_1, \ldots, y_\ell$. If the joint distribution is given, we can use $\mathbf{x}$ to compress $\mathbf{y}$ further than without knowing $\mathbf{x}$ (because the occurrence of $x_j = c_i$ implies $y_j = d_i$). However, first describing $P(X, Y)$ in order to get a better compression of $\mathbf{y}$ would not be economical due to the high description complexity of $P(X, Y)$. Despite the statistical dependence, $\mathbf{x}$ does not help for better compressing $\mathbf{y}$.

(3) On the other hand, knowledge of $\mathbf{x}$ could allow us to compress $\mathbf{y}$ even in the case of a product distribution on X and Y. Consider, for instance, the following case. Let X and Y attain values in the space $\{0,1\}^n$, i.e., binary words of length n. Let $P(X)$ and $P(Y)$ both have point mass on the same string $c \in \{0,1\}^n$ and $K(c)$ not negligible (n must be large). Then $I(X;Y) = 0$ because the joint distribution is obviously given by $P(X, Y) = P(X)P(Y)$. After sampling from $P(X, Y)$ we observe algorithmic dependences between the list of x-values and the list of y-values because they coincide. The algorithmic dependences are thus due to the fact that both variables share the same distribution and that the description length of the latter is significant.

The example (3) is only a very simple instance of a probability distribution whose algorithmic information is significant. We now introduce a more sophisticated family of distributions on $\{0,1\}^n$ with high complexity (generalizing the above case) that we need several times throughout the paper:

Definition 5 (product distributions from strings): Let P_0, P_1 be two probability distributions on $\{0,1\}$ and c be a binary string of length n. Then

$$\mathbf{P}_c := P_{c_1} \otimes P_{c_2} \otimes \cdots \otimes P_{c_n}$$

defines a product measure on $\{0,1\}^n$. We will later also need the following generalization: If $P_{00}, P_{01}, P_{10}, P_{11}$ are four distributions on $\{0,1\}$, and $c, d \in \{0,1\}^n$, then

$$\mathbf{P}_{c,d} := P_{c_1,d_1} \otimes P_{c_2,d_2} \otimes \cdots \otimes P_{c_n,d_n}$$

defines also a family of product measures on $\{0,1\}^n$ that is labeled by two strings c, d.

Denote by $\mathbf{P}_c^{\otimes m}$ the m-fold copy of $\mathbf{P}_c$ from Definition 5. It describes a distribution on $\{0,1\}^{nm}$ assigning the probbaility $\mathbf{P}_c^{\otimes m}(x)$ to $x \in \{0,1\}^{nm}$. If

$$Q(x, y) := \mathbf{P}_c^{\otimes m}(x)\, \mathbf{P}_c^{\otimes m}(y)\,,$$

knowledge of x in the typical case provides knowledge of c, provided that m is large enough. Then we can compress y better than without knowing x because we do not have to describe c a second time. Hence the algorithmic mutual information is large and the statistical mutual information is zero because Q is by construction a product distribution. In other words, algorithmic dependences in a setting with i.i.d. sampling can arise from statistical dependences and from algorithmic dependences between probability distributions.

B. Markov condition for algorithmic dependences among individual objects

Now we state the causal Markov condition for individual objects as a postulate that links algorithmic mutual dependences with causal structure:

Postulate: algorithmic causal Markov condition
Let $x_1, \ldots, x_n$ be n strings representing descriptions of observations whose causal connections are formalized by a directed acyclic graph G with $x_1, \ldots, x_n$ as nodes. Let pa_j be the concatenation of all parents of x_j and nd_j the concatenation of all its non-descendants except x_j itself. Then

$$x_j \perp\!\!\!\perp nd_j \,|pa_j^*\,. \tag{8}$$

As in Definition 4, the appropriate cut-off rate for rejecting G when $I(x_j : nd_j|pa_j^*) > 0$ will not be specified here. Note that the order of concatenating strings into nd_j and pa_j is irrelevant for the above statetements because n is considered constant and there is thus only a constant overhead to describe an ordering of a set of nodes. The postulate is a natural interpretation of the abstract causal Markov condition in terms of algorithmic independences. The only point that remains to be justified is why we condition on pa_j^* instead of pa_j, i.e., why we are given the optimal joint compression of the parent strings. The main reason is that this turns out to yield nice statements on the equivalence of different Markov conditions (in analogy to Lemma 1). Since the differences between $I(x_j : nd_j|pa_j)$ and $I(x_j : nd_j|pa_j^*)$ can only be logarithmic in the string lengths[3] we will not focus on this issue any further.

[3]this is because $K(x|y) - K(x|y^*) = O(\log |y|)$, where $|y|$ denotes the length of y (see [22])

If we apply Postulate 8 to a trivial graph consisting of two disconnected nodes, we obtain the following statement.

Lemma 5 (no algorithmic mutual inf. without causation): If the mutual information $I(x:y)$ between two objects x, y is significantly greater than zero they have some kind of common past.

Here, common past between two objects means that one has causally influenced the other or there is a third one influencing both. The statistical version of this principle is part of Reichenbach's principle of the common cause [32] stating that statistical dependences between random variables[4] X and Y are always due to at least one of the following three types of causal links: (1) X is a cause of Y, or (2) vice versa, or (3) there is a common cause Z. For objects, the term "common past" includes all three types of causal relations. For two texts x and y, for instance, it reads: similarities of x, y indicate that one author has been influenced by the other or that both have been influenced by a third one.

Before we construct a model of causality that makes it possible to prove the causal Markov condition we want to discuss some examples. If one discovers significant similarities in the genome of two sorts of animals one will try to explain the similarities by relatedness in the sense of evolution. Usually, one would, for instance, assume such a common history if one has identified *long* substrings that both animals have in common. However, the following scenario shows two observations that superficially look similar, but nevertheless we cannot infer a common past since their algorithmic complexity is low (implying that the algorithmic mutual information is low, too).

Assume two persons are instructed to write down a binary string of length 1000 and both decide to write the same string $x = 1100100100001111110...$. It seems straightforward to assume that the persons have communicated and agreed upon this choice. However, after observing that x is just the binary representation of π, one can easily imagine that it was just a coincidence that both subjects wrote the same sequence. In other words, the similarities are no longer significant after observing that they stem from a *simple* rule. This shows that the *length* of the pattern that is common to both observations, is not a reasonable criterion on whether the similarities are significant.

To understand the algorithmic causal Markov condition we will study its implications as well as its justification. In analogy to Lemma 1 we have

Theorem 3 (equiv. of algorithmic Markov conditions): Given the strings $x_1, \ldots, x_n$ and a directed acyclic graph G. Then the following conditions are equivalent:

I. **Recursive form:** the joint complexity is given by the sum of complexities of each node, given the optimal compression of its parents:

$$K(x_1, \ldots, x_n) \stackrel{+}{=} \sum_{j=1}^{n} K(x_j | pa_j^*) \qquad (9)$$

II. **Local Markov condition:** Every node is independent of its non-descendants, given the optimal compression of its parents:

$$I(x_j : nd_j | pa_j^*) \stackrel{+}{=} 0 \,.$$

III. **Global Markov condition:**

$$I(S : T | R^*) \stackrel{+}{=} 0$$

if R d-separates S and T.

Below we will therefore no longer distinguish between the different versions and just refer to "the algorithmic Markov condition". The intuitive meaning of eq. (9) is that the shortest description of all strings generated by the causal model is given by describing how to generate every string from its direct causes. A similar kind of "modularity" of descriptions will also occur later in a different context when we consider description complexity of joint probability distributions.

For the proof of Theorem 3 we will need a Lemma that is an analog of the observation that for any two random variables X, Y the statistical mutual information satisfies $I(f(X);Y) \leq I(X;Y)$ for every measurable function f. The algorithmic version is to consider two strings x, y and one string z that is derived from x^* by a simple rule.

Lemma 6 (monotonicity of algorithmic information): Let x, y, z be three strings such that $K(z|x^*) \stackrel{+}{=} 0$. Then

$$I(z : y) \stackrel{+}{\leq} I(x : y) \,.$$

This lemma is a special case of Theorem II.7 in [23]. We will also need the following result:

Lemma 7 (monotonicity of conditional information): Let x, y, z be three strings. Then

$$K(z|x^*) \stackrel{+}{\geq} K(z|(x,y)^*) \,.$$

Note that $K(z|x^*) \stackrel{+}{\geq} K(z|x^*, y)$ and $K(z|x^*) \stackrel{+}{\geq} K(z|x^*, y^*)$ is obvious but Lemma 7 is non-trivial because the star operation is *jointly* applied to x and y.

Proof of Lemma 7: Clearly the string x can be derived from x, y by a program of length $O(1)$. Lemma 6 therefore implies

$$I(z : x) \stackrel{+}{\leq} I(z : x, y) \,,$$

where $I(z : x, y)$ is shorthand for $I(z : (x, y))$. Hence

$$\begin{aligned} K(z) - K(z|x^*) &\stackrel{+}{=} I(z : x) \stackrel{+}{\leq} I(z : x, y) \\ &\stackrel{+}{=} K(z) - K(z|(x,y)^*) \,. \end{aligned}$$

Then we obtain the statement by subtracting $K(z)$ and inverting the sign. □

The following lemma will only be used in Subsection III-C. We state it here because it is closely related to the ones above.

Lemma 8 (generalized data processing inequality): For any three strings x, y, z,

$$I(x : y | z^*) \stackrel{+}{=} 0$$

implies

$$I(x : y) \stackrel{+}{\leq} I(x : z) \,.$$

[4]The original formulation considers actually dependences between events, i.e., binary variables.

The name "data processing inequality" is justified because the assumption $x \perp\!\!\!\perp y\,|z^*$ may arise from the typical data processing scenario where y is obtained from x via z.

Proof of Lemma 8: Using Lemma 7 we have

$$\begin{aligned} K(x|y^*) &\stackrel{+}{\geq} K(x|(z,y)^*) \\ &\stackrel{+}{=} K(x|z,y,K(y,z)) \\ &\stackrel{+}{=} K(x|z,y,K(z)+K(y|z^*)) \\ &\stackrel{+}{\geq} K(x|z,y,K(z),K(y|z^*)) \\ &\stackrel{+}{=} K(x|z^*,y,K(y|z^*))\,, \end{aligned} \tag{10}$$

where the second inequality holds because $K(z)+K(y|z^*)$ can obviously be computed from the pair $(K(z),K(y|z^*))$ by an $O(1)$ program. The last equality uses, again, the equivalence of z^* and $(z,K(z))$. Hence we obtain:

$$\begin{aligned} I(x:y) &\stackrel{+}{=} K(x)-K(x|y^*) \\ &\stackrel{+}{=} K(x|z^*)+I(x:z)-K(x|y^*) \\ &\stackrel{+}{\leq} K(x|z^*)+I(x:z)-K(x|y,K(y|z^*),z^*) \\ &\stackrel{+}{=} I(x:z)+I(x:y|z^*) \stackrel{+}{=} I(x:z)\,. \end{aligned}$$

The first step is by Definition 2, the second one uses Lemma 7, the third step is a direct application of ineq. (10), the fourth one is due to Definition 3, and the last step is by assumption. □

Proof of Theorem 3: I ⇒ III: Define a probability mass function P on $(\{0,1\}^*)^n$ (which formalizes the n-fold cartesian product of the set of strings $\{0,1\}^*$), as follows. Set

$$P(x_j|pa_j) := \frac{1}{z_j} 2^{-K(x_j|pa_j^*)}\,, \tag{11}$$

where z_j is a normalization factor. In this context, it is important that the symbol pa_j on the left hand side refers to conditioning on the k-tuple of strings x_i that are parents of x_j (in contrast to conditional complexities where we can interpret $K(.|pa_j^*)$ equally well as conditioning on *one* string given by the *concatenation* of all those x_i).

$$K(x_j|pa_j^*) \stackrel{+}{=} -\log_2 P(x_j|pa_j)\,.$$

Then we set

$$P(x_1,\ldots,x_n) := \prod_{j=1}^n P(x_j|pa_j)\,, \tag{12}$$

i.e., P satisfies the factorization property with respect to G. It is easy to see that $K(x_1,\ldots,x_n)$ can be determined from P using eq. (2):

$$\begin{aligned} K(x_1,\ldots,x_n) &\stackrel{+}{=} \sum_{j=1}^n K(x_j|pa_j^*) \\ &\stackrel{+}{=} -\sum_{j=1}^n \log_2 P(x_j|pa_j) \\ &= -\log_2 P(x_1,\ldots,x_n)\,. \end{aligned} \tag{13}$$

Remarkably, we can also determine Kolmogorov complexities of *subsets* of $\{x_1,\ldots,x_n\}$ from the corresponding marginal probabilities. We start by proving

$$K(x_1,\ldots,x_{n-1}) \stackrel{+}{=} -\log_2 \sum_{x_n} 2^{-K(x_1,\ldots,x_n)}\,. \tag{14}$$

Note that Kraft's inequality (see [22], Example 3.3.1) implies

$$\sum_x 2^{-K(x|y)} \leq 1\,,$$

for any two strings x and y. On the other hand,

$$\sum_x 2^{-K(x|y)} \geq 2^{-K(x_0|y)}\,,$$

where x_0 is the shortest string allowed in the prefix code. Hence[5]

$$\sum_x 2^{-K(x|y)} \stackrel{\times}{=} 1\,, \tag{15}$$

where $\stackrel{\times}{=}$ denotes equality up to a positive multiplicative constant.

Eq. (15) entails

$$\begin{aligned} \sum_{x_n} 2^{-K(x_1,\ldots,x_n)} &\stackrel{\times}{=} \sum_{x_n} 2^{-K(x_1,\ldots,x_{n-1})-K(x_n|(x_1,\ldots,x_{n-1})^*)} \\ &\stackrel{\times}{=} 2^{-K(x_1,\ldots,x_{n-1})}\,. \end{aligned}$$

Using eq. (13) we obtain eq. (14):

$$\begin{aligned} K(x_1,\ldots,x_{n-1}) &\stackrel{+}{=} -\log_2 \sum_{x_n} 2^{-K(x_1,\ldots,x_n)} \\ &\stackrel{+}{=} -\log_2 \sum_{x_n} P(x_1,\ldots,x_n) \\ &\stackrel{+}{=} -\log_2 P(x_1,\ldots,x_{n-1})\,. \end{aligned}$$

Since the same argument holds for marginalizing over any other variable x_j we conclude that

$$K(x_{j_1},\ldots,x_{j_k}) \stackrel{+}{=} -\log_2 P(x_{j_1},\ldots,x_{j_k})\,, \tag{16}$$

for every subset of strings of size k with $k \leq n$. This follows by induction over $n-k$.

Now we can use the relation between marginal probabilities and Kolmogorov complexities to show that conditional complexities are also given by the corresponding *conditional* probabilities, i.e., for any two subsets $S,T \subset \{x_1,\ldots,x_n\}$ we have

$$K(S|T^*) \stackrel{+}{=} -\log_2 P(S|T)\,.$$

Without loss of generality, set $S := \{x_1,\ldots,x_j\}$ and $T := \{x_{j+1},\ldots,x_k\}$ for $j < k \leq n$. Using eqs. (6) and (16) we

[5]Note that eq. (14) also follows easily from

$$\sum_{x \text{ with } f(x)=y} 2^{-K(x)} \stackrel{\times}{=} 2^{-K(y)} \quad \text{for } K(f)=O(1)\,,$$

(shown in [33], eq. (11f)) by setting $f(z) := z_1,...,z_{n-1}$ if $z = z_1,...,z_n$ and undefined if z is not of this form.

get

$$\begin{aligned} & K(x_1,\ldots,x_j|(x_{j+1},\ldots,x_k)^*) \\ \stackrel{+}{=} & K(x_1,\ldots,x_k) - K(x_{j+1},\ldots,x_k) \\ \stackrel{+}{=} & -\log_2\Big(P(x_1,\ldots,x_k)/P(x_{j+1},\ldots,x_k)\Big) \\ \stackrel{+}{=} & -\log_2 P(x_1,\ldots,x_j|x_{j+1},\ldots,x_k)\,. \end{aligned}$$

Let S,T,R be three subsets of $\{x_1,\ldots,x_n\}$ such that R d-separates S and T. Then $S \perp\!\!\!\perp T\,|R$ with respect to P because P satisfies the recursion (12) (see Lemma 1)[6]. Hence

$$\begin{aligned} K(S,T|R^*) & \stackrel{+}{=} -\log_2 P(S,T|R) \\ & \stackrel{+}{=} -\log P(S|R) - \log_2 P(T|R) \\ & \stackrel{+}{=} K(S|R^*) + K(S|R^*)\,. \end{aligned}$$

This proves algorithmic independence of S and T, given R^* and thus I $\Rightarrow$ III.

To show that III $\Rightarrow$ II it suffices to recall that nd_j and x_j are d-separated by pa_j. Now we show II $\Rightarrow$ I in strong analogy to the proof for the statistical version of this statement in [3]: Consider first a terminal node of G. Assume, without loss of generality, that it is x_n. Hence all strings $x_1,\ldots,x_{n-1}$ are non-descendants of x_n. We thus have $(nd_n,pa_n) \equiv (x_1,\ldots,x_{n-1})$ where $\equiv$ means that both strings coincide up to a permutation (on one side) and removing those strings that occur twice (on the other side). Due to eq. (6) we have

$$K(x_1,\ldots,x_n) \stackrel{+}{=} K(x_1,\ldots,x_{n-1}) + K(x_n|(nd_n,pa_n)^*)\,. \tag{17}$$

Using, again, the equivalence of $w^* \equiv (w,K(w))$ for any string w we have

$$\begin{aligned} & K(x_n|(nd_n,pa_n)^*) \\ \stackrel{+}{=} & K(x_n|nd_n,pa_n,K(nd_n,pa_n)) \\ \stackrel{+}{=} & K(x_n|nd_n,pa_n,K(pa_n)+K(nd_n|pa_n^*)) \\ \stackrel{+}{\geq} & K(x_n|nd_n,pa_n^*,K(nd_n|pa_n^*)) \\ \stackrel{+}{=} & K(x_n|pa_n^*)\,. \end{aligned} \tag{18}$$

The second step follows from $K(nd_n,pa_n) \stackrel{+}{=} K(pa_n) + K(nd_n|pa_n^*)$. The inequality holds because $nd_n,pa_n,K(pa_n)+K(nd_n|pa_n^*)$ can be computed from $nd_n,pa_n^*,K(nd_n|pa_n^*)$ via a program of length $O(1)$. The last step follows directly from the assumption $x_n \perp\!\!\!\perp nd_n\,|pa_n^*$. Combining ineq. (18) with Lemma 7 yields

$$K(x_n|(nd_n,pa_n)^*) \stackrel{+}{=} K(x_n|pa_n^*)\,. \tag{19}$$

Combining eqs. (19) and (17) we obtain

$$K(x_1,\ldots,x_n) \stackrel{+}{=} K(x_1,\ldots,x_{n-1}) + K(x_n|pa_n^*)\,. \tag{20}$$

Then statement I follows by induction over n. □

[6]Since P is, by construction, a discrete probability function, the density with respect to a product measure is directly given by the probability mass function itself.

To show that the algorithmic Markov condition can be derived from an algorithmic version of the functional model in Postulate 3 we introduce the following model of causal mechanisms.

Postulate: algorithmic model of causality
Let G be a DAG formalizing the causal structure among the strings $x_1,\ldots,x_n$. Then every x_j is computed by a program q_j with length $O(1)$ from its parents pa_j and an additional input n_j. We write formally

$$x_j = q_j(pa_j,n_j)\,, \tag{21}$$

meaning that the Turing machine computes x_j from the input pa_j,n_j using the additional program q_j and halts. The inputs n_j are jointly independent in the sense

$$n_j \perp\!\!\!\perp n_1,\ldots,n_{j-1},n_{j+1},\ldots,n_n\,.$$

We could also have assumed that x_j is a *function* f_j of all its parents, but our model is more general since the map defined by the input-output behavior of q_j need not be a total function [22], i.e., the Turing machine simulating the process would not necessarily halt on *all* inputs pa_j,n_j.

The idea to represent causal mechanisms by programs written for some universal Turing machine is basically in the spirit of various interpretations of the Church-Turing thesis. One formulation, given by Deutsch [34], states that every process taking place in the real world can be simulated by a Turing machine. Here we assume that the way different systems influence each other by physical signals can be simulated by computation processes that exchange messages of bit strings.[7]

Note that mathematics also allows us to construct strings that are linked to each other in an *uncomputable* way. For instance, let x be an arbitrary binary string and y be defined by $y := K(x)$. However, it is hard to believe that a real causal mechanism could create such kind of relations between objects given that one believes that real processes can always be simulated by algorithms. These remarks are intended to give sufficient motivation for our model.

The algorithmic model of causality implies the algorithmic causal Markov condition:

Theorem 4 (algorithmic model implies Markov):
Let $x_1,\ldots,x_n$ be generated by the model in eq. (21). Then they satisfy the algorithmic Markov condition with respect to G.

Proof : First we observe that the model class defined by our algorithmic model of causality becomes larger if we assume that every node is computed from $(pa_j,n_j)^*$ instead

[7]Note, however, that sending quantum systems between the nodes could transmit a kind of information ("quantum information" [35]) that cannot be phrased in terms of bits. It is known that this enables completely new communication scenarios, e.g. quantum cryptography. The relevance of quantum information transfer for causal inference is not yet fully understood. It has, for instance, been shown that the violation of Bell's inequality in quantum theory is also relevant for causal inference [36]. This is because some causal inference rules between classical variables break down when the latent factors are represented by *quantum* states rather than being classical variables.

of (pa_j, n_j). While (21) seems more natural from the perspective of interpretation (why should nature have access to the *shortest compression* of (pa_j, n_j) ?), it is remarkable from the mathematical point of view that the proof below only uses the weaker assumption

$$x_j = q'_j((pa_j, n_j)^*)\,, \tag{22}$$

for some $O(1)$-program q'_j.

The arguments below are similar to the proof of Lemma 2: Extend G to a causal structure $\tilde{G}$ with nodes $x_1, \ldots, x_n, n_1, \ldots, n_n$. To see that the extended set of nodes satisfy the local Markov condition w.r.t. $\tilde{G}$, observe first that

$$K(x_j|\widetilde{pa}_j^*) \stackrel{+}{=} 0\,,$$

where $\widetilde{pa}_j := (pa_j, n_j)$ denotes the parents of x_j with respect to $\tilde{G}$. This follows from (22). Hence,

$$x_j \perp\!\!\!\perp \widetilde{nd}_j \,|\widetilde{pa}_j^*\,,$$

if $\widetilde{nd}_j$ denotes the non-descendants of x_j with respect to $\tilde{G}$. Since every n_j is parentless, it remains to show that

$$n_j \perp\!\!\!\perp \widehat{nd}_j\,,$$

if $\widehat{nd}_j$ denotes the non-descendants of n_j. Introducing the notation

$$n_{-j} := n_1, \ldots, n_{j-1}, n_{j+1}, \ldots, n_n\,,$$

we have assumed

$$n_j \perp\!\!\!\perp n_{-j}\,. \tag{23}$$

We now show that the non-descendants of n_j can be obtained from n^*_{-j} via an $O(1)$-program, i.e.,

$$K(\widehat{nd}_j|n^*_{-j}) \stackrel{+}{=} 0\,. \tag{24}$$

Let

$$x_{k_1}, \ldots, x_{k_\ell} := \widehat{nd}_j$$

denote the non-descendants of n_j apart from n_{-j}, written in a causal order. Then every x_{k_i} for $i = 1, \ldots, k_\ell$ is computed from its parents and n_{k_i} via the program q_{k_i}. Hence,

$$K(x_{k_i}|(pa_{k_i}, n_{k_i})^*) \stackrel{+}{=} 0\,.$$

Due to Lemma 7, this implies

$$K(x_{k_i}|(x_{k_1}, \ldots, x_{k_{i-1}}, n_{k_i})^*) \stackrel{+}{=} 0\,. \tag{25}$$

Hence,

$$\begin{aligned} K(x_{k_i}|n^*_{-j}) &\stackrel{+}{\leq} K(x_{k_1}, \ldots, x_{k_i}|n^*_{-j}) \\ &\stackrel{+}{=} K(x_{k_i}|(x_{k_1}, \ldots, x_{k_{i-1}}, n_{k_i})^*) \\ &\quad + K(x_{k_1}, \ldots, x_{k_{i-1}}|n^*_{-j}) \\ &\stackrel{+}{\leq} \sum_{r=1}^{i-1} K(x_{k_r}|n^*_{-j})\,. \end{aligned}$$

The first inequality is obvious, the equality uses eqs. (25) and (6). By induction over i, we thus obtain

$$K(x_{k_i}|n^*_{-j}) \stackrel{+}{=} 0\,,$$

which implies eq. (24). Together with (23), this yields

$$n_j \perp\!\!\!\perp \widehat{nd}_j\,,$$

due to Lemma 6. Hence, the extended set of strings $x_1, \ldots, x_n, n_1, \ldots, n_n$ satisfies the local Markov condition with respect to $\tilde{G}$. By Theorem 3, the extended set of nodes is also *globally* Markovian w.r.t. $\tilde{G}$. The parents pa_j d-separate x_j and nd_j in $\tilde{G}$ (here the parents pa_j are still defined with respect to G). This implies the local Markov condition for G. □

It is trivial to construct examples where the causal Markov condition is violated if the programs n_j are mutually dependent (for instance, the trivial graph with two nodes x_1, x_2 and no edge would satisfy $I(x_1 : x_2) > 0$ if the programs n_1, n_2 computing x_1, x_2 from an empty input are dependent).

The model given by equation (21) can also be interpreted as follows. Each n_j is the description of the mechanism that generates x_j from its parents. This perspective makes apparent that the *mechanisms* that generate causal relations are assumed to be independent. This is essential for the general philosophy of this paper. To see that such a mutual independence of mechanisms is a reasonable assumption we recall that the causal graph is meant to formalize *all* relevant causal links between the objects. If we observe, for instance, that two nodes are generated from their parents by the same complex rule we postulate another causal link between the nodes that explains the similarity of mechanisms.

C. Relation between the postulates

In our presentation the algorithmic Markov condition plays the role of the fundamental postulate. It provides the basis for all causal conclusions that we discuss later. The algorithmic model of causality has only been described to provide an additional justification for the former. Rather than postulating the (algorithmic) causal Markov condition one could also develop the theory as follows. One states the Causal Principle in Lemma 5 as *postulate* and additionally postulates that every causal mechanism is Turing computable in the sense that every effect is computed from its causes pa_j by some program n_j. Recall that the strings n_j then describe the causal mechanisms – if these mechanisms have been designed independently, the joint independence of $n_1, \ldots, n_n$ then follows from the Causal Principle.

All these postulates referring to algorithmic information imply, in an appropriate limit, the corresponding postulates for statistical causal inference: Assume that all strings x_j and n_j represent lists of X_j or N_j values, respectively, after repeated i.i.d. sampling. Assume, moreover, that the Kolmogorov complexity of the joint distribution of $X_1, \ldots, X_n, N_1, \ldots, N_n$ is negligible. Then the algorithmic mutual information reduces to the statistical mutual information (consider conditional versions of Theorems 1 and 2) and the statistical versions of the postulates are implied by the algorithmic ones .

In particular, the algorithmic Markov condition then reduces to the statistical causal Markov condition. However, the former is more fundamental than the latter since it is also applicable

to the case where the i.i.d. assumption is violated or the complexity of the distribution becomes significant.

D. Relative causality

This subsection explains why it is sensible to define algorithmic dependence and the existence or non-existence of causal links *relative* to some background information. To this end, we consider genetic sequences s_1, s_2 of two persons that are not relatives. We certainly find high similarity that leads to a significant violation of $I(s_1 : s_2) = 0$ due to the fact that both genes are taken from humans. However, given the background information "s_1 is a human genetic sequence", s_1 can be further compressed. The same applies to s_2. Let h be a code that is particularly adapted to the human genome in the sense that it minimizes the expected complexity of a randomly chosen human genome, i.e., h minimizes

$$\frac{1}{N}\sum_s K(s|h)\,,$$

where s runs over the genomes of all humans and N is the total size of the human population. Then it would make sense to consider $I(s_1 : s_2|h) > 0$ as a hint for a relation that goes beyond the fact that both persons are human. In contrast, for the unconditional mutual information we expect $I(s_1 : s_2) \geq K(h)$. We will therefore infer some causal relation (here: common ancestors in the evolution) using the Causal Principle in Lemma 5 (cf. [29]).

The common properties between different and unrelated individuals of the same species can be screened off by providing the relevant background information. Given this causal background, we can detect further similarities in the genes by the conditional algorithmic mutual information and take them as an indicator for an additional causal relation that goes beyond the common evolutionary background. For this reason, every discussion on whether there exists a causal link between two objects (or individuals) requires a specification of the background information. In this sense, causality is a relative concept.

One may ask whether such a relativity of causality is also true for the statistical version of the causality principle, i.e., Reichenbach's principle of the common cause. In the statistical version of the link between causality and dependence, the relevance of the background information is less obvious because it is evident that statistical methods are always applied to a *given statistical sample*. If we, for instance, ask whether there is a causal relation between the height and the income of a person without specifying whether we refer to people of a certain age, we observe the same relativity with respect to additionally specifying the "background information", which is here given by referring to a specific sample.

In the following sections we will assume that the relevant background information has been specified and it has been clarified how to translate the relevant aspects of a real object into a binary string such that we can identify every object with its binary description.

III. Novel statistical inference rules from the algorithmic Markov condition

A. Algorithmic independence of Markov kernels

To describe the implications of the algorithmic Markov condition for statistical causal inference, we consider random variables X and Y where X causally influences Y. We can think of $P(X)$ as describing a source S that generates x-values and sends them to a "machine" M that generates y-values according to $P(Y|X)$. Assume we observe that

$$I(P(X) : P(Y|X)) \gg 0\,.$$

Then we conclude that there must be a causal link between S and M that goes beyond transferring x-values from S to M. This is because $P(X)$ and $P(Y|X)$ are inherent properties of S and M, respectively, which do not depend on the current value of x that has been sent. Hence there must be a causal link that explains the similarities in the *design* of S and M. Here we have assumed that we know that $X \to Y$ is the correct causal structure on the *statistical level*. Then we have to accept that a further causal link on the higher level of the *machine design* is present.

If the causal structure on the statistical level is unknown, we would prefer causal hypotheses that explain the data without needing a causal connection on the higher level provided that the hypotheses are consistent with the statistical Markov condition. Given this principle, we thus prefer causal graphs G for which the Markov kernels $P(X_j|PA_j)$ become algorithmically independent. This is equivalent to saying that the shortest description of $P(X_1, \ldots, X_n)$ is given by concatenating the descriptions of the Markov kernels, a postulate that has already been formulated by Lemeire and Dirkx [37] (see also [38]) in a similar form:

Postulate: algorithmic independence of conditionals
A causal hypothesis G (i.e., a DAG) is only acceptable if the shortest description of the joint density P is given by a concatenation of the shortest description of the Markov kernels, i.e.

$$K(P(X_1, \ldots, X_n)) \stackrel{+}{=} \sum_j K(P(X_j|PA_j))\,. \qquad (26)$$

If no such causal graph exists, we reject every possible DAG and assume that there is a causal relation of a different type, e.g., a latent common cause, selection bias, or a cyclic causal structure.

Here we have implicitly assumed that $P(X_1, \ldots, X_n)$ is computable is the sense that it as has computable function $P(x_1, \ldots, x_n)$ as density. We will keep this assumption unless it is explicitly stated. The sum on the right hand side of eq. (26) will be called the *total complexity* of the causal model G. Note that the postulate of algorithmic independence of conditionals implies that we have to reject every causal hypothesis for which the total complexity is not minimal because a model with shorter total complexity already provides a shorter description of the joint distribution. Inferring causal directions by minimizing this expression (or actually a computable modification) could also be interpreted in a Bayesian way if we

consider $K(P(X_j|PA_j))$ as the negative log likelihood for the prior probability for having the conditional $P(X_j|PA_j)$ (after appropriate normalization). However, postulating eq. (26) has implications that go beyond known Bayesian approaches to causal discovery because one can get hints on the incompleteness of the class of models under consideration (in addition to providing rules for giving preference *within* the class). If eq. (26) is violated for all DAGs, none of them can be accepted. One possible explanation is that the set of variables is not causally sufficient because there is a latent common cause.

Lemeire and Dirkx [37] already sketched a relation between causal faithfulness and postulating (26) saying basically that, under appropriate conditions, the algorithmic independence of conditionals implies the causal faithfulness principle. Even though a detailed specification of the conditions has been left to future work, the idea is as follows: Unless the conditionals are very specific (e.g., deterministic relations between causes and effects), violations of faithfulness require mutual adjustments of conditionals in the sense that they jointly satisfy equations that would not hold for generic choices. Therefore, they are algorithmically dependent. Now we want to show that (26) implies causal inference rules that go beyond the known ones.

To this end, we focus again on the example in Subsection I-B with a binary variable X and a continuous variable Y. The hypothesis $X \to Y$ is not rejected because $I(P(X) : P(Y|X)) \stackrel{+}{=} 0$. For the equally weighted mixture of two Gaussians this already follows[8] from $K(P(X)) \stackrel{+}{=} 0$. On the other hand, $Y \to X$ violates (26). Elementary calculations show that the conditional $P(X|Y)$ is given by the sigmoid function

$$P(X = 1|y) = \frac{1}{2}\Big(1 + \tanh \frac{\lambda(y-\mu)}{\sigma^2}\Big)\,. \qquad (27)$$

We observe that the same parameters σ, λ, μ that occur in $P(Y)$ (see eq. (4), also occur in $P(X|Y)$ and both $P(Y)$ and $P(X|Y)$ are complex, see eq. (27). This already shows that the two Markov kernels are algorithmically dependent. To be more explicit, we observe that μ, λ, and σ are required to specify $P(Y)$. To describe $P(X|Y)$, we need λ/σ^2 and μ. Hence we have

$$\begin{aligned} K(P(Y)) &\stackrel{+}{=} K(\mu,\lambda,\sigma) \\ &\stackrel{+}{=} K(\mu) + K(\lambda) + K(\sigma) \\ K(P(X|Y)) &\stackrel{+}{=} K(\mu,\lambda/\sigma^2) \\ &\stackrel{+}{=} K(\mu) + K(\lambda/\sigma^2) \\ K(P(X,Y)) &\stackrel{+}{=} K(P(Y),P(X|Y)) \stackrel{+}{=} K(\mu,\lambda,\sigma) \\ &\stackrel{+}{=} K(\mu) + K(\lambda) + K(\sigma)\,, \end{aligned}$$

where we have assumed that the strings μ, λ, σ are jointly independent. Note that the information that $P(Y)$ is a mixture of two Gaussians and that $P(X|Y)$ is a sigmoid counts as a constant because its description complexity does not depend on the parameters.

We thus get

$$I(P(Y) : P(X|Y)) \stackrel{+}{=} K(\mu) + K(\lambda/\sigma^2)\,.$$

Therefore we reject the causal hypothesis $Y \to X$ because eq. (26) is violated. The interesting point is that we need not look at the alternative hypothesis $X \to Y$. In other words, we do not reject $Y \to X$ *only* because the converse direction leads to simpler expressions. We can reject it alone on the basis of observing algorithmic dependences between $P(Y)$ and $P(X|Y)$ making the causal model suspicious.

The following thought experiment shows that $Y \to X$ would become plausible if we "detune" the sigmoid $P(X|Y)$ by changing slope or offset, i.e., choosing $\tilde{\lambda}, \tilde{\mu}, \tilde{\sigma}$ independently of λ and μ, and σ. Then $P(Y)$ and $P(X|Y)$ are by definition algorithmically independent and therefore we obtain a more complex joint distribution:

$$K(P(X,Y)) = K(\lambda)+K(\mu)+K(\sigma)+K(\tilde{\lambda}/\tilde{\sigma}^2)+K(\tilde{\mu}/\tilde{\sigma}^2)\,.$$

The fact that the set of mixtures of two Gaussians does not have five free parameters already shows that $P(X,Y)$ must be a more complex distribution than the one above. Fig. 4 shows an example of a joint distribution obtained for the "detuned" situation.

As already noted by [37], the independence of mechanisms is related to Pearl's thoughts on the stability of causal statements: the causal mechanism $P(X_j|PA_j)$ does not change if one changes the input distribution $P(PA_j)$ by influencing the variables PA_j. The same conditional can therefore occur, under different background conditions, with different input distributions.

An equivalent of eq. (26) naturally occurs in the probability-free version of the causal Markov condition. To explain this, assume we are given two strings $\mathbf{x}$ and $\mathbf{y}$ of length n (describing two real-world observations) and notice that $\mathbf{x} = \mathbf{y}$. Now we consider two alternative scenarios:

(I) Assume that every pair (x_j, y_j) of digits ($j = 1, \ldots, n$) has been independently drawn from the same joint distribution $P(X,Y)$ of the *binary* random variables X and Y. Hence, X and Y both take values in $\{0,1\}$

(II) Let $\mathbf{x}$ and $\mathbf{y}$ be single instances of string-valued random variables X and Y, i.e., both X and Y take values in $\{0,1\}^n$.

The difference between (I) and (II) is crucial for statistical causal inference: In case (I), statistical independence is rejected with high confidence proving the existence of a causal link. In contrast, there is no evidence for statistical dependence in case (II) since the underlying joint distribution on $\{0,1\}^n \times \{0,1\}^n$ could, for instance, be the point mass on the pair $(\mathbf{x}, \mathbf{y})$, which is a product distribution, i.e.,

$$P(X,Y) = P(Y)P(X)\,.$$

Hence, statistical causal inference would not infer a causal connection in case (II).

Algorithmic causal inference, on the other hand, infers a causal link in both cases because the equality $\mathbf{x} = \mathbf{y}$ requires

[8] for the more general case $P(X = 1) = p$ with $K(p) \gg 0$, this also follows if we assume that p is algorithmically independent of the parameters that specify $P(Y|X)$.

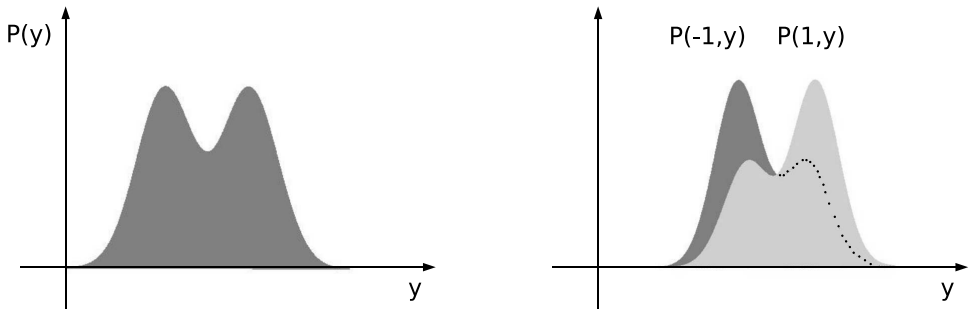

Fig. 4. Left: a source generates the bimodal distribution $P(Y)$. A machine generates x-values according to a conditional $P(X|Y)$ given by the sigmoid (27). If the slope and the position parameters of the sigmoid are not correctly adjusted to the distance, the position, and the width of the two Gaussian modes, the generated joint distribution no longer consists of two Gaussians (right).

an explanation. The relevance of switching between (I) and (II) then consists merely in shifting the causal connection to another level: In the i.i.d setting, every x_j must be causally linked to y_j. In case (II), there must be a connection between the two *mechanisms* that have generated the entire strings because $I(P(X) : P(Y|X)) = I(P(X) : P(Y)) \gg 0$. This can, for instance, be due to the fact that two machines emitting the same string were designed by the same engineer. A detailed discussion of the relevance of translating the i.i.d. assumption into the setting of algorithmic causal inference will be given in Subsection III-B.

Examples with large probability spaces

In the preceding subsection we have ignored a serious problem with defining the Kolmogorov complexity of (conditional) probability distributions that even occurs in finite probability spaces. First of all the "true" probabilities may not be computable. For instance, a coin may produce "head" with probability p where p is some *uncomputable* number, i.e., $K(p) = \infty$. And even if it were some computable value p with large $K(p)$ it is not clear whether one should call the probability distribution $(p, 1-p)$ "complex" because $K(p)$ is high and "simple" if we have, for instance $p = 1/\pi$. A more reasonable notion of complexity can be obtained by describing the probabilities only up to a certain accuracy ϵ. If ϵ is not too small we obtain small complexity values for the distribution of a binary variable, and also low complexity for a distribution on a larger set that is ϵ-close to the values of some simple analytical expression like a Gaussian distribution. There will still remain some unease about the concept of Kolmogorov complexity of "the true distribution". We will subsequently develop a formalism that avoids this concept. However, Kolmogorov complexity of distributions is a useful idea to start with since it provides an intuitive understanding of the roots of the asymmetries between cause and effects that we will describe in Subsection III-B.

Below, we will describe a thought experiment with two random variables X, Y linked by the causal structure $X \to Y$ where the total complexities of the causal models $X \to Y$ and $Y \to X$ both are well-defined and, in the generic case, different. First we will show that they can at most differ by a factor two.

Lemma 9 (maximal complexity quotient):
For every joint distribution $P(X, Y)$ we have

$$K(P(Y)) + K(P(X|Y)) \stackrel{+}{\leq} 2\Big(K(P(X)) + K(P(Y|X))\Big)\,.$$

Proof: Since marginals and conditionals both can be computed from $P(X, Y)$ we have

$$K(P(Y)) + K(P(X|Y)) \stackrel{+}{\leq} 2K(P(X, Y))\,.$$

Then the statement follows because $P(X, Y)$ can be computed from $P(X)$ and $P(Y|X)$. □

To construct examples where the bound in Lemma 9 is attained we first introduce a method to construct conditionals with well-defined complexity:

Definition 6 (Conditionals and joint distributions from strings):

Let M_0, M_1 be two stochastic matrices that specify transition probabilities from $\{0, 1\}$ to $\{0, 1\}$. Then

$$\mathbf{M}_c := M_{c_1} \otimes M_{c_2} \otimes \cdots \otimes M_{c_n}$$

defines transition probabilities from $\{0, 1\}^n$ to $\{0, 1\}^n$.

We also introduce the same construction for double indices: Let $M_{00}, M_{01}, M_{10}, M_{11}$ be stochastic matrices describing transition probabilities from $\{0, 1\}$ to $\{0, 1\}$. Let $c, d \in \{0, 1\}^n$ be two strings. Then

$$\mathbf{M}_{c,d} := M_{c_1,d_1} \otimes M_{c_2,d_2} \otimes \cdots \otimes M_{c_n,d_n}$$

defines a transition matrix from $\{0, 1\}^n$ to $\{0, 1\}^n$. If the matrices M_j or M_{ij} denote joint distributions on $\{0, 1\} \times \{0, 1\}$ the objects $\mathbf{M}_c$ and $\mathbf{M}_{c,d}$ define joint distributions on $\{0, 1\}^n \times \{0, 1\}^n$ in a canonical way.

Let X, Y be variables whose values are binary strings of length n. To define $P(X, Y)$ we first define distributions $P_0(U), P_1(U)$ of a binary random variable U. Moreover, we introduce stochastic matrices A_0, A_1 describing transition probabilities $P_0(V|U)$ and $P_1(V|U)$, respectively, where V is also binary. Then a string $c \in \{0, 1\}^n$ determines, together with P_0 and P_1 given above, a distribution $P(X) := \mathbf{P}_c$ (using Definition 5) that has well-defined Kolmogorov complexity $K(c)$ if the description complexity of P_0 and P_1 is neglected.

Furthermore, for an arbitrary random string $d \in \{0,1\}^n$, we set $P(Y|X) := \mathbf{A}_d$ as in Definition 6, where we have used the canonical identification between stochastic matrices and conditional probabilities. The joint distribution $P(X,Y)$ is then determined by c and d.

To investigate the description length of $P(Y)$ and $P(X|Y)$ we introduce the following notions. Let R_{ij} be short hand for the joint distribution of U, V defined by

$$P_{ij}(U,V) := P_i(U)P_j(V|U)\,.$$

Let Q_{ij} denote the corresponding marginal distribution of V and B_{ij} short hand for the conditional

$$P_{ij}(U|V) = \frac{P_{ij}(U,V)}{P_{ij}(V)}\,.$$

Using these notations and the ones in Definition 6, we obtain

$$\begin{aligned} P(X) &= \mathbf{P}_c \\ P(Y|X) &= \mathbf{A}_d \\ P(X,Y) &= \mathbf{R}_{c,d} \\ P(Y) &= \mathbf{Q}_{c,d} \\ P(X|Y) &= \mathbf{B}_{c,d} \end{aligned} \tag{28}$$

It is noteworthy that $P(Y)$ and $P(X|Y)$ are labeled by both strings while $P(X)$ and $P(Y|X)$ are described by only one string each. This already suggests that the latter are more complex in the generic case.

Now we compare the sum $K(P(X)) + K(P(Y|X))$ to $K(P(Y)) + K(P(X|Y))$ for the case $K(c) \stackrel{+}{=} K(d) \stackrel{+}{=} n$ and $I(c:d) \stackrel{+}{=} 0$. We assume that P_i and A_j are computable and their complexity is counted as $O(1)$ because it does not depend on n. Nevertheless, we assume that P_i and A_j are "generic" in the following sense: All marginals Q_{ij} and conditionals B_{ij} are different whenever $P_0 \neq P_1$ and $A_0 \neq A_1$. If we impose one of the conditions $P_0 = P_1$ and $A_0 = A_1$ or both, we assume that only those marginals Q_{ij} and conditionals B_{ij} coincide for which the equality follows from the conditions imposed. Consider the following cases:

Case 1: $P_0 = P_1$, $A_0 = A_1$. Then all the complexities vanish because the joint distribution does not depend on the strings c and d.

Case 2: $P_0 \neq P_1$, $A_0 = A_1$. Then the digits of c are relevant, but the digits of d are not. Those marginals and conditionals in table (28) that formally depend on c and d, as well as those that depend on c, have complexity n. Those depending on d have complexity 0.

$$\begin{aligned} K(P(X)) + K(P(Y|X)) &\stackrel{+}{=} n+0 = n \\ K(P(Y)) + K(P(X|Y)) &\stackrel{+}{=} n+n = 2n \quad . \end{aligned}$$

Case 3: $P_0 = P_1$, $A_0 \neq A_1$. Only the dependence on d contributes to the complexity. This implies

$$\begin{aligned} K(P(X)) + K(P(Y|X)) &\stackrel{+}{=} 0+n = n \\ K(P(Y)) + K(P(X|Y)) &\stackrel{+}{=} n+n = 2n \quad . \end{aligned}$$

Case 4: $P_0 \neq P_1$ and $A_0 \neq A_1$. Every formal dependence of the conditionals and marginals on c and d in table (28) is a proper dependence. Hence we obtain

$$\begin{aligned} K(P(X)) + K(P(Y|X)) &\stackrel{+}{=} n+n = 2n \\ K(P(Y)) + K(P(X|Y)) &\stackrel{+}{=} 2n+2n = 4n \quad . \end{aligned}$$

The general principle of the above example is very simple. Given that $P(X)$ is taken from a model class that consists of N different elements and $P(Y|X)$ is taken from a class with M different elements. Then the class of possible $P(Y)$ and the class of possible $P(X|Y)$ both can contain $N \cdot M$ elements. If the simplicity of a model is quantified in terms of the size of the class it is taken from (within a hierarchy of more and more complex models), the statement that $P(Y)$ and $P(X|Y)$ are typically complex is just based on this simple counting argument.

Detecting common causes via dependent Markov kernels

The following model shows that latent common causes can yield joint distributions whose Kolmogorov complexity is smaller than $K(P(X)) + K(P(Y|X))$ and $K(P(Y)) + K(X|Y))$. Let X, Y, Z have values in $\{0,1\}^n$ and let $P(Z) := \delta_c$ be the point mass on some random string $c \in \{0,1\}^n$. Let $P(X|Z)$ and $P(Y|Z)$ both be given by the stochastic matrix $A \otimes A \otimes \cdots \otimes A$. Let $P_0 \neq P_1$ be the probability vectors given by the columns of A. Then

$$P(X) = P(Y) = \mathbf{P}_c\,,$$

with $\mathbf{P}_c$ as in Definition 5. Since $P(Z)$ is supported by the singleton set $\{c\}$, the hidden common cause Z does not generate any statistical dependence and we obtain $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$. Thus

$$\begin{aligned} K(P(X)) + K(P(Y|X)) &\stackrel{+}{=} K(P(X|Y)) + K(P(Y)) \\ &\stackrel{+}{=} K(P(X)) + K(P(Y)) \stackrel{+}{=} 2n\,. \end{aligned}$$

On the other hand, we have

$$K(P(X|Z)) + K(P(Y|Z)) + K(P(Z)) \stackrel{+}{=} 0+0+n = n\,.$$

By observing that there is a third variable Z such that

$$K(P(X|Z)) + K(P(Y|Z)) + K(P(Z)) \stackrel{+}{=} K(P(X,Y))\,,$$

we thus have obtained a hint that the latent model is the more appropriate causal hypothesis.

Note that this is an example where both $X \to Y$ and $Y \to X$ can be rejected without checking the alternative hypothesis. Given that we have observed

$$I(P(X) : P(X|Y)) \gg 0 \quad \text{and} \quad I(P(Y) : P(X|Y)) \gg 0$$

we know that both DAGs are wrong even though we may not have observed that the causal structure

$$X \leftarrow Z \rightarrow Y$$

does satisfy the independence condition.

Analysis of the required sample size

The following arguments show that the above algorithmic dependences between the Markov kernels corresponding to the wrong causal hypotheses can already be observed for moderate sample size. Readers who are not interested in technical details may skip the remaining part of the subsection.

Consider first the sampling required to estimate c by drawing i.i.d. from $\mathbf{P}_c$ as in Definition 5. By counting the number of symbols 1 that occur at position j we can guess whether c_j is 0 or 1 by choosing the distribution for which the relative frequency is closer to the corresponding probability. To bound the error probabilities from above set

$$\mu := |P_0(1) - P_1(1)| .$$

Then the probability q that the relative frequency deviates by more than $\mu/2$ decreases exponentially in the number m of copies, i.e., $q \leq e^{-\mu m \alpha}$ where α is an appropriate constant. The probability to have no error for any digit is then bounded from below by $(1 - e^{-\mu m \alpha})^n$. We want to increase m such that the error probability tends to zero. To this end, choose m such that $e^{-\mu m \alpha} \leq 1/n^2$, i.e., $m \geq \ln n^2/(\mu\alpha)$. Hence

$$\left(1 - e^{-\mu m \alpha}\right)^n \geq \left(1 - \frac{1}{n^2}\right)^n \to 1 .$$

Given the information that the probability distribution $P(X)$ factorizes with respect to the digits, the sample size required to estimate it (and determine the string d) thus grows only logarithmically in n.

In the same way, one shows that the sample size needed to distinguish between different conditionals $P(Y|X) = \mathbf{A}_d$ increases only with the logarithm of n provided that $P(X)$ is a strictly positive product distribution on $\{0,1\}^n$. This shows that the high description complexity of a distribution can get relevant even for moderate sample size.

B. Resolving statistical samples into individual observations

The assumption of independent identically distributed random variables is one of the cornerstones of standard statistical reasoning. In this section we show that the *independence* assumption in a typical statistical sample is often due to prior knowledge on causal relations among single objects which can nicely be represented by a DAG. We will see that the algorithmic causal Markov condition then leads to non-trivial implications.

Assume we describe a biased coin toss, m times repeated, and obtain the binary string $x_1, \ldots, x_m$ as result. This is certainly one of the scenarios where the i.i.d. assumption is well justified because we do not believe that the coin changes or that the result of one coin toss influences the other ones. The only relation between the coin tosses is that they refer to the same coin. We will thus draw a DAG representing the relevant causal relations for the scenario where C (the coin) is the common cause of all x_j (see fig. 5).

Given the relevant information on C (i.e., given the probability p for "head"), we have conditional algorithmic independence between the x_j when applying the Markov condition

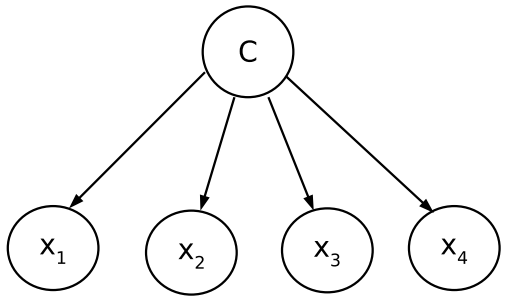

Fig. 5. Causal structure of the coin toss. The statistical properties of the coin C define the common cause that links the results of the coin toss.

to this causal graph.[9] However, there are two problems: (1) it does not make sense to consider algorithmic mutual information among binary strings of length 1. (2) Our theory developed so far (Theorems 3 and 4) considered the number of strings (which is $m+1$ here) as constant and thus even the complexity of $x_1, \ldots, x_m$ is considered as $O(1)$. To solve this problem, we define a new structure with three nodes as follows. For some arbitrary $k < m$ set $\mathbf{x}^1 := x_1, \ldots, x_k$ and $\mathbf{x}^2 := x_{k+1}, \ldots, x_m$. Then C is the common cause of $\mathbf{x}^1$ and $\mathbf{x}^2$ and $I(\mathbf{x}^1; \mathbf{x}^2|C) = 0$ because every similarity between $\mathbf{x}^1$ and $\mathbf{x}^2$ is due to their common source (note that the information that the strings $\mathbf{x}^j$ have been obtained by combining k and $n-k$ results, respectively, is here implicitly considered as background information in the sense of relative causality in Subsection II-D). We will later discuss examples where a source generates symbols from a larger probability space. Then every x_j is a string and it is important to keep in mind the sample size such that the string defined by the concatenation of $x_1, x_2, \cdots, x_m$ cen be decomposed into the original sequence of m strings x_j again. This information will always be considered as background, too.

Of course, we may also consider partitions into more than two substrings keeping in mind that their number is considered as $O(1)$. When we consider causal relations between *short* strings we will thus always apply the algorithmic causal Markov condition to groups of strings rather than applying it to the "small objects" itself. The DAG that formalizes the causal relations between instances or groups of instances of a statistical sample and the source that determines the statistics in the above sense will be called the "resolution of statistical samples into individual observations".

The resolution gets more interesting if we consider causal relations between two random variables X and Y. Consider the following scenario where X is the cause of Y. Let S be a source generating x-values $x_1, \ldots, x_m$ according to a fixed probability distribution $P(X)$. Let M be a machine

[9] This is consistent with the following Bayesian interpretation: if we define a non-trivial prior on the possible values of p, the individual observations are statistically dependent when marginalizing over the prior, but knowing p renders them independent.

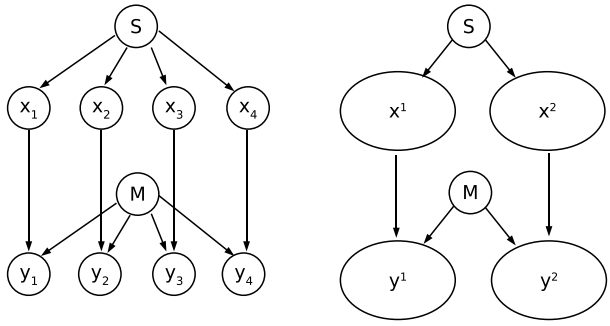

Fig. 6. Left: causal structure obtained by resolving the causal structure $X \to Y$ between the random variables X and Y into causal relations among single events. Right: causal graph obtained by combining the first k observations to $\mathbf{x}^1$ and the remaining $m-k$ to $\mathbf{x}^2$ and the same for Y. We observe that $\mathbf{x}^2$ d-separates $\mathbf{x}^1$ and $\mathbf{y}^2$, while $\mathbf{y}^2$ does not d-separate $\mathbf{y}^1$ and $\mathbf{x}^2$. This asymmetry distinguishes causes from effects.

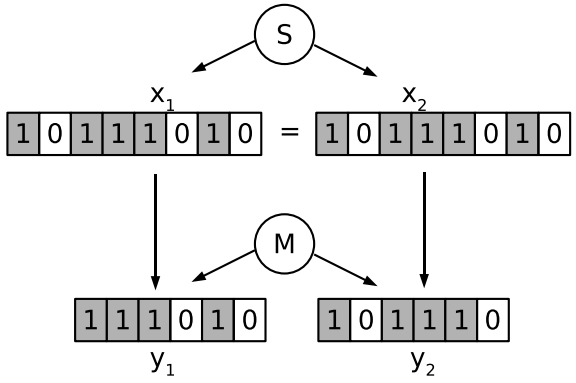

Fig. 7. Visualization of the truncation process: The source S generates always the same string, the machine truncates either the left or the right end. Given only the four strings x_1, x_2 and y_1, y_2 as observations, we can reject the causal hypothesis $Y \to X$. This is because $I(x_1 : y_2|y_1^*)$ can be significantly greater than zero provided that the substrings missing in y_1, y_2 at the left or at the right end, respectively, are sufficiently complex.

that receives these values as inputs and generates y-values $y_1, \ldots, y_m$ according to the conditional $P(Y|X)$. Fig 6 (left) shows the causal graph for $m = 4$.

In analogy to the procedure above, we divide the string $\mathbf{x} := x_1, \ldots, x_m$ into $\mathbf{x}^1 := x_1, \ldots, x_k$ and $\mathbf{x}^2 := x_{k+1}, \ldots, x_m$ and use the same grouping for the y-values. We then draw the causal graph in fig. 6 (right) showing causal relations between $\mathbf{x}^1, \mathbf{x}^2, \mathbf{y}^1, \mathbf{y}^2, S, M$. Now we assume that $P(X)$ and $P(Y|X)$ are not known, i.e., we don't have access to the relevant properties of S and M. Thus we have to consider S and M as "hidden objects" (in analogy to hidden variables in the statistical setting). Therefore we have to apply the Markov condition to the causal structure in such a way that only the observed objects $\mathbf{x}^1, \mathbf{x}^2, \mathbf{y}^1, \mathbf{y}^2$ occur. One checks easily that $\mathbf{x}^2$ d-separates $\mathbf{x}^1$ and $\mathbf{y}^2$ and $\mathbf{x}^1$ d-separates $\mathbf{x}^2$ and $\mathbf{y}^1$. Exhaustive search over all possible triples of subsets of $\mathbf{x}^1, \mathbf{x}^2, \mathbf{y}^1, \mathbf{y}^2$ shows that these are the only non-trivial d-separation conditions. We conclude

$$I(\mathbf{x}^1; \mathbf{y}^2|(\mathbf{x}^2)^*) \stackrel{+}{=} 0 \quad \text{and} \quad I(\mathbf{x}^2; \mathbf{y}^1|(\mathbf{x}^1)^*) \stackrel{+}{=} 0\,. \quad (29)$$

The most remarkable property of eq. (29) is that it is asymmetric with respect to exchanging the roles of X and Y since, for instance, $I(\mathbf{y}^1; \mathbf{x}^2|(\mathbf{y}^2)^*) \stackrel{+}{=} 0$ can be violated. Intuitively, the reason is that given $\mathbf{y}^2$, the knowledge of $\mathbf{x}^2$ provides better insights into the properties of S and M than knowledge of $\mathbf{x}^1$ would do, which can be an advantage when describing $\mathbf{y}^1$. The following example shows that this asymmetry can even be relevant for sample size $m = 2$ provided that the probability space is large.

Let S be a source that always generates the same string $a \in \{0,1\}^n$. Assume furthermore that a is algorithmically random in the sense that $K(a) \stackrel{+}{=} n$. For sample size $m = 2$ we then have $\mathbf{x} = (x_1, x_2) = (a, a)$. Let M be a machine that randomly removes ℓ digits either at the beginning or the end from its input string of length n. By this procedure we obtain a string $y_j \in \{0,1\}^{\tilde{n}}$ with $\tilde{n} := n - \ell$ from x_j.

For sample size 2 it is likely that y_1 and y_2 contain the last $n-\ell$ and the first $n-\ell$ digits of a, respectively, or vice versa. This process is depicted in fig. 7 for $n = 8$ and $\ell = 2$. Since the sample size is only two, the partition of the sample into two halves leads to single observations, i.e., $\mathbf{x}^j = x_j$ and $\mathbf{y}^j = y_j$ for $j = 1, 2$.

In short-hand notation, $\mathbf{y}^1 = a_{[1..n-\ell]}$ and $\mathbf{y}^2 = a_{[\ell+1..n]}$. We then have

$$I(\mathbf{x}^1; \mathbf{y}^2|(\mathbf{x}^2)^*) \stackrel{+}{=} 0 \quad \text{and} \quad I(\mathbf{y}^1; \mathbf{x}^2|(\mathbf{x}^1)^*) \stackrel{+}{=} 0\,,$$

but

$$I(\mathbf{y}^1 : \mathbf{x}^2|(\mathbf{y}^2)^*) \stackrel{+}{=} \ell \quad \text{and} \quad I(\mathbf{x}^1 : \mathbf{y}^2|(\mathbf{y}^1)^*) \stackrel{+}{=} \ell\,,$$

which correctly lets us prefer the causal direction $X \to Y$. This is because these dependences violate the global algorithmic Markov condition in Theorem 3 when applied to a hypothetical graph where $\mathbf{y}^1$ and $\mathbf{y}^2$ are the outputs of the source and $\mathbf{x}^1$ and $\mathbf{x}^2$ are the outputs of a machine that has received $\mathbf{y}^1$ and $\mathbf{y}^2$.

Even though the condition in eq. (29) does not explicitly contain the notion of complexities of Markov kernels it is closely related to the algorithmic independence of Markov kernels. To explain this, assume we would generate algorithmic dependences between S and M by adding an arrow $S \to M$ or $S \leftarrow M$ or by adding a common cause. Then $\mathbf{x}^2$ would no longer d-separate $\mathbf{x}^1$ from $\mathbf{y}^2$. The possible violation of eq. (29) could then be an observable result of the algorithmic dependences between the hidden objects S and M (and their statistical properties $P(X)$ and $P(Y|X)$, respectively).

C. Conditional density estimation on subsamples

Now we develop an inference rule that is even closer to the idea of checking algorithmic dependences of Markov kernels than condition (29), but still avoids the notion of *Kolmogorov complexity of the "true" conditional distributions* by using finite sample estimates instead. Before we explain the idea

we mention two simpler approaches for doing so and describe their potential problems. It would be straightforward to check (26) for the finite sample estimates of the conditionals. In particular, minimum description length (MDL) approaches [39] appear promising from the theoretical point of view due to their close relation to Kolmogorov complexity. We rephrase the minimum complexity estimator described by Barron and Cover [40]: Given a string-valued random variable X and a sample $x_1,\ldots,x_m$ drawn i.i.d. from $P(X)$, set

$$\hat{P}_m := \mathrm{argmin}_Q \Big\{ K(Q) - \sum_{j=1}^{m} \log Q(x_j) \Big\} , \qquad (30)$$

where Q runs over all computable probability densities on the probability space under consideration. If the data is sampled from a computable distribution, then $\hat{P}_m(X)$ converges almost surely to $P(X)$ [40]. Let us define a similar estimator $\hat{P}_m(Y|X)$ for the conditional density $P(Y|X)$. Could we reject the causal hypothesis $X \to Y$ after observing that $\hat{P}_m(X)$ and $\hat{P}_m(Y|X)$ are mutually dependent? In the context of the true probabilities, we have argued that $P(X)$ and $P(Y|X)$ represent independent mechanisms. However, for the estimators we do not see a justification for independence because the relative frequencies of the x-values influence the estimation of $\hat{P}_m(X)$ *and* $\hat{P}_m(Y|X)$. Moreoever, the empirical frequencies $\hat{p}(x)$ and $\hat{p}(y|x)$ are related by the fact that for any fixed x, $\hat{p}(x)$ and all $\hat{p}(y|x)$ are divisible by the number of occurrences of x. Whether similar dependences also hold for the MDL-estimators is unclear.

This problem, however, becomes certainly irrelevant if the sample size is such that the complexities of the estimators coincide with the complexities of the true distributions, but if we assume that the latter are typically uncomputable (because generic real numbers are uncomputable) this sample size will never be attained.

The general idea of MDL [39] also suggests the following causal inference principle: Assume we are given the data points (x_j, y_j) with $j = 1,\ldots,m$ and consider the MDL estimators $\hat{P}_m(X)$ and $\hat{P}_m(Y|X)$. They define a joint distribution that we denote by $\hat{P}_{X\to Y}(X,Y)$ (where we have dropped m for convenience). The total description length

$$\begin{aligned} C_{X\to Y} &:= K(\hat{P}_m(X)) + K(\hat{P}_m(Y|X)) \qquad (31) \\ &- \sum_{j=1}^{m} \log \hat{P}_{X\to Y}(x_j, y_j) \end{aligned}$$

measures the complexity of the probabilistic model plus the complexity of the data, given the model.

The following remarks may provide a better intuition about the expression (31). In "classical MDL", the complexity of an element in a continuously parameterized family of distributions with d parameters is measured as $(d/2)\log m + O(1)$ bits. The first two terms in eq. (31) correspond to the right hand side of eq. (26) and the last term to the logarithm of eq. (2). Then we compare $C_{X\to Y}$ to $C_{Y\to X}$ (defined correspondingly) and prefer the causal direction with the smaller value. If the optimization in eq. (30) and the corresponding one for conditionals is restricted to a set of marginals and conditionals with zero complexity, the method thus reduces to a maximum likelihood fit. On the other hand, if the estimators $\hat{P}(Y|X)\hat{P}(X)$ and $\hat{P}(X|Y)\hat{P}(Y)$ coincide, the method reduces to choosing the direction with smaller algorithmic dependences of conditionals, i.e., the direction that is closer to satisfying (26).

It would be interesting to know whether MDL-based causal inference could also be derived from the algorithmic Markov condition. An interesting conceptual difference between the algorithmic Markov condition and MDL is that the former is in principle able to reject all causal DAGs if all of them violate the algorithmic independence of conditionals.

For this paper, we want to infer causal directions only on the basis of the algorithmic Markov condition and construct an inference rule that uses estimators in a more sophisticated way. Its justification is directly based on applying the algorithmic Markov condition to the resolution of samples as introduced in Subsection III-B. The idea of our strategy is that we do not use the full data set to estimate $P(Y|X)$. Instead, we apply the estimator to a subsample of (x,y) pairs that no longer carries significant information about the relative frequencies of x-values in the full data set. As we will see below, this leads to algorithmically independent *finite sample* estimators for the Markov kernels if the causal hypothesis is correct.

Let $X \to Y$ be the causal structure that generated the data $(\mathbf{x},\mathbf{y})$, with $\mathbf{x} := x_1,\ldots,x_m$ and $\mathbf{y} := y_1,\ldots,y_m$ after m-fold i.i.d. sampling from $P(X,Y)$. The resolution of the sample is the causal graph in fig. 8, left.

According to the model in eq. (21), there are mutually independent programs p_j computing x_j from the description of S. Likewise, there are mutually independent programs q_j computing y_j from M and x_j. Assume we are given a rule how to generate a subsample of $x_1,\ldots,x_m$ from $\mathbf{x}$. It is important that this selection rule does not refer to $\mathbf{y}$ but only uses $\mathbf{x}$ (as well as some random string as additional input) and that the selection can be performed by a program of length $O(1)$. Denote the subsample by

$$\tilde{\mathbf{x}} = \tilde{x}_1,\ldots,\tilde{x}_l := x_{j_1},\ldots,x_{j_l} ,$$

with $l < m$. The above selection of indices defines also a subsample of y-values

$$\mathbf{y} := y_{j_1},\ldots,y_{j_l} := \tilde{y}_1,\ldots,\tilde{y}_l .$$

By construction, we have

$$\tilde{y}_i = q_{j_i}(\tilde{x}_i, M) .$$

Hence we can draw the causal structure depicted in fig. 8, right.

Let now D_X be any string that is derived from $\mathbf{x}$ by some program of length $O(1)$. D_X may be the full description of relative frequencies or any *computable* density estimator $\hat{P}(X)$, or some other description of interesting properties of the relative frequencies. Similarly, let $\tilde{D}_{YX}$ be a description that is derived from $\mathbf{x},\mathbf{y}$ by some simple algorithmic rule. The idea is that it is a computable estimator $\tilde{P}(Y|X)$ for the conditional distribution $P(Y|X)$ or any relevant property of the latter. Instead of estimating conditionals, one may also consider an estimator of the *joint* density of the subsample. We augment the causal structure in fig. 8, right, with D_X and

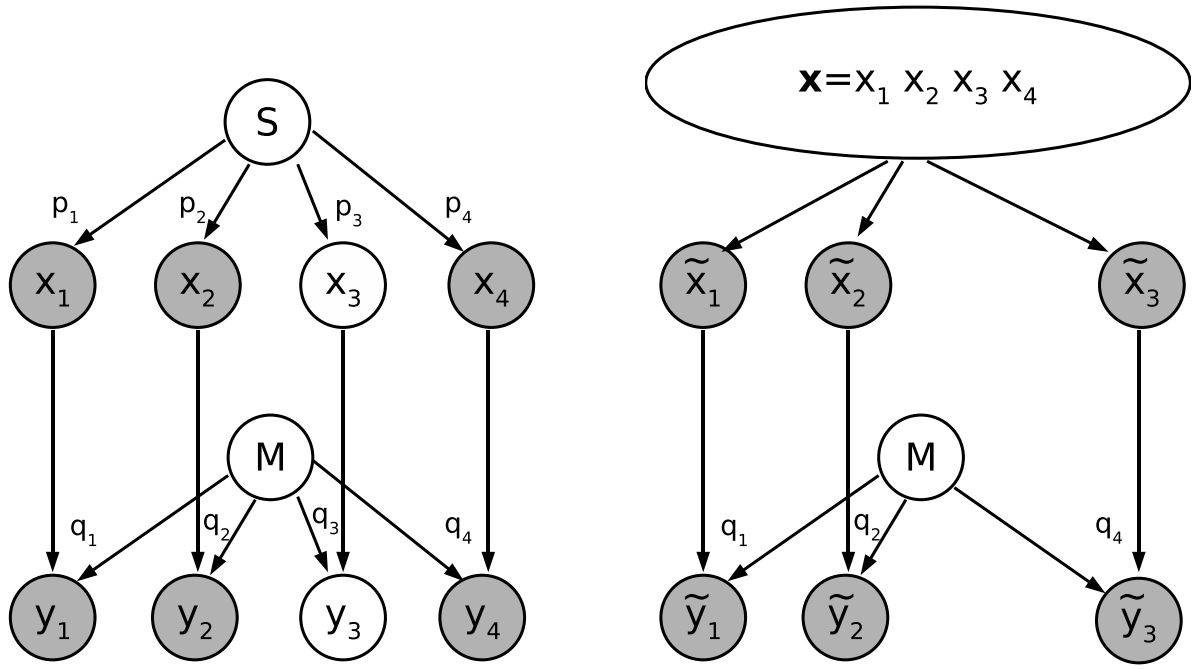

Fig. 8. (left) Causal structure between single observations $x_1, \ldots, x_m$,$y_1, \ldots, y_m$ for sampling from $P(X,Y)$, given the causal structure $X \to Y$. The programs p_j compute x_j from the description of the source S. The programs q_j compute y_j from x_j and the description of the machine M, respectively. The grey nodes are those that are selected for the subsample (see text). Right: Causal structure relating $\mathbf{x}$, $\tilde{x}_j$, and $\tilde{y}_j$. Note that the causal relation between $\tilde{x}_j$ and $\tilde{y}_j$ is the same as the one between the corresponding pair x_j and y_j. Here, for instance, $\tilde{x}_3 = x_4$ and $\tilde{y}_3 = y_4$ and it is thus still the same program q_4 that computes y_4 from x_4 and M. Hence, the causal model that links M with the selected values $\tilde{x}_j$ and $\tilde{y}_j$ is the subgraph of the graph showing relations between x_j, y_j and M. This kind of robustness of the causal structure with respect to the selection procedure will be used below.

$\tilde{D}_{YX}$. The structure can be simplified by merging nodes in the same level and we obtain the structure in fig. 9.

To derive testable implications of the causal hypothesis, we observe that every information between D_X and $\tilde{D}_{YX}$ is processed via $\tilde{\mathbf{x}}$. We thus have

$$\tilde{D}_{YX} \perp\!\!\!\perp D_X \,|\tilde{\mathbf{x}}^* \,, \tag{32}$$

which formally follows from the global Markov condition in Theorem 3. Using Lemma 8 and eq. (32) we conclude

$$I(D_X : \tilde{D}_{YX}) \stackrel{+}{\leq} I(\tilde{\mathbf{x}} : D_X) \,. \tag{33}$$

The intention behind generating the subsample $\tilde{\mathbf{x}}$ is to "blur" the distribution of X in the sense that the subsample does not contain any noteworthy amount of algorithmic information on $P(X)$. If we have a density estimator $\hat{P}(X)$ we try to choose the subsample such that the algorithmic mutual information between $\tilde{\mathbf{x}}$ and $\hat{P}(X)$ is small. Otherwise we have not sufficiently blurred the distribution of X. Then we apply an arbitrary conditional density estimator $\hat{P}(Y|X)$ to the subsample. If there still is a non-negligible amount of mutual information between $\hat{P}_X$ and $\hat{P}(Y|X)$, the causal hypothesis in fig. 6, left, cannot be true and we reject $X \to Y$.

To show that the above procedure can also be applied to data sampled from *uncomputable* probability distributions, let P_0 and P_1 be uncomputable distributions on $\{0,1\}$ and A_0, A_1 uncomputable stochastic maps from $\{0,1\}$ to $\{0,1\}$. Define a string-valued random variable X with distribution $P(X) := \mathbf{P}_c$ as in Definition 5 and the conditional distribution of a string-valued variable Y by $P(Y|X) := \mathbf{A}_d$ as in Definition 6 for strings $c, d \in \{0,1\}^n$. Let P_0 and P_1 as well as A_0 and A_1 be known up to an accuracy that is sufficient to distinguish between them. We assume that all this information (including n) is given as background knowledge, but c and d are unknown. Let $D_X =: \hat{c}$, where $\hat{c}$ is the estimated value of c computed from the finite sample $\mathbf{x}$ of size m. Likewise, let $\tilde{D}_{XY} := \hat{d}$ be the estimated value of d derived from the subsample $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ of size $\tilde{m}$. If m is large enough (such that also $\tilde{m}$ is sufficiently large) we can estimate c and d, i.e, $\hat{c} = c$ and $\hat{d} = d$ with high probability. The most radical method to ensure that $\tilde{\mathbf{x}}$ shares little information with $\mathbf{x}$ and $P(X)$ is the following. Choose some r such that every possible x-value occurs at least r times in $\mathbf{x}$. If $x^1, \ldots, x^\ell$ is the lexicographic order of the ℓ possible x-values, we define

$$\tilde{\mathbf{x}} := \underbrace{x^1 \cdots x^1}_{r} \underbrace{x^2 \cdots x^2}_{r} \cdots \underbrace{x^\ell \cdots x^\ell}_{r} \,.$$

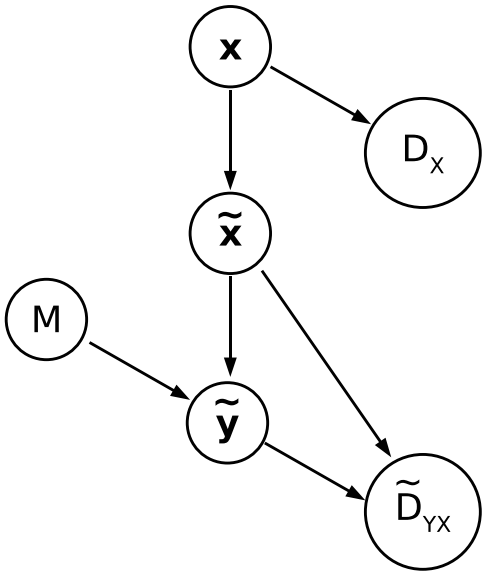

Fig. 9. D_X is some information derived from $\mathbf{x}$. The idea is that it is a density estimator for $P(X)$ or that it describes properties of the empirical distribution of x-values. If the selection procedure $\mathbf{x} \to \tilde{\mathbf{x}}$ has sufficiently blurred this information, the mutual information between $\tilde{\mathbf{x}}$ and D_X is low. D_{XY} on the other hand, is a density estimator for $P(Y|X)$ or it encodes some desired properties of the empirical joint distribution of x- and y-values in the subsample. If the mutual information between D_X and $\tilde{D}_{YX}$ exceeds the one between $\tilde{\mathbf{x}}$ and D_X, we reject the hypothesis $X \to Y$.

For each $j = 1, \ldots, \ell$ we randomly assign each copy of x^j with some index i for which $x_i = x^j$. The string $\tilde{\mathbf{y}}$ is then defined by concatenating the corresponding values y_i. Given the set of possible x-values as background knowledge, the only algorithmic information that $\tilde{\mathbf{x}}$ then contains is the description of r, i.e., $\log_2 r$ bits. Hence we have

$$I(D_X : \tilde{\mathbf{x}}) \stackrel{+}{\leq} \log_2 r\,.$$

Assume now that $c = d$. Then

$$I(D_X : \tilde{D}_{XY}) \stackrel{+}{=} n\,,$$

provided that the estimation was correct. As shown at the end of Subsection III-A, this is already possible for $r = O(\log n)$, i.e.,

$$I(D_X : \tilde{\mathbf{x}}) \in O(\log_2 n)\,,$$

which violates ineq. (33). The importance of this example lies in the fact that $I(P(X) : P(Y|X))$ is not well-defined here because $P(X)$ and $P(Y|X)$ both are uncomputable. Nevertheless, $P(X)$ and $P(Y|X)$ have a computable aspect, i.e, the strings c and d characterizing them. Our strategy is therefore suitable to detect algorithmic dependences between *computable* aspects of *uncomputable* probability distributions.

It is remarkable that the above scheme is general enough to include also strategies for very small sample sizes provided that the probability space is large. To describe an extreme case, we consider again the example with the truncated strings in fig. 7 with the role of X and Y reversed. Let Y be a random variable whose value is always the constant string $a \in \{0,1\}^n$. Let $P(X|Y)$ be the mechanism that generates X by truncating either the l leftmost digits or the l rightmost digits of Y (each with probability $1/2$). We denote these strings by $a_{\rm left}$ and $a_{\rm right}$, respectively. Assume we have two observations $x_1 = a_{\rm left}$, $y_1 = c$ and $x_2 = a_{\rm right}$, $y_2 = a$. We define a subsample by selecting only the first observation $\tilde{x}_1 := x_1 = a_{\rm left}$ and $\tilde{y}_1 := y_1 = a$. Then we define $D_X := x_1, x_2$ and $\tilde{D}_{XY} := y_1$. We observe that the mutual information between $\tilde{D}_{XY}$ and D_X is $K(a)$, while the mutual information between D_X and $\tilde{\mathbf{x}}$ is only $K(a_{\rm left})$. Given generic choices of a, this violates condition (33) and we reject the causal hypothesis $X \to Y$.

D. Plausible Markov kernels in time series

Time series are interesting examples of causal structures where the time order provides prior knowledge on the causal direction. Since there is a large number of them available from all scientific disciplines they can be useful to test causal inference rules on data with known ground truth. Let us consider the following example of a causal inference problem. Given a time series and the prior knowledge that it has been generated by a first order Markov process, but the direction is unknown. Formally, we are given observations $x_1, x_2, x_3, \ldots, x_m$ corresponding to random variables $X_1, X_2, \ldots, X_m$ such that the causal structure is either

$$\cdots \to X_1 \to X_2 \to X_3 \cdots \to X_n \to \cdots\,, \qquad (34)$$

or

$$\cdots \leftarrow X_1 \leftarrow X_2 \leftarrow X_3 \cdots \leftarrow X_n \leftarrow \cdots\,, \qquad (35)$$

where we have extended the series to infinity in both directions.

The question is whether our theory also helps to infer the time direction by using some asymmetry of the joint distribution.[10] Let us assume now that the graph (34) corresponds to the true time direction. Then the hope is that $P(X_{j+1}|X_j)$ is simpler, in some reasonable sense, than $P(X_j|X_{j+1})$. At first glance this seems to be a straightforward extension of the principle of plausible Markov kernel discussed in Subsection III-A. However, there is a subtlety with the justification when we apply our ideas to stationary time series:

Recall that the principle of minimizing the total complexity of all Markov kernels over all potential causal directions has been derived from the independence of the true Markov kernels (remarks after eq. (26)). However, the algorithmic independence of $P(X_j|PA_j) = P(X_j|X_{j-1})$ and $P(X_i|PA_i) = P(X_i|X_{i-1})$ fails spectacularly because stationarity implies that these Markov kernels *coincide* and represent a causal mechanism that is constant in time. This shows that the justification of minimizing total complexity breaks down for stationary time series.

The following argument shows that not only the justification breaks down but also the principle as such: Consider the

[10] [41] describes an asymmetry that sometimes helped to identify the direction in empirical time series. [42] describes a physical toy model that provides a thermodynamical justification, but the relation of these results to the algorithmic Markov condition is not obvious.

case where $P(X_j)$ is the unique stationary distribution of the Markov kernel $P(X_{j+1}|X_j)$. Then we have

$$K(P(X_j|X_{j+1})) \stackrel{+}{\leq} K(P(X_{j+1},X_j)) \stackrel{+}{=} K(P(X_{j+1}|X_j))\,.$$

Because the forward time conditional describes uniquely the backward time conditional (via implying the description of the unique stationary marginal) the Kolmogorov complexity of the latter can exceed the complexity of the former only by a constant term.

We now focus on *non*-stationary time series. To motivate the general idea we first present an example described in [43]. Consider a random walk of a particle on $\mathbb{Z}$ starting at $z \in \mathbb{Z}$. In every time step the probability is q to move one site to the right and $(1-q)$ to move to the left. Let X_j with $j = 0, 1, \ldots$ be the random variable describing the position after step j. Then we have $P(X_0 = z) = 1$. The forward time conditional reads

$$P(x_{j+1}|x_j) = \begin{cases} q & \text{for} \quad x_{j+1} = x_j + 1 \\ 1-q & \text{for} \quad x_{j+1} = x_j - 1 \\ 0 & \text{otherwise}\ . \end{cases}$$

To compute the backward time conditional we first compute $P(X_j)$ which is given by the distribution of a Bernoulli experiment with j steps. Let k denote the number of right moves, i.e., $j-k$ is the number of left moves. With $x_j = k - (j-k) + z = 2k - j + z$ we thus obtain

$$\begin{aligned} P(x_j) &= q^k (1-q)^{j-k} \binom{j}{k} \\ &= q^{(j+x_j-z)/2} (1-q)^{(j-x_j+z)/2} \binom{j}{(j+x_j-z)/2}\,. \end{aligned}$$

Elementary calculations show

$$\begin{aligned} P(x_j|x_{j+1}) &= P(x_{j+1}|x_j) \frac{P(x_j)}{P(x_{j+1})} \\ &= \begin{cases} \frac{(j+x_j-z)/2+1}{j+1} & \text{for} \quad x_j = x_{j+1} - 1 \\ \frac{(j-x_j+z)/2+1}{j+1} & \text{for} \quad x_j = x_{j=1} + 1 \\ 0 & \text{otherwise}\,. \end{cases} \end{aligned}$$

The forward time process is specified by the initial condition $P(X_0)$ (given by z) and the transition probabilities $P(X_j, \ldots, X_1|X_0)$ (given by p). A priori, these two "objects" are mutually unrelated, i.e.,

$$\begin{aligned} &K(P(X_0), P(X_j, X_{j-1}, \ldots, X_1|X_0)) \stackrel{+}{=} \\ &K(P(X_0)) + K(P(X_j, X_{j-1}, \ldots, X_1|X_0)) \stackrel{+}{=} \\ &K(z) + K(q)\,. \end{aligned}$$

On the other hand, the description of $P(X_j)$ (the "initial condition" of the backward time process) alone already requires the specification of *both* z and q. The description of the "transition rule" $P(X_1, \ldots, X_{j-1}|X_j)$ refers only to z. Assuming $z \perp\!\!\!\perp q$, we thus have

$$K(P(X_j)) + K(P(X_0, X_1, \ldots, X_{j-1}|X_j)) \stackrel{+}{=} 2K(z) + K(q)\,.$$

Here, we have set $K(j) \stackrel{+}{=} 0$ because the number of nodes is always considered constant throughout the paper. Hence

$$I(P(X_j) : P(X_0, X_1, \ldots, X_{j-1}|X_j)) \stackrel{+}{=} K(z)\,.$$

The fact that the initial distribution of the hypothetical process

$$X_j \to X_{j-1} \to \cdots \to X_0$$

shares algorithmic information with the transition probabilities makes the hypothesis suspicious.

Resolving time series

We have seen that the algorithmic dependence between "initial condition" and "transition rule" of the backward time process (which would be surprising if it occurred for the forward time process) represents an asymmetry of non-stationary time-series with respect to time reflection. We will now discuss this asymmetry after resolving the statistical sample into individual observations.

Assume we are given m instances of n-tuples $x_1^{(i)}, \ldots, x_n^{(i)}$ with $i = 1, \ldots, m$ that have been i.i.d. sampled from $P(X_1, \ldots, X_n)$ and $X_1, \ldots, X_n$ are part of a time series that can be described by a first order stationary Markov process. Our resolution of a statistical sample generated by $X \to Y$ contained a source S and a machine M. The source generates x-values and the machine generates y-values from the input x. The algorithmic independence of S and M was essential for the asymmetry between cause and effect described in Subsection III-B. For the causal chain

$$\cdots \to X_1 \to X_2 \to X_3 \to \cdots$$

we would therefore have machines M_j generating the x_j-value from x_{j-1}. However, for stationary time-series all M_j are the *same* machine. The causal structure of the resolution of the statistical sample for $m = 2$ is shown in fig. 10, left.

This graph entails no independence constraint that is asymmetric with respect to reversing the time direction. To see this, recall that two DAGs entail the same set of independences if and only if they have the same skeleton (i.e. the corresponding undirected graphs coincide) and the same set of unshielded colliders (v-structures), i.e., substructures $A \to C \leftarrow B$ where A and B are non-adjacent (Theorem 1.2.8 in [1]). Fig. 10 has no such v-structure and the skeleton is obviously symmetric with respect to time-inversion.

The initial part is, however, asymmetric (in agreement with the asymmetries entailed by fig. 6, left) and we have

$$I(x_0^{(1)} : x_1^{(2)} | (x_0^{(2)})^*) \stackrel{+}{=} 0\,.$$

This is just the finite-sample analog of the statement that the initial distribution $P(X_0)$ and the transition rule $P(X_j|X_{j-1})$ are algorithmically independent.

IV. Decidable modifications of the inference rule

To use the algorithmic Markov condition in practical applications we have to replace it with *computable* notions of complexity. The following two subsections discuss different directions along which practical inference rules can be developed.

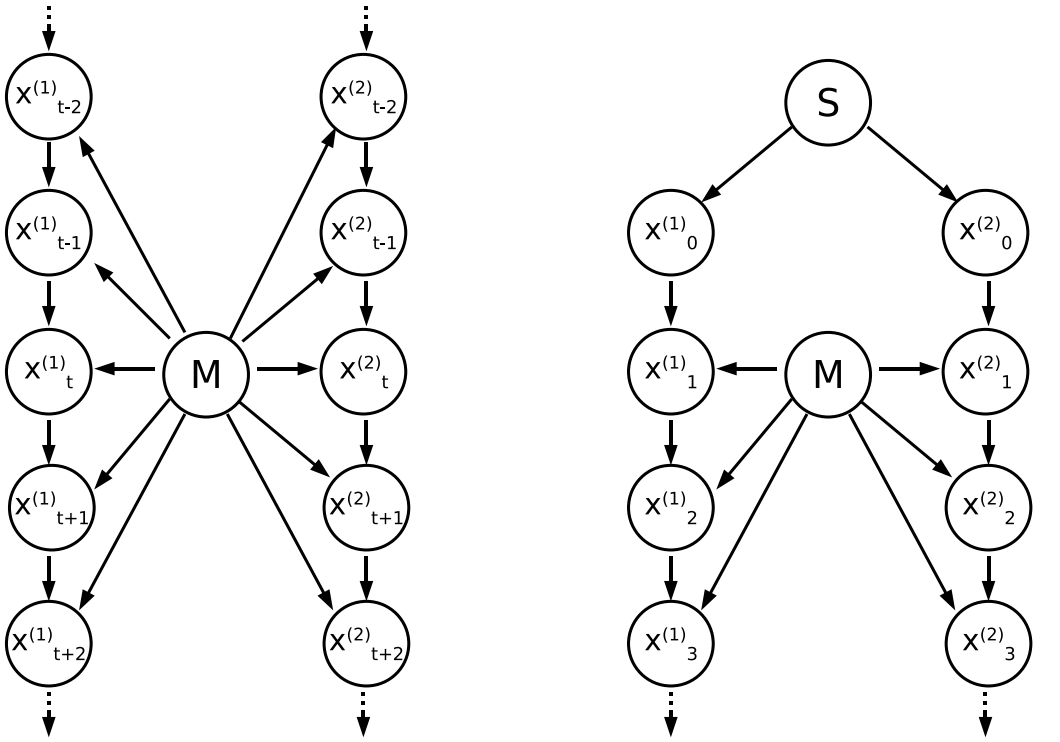

Fig. 10. Left: causal graph of a time series. The values $x_i^{(j)}$ corresponds to the jth instance at time i. Right: the initial part of the time-series is asymmetric with respect to time-inversion.

A. Causal inference using symmetry constraints

We have seen that the algorithmic causal Markov condition implies that the the sum of the Kolmogorov complexities of the Markov kernels must be minimized over all possible causal graphs. In practical applications, it is natural to replace the minimization of Kolmogorov complexity with a decidable simplicity criterion even though this makes the relation to the theory developed so far rather vague. In this subsection we sketch some ideas on how to develop empirically decidable inference rules whose relation to Kolmogorov complexity of conditionals is closer than it may seem at first glance.

Moreover, the example below shows a scenario where the causal hypothesis $X \to Y$ can already be preferred to $Y \to X$ by comparing only the *marginal* distributions $P(X)$ and $P(Y)$ and observing that a simple conditional $P(Y|X)$ leads from the former to the latter but no *simple* conditional leads into the opposite direction. The example will furthermore show why the identification of causal directions is often easier for *probabilistic* causal relations than for *deterministic* ones, a point that has also been pointed out by Pearl [1] in a different context.

Consider the discrete probability space $\{1, \dots, N\}$. Given two distributions $P(X), P(Y)$ like the ones depicted in fig. 11 for $N = 120$. The marginal $P(X)$ consists of k sharp peaks of equal height at positions $n_1, \dots, n_k$ and $P(Y)$ also has k modes centered at the same positions, but with greater width. We assume that $P(Y)$ can be obtained from $P(X)$ by repeatedly applying a doubly stochastic matrix $A = (a_{ij})_{i,j=1,\dots,N}$ with $a_{ii} = 1 - 2p$ for $p \in (0, 1/2)$ and $a_{ij} = p$ for $i = j \pm 1 (\mathrm{mod}\, N)$. The stochastic map A thus defines a random walk and we have by assumption

$$P(Y) = A^m P(X)$$

for some $m \in \mathbb{N}$.

Now we ask which causal hypothesis is more likely: (1) $P(Y)$ has been obtained from $P(X)$ by some stochastic map M. (2) $P(X)$ has been obtained from $P(Y)$ by some stochastic map $\tilde{M}$. Our assumptions already contain an example M that corresponds to the first hypothesis ($M := A^m$). Clearly, there also exist maps $\tilde{M}$ for hypothesis (2). One example would be

$$\tilde{M} := [P(X), P(X), \dots, P(X)]\,, \tag{36}$$

i.e. M has the probability vector $P(X)$ in every column.

To describe in which sense $X \to Y$ is the simpler hypothesis we observe that $\tilde{M}$ in eq. (36) already contains the description of the positions $n_1, \dots, n_k$ whereas $M = A^m$ is rather simple. The Kolmogorov complexity of $\tilde{M}$ as chosen above is for a generic choice of the positions $n_1, \dots, n_k$ given by

$$K(\tilde{M}) \stackrel{+}{=} K(P(Y)) \stackrel{+}{=} \log \binom{N}{k}\,,$$

where $\stackrel{+}{=}$ denotes equality up to a term that does not depend on N. This is because different locations $n_1, \dots, n_k$ of the

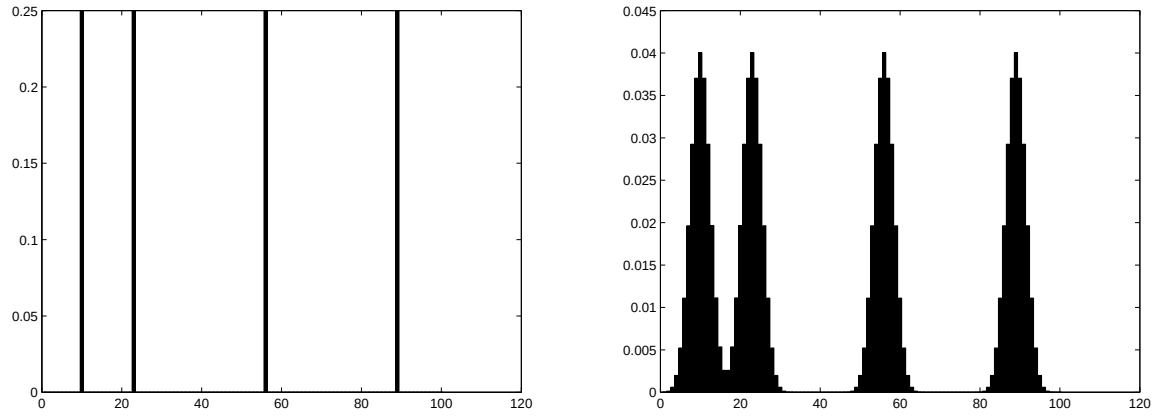

Fig. 11. Two probability distributions $P(X)$ (left) and $P(Y)$ (right) on the set $\{1,\ldots,120\}$ both having 4 peaks at the positions $n_1,\ldots,n_4$, but the peaks in $P(X)$ are well-localized and those of $P(Y)$ are smeared out by a random walk

original peaks lead to different distributions $P(Y)$ and, conversely, every such $P(Y)$ is uniquely defined by describing the positions of the corresponding sharp peaks and M.

However, we want to prove that also other choices of $\tilde{M}$ necessarily have high values of Kolmogorov complexity. We first need the following result.

Lemma 10 (average complexity of stochastic maps): Let $(Q_j(X))_{j=1,\ldots,\ell}$ and $(Q_j(Y))_{j=1,\ldots,\ell}$ be two families of marginal distributions of X and Y, respectively. Moreover, let $(A_j)_{j=1,\ldots\ell}$ be a family of not necessarily different stochastic matrices with $A_jQ_j(Y) = Q_j(X)$. Then

$$\frac{1}{\ell}\sum_{j=1}^{\ell} K(A_j) \geq I(X;J) - I(Y;J)\,, \tag{37}$$

where the information that X contains about the index j is given by

$$I(X;J) := H\Big(\frac{1}{\ell}\sum_j Q_j(X)\Big) - \frac{1}{\ell}\sum_j H(Q_j(X))\,,$$

J denotes the random variable with values j. Here, $H(.)$ denotes the Shannon entropy and $I(Y;J)$ is computed in a similar way as $I(X;J)$ using $Q_j(Y)$ instead of $Q_j(X)$.

Proof: The intuition is the following. If $I(X;J)$ is properly greater than $I(Y;J)$, it is not possible that all A_j coincide because applying a fixed stochastic matrix cannot increase the information about the index variable J due to the data processing inequality. Applying A_j can only increase the information on J by the amount of information that the matrices A_j contain about J. Then the statement follows because the average Kolmogorov complexity of a set of objects cannot be smaller than the entropy of the probability distribution of their occurrence.

To show this formally, we define a partition of $\{1,\ldots,\ell\}$ into d sets $S_1,\ldots,S_d$ for which the A_j coincide. In other words, we have $A_j = B_r$ if $j \in S_r$ and the matrices $B_1,\ldots,B_d$ are chosen appropriately. We define a random variable R whose value r indicates that j lies in the rth equivalence class. The above "data processing argument" implies

$$I(X;J|R) \leq I(Y;J|R)\,. \tag{38}$$

Then we have:

$$\begin{aligned} I(X;J) &= I(X;J,R) = I(X;R) + I(X;J|R) \\ &\leq H(R) + I(Y;J|R) \\ &\leq H(R) + I(Y;J)\,. \end{aligned}$$

The first equality follows because R contains no *additional* information on X (when J is known) since it describes only from which equivalence class j is taken. The second equality is a general rule for mutual information [4]. The first inequality uses $I(X:R) \leq H(R)$. The last inequality follows similar as the equalities in the first line. Let now n_r denote the number of times B_r occurs in the set $\{A_j\}_{j=1,\ldots,\ell}$. The distribution of R is formally defined via $p(r) := n_r/\ell$. Then we have

$$\sum_r p(r)K(B_r) \geq H(R)\,. \tag{39}$$

This follows easily from Jensen's inequality [4] via

$$\sum_r p(r)\log\frac{2^{-K(B_r)}}{p(r)} \leq \log\sum_r 2^{-K(B_r)} \leq 0\,,$$

where the last step uses Kraft's inequality. Recalling that the left hand sides of eq. (39) and ineq. (37) coincide by definition completes the proof. □

To apply Lemma 10 to the above example we define families of $\ell := \binom{N}{k}$ distributions $P_j(X)$ having peaks of equal height at the positions $n_1,\ldots,n_k$ and also their smoothed versions $P_j(Y)$. Mixing all probability distributions will generate the entropy $\log N$ for $P_j(X)$ because we then obtain the uniform distribution. Since we have assumed that $P_j(Y)$ is obtained from $P_j(X)$ by a doubly stochastic map, mixing all $P_j(Y)$ also yields the uniform distribution. Hence the difference between $I(X:J)$ and $I(Y:J)$ is simply given by the average

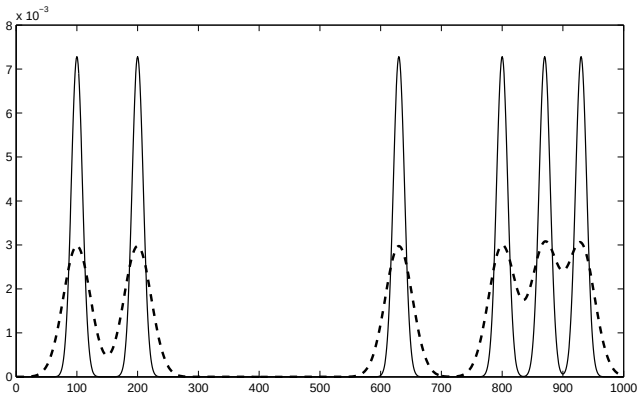

Fig. 12. Two probability distributions $P(X)$ (solid) and $P(Y)$ (dashed) where $P(Y)$ can be obtained from $P(X)$ by convolution with a Gaussian distribution

entropy difference

$$\Delta H := \frac{1}{\ell}\sum_{j=1}^{\ell}\Big(H(P_j(Y)) - H(P_j(X))\Big)\,.$$

The Kolmogorov complexity required to map $P_j(Y)$ to $P_j(X)$ is thus, on average over all j, at least the entropy generated by the double stochastic random walk. Hence we have shown that a typical example of two distributions with peaks at arbitrary positions $n_1,\ldots,n_k$ needs a process $\tilde{M}$ whose Kolmogorov complexity is at least the entropy difference.

One may ask why to consider distributions with several peaks even though the above result will formally also apply to distributions $P_j(X)$ and $P_j(Y)$ with only *one* peak. The problem is that the statement "two distributions have a peak at the same position" does not necessarily make sense for empirical data. This is because the definition of variables is often chosen such that the distribution becomes centralized. The statement that *multiple* peaks occur on seemingly random positions seems therefore more sensible than the statement that *one* peak has been observed at a random position.

We have above used a finite number of discrete bins in order or keep the problem as much combinatorial as possible. In reality, we would rather expect a scenario like the one in fig. 12 where two distributions on $\mathbb{R}$ have the same peaks, but the peaks in the one distribution have been smoothed, for example by an additive Gaussian noise.

As above, we would rather assume that X is the cause of Y than vice versa since the smoothing process is simpler than any process that leads in the opposite direction. We emphasize that *denoising* is an operation that cannot be represented by a *stochastic* matrix, it is a linear operation that can be applied to the whole data set in order to reconstruct the original peaks, the corresponding matrix contains also negative entries. The statement is thus that no simple *stochastic process* leads in the opposite direction. To further discuss the rationale behind this way of reasoning we introduce another notion of simplicity that does not refer to Kolmogorov complexity. To this end, we introduce the notion of translation covariant conditional probabilities:

Definition 7 (translation covariance):
Let X, Y be two real-valued random variables. A conditional distribution $P(Y|X)$ with density $P(y|x)$ is called translation covariant if

$$P(y|x+t) = P(y-t|x) \qquad \forall t \in \mathbb{R}\,.$$

Translation covariant conditionals are always given by convolutions with some probability measure. Therefore, they can never decrease the entropy. Increase of entropy can therefore quantify the amount of non-covariance. We want to describe further options for quantifying non-covariance. To this end, we introduce a concept from statistical estimation theory [44]:

Definition 8 (Fisher information):
Let $P(x)$ be a continuously differentiable probability density of $P(X)$ on $\mathbb{R}$. Then the Fisher information with respect to the translation is defined as

$$F(P(X)) := \int \left(\frac{d}{dx}\ln P(x)\right)^2 P(x)dx\,.$$

Actually, Fisher information is defined for a *family* of distributions. The above expression is obtained by defining the family of shifted densities via $P_t(x) := P(x-t)$.

Then we have the following Lemma (see Lemma 1 in [45] showing the statement in a more general setting that includes also quantum stochastic maps):

Lemma 11 (monotonicity under covariant maps):
Let $P(X,Y)$ be a joint distribution such that $P(Y|X)$ is translation covariant. Then

$$F(P(Y)) \leq F(P(X))\,.$$

The intuition is that F quantifies the degree to which a distribution is non-invariant with respect to translations and that no translation covariant process is able to increase this measure. The convolution with a non-degenerate distribution decreases the Fisher information. Hence there is no translation covariant stochastic map in backward direction.

We can also consider more general symmetries:

Definition 9 (general group covariance):
Let X, Y be random variables with equal range S. Let G be a group of bijections $g : S \to S$ and X^g and Y^g denoting the random variables obtained by permuting the outcomes of the corresponding random experiment according to g. Then we call a conditional $P(Y|X)$ G-covariant if

$$P(Y^g|X)) = P(Y|X^{g^{-1}}) \qquad \forall g \in G\,.$$

It is easy to see that covariant stochastic maps define a quasi-order of probability distributions on S by defining $P \geq \tilde{P}$ if there is a covariant stochastic map A such that $AP = \tilde{P}$. This is transitive since the concatenation of covariant maps is again covariant.

If a G-invariant measure μ ("Haar measure") exists on G we can easily define an information theoretic quantity that measures the degree of non-invariance with respect to G:

Definition 10 (reference information):
Let $P(X)$ be a distribution on S and G be a group of measure

preserving bijections on S with Haar measure μ. Then the reference information is given by:

$$\begin{aligned} I_G &:= H\left(\int_G P\left[X^g\right] d\mu(g)\right) - \int_G H\Big(P(X^g)\Big) d\mu(g) \\ &= H\left(\int_G P\left[X^g\right] d\mu(g)\right) - H(P(X))\,. \end{aligned} \tag{40}$$

The name "reference information" has been used in [46] in a slightly different context where this information occurred as the value of a physical system for communicating a reference system (e.g. spatial or temporal) where G describes, for instance, translations in time or space. For non-compact groups (translations of aperiodic functions), however, there is no averaging operation. For the group $\mathbb{R}$, the Fisher information thus provides a better concept to quantify the non-invariance under the group operation. The quantity I_G can easily be interpreted as mutual information $I(X : Z)$ if we introduce a G-valued random variable Z whose values indicate which transformation g has been applied. One can thus show that I_G is non-increasing with respect to every G-covariant map [46], [47].

The following model describes a link between inferring causal directions by preferring covariant conditionals over non-covariant ones to preferring directions with algorithmically independent Markov kernels. Consider first the probability space $S := \{0,1\}$. We define the group $G := \mathbb{Z}_2 = (\{0,1\}, \oplus)$, i.e., the additive group of integers modulo 2, acting on S as bit-flips or identity. For any distribution on P on $\{0,1\}$, the reference information $I_G(P)$ then measures the asymmetry with respect to bit-flips. Here, the first term on the right hand side of eq. (40) here is the entropy of the uniform distribution and we obtain:

$$I_G(P) = \log 2 - H(P)\,.$$

Hence, the decrease of reference information coincides here with an increase of entropy.

More generally speaking, it can happen for two distributions P and $\tilde{P}$ that a G-symmetric stochastic matrix leads from P to $\tilde{P}$, but only asymmetric stochastic maps convert $\tilde{P}$ into P. To give a more interesting example where the relation to algorithmic information becomes more evident is the following. We consider the group $\mathbb{Z}_2^n$ acting on strings of length n by independent bit-flips. Assume we have a distribution on $\{0,1\}^n$ of the form $\mathbf{P}_c$ in Definition 5 for some string c and generate the distribution $\tilde{\mathbf{P}}_c$ by applying M to $\mathbf{P}_c$ where

$$\begin{aligned} M &:= \begin{pmatrix} 1-\epsilon_1 & \epsilon_1 \\ \epsilon_1 & 1-\epsilon_1 \end{pmatrix} \otimes \begin{pmatrix} 1-\epsilon_2 & \epsilon_2 \\ \epsilon_2 & 1-\epsilon_2 \end{pmatrix} \otimes \cdots \\ &\otimes \begin{pmatrix} 1-\epsilon_n & \epsilon_n \\ \epsilon_n & 1-\epsilon_n \end{pmatrix}, \end{aligned}$$

with $\epsilon_j \in (0,1)$. Then M is G-symmetric, but no G-symmetric process leads backwards. This is because every such stochastic map would be asymmetric in a way that encodes c, i.e., the map would have "to know" c because M has destroyed some amount of information about it.

We summarize the general idea of this subsection as follows. If we observe that $P(Y|X)$ is group-covariant but $P(X|Y)$ is not, we tend to prefer the causal hypothesis $X \to Y$ to $Y \to X$. Rather than justifying such a conclusion by Occam's Razor only, we have described the link to the algorithmic independence of conditionals. For our examples the non-covariance of $P(X|Y)$ expressed the fact that this conditional was adapted to the specific instance of $P(Y)$.

B. Resource-bounded complexity

The problem that the presence or absence of mutual information is undecidable (when defined via Kolmogorov complexities) is similar to statistics, but also different in other respects. Let us first focus on the analogy. Given two real-valued random variables X, Y, it is impossible to show by finite sampling that they are statistically independent. $X \perp\!\!\!\perp Y$ is equivalent to $E(f(X)g(Y)) = E(f(X))E(g(Y))$ for every pair (f,g) of measurable functions. If we observe significant correlations between $f(X)$ and $g(Y)$ for some pair defined in advance, it is well-justified to reject independence. The same holds if such correlations are detected for f, g in some sufficiently small set of functions (cf. [48]) that was defined in advance. However, if this is not the case, we can never be sure that there is not some pair of arbitrarily complex functions f, g that are correlated with respect to the true distribution. Likewise, if we have two strings x, y and find no simple program that computes x from y this does not mean that there is no such a rule. Hence, we also have the statement that there can be an algorithmic dependence even though we do not find it.

However, the difference to the statistical situation is the following. Given that we have found functions f, g yielding correlations it is only a matter of the statistical significance level whether this is sufficient to reject independence. For algorithmic dependences, we do not even have a decidable criterion to reject *in*dependence. Given that we have found a simple program that computes x from y, it still may be true that $I(x : y)$ is small because there may also be a simple rule to generate x (which would imply $I(x : y) \approx 0$) that we were not able to find. This shows that we can neither show dependence nor independence.

One possible answer to these problems is that Kolmogorov complexity is only an idealization of empirically decidable quantities. Developing this idealization only aims at providing hints in which directions we have to develop practical inference rules. Compression algorithms have already been developed that are intended to approximate, for instance, the algorithmic information of genetic sequences [49], [50]. Chen et al. [50] constructed a "conditional compression scheme" to approximate conditional Kolmogorov complexity and applied it to the estimation of the algorithmic mutual information between two genetic sequences. To evaluate to which extent methods of this kind can be used for causal inference using the algorithmic Markov condition is an interesting subject of further research.

It is also noteworthy that there is a theory on *resource-bounded* description complexity [22] where compressions of x are only allowed if the decompression can be performed within a previously defined number of computation steps and

on a tape of previously defined length. An important advantage of resource-bounded complexity is that it is computable. The disadvantage, on the other hand, is that the mathematical theory is more difficult. Parts of this paper have been developed by converting statements on statistical dependences into their algorithmic counterpart. The strong analogy between statistical and algorithmic mutual information occurs only for complexity with unbounded resources. For instance, the symmetry $I(x:y) \stackrel{+}{=} I(y:x)$ breaks down when replacing Kolmogorov complexity with resource-bounded versions [22]. Nevertheless, to develop a theory of inferred causation using *resource-bounded* complexity could be a challenge for the future. There are several reasons to believe that taking into account computational complexity can provide additional hints on the causal structure:

Bennett [51], [52], [53], for instance, has argued that the *logical depth* of an object echoes in some sense its history. The former is, roughly speaking, defined as follows. Let x be a string that describes the object and s be its shortest description. Then the logical depth of x is the number of time steps that a parallel computing device requires to compute x from s. According to Bennett, large logical depth indicates that the object has been created by a process that consisted of many non-trivial steps. This would mean that there also is some causal information that follows from the time-resources required to compute a string from its shortest description.

The time-resources required to compute one observation from the other also plays a role in the discussion of causal inference rules in [43]. The paper presents a model where the conditional

$$P(\text{effect}|\text{cause})$$

can be *efficiently* computed, while computing

$$P(\text{cause}|\text{effect})$$

is NP-hard. This suggests that the computation time required to use information of the cause for the description of the effect can be different from the time needed to obtain information on the cause from the effect. However, the goal of the present paper was to describe asymmetries between cause and effect that even occur when computational complexity is ignored.

V. Conclusions

We have shown that our algorithmic causal Markov condition links algorithmic dependences between single observations (i.e., when the statistical sample size is one) with the underlying causal structure. This is similar to the way the statistical causal Markov condition links statistical dependences among random variables to the causal structure. The algorithmic Markov condition has implications on different levels:

(1) In conventional causal inference one can drop the assumption that observations

$$(x_1^{(i)}, \ldots, x_n^{(i)})$$

have been generated by *independent* sampling from a constant joint distribution

$$P(X_1, \ldots, X_n)$$

of n random variables $X_1, \ldots, X_n$. Algorithmic information theory thus replaces statistical causal inference with a probability-free formulation.

(2) Causal relations among individual objects can be inferred provided their shortest descriptions are sufficiently complex. Then the Kolmogorov complexities must be estimated, e.g., by the compression length with respect to appropriate compression schemes.

(3) New statistical causal inference rules follow because causal hypotheses are suspicious if the corresponding Markov kernels are algorithmically dependent. Remarkably, this criterion can in principle be used to reject a causal hypothesis without comparing it to another hypothesis. If all causal DAGs under consideration are rejected because independence of conditionals holds for none of the causal directions, the model class is too small (e.g. one has to account for hidden common causes rather than assuming that the observed variables are causally sufficient).

Since algorithmic mutual information is uncomputable because Kolmogorov complexity is uncomputable, we have discussed some ideas on how to develop inference rules that are motivated by the uncomputable idealization.

Acknowledgements

The authors would like to thank Bastian Steudel for helpful comments.

References

[1] J. Pearl. *Causality*. Cambridge University Press, 2000.
[2] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Lecture Notes in Statistics. Springer, New York, 1993.
[3] S. Lauritzen, A. Dawid, B. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–505, 1990.
[4] T. Cover and J. Thomas. *Elements of Information Theory*. Wileys Series in Telecommunications, New York, 1991.
[5] S. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, New York, Oxford Statistical Science Series edition, 1996.
[6] C. Meek. Strong completeness and faithfulness in Bayesian networks. Proceedings of *11th Uncertainty in Artificial Intelligence (UAI), Montreal, Canada*, Morgan Kaufmann, pages 411–418, 1995.
[7] D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 141–165, Cambridge, MA, 1999. MIT Press.
[8] Omnès, R. *The interpretation of quantum mechanics*. Princeton Series in Physics. Princeton University Press, 1994.
[9] J. Mooij and D. Janzing. Distinguishing between cause and effect. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 6:147–146, 2010.
[10] X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 1–11, Fort Lauderdale, FL, 2006.
[11] X. Sun, D. Janzing, and B. Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing*, 71:1248–1256, 2008.
[12] H. Reichenbach. *The Philosophy of Space and Time*. Dover, 1958.
[13] W. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, 1984.
[14] Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, Tokyo, Japan, 2003.
[15] C.-H. Bennett, M. Li, and B. Ma. Chain letters and evolutionary histories. *Scientific American*, 288(6):76–81, 2003.

[16] D. Hofheinz, J. Müller-Quade, and R. Steinwandt. On IND-CCA security modeling in cryptographic protocols. *Tatra Mt. Meth. Publ.*, 33:83–97, 2006.

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

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

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

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

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

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

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

[24] G. Chaitin. Toward a mathematical definition of life. In R. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 477–498. MIT Press, Cambridge, MA, 1997.

[25] P. Grünwald and P. Vitányi. Kolmogorov complexity and information theory. With an interpretation in terms of questions and answers . *J. Logic, Language, and Information*, 12(4):497–529, 2003.

[26] C.-H. Bennett, P. Gács, M. Li, P. Vitányi, and W. Zurek. Information distance. *IEEE Trans. Inf. Th.*, IT-44:4:1407–1423, 1998.

[27] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi. The similarity metric. *IEEE Trans. Inf. Th.*, IT-50:12:3250–3264, 2004.

[28] M. Mahmud. On universal transfer learning. *Theoretical Computer Science*, 410(19):1826–1846, 2009.

[29] A. Milosavljević and J. Jurka. Discovery by minimal length encoding: a case study in molecular evolution. *Machine Learning*, 12:69–87, 1993.

[30] P. Hanus, Z. Dawy, J. Hagenauer, and J. Mueller. DNA classification using mutual information based compression distance measures. In *Proceedings of the 14th International Conference of Medical Physics, Biomedizinische Technik, vol. 50 supp. vol. 1, part 2*, pages 1434–1435, Nürnberg, Germany, 2005.

[31] A. Brudno. Entropy and the complexity of the trajectories of a dynamical system. *Transactions of the Moscow Mathematical Society*, 2:127–151, 1983.

[32] H. Reichenbach. *The direction of time*. University of California Press, Berkeley, 1956.

[33] M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.

[34] D. Deutsch. Quantum theory, the Church-Turing Principle and the universal quantum computer. *Proceedings of the Royal Society*, Series A(400):97–117, 1985.

[35] M. Nielsen and I. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[36] D. Janzing and T. Beth. On the potential influence of quantum noise on measuring effectiveness of drugs in clinical trials. *Int. Journ. Quant. Inf.*, 4(2):347–364, 2006.

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

[38] J. Lemeire and K. Steenhaut. Inference of graphical causal models: Representing the meaningful information of probability distributions. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 6:107–120, 2010.

[39] P. Grünwald. *The minimum description length principle*. MIT Press, Cambridge, MA, 2007.

[40] A. Barron and T. Cover. Minimum complexity density estimation. *IEEE Trans. Inf. Theory*, 37(4):1034–1054, 1991.

[41] J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. Detecting the direction of causal time series. In *Proceedings of the International Conference on Machine Learning, Montreal, ACM International Conference Proceeding Series*, volume 382, pages 801–808, New york, NY, USA, 2009. http://www.cs.mcgill.ca/∼icml2009/papers/503.pdf,http://portal.acm.org/citation.cfm?doid=1553374.1553477.

[42] D. Janzing. On the entropy production of time series with unidirectional linearity. *Journ. Stat. Phys.*, 138:767–779, 2010.

[43] D. Janzing. On causally asymmetric versions of Occam's Razor and their relation to thermodynamics. `http://arxiv.org/abs/0708.3411v2`, 2008.

[44] H. Cramér. *Mathematical methods of statistics*. Princeton University Press, Princeton, 1946.

[45] D. Janzing and T. Beth. Quasi-order of clocks and their synchronism and quantum bounds for copying timing information. *IEEE Trans. Inform. Theor.*, 49(1):230–240, 2003.

[46] J. Vaccaro, F. Anselmi, H. Wiseman, and K. Jacobs. Complementarity between extractable mechanical work, accessible entanglement, and ability to act as a reference frame, under arbitrary superselection rules. `http://arxiv.org/abs/quant-ph/0501121`.

[47] D. Janzing. Quantum thermodynamics with missing reference frames: Decompositions of free energy into non-increasing components. *J. Stat. Phys.*, 125(3):757–772, 2006.

[48] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[49] S. Grumbach and F. Tahi. A new challenge for compression algorithms: genetic sequences. *Information Processing & Management*, 30(6), 1994.

[50] X. Chen, Kwong X., and M. Li. A compression algorithm for DNA sequences and its applications in genome comparison. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, RECOMB 2000, Tokyo, Japan. ACM*, page 107, 2000.

[51] C.-H. Bennett. How to define complexity in physics and why. In W. Zurek, editor, *Complexity, Entropy, and the Physics of Information*, volume VIII of *Santa Fee Studies of Complexity*, pages 137–148. Adosin-Wesley, 1990.

[52] C.-H. Bennett. On the nature and origin of complexity in discrete, homogeneous, locally-interacting systems. *Foundations of Physics*, 16(6):585–592, 1986.

[53] C.-H. Bennett. Logical depth and physical complexity. In R. Herken, editor, *The Universal Turing Machine - a Half-Century Survey*, pages 227–257. Oxford University Press, 1988.