

On causally asymmetric versions of Occam's Razor and their relation to thermodynamics

Dominik Janzing*

School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816-2362

August 24, 2007

Abstract

In real-life statistical data, it seems that conditional probabilities for the *effect given their causes* tend to be less complex and smoother than conditionals for causes, given their effects. We have recently proposed and tested methods for causal inference in machine learning using a formalization of this principle.

Here I try to provide some theoretical justification for causal inference methods based upon such a “causally asymmetric” interpretation of Occam's Razor. To this end, I discuss toy models of cause-effect relations from classical and quantum physics as well as computer science in the context of various aspects of complexity, such as exponential hierarchies of probability distributions and computational complexity.

I argue that this asymmetry of the statistical dependences between cause and effect has a thermodynamic origin. The essential link is the tendency of the environment to provide independent background noise realized by physical systems that are *initially* uncorrelated with the system under consideration rather than being *finally* uncorrelated.

*e-mail: janzing@ira.uka.de

This link extends ideas from the literature relating Reichenbach’s principle of the common cause to the second law.

1 Causal reasoning from statistical data

The investigation of non-deterministic causal relation between observed quantities relies on the evaluation of statistical dependences and correlations in empirical data. Two types of statistical data have to be carefully distinguished: in so-called *experimental* data, one observes the change of the distribution of one variable after interventions that control the value of the other. More often, one has to evaluate *non-experimental* data where no controlling intervention by the researcher is possible and he tries to draw causal conclusions merely from observed dependences in the statistics. Causal reasoning that relies on non-experimental data is likely to lead to serious misconclusions. The main obstacle is that statistical dependences between two random variables X and Y may stem from three types of (non-exclusive) causal relations. First, X may be a cause of Y , second, Y may be a cause of X , or third, there may be a (latent) common cause, i.e., a hidden variable Z effecting X and Y (“principle of the common cause” [1]). If the variables X and Y are time-ordered and X refers to observations that precedes the observation of Y it is still hard to decide whether X effects Y or there is a so-called confounding variable Z . However, it is known that the joint distribution of at least 3 variables provides some hints on causal directions [1, 2, 3] by analyzing *conditional* independences among variables. For instance, if the stochastic dependence between X and Y is only generated by some common cause Z (see Fig. 1, left), the variable X and Y must be independent with respect to the conditional probability given Z . On the other hand, if Z is a common effect of X and Y (see Fig. 1, right), the role of *unconditional* probabilities and *conditional* probabilities is in some sense reversed: the conditional probability given Z would then, in the generic case, generate dependences between the (actually independent) variables X and Y . Common effects cannot be accepted as an explanation for (unconditional) dependences but common causes can. Already Reichenbach [1] argued that this stochastic asymmetry with respect to reversing causal arrows is linked to the thermodynamic arrow of time. Before I describe another asymmetry between cause and effect that we have observed to be useful in causal reasoning and discuss

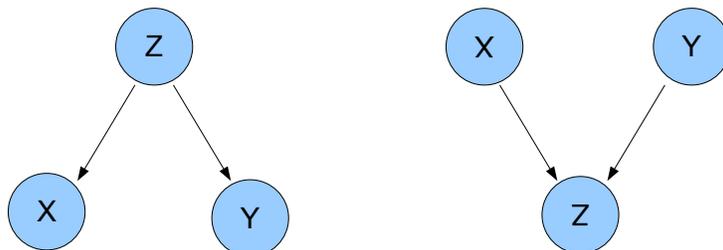


Figure 1: Causal fork (left) and a collider (right). The simplest example of the statistical asymmetry between cause and effect (see text).

its relation to statistical physics I first sketch the known approaches to causal inference from empirical data.

Following [4, 2] we restrict our attention to causal structures without feedback loops and describe a causal structure as a directed acyclic graph (DAG) with random variables X_1, \dots, X_n as nodes. An arrow from X_i to X_j indicates that X_i directly effects X_j , where “directly” means that the effect is not intermediated by some of the other $n - 2$ variables occurring in the model. I am aware of the fact that a definition of cause and effect would require deep philosophical discussions [3, 5]. However, here I will define causal relations by referring to *hypothetical* interventions. The variables X influences Y whenever adjusting X (by external control) to some different values x changes the distribution of Y (here and in the following we will capitalize random variables and denote their values by small letters). The influence from X_i on X_j is *direct* (relative to the set X_1, \dots, X_n) whenever the change of the distribution of X_j caused by different adjustments of X_i occurs also when all the other variables are fixed by an additional intervention. This definition, even though philosophically disputable, makes clear that causal inference from non-experimental data infers probability distributions of *hypothetical experimental data*. The essential link between the statistics of non-experimental observations and the causal graph (encoding information about the effect of hypothetical interventions) is the causal Markov condition.

Definition 1 (Causal Markov condition)

Let G be a DAG with n random variables X_1, \dots, X_n as nodes. A joint distribution P on these variables satisfies the Markov condition with respect to G if each variable is, given its parents, conditionally independent from all the other variables except from its descendants.

We can then factorize P into conditional probabilities for each variable, given its parents:

$$P(x_1, \dots, x_n) = \prod_{j=1}^n P(x_j | pa_j), \quad (1)$$

where pa_j is a short notation for the subset of values x_1, \dots, x_n that correspond to the parents of X_j (this set of variables is denoted by PA_j). Conversely, an arbitrary choice of the "Markov kernels" $P(X_j | PA_j)$ leads to a Markovian distribution.

Following [3] we accept a causal hypothesis only if the observed statistical dependences are consistent with the Markov condition and mention also that this can be justified by so-called *functional models*: For each node X_j we introduce an additional noise variable S_j and assume that the actual value x_j of X_j is a deterministic function of S_j and all parents of X_j . Then the Markov condition follows given that the noise variables are assumed to be stochastically independent. *Dependent* noise variables must explicitly be included as common causes.

There are at least $n!$ causal graphs for which all measures P are Markovian, namely every *complete* DAG (that is, a graph where each X_i has an arrow to X_j for every $j > i$ if some arbitrary order of nodes is given). One needs therefore additional inference rules. So-called constraint-based approaches to causal inference [2, 3] are based upon the assumption that the observed probability distribution should be faithful with respect to the causal graph. This condition is defined as follows.

Definition 2 (Causal faithfulness condition)

A joint probability distribution P on n random variables X_1, \dots, X_n is faithful with respect to G if only those statistical independences are true which are implied by the Markov condition.

The idea is the following: Given a DAG G and an independent choice of values for the free parameters $P(x_j | pa_j)$ it is unlikely to obtain a non-faithful graph.

It is more natural to assume that an independence relation holds because it is entailed by the causal structure than that it is due to *specific adjustments* of the parameters $P(x_j|pa_j)$. Arguments of this kind are justified by referring to “Occam’s Razor” [3]. However, faithfulness leads rarely to a unique causal graph. More often there are still several possible causal hypotheses. One example is to determine the causal direction between two variables.

Causal inference based on faithfulness is impossible if no conditional independence is true. Therefore, additional inference rules are desirable. In seeking new methods one must be aware of the fact that no inference principle can always lead to correct results since there is in principle no method to infer causal relations from non-experimental data that is always reliable. This is seen from the following argument. Given a DAG G and a Markovian joint distribution P , we consider n random generators, each corresponding to one node of a causal graph. We assume that the output distributions of these generators are individually tunable in the following sense: The number of inputs of generator j coincides with the number of parents of node j and whenever the input is pa_j the generator samples random values according to the distribution $P(X_j|pa_j)$. When the inputs are connected according to the causal graph the network will generate data according to the joint distribution P .

Bayesian methods define another approach to causal discovery [6]. One defines priors on the causal graphs G and the corresponding Markov kernels and computes a posteriori probabilities for each G . Even though the faithfulness principle is not used in an explicit way, one obtains an implicit preference for faithful structures provided that the priors are positive densities on the space of all $P(x_j|pa_j)$ [7] (for more general remarks on the implicit tendency of Bayesian methods to respect Occam’s Razor see [8]).

Recent proposals for alternative causal inference methods are based on the observations that in many cases $P(x_j|pa_j)$ are quite complex functions for some hypothetical causal directions and simple for others [9, 10]. These approaches have successfully generated causal hypotheses in cases where constrained-based approaches failed. The idea is, roughly speaking, that conditional probability of the effect, given all its causes has in general less complex and has a smoother shape than the probability of one of the causes given the effect and the other causes. We have proposed [11] to use this approach for post selection of causal hypotheses after constraint-based algorithms have already reduced the set of potential causal graphs. But also

Bayesian approaches could benefit from evaluating smoothness of conditionals since this could provide better priors for $P(X_j|PA_j)$.

The intention of this paper is to further support “asymmetric Occam’s Razor principles” by discussing simple models from quantum and classical physics where simple interactions among physical systems tend to generate simple conditionals $P(\text{effect}|\text{cause})$ rather than generating simple expressions for $P(\text{cause}|\text{effect})$ and show that this asymmetry is closely related to the arrow of time in thermodynamically irreversible processes.

The paper is organized as follows. In Section 2 we sketch the inference rule proposed in [10] and our approach to define smoothness of probability distributions by constrained maximization of conditional entropy. Section 3 describes two simple physical experiments that are consistent with our inference rules. This provides a first intuition about what kind of assumptions lead to the statistical asymmetry. Section 4 recalls how to define a hierarchy of conditional distributions with increasing complexity and describes a toy model of a causal relation between thermodynamic systems where simplicity of causal conditionals (in the sense of this hierarchy) follow from the simplicity of interaction Hamiltonians. In Section 5 we consider a thermal non-equilibrium system consisting of two subsystems with different temperatures and a separation of time scales where one system influences the other with almost no back action. In agreement with our inference rule, we obtain simple conditionals $P(\text{effect}|\text{cause})$ but not vice versa. In this example, a definite causal direction between the two coexisting systems requires a temperature gradient. Motivated by this insight, Section 6 discusses whether causal unidirectionality necessarily requires non-equilibrium systems. In Section 7 I discuss the relation between the asymmetry of cause and effect in statistics and the arrow of time in the statistical physics of mixing processes. For doing so, I describe models where the microphysical dynamics generating the causal relations is reversible and the time asymmetry of the whole scenario stems only from the assumption that the relevant systems *start* in a product distribution but *end up* with statistical dependences rather than starting with statistical dependences which are exactly resolved by the reversible dynamics. In all these scenarios, the asymmetry in the statistical dependence between cause and effect stems exactly from this asymmetry. In Section 8 I consider the computational complexity of conditional probabilities connecting input and output of a boolean circuit with additional noise and argue that $P(\text{effect}|\text{cause})$ can efficiently be computed but $P(\text{cause}|\text{effect})$

cannot provided that the inputs are independent. I describe how this asymmetry is linked to the thermodynamics of computation. In Section 9 the de Finetti theorem is revisited from the perspective of justifying the abundance of stochastically independent noise variables that are crucial in this paper and I discuss to what extent this idea suggests to modify the principle of the common cause.

The idea that models in forward (time and causal) direction tend to be simpler than in backward direction, is certainly not new. The underlying intuition has influenced human and automated reasoning [12] since a long time. Psychological studies indicate that human intuition is better in estimating the strength of causal links (which is encoded in causal conditionals $P(\text{effect}|\text{causes})$) than in inferring non-causal conditionals [13] (see also remarks in Section 2). For this reason, it is straightforward that simplicity principles (“Occam’s Razor”) are automatically interpreted as simplicity of a model when described in the correct causal direction. However, I am not aware of any systematic exploration of the theoretical background of causally asymmetric interpretations of Occam’s Razor from a statistical physics point of view.

2 The principle of plausible Markov kernels and its motivation

To explain our inference principle, we consider (without loss of generality) *complete* DAGs. They are given by an arbitrary ordering of the n variables and drawing an arrow from each variables to every other that appears later in the order. Then the causal hypotheses are uniquely characterized by one out of $n!$ possible orderings of the nodes (“causal ordering”). The *Markov kernels* corresponding to such a hypothetical causal order X_1, \dots, X_n are defined as the conditional probabilities $P(X_j|X_1, X_2, \dots, X_{j-1})$. Here and in the following we will not explicitly distinguish between probabilities and probability densities since the distinction should be clear from the context. The venue of our discussion will be the following vague formulation of our inference rule.

Definition 3 (Principle of plausible Markov kernels (PplMk)) Prefer the hypothetical causal order X_1, \dots, X_n for which the corresponding

Markov kernels $P(X_j|X_1, \dots, X_{j-1})$ are as simple and smooth as possible.

How to define smoothness and simplicity in a reasonable way is, however, a difficult problem. As a first attempt, which provided some encouraging results, we have chosen the following definition [10].

Definition 4 (Simplest conditional)

The simplest non-trivial conditionals $P(X_j|X_1, \dots, X_{j-1})$ are those that maximize the conditional entropy of X_j given X_1, \dots, X_{j-1} subject to some given first and second moments $E(X_i)$ and $E(X_j X_i)$ for $i = 1, \dots, j$.

Here $E(\cdot)$ denotes the expectation of a random variable. The definition generalizes in a straightforward way to vector-valued variables [10] using first and second moments of all coefficients $X_i^{(l)}$ of the vector X_i . Then our inference rule reads:

Definition 5 (First implementation of PplMk)

Estimate the first and second moments $E(X_i)$ and $E(X_i X_j)$ from the data set. For all hypothetical causal orders X_1, \dots, X_n compute the simplest conditionals $P(X_j|X_1, \dots, X_{j-1})$ in the sense of Definition 3 by maximizing conditional entropies subject to these moments. Decide by appropriate statistical tests for which ordering the obtained joint measure provides the best fit to the observed data.

Now we describe two instances where the principle is very intuitive. First we consider two random variables X and Y where X is binary, i.e., its value set is $\{0, 1\}$ and the value set of Y is \mathbb{R} . Assume that X effects Y . The most plausible distribution $P(X)$ is just the distribution given by the observed relative frequencies because the distribution of a binary variable is already determined by its expected value. The smoothest Markov kernel $P(Y|X)$ according to our definition can be shown to be (after constrained entropy maximization using Lagrange multipliers):

$$P(y|x) = \exp \left(- (y - a - bx)^2 - c \right), \tag{2}$$

with appropriate constants a, b, c . In other words, we have for each x a Gauss distribution where the two values of x determine the position of the Gauss functions [10]. Indeed, after having observed that the marginal distribution

of Y is a mixture of two Gaussians and that the conditionals $P(y|x)$ are simple Gaussians it seems very plausible to assume that X effects Y and not vice versa.

Now we consider the reverse situation where Y effects X . The most plausible marginal distribution of Y is, according to our principle, the Gauss distribution. For $P(X|Y)$ we obtain [10]

$$P(x|Y = 1) = \frac{1}{2} \left(1 + \tanh(ax + b) \right) \quad (3)$$

with appropriate $a, b \in \mathbb{R}$. The principle is less trivial when the variables are vector-valued and the value set of a variable is some subset of \mathbb{R}^k . In section 3 we will analyze examples from physics with one binary and one continuous variable that are consistent with the above plausible Markov kernels. I have decided to choose examples from quantum mechanics for two reasons. First, the quantum world provides us with natural realizations of binary variables. Second, the simplicity of the models under consideration is intriguing. Nevertheless, quantum *superpositions* are not relevant for the arguments in the next subsections.

However, the inference rule in [10] should only be considered as a preliminary attempt to formalize simplicity. In section 4 we will discuss more flexible definitions. But first I should mention which different aspects of simplicity may be of interest.

Aspects of simplicity

The plausible Markov kernels defined above are simple with respect to the following two criteria:

- **Simplicity of the dependence on the effect.** Each function $x_j \mapsto P(x_j|x_1, \dots, x_{j-1})$ is simple, e.g., in the sense of being a smooth function.
- **Simplicity of the dependence on the causes.** Each function $x_i \mapsto P(x_j|x_1, \dots, x_{j-1})$ is simple for all $i = 1, \dots, j - 1$.

Now we will describe another aspect of simplicity that does not fit into these two categories. Consider a random walk on \mathbb{Z} (the set of integers)

starting at position 0. In every step we move either one site to the left or one site to the right with probability $1/2$ each and stop after n steps. Accordingly, we define the random variables X_1, \dots, X_n with values in \mathbb{Z} describing the position after step $1, \dots, n$. The causal structure of the walk is certainly given by the linear directed graph

$$X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n. \quad (4)$$

The corresponding conditionals for every variable, given its parent node read:

$$P(X_1 = \pm 1) = 1/2 \quad P(x_j | x_{j-1}) = \begin{cases} 1/2 & \text{for } |x_j - x_{j-1}| = 1 \\ 0 & \text{otherwise} \end{cases}.$$

The conditional independences entailed by the causal structure (4) are also consistent with the reverse causal hypothesis

$$X_n \rightarrow X_{n-1} \rightarrow \dots \rightarrow X_1. \quad (5)$$

To see this, recall that we only need to check the Markov condition (Definition 1). Given its parent X_{j+1} , every X_j must be conditionally independent of all its non-descendants (except from its parent), i.e., the variables $X_{j+2}, X_{j+3}, \dots, X_n$. Using the d-separation criterion in [3] one can easily show that this follows from the Markov condition corresponding to the true causal structure (4). Thus, the joint probability admits the factorization

$$P(x_1, \dots, x_n) = P(x_n)P(x_{n-1}|x_n)P(x_{n-2}|x_{n-1}) \cdots P(x_1|x_2).$$

The conditionals $P(x_{j-1}|x_j)$ are of course “simple” in the sense that they vanish for every pair x_{j-1}, x_j for which $|x_{j-1} - x_j| \neq 1$. Smoothness of $x_{j-1} \mapsto P(x_{j-1}|x_j)$ and smoothness of $x_j \mapsto P(x_{j-1}|x_j)$ makes little sense for functions that are only non-zero for two arguments. However, the conditionals are less simple in the sense that $P(x_{j-1}|x_j)$ *depends on j* . To see this, assume that we are on position ℓ after ℓ steps. Then the position after step $\ell - 1$ was definitely $\ell - 1$. In other words, the two cases $x_{j-1} - x_j = 1$ and $x_{j-1} - x_j = -1$ are not equally likely for the backward time conditional and the bias depends on j .

From the psychological point of view, it is remarkable that one is tempted to think that the backward time conditional would be the same for this example as the forward time conditional. This has been confirmed by conversations with my students. This is consistent with a remark in the introduction:

Our intuition seem to evaluate the simplicity of a model according to the simplicity of *causal* conditionals because we do not even recognize when a model is complex in the converse direction.

The random walk represents another aspect of simplicity hat is not taken into account in any of our inference rules proposed so far. Nevertheless, I state it for the sake of completeness:

- **Simplicity of the dependence on the nodes.** The function $j \mapsto P(X_j|X_1, \dots, X_{j-1})$ is simple.

Due to the thermodynamic spirit of this paper it is worth mentioning that the discussed time asymmetry is “fading away” after many steps. This is because

$$P(X_j = \ell, X_{j-1} = m) = \begin{cases} P(X_{j-1} = m)/2 & \text{for } |\ell - m| = 1 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Since we have $P(X_{j-1} = m) \approx P(X_{j-1} = m \pm 1)$ for large j the expression on the left hand side of Eq. (6) becomes asymptotically symmetric with respect to exchanging ℓ and m . To obtain a strictly time-symmetric analogue, consider a random walk on a cycle consisting of N sites. If the initial position is completely unknown, i.e., $P(x_1) = 1/N$ for all $x_1 \in \{0, 1 \dots, N - 1\}$, the process is perfectly symmetric with respect to time inversion.

In the next section we will focus on physical situations that make apparent in what sense smoothness of conditionals has its origins in (1) the smoothness of dynamical laws, (2) the fact that the energy is a smooth function in the natural physical coordinates, and (3) the fact that the physical system which provides background noise and the system under consideration are more likely to be *initially* uncorrelated than to be *finally* uncorrelated.

3 Examples from real physics

Stern-Gerlach experiment

Consider first an experiment like the one designed by Stern and Gerlach in 1922 [14] to prove the quantization of the magnetic moment. A beam of atoms is emitted from a furnace and enters an inhomogeneous magnetic field perpendicular to the beam (see Fig. 3, here the field is in vertical direction¹),

¹Diagram drawn by Theresa Knott, taken from the free encyclopedia *wikipedia*

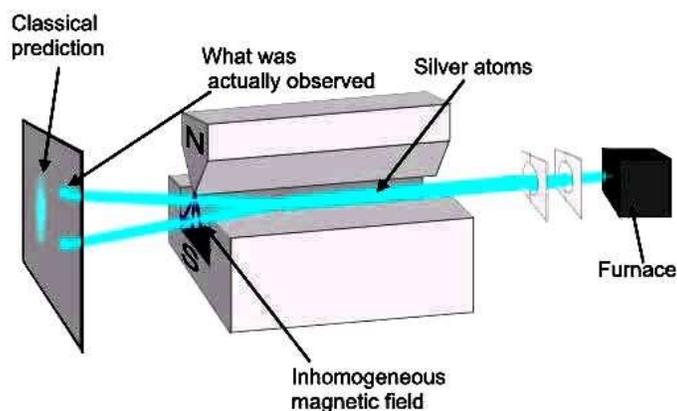


Figure 2: Stern-Gerlach experiment. The atom beam emitted by the furnace splits up into two beams according to the spin.

The field induces a force in the direction of its gradient which is proportional to the magnetic moment of the particles. For spin-1/2 particles, for instance, the magnetic moment can attain the values $+1/2, -1/2$ causing forces in opposite vertical directions. This effect can be used as a measurement apparatus for the quantum observable *magnetic moment* since it separates the beam into two parts that hit the screen at different vertical positions. We consider the values $\pm 1/2$ as the two values of a binary variable X . Even though quantum mechanical observables are in general not random variables on a probability space, this is well-justified because the quantum superpositions is already incoherent (by creating entanglement with the position degree of freedom) from the moment on where the beam *begins* to split up. We define furthermore a random variable Y for the vertical coordinate of the point where the atom hits the screen. It is natural to assume that the conditional probabilities $P(Y|X = \pm 1/2)$ are both not too different from normal distributions. The following extremely simplified model would, for instance, lead to such a normal distribution. Before the particles have left the source they are subjected to some focusing forces. For simplicity we re-

strict our attention to the focus in vertical direction and assume that the forces are induced by a harmonic potential in vertical direction. In thermal equilibrium, the probability distribution of momenta in a classical as well as in a quantum harmonic oscillator is Gaussian (see Section 4 and [15], respectively). Assuming that the probability distribution of the particle momenta is still Gaussian when they leave the source we obtain for both spin values Gauss distributions for Y with different expected values.

Even though the terms *cause* and *effect* are even more philosophically problematic when quantum effects come into play, we claim that X influences Y : if we subject a spin measurement to the particles before the beam passes the inhomogeneous field and remove all atoms with spin down, for instance, we get only one branch of the beam. Given the simplified assumptions above, the Markov kernel $P(Y|X)$ coincides with the plausible kernel in eq. (2).

Assume now, we had observed a Gaussian *marginal* distribution for Y instead of Gaussian *conditionals* and a conditional $P(X|Y)$ as in eq. (3) Our inference rule would then assume that Y is the cause. For this reason we want to check whether there are modifications of the Stern-Gerlach experiment which keep the causal direction but generate such a distribution. We could, for instance, assume that the transversal potential in the furnace is strongly anharmonic such that the particle momenta p are distributed according to some probability measure $Q(Z)$ after the particles have left the source (when referring to p as a random variable we describe it by Z in order to avoid the variable name “ P ”). Due to the laws of motion, we assume that y will for both spin values be a linear function in z :

$$Y = aZ + b_+ \quad \text{and} \quad Y = aZ + b_- .$$

Here b_{\pm} denotes the shift of the expected values caused by the magnetic moments of particles with spin $X = \pm 1/2$ and $a \in \mathbb{R}$ is some constant. In order to obtain Gaussian marginals for Y , the distribution $Q(Z)$ would be such that the convex sum of $Q(Z)$ and its shifted copy is Gaussian. To see that this is impossible we recall that the Gauss measure could then be written as a convolution of $Q(Z)$ with a measure μ that is supported by two points. Hence the Fourier transform of μ multiplied with the Fourier transform of $Q(Z)$ would be the Fourier transform of a Gaussian which is again a Gaussian (up to a phase function). But this is in contradiction to the fact that the Fourier transform of μ has zeros. This shows that Gaussian marginals for

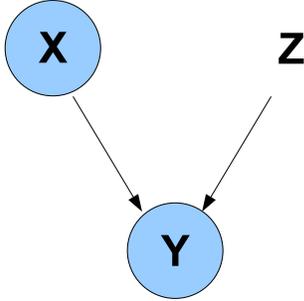


Figure 3: Graphical model of the causal structure of the Stern-Gerlach experiment (see text), where the initial momentum represents the noise (that is not explicitly considered in the causal structure $X \rightarrow Y$).

Y cannot be obtained by choosing an “odd” potential only. We would also have to modify the laws of motion given by the magnetic field. One could, for instance, have a field with strongly inhomogeneous field gradient such that atoms with different transversal momenta enter locations with different field gradient.

Fig. 3 shows a simplified graphical model of the causal structure: the position Y is here assumed to be a deterministic function $f(x, z)$ of the binary variable spin X and a “noise” variable Z , the initial momentum. Smoothness of the conditional $P(Y|X)$ is here due to the smoothness of f and the smoothness of the distribution of momenta. Last but not least, we should stress the decisive assumption that spin and *initial* momenta are stochastically independent.

Spin in a stationary magnetic field

Now we present an example where a continuous classical variable Y influences a discrete variable X . Given a spin-1/2 particle subjected to a field in z -direction whose (randomly fluctuating) strength is represented by the random variable Y . The binary variable X represents here the possible outcomes $X = \pm 1/2$ for a spin measurement in z direction. They occur in thermal

equilibrium with the Boltzmann probabilities, i.e., we have

$$P(X = 1/2 | y) = \frac{\exp(-ay)}{\exp(-ay) + 1} = \frac{1}{2} \left(1 + \tanh \frac{-ay}{2} \right), \quad (7)$$

with an appropriate constant a containing Boltzmann's constant k , temperature and the magnetic moment. This is because the density operator of a quantum system with Hamiltonian H and temperature T is given by

$$\rho = \frac{1}{\text{tr}(e^{-\frac{1}{kT}H})} e^{-\frac{1}{kT}H}. \quad (8)$$

The conditional probability for the effect X given the cause Y is then exactly the one that we have considered as the most plausible one (see eq. (3)).

For a $d/2$ -spin system having the $d + 1$ possible values $j = -d/2, -d/2 + 1, \dots, d/2$ for the spin in a given direction, our approach for the “most plausible kernel” yields conditionals of the form

$$P(X = j | y) = \frac{\exp(-ajy + bj^2)}{\sum_{l=-d/2}^{d/2} \exp(-aly + bl^2)},$$

(with appropriate constants a, b) as “plausible” conditional distributions. This parametric family contains the physically correct Boltzmann probabilities

$$P(X = j | y) = \frac{\exp(-ajy)}{\sum_{l=-d/2}^{d/2} \exp(-alx)},$$

by setting $b = c = 0$.

As in the Stern-Gerlach experiment we try to modify the setup (for spin $1/2$) such that the same causal mechanism leads to a distribution that would actually be the most plausible one for the *opposite* causal direction. Then $P(Y)$ would be a mixture of two Gaussians with equal variance σ^2 but different expected values m_{\pm} , i.e.,

$$P(y) = \frac{1}{\sqrt{2\pi}\sigma} \times \left[P(X = 1/2) \exp\left(-\frac{(y - m_+)^2}{2\sigma^2}\right) + P(X = -1/2) \exp\left(-\frac{(y - m_-)^2}{2\sigma^2}\right) \right],$$

To have a field strength that is a mixture of two Gaussians is not really unphysical even though it is certainly less natural than having a unimodal field strength. In order to generate the corresponding Gaussian conditionals $P(Y|X = \pm 1/2)$ we had to choose the constant a in eq. (7) such that

$$\frac{P(X = 1/2|y)}{P(X = -1/2|y)} = \frac{P(X = 1/2)}{P(X = -1/2)} \exp\left(\frac{-(y - m_+)^2 - (y - m_-)^2}{2\sigma^2}\right).$$

Comparing this to the Boltzmann probabilities in eq. (7) we conclude that the temperature has to be chosen such that the constant a satisfies $a/2 = 2(m_- - m_+)$. In contrast to the modifications in the Stern-Gerlach experiment that were required to “outsmart our principle” the causal mechanism as such has not to be modified here. There is nevertheless a constraint that makes the described situation unlikely to occur unless the setup as designed *by hand* in order to construct objections: The fact that the temperature value has to be adjusted to *one specific* value that is derived from m_{\pm} (even though there is no physical reason that makes this coincidence likely) shows that the counterexample is *non-generic*.

Here, the reason why thermodynamics predicts a smooth conditional probability for the effect given the cause is, abstractly speaking, the following. The equilibrium states maximize entropy subject to the energy. Here the energy depends smoothly (just linearly) on the cause (i.e. the field strength). The following section will describe a systematic way to define smoothness of conditionals via smoothness of the constraints of entropy maximization.

4 Exponential hierarchies and simplicity of distributions

The causal inference principle in Definition 4 should only be considered as a first attempt to formalize the vague idea of preferring smooth conditionals. When analyzing real data one will certainly not expect to always find conditionals of this type. Instead of defining *the simplest (conditional) probability measure* it is therefore desirable to define a hierarchy of more and more complex measures. Such notions of complexity of probability measures are also important in non-parametric statistics if a distribution should be estimated from observed relative frequencies. To avoid overfitting one has to

give preference to distributions that are considered simpler than others in an appropriate sense [16]. Before we discuss notions of simplicity for *conditional* distributions we first focus on unconditional measures. For the distribution of a real-valued variable one will, for instance, prefer densities which are smooth in the sense that they do not have too many modes unless a large number of observed data provided enough evidence for a less smooth distribution. We define probability measures of increasing complexity by an exponential hierarchy

$$P(x) = \exp\left(\sum_{j=1}^k c_j x^j + \ln z\right) \quad \text{with} \quad z = \int \exp\left(\sum_{j=1}^k c_j x^j\right) dx,$$

where $c_j \in \mathbb{R}$ and k is some natural number defining the complexity of the class considered.

Another interesting hierarchy of more and more complex measures is given by log-linear models. A distribution on n variables X_1, \dots, X_n is called a log-linear model of order k [17, 18] if its logarithm can be written as

$$\ln P(x_1, \dots, x_n) = \sum_l^k \sum_{|S|=l} \theta_S f_S(\mathbf{x}_S),$$

where S denotes subsets of $\{1, \dots, n\}$ and θ_S a coefficient corresponding to this subset. f_S is some function with argument \mathbf{x}_S . Here \mathbf{x}_S denotes the vector of dimension $|S|$ having all the values x_j with $j \in S$ as entries. Since these models are often used for categorical data where the numerical values x represent only nominal categories, there are no additional restrictions for the functions f_S . When dealing with variables whose values represent numeric variables it is natural to prefer *smooth* functions f_S . We define then a hierarchy of models as follows:

Definition 6 (Exponential hierarchy)

A probability distribution on n random variables is an exponential distribution of order k if it can be written as

$$\ln P(x_1, \dots, x_n) = \sum_l^k \sum_{|S|=l} \theta_S \prod_{j \in S} x_j.$$

This multivariate hierarchy unifies the complexity aspects of log-linear models and the smoothness condition explained above for the exponential hierarchy for distributions on \mathbb{R} . Distributions of this kind can be obtained by maximizing the entropy of P subject to the constraint that the expected values of the random variables $\prod_{j \in S} X_j$ attain some given values c_S [19, 20]. This gives some justification for defining *simplicity* of a distribution P by the *simplicity of the logarithm* of P .

If the variables X_1, \dots, X_n represent physical coordinates of a classical system thermodynamics provides very obvious arguments to prefer distributions lying in a *low* class of the exponential hierarchy. If the system is in its Gibbs equilibrium state for temperature T we have, as the classical analogue of eq. (8),

$$P(x_1, \dots, x_n) = \frac{1}{z} e^{-\beta H(x_1, \dots, x_n)},$$

where $\beta = 1/(kT)$ and z is a normalization constant [21], and $H(x_1, \dots, x_n)$ in the energy function, i.e., the Hamiltonian of the system.

The set of Hamiltonians that are present in nature, is quite restricted even though the set of *effective* Hamiltonians is huge. But also for effective Hamiltonians it is natural to assume that they are not too complex functions in the physical variables. An example is the energy of a classical harmonical oscillator which is, expressed in terms of the variables X, P , given by

$$H(x, p) = Dx^2 + \frac{m}{2}p^2,$$

with spring constant D and mass m . Hence the Gibbs distribution

$$P(x, p) = \frac{1}{z} e^{-\beta(Dx^2 + \frac{m}{2}p^2)}$$

is then an exponential model of order 2.

An example with n binary variables is given by a system with interacting classical spin variables X_1, \dots, X_n having the value set $\{-1, 1\}$. The physically reasonable assumption of having only pair-interactions (see e.g., the Ising model [22]) leads to a Hamiltonian of the form

$$H(x_1, \dots, x_n) = \sum_{j=1}^n a_j x_j + \sum_{1 \leq i < j \leq n} b_{ij} x_i x_j.$$

It induces a probability distribution of order $k = 2$ in the exponential hierarchy.

Clearly, also more general notions of equilibrium states are important in physics. One may, for instance, control the expected value of some observable $N(x_1, \dots, x_n)$. According to Jaynes' principle [23], the equilibrium is then given by a constrained entropy maximization subject to the expected value of the energy H and the additional variable N . The set of distributions that can be obtained in an experimental setup is then determined by the set of observables whose expectation can be controlled or is given by prior knowledge. For many-particle systems, it is rather feasible to control *macroscopic* variables than microscopic ones. In order to show why this, again, imposes smoothness conditions we consider a system with n spins (considered as classical dichotomic variables) and define the average spin by

$$A(x_1, \dots, x_n) := \frac{1}{n} \sum_{j=1}^n x_j.$$

For large n , this can certainly be considered as a macroscopic observable [24]. We could also think of controlling the expected value of $f(A)$ for some function f instead of controlling the expected value of A itself. Then the equilibrium distribution is given by

$$P(x_1, \dots, x_n) = \frac{1}{Z} \exp \left(-\beta H(x_1, \dots, x_n) + \lambda f[A(x_1, \dots, x_n)] \right),$$

with some additional Lagrange parameter λ apart from the inverse temperature given β . If f is given by an infinite Taylor series, P will not be an element of any class in the exponential hierarchy. However, for a general function f we should not consider $f(A)$ as a *macroscopic* observable. This becomes more obvious when we replace the classical variable A with its quantum analogue \hat{A} given by an average over Pauli matrices:

$$\hat{A}_\alpha := \frac{1}{n} \sum_j \sigma_\alpha^{(j)}$$

where $\sigma_\alpha^{(j)}$ for $\alpha = x, y, z$ denote the Pauli matrix σ_α acting on spin j . First of all we observe that the commutator between the operators \hat{A}_α for different α is at most of norm $2/n$ and can hence be neglected for large n . This justifies to

consider them as classical macroscopic variables. However, the commutator between $f(\hat{A}_\alpha)$ for different α may be considerably larger and the observables are no longer approximately compatible. In other words, for finite but large n it may be well justified to consider \hat{A}_α as a macroscopic observable but not any function of it unless it satisfies appropriate smoothness conditions. Whenever f can be well approximated by polynomials of small degree, the observables $f(\hat{A}_\alpha)$ are almost compatible and we may consider them also as macroscopic [24]. This suggests that it is easier to obtain a state where the exponent is approximately given by a polynomial of not too large degree, even in the case where additional observables are controlled.

Causally asymmetric versions of Occam's Razor require a notion of simplicity for *conditional* distributions (which contain unconditional measures as a special case). In order to give an impression of how simplicity could be defined in a reasonable way we generalize the exponential hierarchy for conditionals:

Definition 7 (Exponential hierarchy for conditionals)

Let $X_1, \dots, X_n, Y_1, \dots, Y_m$ be random variables. The conditional probability $P(Y|X)$ is a model of order k if

$$\ln P(y|x) = \sum_{l=0}^k \sum_{|S|=l} \theta_S f_S((\mathbf{x}, \mathbf{y})_S) + \ln z(x),$$

where we have introduced the following notations: Each S denotes some subset of $\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$ and $(\mathbf{x}, \mathbf{y})_S$ is the collection of the values of the variables in S . The function f_S is the monom given by the product of all these values.

The parameters θ_S are the corresponding coefficients. The partition function $z(x)$ is given by

$$z(\mathbf{x}) = \sum_{\mathbf{y}} \exp \left(\sum_{l=0}^k \sum_{|S|=l} \theta_S f_S((\mathbf{x}, \mathbf{y})_S) \right).$$

It is important to note that the partition function $z(\mathbf{x})$ can depend on \mathbf{x} in a *complex* way. Thus, simple conditionals do not necessarily generate simple joint distributions. The motivating remarks in Section 2 provide already a nice example: Let Y be a variable with standard normal distribution and X

binary and assume that $P(X|Y)$ is of the order 2, i.e., is of the form eq. (3). Elementary calculations show that its joint distribution satisfies

$$\ln P(X = 1, y) = -\frac{y^2}{2} + ay + b - \ln(e^{ay+b} + e^{-ay-b}).$$

This expression is a function with an infinite Taylor series. The joint probability is therefore not a member of the exponential hierarchy for any finite k .

On the other hand, simple joint distributions do not necessarily induce simple marginals. Let the joint distribution of the real variable Y and the binary variable X with values $\{0, 1\}$ be given by

$$P(x, y) = \frac{\exp(y^2 + xy)}{\sum_{\tilde{y}, \tilde{x}} \exp(\tilde{y}^2 + \tilde{x}\tilde{y})}.$$

We have

$$P(y) = \frac{\exp(y^2 + y) + \exp(y^2)}{\sum_{\tilde{x}, \tilde{y}} \exp(\tilde{y}^2 + \tilde{y}\tilde{x})},$$

which is a sum of two Gaussians and its logarithm has also an infinite Taylor series. In [11] we describe experiments where we have indeed obtained a reasonable causal order among binary variables in demographic data by preferring conditionals lying in a lower exponential hierarchy.

Another definition for simplicity has been used in [25] where the complexity of a conditional was defined as follows. Write

$$P(x_j|x_1, \dots, x_{j-1}) = \exp(-f(x_1, \dots, x_j) - \ln z_f(x_1, \dots, x_{j-1})), \quad (9)$$

where f is a function in some Hilbert space \mathcal{H} and z_f is the partition function. Then we consider the minimum of all $\|f\|^2$ over all f satisfying eq. (9) as the complexity of the conditional. Here $\|\cdot\|$ denotes an appropriate seminorm in \mathcal{H} . It has to be emphasized that the logarithm of $P(x_j|x_1, \dots, x_{j-1})$ itself can be arbitrarily complex, only the complexity (here defined as a Hilbert space seminorm) of f is considered. When the Hilbert-space seminorm penalizes in particular polynomials with high degree we obtain a complexity measure that is closely related to the exponential hierarchy in Definition 7.

We summarize the remarks above:

Definition 8 (Complexity of Conditionals)

Let X_1, \dots, X_j be random variables and C be a complexity measure on the space of functions of the form $f(x_1, \dots, x_j)$. Then we define the complexity of $P(x_j|x_1, \dots, x_{j-1})$ as the minimum of $C(f)$ over all f satisfying (9).

The above remarks show that thermodynamics leads very naturally to simple joint distributions if we assume that the Hamiltonians are functions with low complexity values C . The main question of this article is to explore under which circumstance physical processes generate simple *conditionals* for the effects given their causes leading to joint distributions that are not necessarily simple.

The following setup may be a bit artificial from the physics point of view. However, it provides a first impression on the link between the causal direction and the order of maximizing the entropies of subsystems that is essential for our first implementation of the pMK principle. Given two classical systems described by discrete-valued variables X, Y and a joint Hamiltonian of the form

$$H(x, y) = H_1(x) + H_2(y) + H_{12}(x, y).$$

Consider the following three hypothetical experiments. For reasons of convenience, we will identify the systems with the variables representing the physical states.

(1) System X and Y influence each other

Subject the joint system to a thermal bath with inverse temperature β . If it is thermalized, its statistical state is given by

$$P_{\leftrightarrow}(x, y) := \frac{1}{z} \exp \left(-\beta(H_1(x) + H_2(y) + H_{12}(x, y)) \right),$$

where z is the partition sum.

(2) System X influences Y

Remove the interaction term H_{12} , subject system 1 to the bath, adjust the state of system 1, i.e., fix the actual value x of the variable X . Couple both

systems by the interaction H_{12} and subject the joint system (or only system 2) to the bath. Thermalization leads to

$$P_{\rightarrow}(x, y) = P_{\rightarrow}(x)P_{\rightarrow}(y|x),$$

with

$$P_{\rightarrow}(x) := \frac{1}{z_1} \exp\left(-\beta H_1(x)\right),$$

where z_1 is the corresponding partition sum and

$$P_{\rightarrow}(y|x) := \frac{1}{z_2(x)} \exp\left(-\beta(H_2(y) + H_{12}(x, y))\right),$$

with the partition function

$$z_2(x) := \sum_y \exp\left(-\beta(H_2(y) + H_{12}(x, y))\right).$$

(3) System Y influences X

The same as (2) with interchanging the roles of system 1 and 2.

In experiment 2 the variable X represents the cause and Y the effect and in experiment 3 we have the reversed case. Obviously, the conditional distributions in experiment (2) and (3) are simple (when calculated with respect to the causal direction) provided that the Hamiltonians are simple. Likewise, the marginal distribution of the cause variable is simple. One checks easily that $P_{\leftarrow}(Y|X) = P_{\rightarrow}(Y|X)$ and $P_{\leftarrow}(X|Y) = P_{\rightarrow}(X|Y)$. Hence the difference between P_{\leftarrow} and P_{\rightarrow} is only caused by different marginal distributions for X . Whereas $P_{\rightarrow}(X)$ is directly determined by the free Hamiltonian $H_1(X)$, the computation of $P_{\leftarrow}(X)$ involves the partition function $z_2(x) = \sum_y \exp(-\beta(H_{12}(x, y) + H_2(y)))$. We obtain:

$$\ln P_{\leftarrow}(x, y) - \ln P_{\rightarrow}(x, y) = -\ln z + \ln z_1 + \ln z_2(x) =: f(x), \quad (10)$$

We see that here the partition functions are “responsible” for the fact that different causal directions lead to different joint distributions. Note, as an aside, that the Kullback Leibler distance between these distributions are

directly obtained from the difference in eq. (10). Interpreting the difference as a random variable $f(X)$ we have, for instance,

$$D(P_{\leftrightarrow}||P_{\rightarrow}) = E_{P_{\leftrightarrow}}(f(X)), \quad \text{and} \quad D(P_{\rightarrow}||P_{\leftrightarrow}) = E_{P_{\rightarrow}}(f(X)),$$

where E_P denotes the expected value with respect to P .

The coupling of two physical systems typically leads to a mutual influence between them. To understand under which circumstances the causal effect of one of the systems is dominating compared to the back action (such that we consider one as the cause and the other as the effect) is an interesting open question. In the following section I describe a bipartite system where a temperature gradient leads to a dominating causal direction that coincides with the thermodynamic entropy flow. Another possibility to obtain a definite causal direction is to consider the relations between the state of system 1 at time t_0 and the state of system 2 at $t_1 > t_0$. In Section 7 we will consider causal relations of this kind from a thermodynamic point of view.

5 Stationary process with temperature gradient

The scenario described in the last section was a bit artificial since it was based on *adjusting* states of one system before the interactions were switched on. In other words, the well-defined causal arrow was put in by hand. In [26] we have considered a model with two interacting systems, one with fast dynamics and one with slow dynamics, where a well-defined causal arrow emerges whenever the interaction is sufficiently weak compared to the free Hamiltonian of the system that acts as a cause. In this limiting case, the stationary of the joint system has the following properties. The state of the system representing the *cause* was given by a microcanonical distribution of its free Hamiltonian and the state of the system representing the *effect* by a microcanonical distribution of its conditional Hamiltonian. Apart from this, it turned out that in the described limit the thermodynamics of the “cause-system” is a well-behaved thermodynamic system whose coarse-grained entropy is only increasing but never decreasing. Hence the conditions to have well-defined thermodynamic properties of subsystems turned out to be related to having well-defined causal directions.

The model considered here is closely related to the one in [26]. In order to be consistent with our definition of simplicity relying on exponential hierarchies we consider Gibbs equilibrium distributions rather than micro-canonical equilibrium states. To this end, we present a model consisting of two baths with different temperatures where entropy maximization for the cause is followed by conditional entropy maximization for the effect.

Following [27] we consider two classical systems 1 and 2, described by variables X and Y , respectively, and a Hamiltonian $H(x, y)$. System j is subjected to temperature T_j . Then [27] describes the coupled Langevin equations

$$\Gamma_1 \dot{x} = \delta_x H(x, y) + \eta_1(t) \quad (11)$$

$$\Gamma_2 \dot{y} = \delta_y H(x, y) + \eta_2(t) \quad (12)$$

with stochastic forces η_j whose product satisfies

$$E(\eta_i(t)\eta_j(t')) = 2\Gamma_i T_i \delta_{ij} \delta(t - t'),$$

where Γ_j is the damping constant for system j and δ_x and δ_y denote partial derivatives.

Now x is assumed to change more slowly than y which is ensured by the condition $\Gamma_1 \ll \Gamma_2$. Then it is argued that one may keep x fixed and solve equation (12) for Y and obtain the x -dependent equilibrium

$$P(y|x) = \frac{1}{z(x)} \exp\left(-\beta_1 H(x, y)\right),$$

with the partition function

$$z(x) := \int \exp\left(-\beta_1 H(x, y)\right) dy,$$

and the inverse temperatures $\beta_j := 1/(kT_j)$. In order to calculate $P(X)$ we average the energy value $H(x, y)$ according to the above conditional distribution $P(X|Y)$ and obtain from eq. (11) the Langevin equation

$$\Gamma_1 \dot{x} = \delta_x H_{eff}(x) + \eta_1(t)$$

with the effective Hamiltonian

$$H_{eff}(x) := -kT_2 \ln \int \exp(-\beta_2 H(x, y)) dy.$$

Then we obtain

$$P(x) = \frac{1}{z} e^{-\beta_1 H_{eff}(x)}$$

and compute the joint measure using

$$P(X, Y) := P(X)P(Y|X).$$

As has been shown in [27] that P can be obtained by maximizing $T_1 S(X) + T_2 S(Y|X)$ subject to

$$\sum_{x,y} P(x, y) H(x, y) = E,$$

for an appropriate value E . Here $S(X)$ and $S(Y|X)$ denote entropy, respective conditional entropy. This indicates that the limit $T_1 \gg T_2$ yields a joint measure that is obtained by *first* maximizing the entropy of system 1 and *then* maximizing the conditional entropy of system 2. To study this limit we write

$$H(x, y) = H_1(x) + H_2(y) + H_{12}(x, y).$$

Obviously, we have

$$H_{eff}(x) = H_1(x) - T_2 \int e^{-\beta_2(H_{12}(x,y)+H_2(y))} dy. \quad (13)$$

In order to be consistent with the preceding section we discretize the system and replace integrals with sums. We introduce the truncated effective Hamiltonian

$$H_t := -T_2 \sum_y e^{-\beta_2(H_{12}(x,y)+H_2(y))}. \quad (14)$$

An interesting limiting case is when the interaction is small compared to kT_1 but not small compared to kT_2 . Then, intuitively speaking, system 1 does not feel the interaction H_{12} , but influences system 2 via H_{12} . To explain the consequences of this assumption formally we consider a sequence of temperatures $T_1^{(n)} := nT_1$ and rescale the free Hamiltonian of system 1 by defining $H_1^{(n)}(x) := nH_1(x)$. The interaction energy and T_2 will be kept constant. Using (14) we obtain

$$H_t^{(n)} = -kT_2^{(n)} \ln \sum_y \exp(-\beta_2(H_{12}(x, y) + H_2(y))) \rightarrow 0.$$

We conclude: If kT_1 is large compared to the interaction energy the joint measure of the bipartite system is obtained by (1) maximizing the entropy of system 1 subject to the energy corresponding to its *free* Hamiltonian and then maximizing the conditional entropy of system 2 subject to the total energy. We obtain the same statement by decreasing H_{12} , H_2 and T_2 according to a common scaling factor. The fact that in these limits the distribution of X is determined by the free Hamiltonian alone is, from an intuitive perspective, already a good indicator for the fact that the influence of system 2 on system 1 goes to zero. But our intention is to *support* this way of reasoning, not to take it for granted. For this purpose, we will now recall the hypothetical experiments in Section 4 where we have adjusted one system to a fixed value and let the other return to equilibrium.

In order to show that we may indeed consider the variable X as the cause and variable Y as the effect (in the above limits), we show that system 1 is quite insensitive with respect to adjusting system 2 to different values. To see to what extent variable Y influences variable X we derive an upper bound on the relative entropy distance between the following two measures (1) measure $P_0(X)$ that would be obtained for system 1 without interaction and (2) the distribution $P_{\leftarrow}(X|y)$ that system 2 induces when it is adjusted to some specific value y :

Lemma 1 (Bound on the back action)

Let P_{\leftarrow} be defined as in Section 4 and y be arbitrary. Define

$$P_0(x) = \frac{1}{z} \exp(-\beta_1 H_1(x)).$$

If the interaction energy $H_{12}(x, y) \geq 0$ for all x we have

$$D(P_0(X)||P_{\leftarrow}(X|y)) \leq \sum_x P_0(x) \beta_1 H_{12}(x, y),$$

This shows that the back action converges indeed to zero for $\beta_1 \rightarrow 0$.

Proof: The relative entropy distance reads

$$\begin{aligned} D(P_0(X)||P_{\leftarrow}(X|y)) &= \sum_x P_0(x) (\ln P_0(x) - \ln P_{\leftarrow}(x|y)) \\ &= \sum_x P_0(x) \left(-\beta_1 H_1(x) - \ln \sum_x e^{-\beta_1 H_1(x)} \right) \end{aligned}$$

$$\begin{aligned}
& + \ln(\beta_1(H_1(x) + H_{12}(x, y))) + \ln \sum_x e^{-\beta_1(H_1(x) + H_{12}(x, y))} \\
= & \sum_x P_0(x) \beta_1 H_{12}(x) + \ln \sum_x P_0(x) e^{-\beta_1 H_{12}(x, y)} \\
\leq & \sum_x P_0(x) \beta H_{12}(x, y).
\end{aligned}$$

It is natural to ask whether one could also construct a limit where the fast system influences the slow one without significant back action by assuming $T_2 \gg T_1$. However, the rescaling $T_2^{(n)} := nT_2$ and $H_2^{(n)}(y) := nH_2(y)$ leads for $n \rightarrow \infty$ to a conditional $P(y|x) = \exp(-H_2(y))/z$, i.e., X and Y become independent. \square

Note that there is a nice way to quantify action and back action in the above “generalized equilibrium” by a hypothetical sender/receiver protocol. Assume a sender having access to system 1 adjusts his system to one value x according to the marginal $P(Y)$ that corresponds to the measure P above. Then the receiver observes values y with probability $P_{\rightarrow}(y|x)$. His information about X is given by

$$\begin{aligned}
I_{\leftarrow}(X : Y) &= \sum_{xy} P(x) P_{\rightarrow}(y|x) \ln P_{\rightarrow}(y|x) \\
&\quad - \sum_{xy} P(x) P_{\rightarrow}(y|x) \ln \sum_x (P(x) P_{\rightarrow}(y|x)) \\
&= \sum_x P(x) D \left(\sum_{x'} P_{\rightarrow}(Y|x') \parallel P_{\rightarrow}(Y|x) \right) \\
&\leq \sum_{x, x'} P(x) P(x') D(P_{\rightarrow}(Y|x) \parallel P(Y|x')),
\end{aligned}$$

where we have used that mutual information can be rewritten as an average relative entropy distance between conditionals and marginals [28] and that relative entropy is convex. If we define $I_{\leftarrow}(X : Y)$ in an analogue way, it follows that the “back action”-information $I_{\leftarrow}(X : Y)$ tends to zero (in the limit $n \rightarrow \infty$) because calculations similar to the proof of Lemma 1 show that $D(P(X|y) \parallel D(X|y'))$ converge to zero for all y, y' . On the other hand, the “forward information” $I_{\rightarrow}(X : Y)$ converges to a non-zero value because

the joint distribution on X, Y obtained by the forward sender/receiver scenario coincides exactly with the natural equilibrium P where we have indeed statistical dependences.

6 Are definite causal directions a non-equilibrium phenomenon?

The fact that we needed two systems with *different* temperatures in order to have one system influencing the other but not vice versa seems to have a deeper reason. This is at least true within the setting of Section 5. This should be discussed here due to the folklowig motivation. If there are indeed relations between the presence of well-defined causal directions on the one hand and irreversibility on the other hand this would further support the conjecture that typical statistical asymmetries between cause and effect stem from some kind of “thermodynamics of causal directions” that is yet to be discovered. Since I can give no general result on such a link, we restrict the attention to the models described earlier.

Assume that the two interacting systems of the preceding section are in equilibrium with a common temperature $T = 1/(\beta k)$. Then their joint distribution reads

$$P(x, y) = \frac{e^{-\beta H(x, y)}}{\sum_{x, y} e^{-\beta H(x, y)}}.$$

Consider now the same sender/receiver protocol where system 1 is adjusted randomly to some value x according to the marginal distribution $P(X)$. The conditional $P_{\rightarrow}(Y|X)$ will then coincide with the usual equilibrium conditional $P(Y|X)$. Hence the intervention preserves the usual equilibrium state. For symmetry reasons, this holds clearly for adjusting system 2, too. But then the backward and the forward information coincide exactly. Hence, different temperatures were really needed in Section 5 to obtain a definite causal direction.

We want to revisit the second example in Section 2 (with the spins in a magnetic field) in light of this result. The interaction between the field and the spin cannot be an interaction between two systems in *Gibbs equilibrium with a common temperature*, otherwise the classical field would feel the field of the probe spin in the same way as vice versa, in contradiction

to our assumption on the definite causal direction. To see how a common Gibbs equilibrium of field and probe spin would change the statistics of their common state, we assume that the field is generated by n spin $1/2$ particles. Let

$$S_z := \sum_{j=1}^n \sigma_z^{(j)}.$$

be their total spin in z direction, where $\sigma_z^{(j)}$ denotes the Pauli matrix σ_z on spin j . The free Hamiltonian of the n -spin system when subjected to a magnetic field B in z direction is given by

$$H := BS_z \otimes \mathbf{1}.$$

The free Hamiltonian of the probe spin system is $B(\mathbf{1} \otimes \sigma_z)$.

In its thermal equilibrium, the total spin S_z follows a binomial distribution $B_p(k)$ with $p/(1-p) = \exp(-1/kT)$. For large n , the total magnetic moment fluctuates on the scale \sqrt{n} . Then we introduce an interaction H_i by

$$H_i := c \frac{1}{\sqrt{n}} S_z \sigma_z,$$

with a constant c determining the interaction strength. The scaling factor $1/\sqrt{n}$ is chosen such that the total field strength “felt” by the probe spin system follows a well-defined distribution in the limit $n \rightarrow \infty$. Now we consider the conditional probability for k spins up given that the probe spin is in its upper state. The total spin of the n -particle system defines an integer-valued random variable Y . We have

$$P(Y = k | X = 1/2) = B_{p^+}(k)$$

and

$$P(Y = k | X = -1/2) = B_{p^-}(k)$$

with p^\pm are given by $p^\pm/(1-p^\pm) = \exp((1 \pm \sqrt{1/n})/kT)$. In the limit of large n the binomial distributions can be replaced with two Gaussians with standard deviation in the order of \sqrt{n} . Their mean value differ also on the scale \sqrt{n} . This shows that we obtain a mixture of two Gaussians for the distribution of magnetic moments of the n -particle system. Moreover, there is an influence of in both directions in agreement with our findings in the beginning of this section.

This result raises the question to what extent non-equilibrium is crucial for the other examples provided by this paper. Let us now consider the Stern-Gerlach experiment. The following extremely simplified model of the dynamics may clarify the role of thermal equilibrium and non-equilibrium. Define the Hilbert space

$$\mathcal{H} := L^2(\mathbb{R}) \otimes \mathbb{C}^2,$$

where the set of square integrable function encodes the momentum degree of freedom in transversal direction and the two-dimensional component represents the spin. We assume, for simplicity, that the only Hamiltonian that is relevant inside the furnace is the Hamiltonian H of the harmonic oscillator corresponding to the confining potential. Hence, the joint Hamiltonian of spin and momentum is then given by $H \otimes \mathbf{1}$. In thermal equilibrium we have then the state

$$\rho := \sum p_n |n\rangle \langle n| \otimes \mathbf{1},$$

where $|n\rangle$ with $0 = 1, 2, \dots$ denotes the eigenstates of the oscillator and p_n the Boltzmann probabilities corresponding to the considered temperature. After the atoms leave the furnace, the oscillator potential is no longer effective and the system is no longer in equilibrium. The inhomogeneous field generates a dynamics that entangles spin and translational degree of freedom. We assume that the position degrees of freedom that correspond to other directions than the transversal direction under consideration are not relevant since we have only free motion in longitudinal direction. When the atom arrives at the screen its state has been transformed to $U(\rho \otimes \mathbf{1})U^\dagger$ with the unitary map

$$U := U_\downarrow \otimes |\downarrow\rangle \langle \downarrow| + U_\uparrow \otimes |\uparrow\rangle \langle \uparrow|,$$

where $|\downarrow\rangle, |\uparrow\rangle$ denote the two possible spin states and U_\downarrow, U_\uparrow are unitary operators that act on the transversal degree of freedom in a way that is controlled by the spin values. The fact that U has this form follows because it acts on the spin degree of freedom only as a measurement apparatus, given that we assume that no further degree of freedom is relevant. When the atoms leave the furnace and enter the field, a unitary dynamics transforms the state, i.e., the system is no longer in equilibrium. The relation between cause and effect in this example is therefore generated in a non-equilibrium dynamics. To see how this is related to the form of the relevant quantum states, we add the following observations. The states

$$U_\uparrow \rho U_\uparrow^\dagger \quad \text{and} \quad U_\downarrow \rho U_\downarrow^\dagger$$

are Gibbs equilibrium states for the transformed Hamiltonians

$$U_{\uparrow}^{\dagger} H U_{\uparrow} \quad \text{and} \quad U_{\downarrow}^{\dagger} H U_{\downarrow}.$$

If we assume that U_{\uparrow} and U_{\downarrow} are *simple* dynamical evolutions like translations, these are, again, simple Hamiltonians. Hence the conditional state of the system representing the effect, given a fixed value of the cause variable, is a Gibbs state for a *simple* Hamiltonian. To be a Gibbs state for a simple formal Hamiltonian can be considered as a quantum generalization of the exponential hierarchy in Section 4.

On the other hand, the marginal state of the effect system itself is given by $(U_{\uparrow} \rho U_{\uparrow} + U_{\downarrow} \rho U_{\downarrow})/2$. This is not an equilibrium state corresponding to the physical Hamiltonian of the system. A formal Hamiltonian, obtained by the logarithm of such a mixture, does not have any direct physical meaning² and need not be simple.

So far, we have identified two different ways how the cause can influence the effect without significant back action. One possibility (discussed in Section 5) is that the system “effect” relaxes to its thermal equilibrium according to a Hamiltonian that is determined by the system “cause”. But then the system “cause” cannot be in Gibbs equilibrium with the same temperature. The other possibility (e.g. our simplified model of the Stern-Gerlach experiment) is that the joint system consisting of “cause” and “effect” is subjected to a Hamiltonian whose unitary evolution leaves the “cause” invariant and evolves the system “effect” conditional on the state of “cause” – a scenario that explicitly involves thermal states. In the following section we will discuss two scenarios where the effect system is driven into a thermal equilibrium according to a fixed Hamiltonian, but the temperature is determined by the “cause”.

The above remarks indicate that there is a close connection between having a definite causal direction and having non-equilibrium even though a profound understanding of this link has to be left to the future. Nevertheless, these preliminary insights suggest that several kinds of statistical asymmetries between cause and effect can be explained using appropriate *thermodynamically irreversible* processes. The following section will elaborate on this.

²Jaynes stated in the context of non-equilibrium thermodynamics [29]: “[...] we must learn how to construct ensembles which describe not only the present values of macroscopic quantities, but also whatever information we have about their past behavior.”

7 Common root of the Markov condition and the principle of plausible Markov kernels

Since causal structure must always be consistent with the time order of events it is straightforward to assume that the postulated asymmetry between cause and effect is related to the asymmetry between past and future. The goal of this section is to show that the common root is due to the tendency of our environment to subject a system to interactions with an abundance of other physical systems that are initially uncorrelated with the former rather than being finally uncorrelated. It is clear, that this tendency is closely related to the usual asymmetry between past and future: We see a scene happening in front of our eyes shortly *after* it has happened because the photons absorbed by the eyes have obtained correlations with the objects at which they were reflected. The physical state of the light beam was uncorrelated with the object *before* it interact with the latter but correlated *afterwards*. The relation to Reichenbach's principle, saying that statistical dependences have to be explained by interactions in the past but not by interactions going to happen in the future, is quite apparent. This examples shows already that some very evident asymmetries between past and future are closely related to the principle of the common cause [30].

First we want to provide microphysical models that give some insight about how the asymmetry in the Markov condition is related to the thermodynamic arrow of time. By a *classical microphysical model* we mean the following. (1) Different random variables represent the state of a different physical system or the state of the same system at a different time. This means that the value set of the variable is identified with the space of *pure* states and the set of measures on the value set is the spaces of *mixed* states. (2) The space of pure states of a composed system is given by the Cartesian product of the spaces of the constituents. (3) A physical process of a closed physical system is a bijective map on its set of pure states. (4) A physical process of an open physical system is a bijective map on the Cartesian product of the set of pure states of the system under consideration and the set of pure states of an additional system, called the environment. Our classical microphysical models are discrete, i.e., one may interpret them as quantum systems with restricted state space.

The simplest models where the asymmetry between cause and effect im-

posed by the Markov condition becomes apparent is the difference between a fork and a collider (see Fig 1): If Z is the common cause of X and Y we have

$$X \perp\!\!\!\perp Y | Z \quad \text{but} \quad X \not\perp\!\!\!\perp Y, \quad (15)$$

where $\perp\!\!\!\perp$ denotes independence and $\cdot \perp\!\!\!\perp \cdot | \cdot$ is conditional independence. If Z is the common effect of two (causally) independent causes we have in the generic case

$$X \not\perp\!\!\!\perp Y | Z \quad \text{but} \quad X \perp\!\!\!\perp Y \quad (16)$$

Reichenbach [1] emphasized that this asymmetry follows from his formulation of the arrow of time where the term “branch systems” plays a crucial role (systems that are isolated from the main system for a certain period of time). The idea is that the observation of a state with untypically low entropy in one branch system deserves an explanation in terms of an interaction with other branch systems in the *past*. He argues that statistical dependences between separated systems define an untypical state of the joint system that deserves thus an explanation in terms of an interaction in the past.

To describe the common thermodynamic root of the asymmetry between (15) and (16) on the one hand and the asymmetry formalized by the pMK principle I will use the same type of models for both asymmetries.

7.1 Microphysical model for the asymmetry between common causes and common effects

To draw the link between causal order and time order we construct models for causal forks and causal colliders where the corresponding dynamical maps are equal up to time inversion. However, as we will see, the assumptions on the initial probability distributions are not symmetric.

Causal fork (common cause): Let $P(X, Y, Z)$ be an arbitrary joint measure satisfying the $X \perp\!\!\!\perp Y | Z$. Now we construct a bijective process acting simultaneously on 6 systems

$$S_Z \times S'_Z \times S_X \times S_Y \times S_{NX} \times S_{NY}.$$

Their role is as follows. The initial state of S_Z represents the variable Z and the final states of S_X and S_Y (after some bijective process has acted jointly on the 6 systems) represent the variables X and Y , respectively. The time order

guarantees that Z can only be a cause and not an effect of X and Y . Systems S_{NX} and S_{NY} represent background noise that prevents X and Y from being deterministic functions of Z (see remarks after Definition 1). The role of S'_Z is a bit more subtle and is easier to explain after the process has been described. Let P be a product distribution on $S_Z \times S_X \times S_Z \times S_{NX} \times S_{NY}$ and let S'_Z be in an arbitrary pure state. Then we construct a process consisting of two steps. In the first step a bijective map F_Z acts on $S_Z \times S'_Z$. Afterwards we apply the maps F_X and F_Y acting on $S_Z \times S_X \times S_{NX}$ and $S'_Z \times S_Y \times S_{NY}$, respectively. Note that F_Z distributes the information such that it (or at least part of it) is afterwards available on S_Z and S'_Z . This “broadcasting” of information into two components ensures that S_Z can have an effect on both S_X and S_Y even though a direct interaction between S_X and S_Y is avoided. If F_X and F_Y both would act (one after another) on S_Z we could not exclude information transfer between them in contradiction to our causal model being a fork. In other words, our structure of concatenated maps ensures that the process is indeed a causal fork on the triple (Z, X, Y) . It is easy to show that processes of the above kind generate a distribution with $X \perp\!\!\!\perp Y | Z$ and that every joint distribution on X, Y, Z satisfying this independence can be generated by a process of this type.

Collider (common effect): Here we consider the same 6 systems and the same bijections, but in time-reversed order. Let P be a product distribution on $S_X \times S_Y \times S_Z \times S_{NX} \times S_{NY} \times S'_Z$. Then implement F_X and F_Y as above and F_Z afterwards. Then we have $X \perp\!\!\!\perp Y$ by assumption, but not necessarily $X \perp\!\!\!\perp Y | Z$.

As already mentioned, scenario 2 is not simply the time-reversed version of scenario 1. If this was the case, we did not have the asymmetry given by equations (15) and (16). The true reason for the asymmetry is that all systems are assumed to be *initially* independent and not *finally* independent. The backward time version of scenario 1 would be the following. The joint system is *initially* correlated in a way that ensures that the application of F_X, F_Y and F_Z makes them stochastically independent! This would be a rather contrived situation. We do not claim that it would be a realistic assumption that every physical system is “initially” uncorrelated from its environment. The essential point is that it is unlikely that the correlations are exactly such that the reversible dynamics resolves them into a product state. This is certainly directly connected with the usual arrow of time

statistical physics where interactions between particles lead typically to an increase of coarse-grained entropy (cp. e.g. [31, 30, 32]).

To be consistent with our model class, we rephrase this asymmetry as follows.

Observation 1 (Arrow of time)

Define a bijective map

$$F : \mathcal{X} \rightarrow \mathcal{X}$$

with

$$\mathcal{X} := \times_{j=1}^k \mathcal{X}_j$$

as representing the dynamical evolution of a composed physical system after some time t . Here \mathcal{X}_j denotes the state space of system j (which is, for simplicity, assumed to be finite). Let P be a probability distribution on \mathcal{X} that formalizes the initial statistical state of the system and $P \circ F$ denote its final state. Given that k is large, it is unlikely that $P \circ F$ is a product state but P is not, unless F has been designed “by hand” in order to transform the non-product state into a product state. The reverse scenario, that P is a product state and $P \circ F$ is not, happens quite often.

A typical permutation of k tuples in the k -fold Cartesian product creates dependences if the initial distribution is a product measure whose entropy is not maximal. This can be considered as a model for increasing correlations being the typical situation in *closed* systems. In open systems, however, we have to take into account the following effect: The restriction of a probability distribution on a Cartesian product to a small fraction of subsystem is typically close to a maximal entropy distribution (hence a product measure) even though the distribution itself may be far away from a product state. In quantum systems, we have even the stronger statement that the restriction of a typical *pure* many-particle state is so strongly entangled that its restriction to a small fraction of subsystems is almost the maximum entropy state [33].

Therefore, it was important for the justification of our way of reasoning that we considered maps on *closed* systems by taking the environment explicitly into account (in form of S_{NX} and S_{NY}). Otherwise we could not justify the remark that increase of dependences is more typical than resolving dependences.

Now we want to understand the asymmetry between cause and effect with respect to the *smoothness* of conditionals within the same setting.

7.2 Asymmetry in the shape of conditionals

Binary variable influencing a continuous variable: Assume we have one two-level system S_X (i.e. the spin degree of freedom for a spin-1/2 particle in a magnetic field in z -direction) representing the binary variable X . The continuous variable Y will be, in an approximate sense, represented by the total spin of n spin-1/2 particles which define our system S_Y . The energy gap E_X of the two-level system S_X is assumed to be an integer multiple of the gap of each of the n two-level systems, i.e., $E_X = \ell E_Z$. We consider the limit that n and ℓ go to infinity with ℓ asymptotically proportional to \sqrt{n} but all the other quantities are constant. Even though we have quantized energy levels, we first discuss the problem in a setting of classical probability theory where every two-level system is represented by a copy of the probability space $\{0, 1\}$. We will later show that the whole discussion leads essentially to the same result within the quantum setting. The quantum model is even simpler in the sense that we can explain mixing without having a noisy environment.

Let us first specify the initial conditions. Let each single spin of the n -spin system be on its upper level with probability $p < 1/2$, i.e., we have Gibbs equilibrium with some positive temperature and S_X starts in its lower state. Then we assume that a weak interaction drives a mixing process on the joint state space $\{0, 1\}^{n+1}$ that can only permute levels with the same total energy. The probability that a randomly chosen permutation drives S_X into its upper state is asymptotically zero since the energy gap of S_X tends to infinity. This is because such a large amount of energy could only be extracted from system S_X using an entropy sink that can absorb more than one bit of entropy. On the other hand, if S_X starts in its upper level, the mixing process will typically drive it to its lower level and thus increase the total energy of S_Y by the amount $E_X = \ell E_Z$. Let $X = 0, 1$ describe the two possible states of S_X *before* the mixing process has started and $Y = 1, \dots, n$ describe the total spin of S_Y *after* the mixing. Asymptotically, $P(Y|X = 0)$ and $P(Y|X = 1)$ are Gauss distributions with variance $n(1 - p)p$ and mean np , respective, $np + \ell$. The two averages of Y differ therefore on the scale \sqrt{n} which is exactly the scale of the standard deviation of Y . Thus the conditional distribution of Y is exactly the most plausible Markov kernel according to Definition 4 (see eq. (2)).

Continuous variable as cause, binary variable as effect: Now we describe a scenario where the continuous variable influences the binary variable. Let S_X and S_Y be as above and assume we are able to control the total spin of the system (the “macroscopic variable”, see Section 4) and set it to some value pn with $0 < p < 1$. Let S_X start in its lower level. This is asymptotically arbitrarily close to its thermal equilibrium state for any constant finite temperature $T > 0$ as we let, again, the gap E_X and n tend to infinity. Since we assume to have no access to the microscopic part of S_Y we describe the initial statistics of S_Y by the uniform mixture over all states with total spin pn . Our mixing process should model a causal influence from Z , i.e., the *initial* rescaled total spin of system S_Z on X , i.e., the *final* total spin of S_X . In order to ensure that a typical mixing process has a non-trivial effect on S_X , the average occupation probability p in S_Y has to get closer to $1/2$ as E_X increases. Otherwise, as standard statistical physics arguments show, the probability to find S_X in its upper state after the mixing would tend to zero. We set

$$p_n := \frac{1}{2} + \frac{y}{2\sqrt{n}},$$

where y denotes a constant. It will later be interpreted as the value of the variable Y for which we compute $P(X|y)$. One checks easily that, asymptotically, this increase of p_n corresponds to increasing the temperature with $kT_n = \sqrt{n}/y$ in a Gibbs equilibrium state. It is therefore natural to expect that a typical mixing process leads to the occupation probability q with $q/(1-q) = \exp(-y)$. To give more formal arguments for this guess we recall that we can estimate the quotient of the number of states of S_Y with total spin $p_n n$ and $p_n n - \ell$ as follows. It is asymptotically given by $(p_n/(1-p_n))^\ell$. Therefore we have

$$\frac{q}{1-q} \approx \left(\frac{p_n}{1-p_n} \right)^\ell = \exp\left(-\ell \ln \frac{p_n}{1-p_n} \right) \rightarrow \exp\left(-\sqrt{n} \frac{y}{\sqrt{n}} \right) = \exp(-y). \quad (17)$$

In order to define a variable with continuous range we rescale the total spin by $1/\sqrt{n}$ to obtain Y . Since the the occupation probabilities q and $1-q$ in (17) are exactly the probabilities $P(X=1|y)$ and $P(X=0|y)$, respectively, we obtain

$$\frac{P(X=1|y)}{P(X=0|y)} = e^{-y},$$

which is exactly the most plausible Markov kernel according to Definition 4 (see eq. (3)).

Spin systems are actually quantum systems. Nevertheless we have described our arguments within a classical setting since the statistical arguments did not require quantum superpositions. In order to further support the general idea we want to briefly describe why the model works even with pure initial states when described in the quantum setting. We will focus on the first model with X effecting Y . Let S_X and S_Y be described by the Hilbert spaces $\mathcal{H}_X := \mathbb{C}^2$ and $\mathcal{H}_Y := (\mathbb{C}^2)^n$, respectively. Let $|\psi\rangle \in \mathcal{H}_Y$ be a state whose total spin in z -direction is $p_n n$ and $|0\rangle \in \mathcal{H}_X$ be the lower level of S_X . Now we discuss what a typical *energy-conserving unitary map* on $\mathcal{H}_X \otimes \mathcal{H}_Y$ would do. The space of states with total energy $pn E_Y$ splits up into the space

$$\mathcal{H}_0 := |0\rangle \otimes \mathcal{G}_{pn},$$

where \mathcal{G}_{pn} consists of all states with total spin pn , and

$$\mathcal{H}_1 := |1\rangle \otimes \mathcal{G}_{pn-l}.$$

Obviously, the quotient of the dimensions coincides with the quotient of the number of states above. After the unitary process has been applied, we have a state of the form

$$|0\rangle \otimes |\psi_0\rangle + |1\rangle \otimes |\psi_1\rangle.$$

The probability that S_X is found in its upper level is then given by $\| |\psi_1\rangle \|^2$. The following lemma shows that in high dimensions almost every state in $\mathcal{H}_0 \oplus \mathcal{H}_1$ has the property that $\| |\psi_1\rangle \|^2 / \| |\psi_0\rangle \|^2$ is close to the quotient of the dimensions of \mathcal{H}_1 and \mathcal{H}_0 .

Lemma 2 *Let \mathcal{H}_1 and \mathcal{H}_2 be two Hilbert spaces of dimension d_1 and d_2 , respectively. Let $|\psi\rangle := |\psi_1\rangle \oplus |\psi_2\rangle$ be a randomly chosen vector according to the Haar measure of $SU(d_1 + d_2)$. Then for every η the probability that $\| |\psi_1\rangle \|^2 / \| |\psi_2\rangle \|^2 - d_1/d_2 > \eta$ tends to zero for $d_1 + d_2 \rightarrow \infty$.*

Sketch of the proof: Define a function f on \mathcal{H} by

$$f(\psi) := \langle \psi | P_1 | \psi \rangle,$$

where P_1 is the projector onto \mathcal{H}_1 . The function f is Lipschitz-constant with $L = 2$. It is easy to show that the average of f over $SU(d_1+d_2)$ is d_1/d_2 . Otherwise the Haar measure would not be invariant with respect to permutations of basis vectors. Then the Lemma follows from Levis Lemma [33] stating that the values of a Lipschitz continuous function on a high-dimensional sphere are everywhere close to its average except from a region whose volume tends to zero for increasing dimension given that L is a constant. \square

This shows that we obtain the same occupation probabilities as in the classical model.

An appropriate quantum scenario for a continuous variable influencing a binary variable required to start with a state in S_X whose total spin is normally distributed in order to obtain the desired distribution of its final energy. This will not be discussed this here.

Also another link between the asymmetry of the causal Markov condition with respect to reversing causal arrows and the asymmetry of the simplicity of conditionals is worth mentioning. Since we have already described this idea in [34] it will only be sketched. According to our pMK principle we expect an asymmetry between the typical shapes of forward vs. backward time conditionals in stochastic processes. Consider a classical physical system whose time evolution is given by a simple Markov chain in discrete time $t \in \mathbb{Z}$. Let $X_1(t), \dots, X_n(t)$ the set of variables describing the system at time t . We assume that the only variables that directly influence the variables at time t are those at time $t-1$, i.e., we exclude instantaneous causal influence among variables within the same time step. We restrict our attention to only two time instants $t, t+1$ (two layers) and denote the variables $X_j(t)$ and $X_j(t+1)$ by C_j and E_j , respectively. The causal structure within these layers is then of the form shown in Fig. 4, i.e., we have only arrows between C_i and E_j for arbitrary i, j (“cause end “effect). The total causal graph is given by continuing this scheme to infinity in both time directions. The causal Markov condition implies that given all variables C_1, \dots, C_n , the variables E_j become stochastically independent., i.e.,

$$P(E_1, \dots, E_n | C_1, \dots, C_n) = \prod_j P(E_j | C_1, \dots, C_n). \quad (18)$$

One can show that the analogue statement for the backward time conditional $P(C_1, \dots, C_n | E_1, \dots, E_n) = \prod_j P(C_j | E_1, \dots, E_n)$ does *not* follow from the

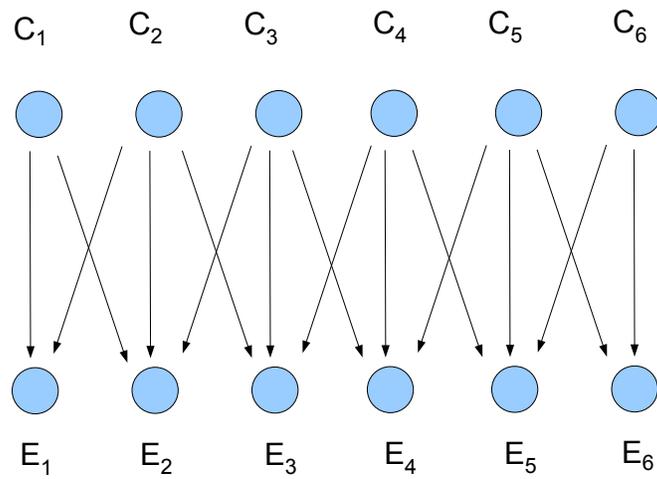


Figure 4: Two time layers of a first order Markovian stochastic process. The first layer represents all relevant variables at time t and the second layer all variables at time $t + 1$.

Markov condition and would therefore even *violate* faithfulness. Note that this is already an asymmetry in the simplicity of forward vs. backward time conditionals: The former has a factorization into simpler terms that is not possible for the latter. In order to make the asymmetry even more obvious we mention that locality conditions like “ j influences only its k neighbors” implies that the conditionals are of low order in the exponential hierarchy (see Section 4), i.e., they become simple in a very well-defined sense. We have thus seen that we can derive simplicity statements for the conditional probability of an effect given its causes if the components of vector-valued cause and effect variables are considered as *separate* variables. On the other hand, there is no such simplicity principle for the backward time conditional. It should be mentioned that this asymmetry does not disappear for stationary processes in contrast to the random walk in Section 2.

8 Computational complexity of Markov kernels and thermodynamic irreversibility

The asymmetry of the complexity of conditional probabilities with respect to causal directions is, in the first place, an intuitive concept. There are several possibilities to find a well-defined meaning. As a first attempt, we have described how to base it on a hierarchy of exponential families. Here we want to describe the same asymmetry with respect to another notion of complexity, namely the complexity classes of computer science.

First we consider a boolean function f with $n + m$ bits input and k bits output. Let $X = (X_1, \dots, X_n)$ be the vector of binary variables that describe the first n input bits and $Z = (Z_1, \dots, Z_m)$ be the vector for the last m input bits. Let furthermore $Y := (Y_1, \dots, Y_k)$ describe the output. Now we interpret X as the *cause*, Y as the *effect* and Z as a *noise* variable that makes the causal mechanism probabilistic. We assume that the components Z_j of the noise variable are stochastically independent, i.e., we have some product probability distribution $P(z_1, \dots, z_k) = P_1(z_1)P_2(z_2)P_k(z_k)$. The following statement is actually almost obvious:

Observation 2 ($P(\text{effect}|\text{cause})$ is Efficiently Computable)

Given a string d including

1. a description of a boolean circuit in terms of elementary gates like AND, NAND, OR, NOR, NOT that computes f and
2. a description of a product probability distribution for Z .

Given some ϵ with $\epsilon = \text{poly}(|d|)$ and some constant $c \in (0, 1)$. The problem to decide for a given pair x, y whether

$$P(y|x) \geq c + \epsilon \quad \text{or} \quad P(y|x) \leq c - \epsilon$$

is in BPP (“Bounded-error, Probabilistic, Polynomial time” [35]). In other words, there is a probabilistic algorithm whose running time increases only polynomially in $|d|$ solving the above decision problem such that the error probability is smaller than some previously specified constant $\delta > 0$.

The “algorithm” for this decision problem is already given by setting the input to x , randomizing the noise variable Z according to the given distribution, simulating the boolean circuit and counting the number of runs with output y .

It should be noted, however, that an exact computation of $P(y|x)$ is not possible in any efficient way provided that the complexity classes $\#P$ (“sharp P”) and BPP do not coincide. To see this, we set $n = 0$ and $k = 1$ i.e., the binary variable Y is only a function of the noise Z . Let the values of the noise variable be uniformly distributed, i.e., $P(z) = 1/2^k$ for all $z \in \{0, 1\}^k$. Then $P(y|x) = P(y)$ is, up to the constant $1/2^k$ simply the number of inputs z for which $f(z) = 1$. The problem to count the number of satisfying inputs for a boolean function (given in so-called conjunctive normal form)

$$f : \{0, 1\}^m \rightarrow \{0, 1\}$$

is complete for the complexity class $\#P$ which is believed to be a huge class containing extremely hard computational problems [35]. However, the hardness of giving *exact* solutions is probably of minor relevance and we will in the sequel, again, consider *approximative* solutions.

We will see that $P(\text{cause}|\text{effect})$ is even hard to compute *approximately*:

Theorem 1 (Computing $P(\text{cause}|\text{effect})$ is NP-Hard)

Let the assumptions and definitions be as in Observation 2 and $P(x)$ be some product measure. The problem to decide whether $P(x|y) \geq 2/3$ or $P(x|y) \leq 1/2$ is NP hard.

Proof: The statement can be proven for $m = 0$, i.e., without introducing a noise variable Z . Let $g : \{0, 1\}^n \rightarrow \{0, 1\}$ be a boolean function. To decide whether there is a binary string x with $g(x) = 1$ is known to be NP complete [35]. We will chose g such that $g(0, \dots, 0) = 0$. It is clear that the restriction to this class of functions g remains NP complete. Then we define a function f by $f(x) = g(x) \vee h(x)$ with $h(x) = 1$ for $x = (0, \dots, 0)$ and $h(x) = 0$ otherwise. Let now the distribution of x be uniform and consider $P(X = (0, \dots, 0) | Y = 1)$. If g has no satisfying input x we have $P(X = (0, \dots, 0) | Y = 1) = 1$ since $x = (0, \dots, 0)$ is the only satisfying input for f . If g has a satisfying input, f has at least two satisfying inputs and $P(X = (0, \dots, 0) | Y = 1) \leq 1/2$. \square

To discuss the reason for the asymmetry between the complexity of computing $P(\text{effect}|\text{cause})$ and $P(\text{cause}|\text{effect})$ we extend the boolean function f to a *bijective* function F . It has been argued [36] that logically irreversible functions lead to energy dissipation and thus thermodynamically reversible computation is only possible by computing only reversible functions [37, 38].

The Toffoli gate [38] provides a useful method to simulate conventional boolean circuits by reversible ones. TOFFOLI is a gate with three inputs a, b, c and three outputs a', b', c' such that $a' = a, b' = b$ and $c' = c \oplus (a \cap b)$ where \oplus denotes the exclusive or (“XOR”). In words, the third bit c is inverted if and only if a and b are true. TOFFOLI can simulate a NAND by setting the third input to $c = 1$. Then we have $c' = \overline{a \cap b}$ and the outputs a', b' can be ignored (note that the existence of “useless” output (“data garbage”) and the need for adjusting certain input bits to fixed values is characteristic for reversible computation). Since NAND is universal and gates like AND, OR, NOR, NOT can be simulated using a small number of NAND gates we can simulate every given boolean circuit with TOFFOLI gates efficiently. Furthermore, a corresponding reversible circuit can efficiently found by substituting every single gate with some TOFFOLI gates. We have then an algorithm to extend f (having $n + m$ input bits and k outputs) to a bijective boolean function F with $\tilde{n} = n + m + r$ inputs and $k + l = \tilde{n}$ outputs (described by Y, W) such that for some additional r -bit string v the restriction of $F(x, z, v)$ to the first k output bits coincides with $f(x, z)$. We may without loss of generality consider the ancilla variables V as additional noise variables since we can specify the corresponding distributions such that the “noise” variable takes always the same value. Hence we obtain a boolean

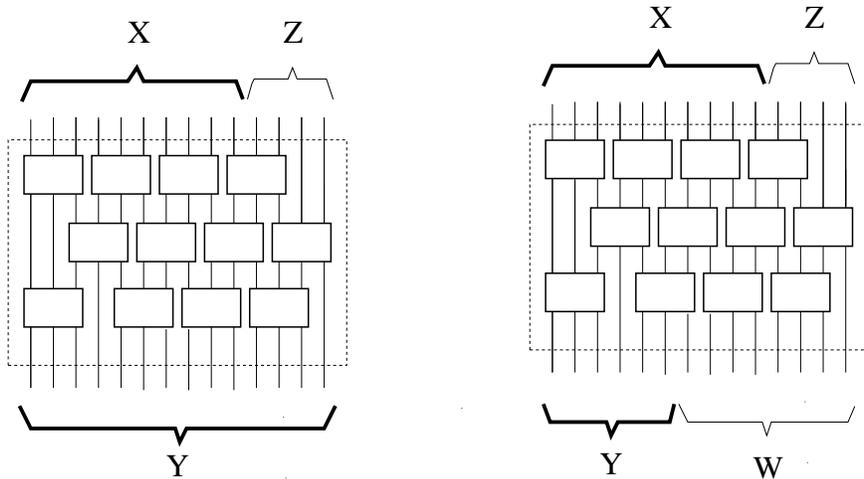


Figure 5: Reversible network as a model for a cause-effect relation. Left: The “cause variable X ” and the noise variable Z determine jointly the effect variable Y and vice versa. Both conditionals $P(Y|X)$ and $P(X|Y)$ are efficiently computable. Right: The effect variable Y does not completely determine the cause variable X and the noise Z . The conditional $P(X|Y)$ is in general not efficiently computable.

function F with $n+m$ input bits and $k+l = m+n$ output bits such that the conditional probabilities for the output y given the input x coincide with the probabilities $P(Y|X)$ generated by the function f . We have then simulated the causal effect from X to Y by a completely reversible process using a noise variable Z and restricting the output Y, W to Y (see Fig. 8).

It is important to note that the inverse function F^{-1} can be efficiently computed: every TOFFOLI gate is its own inverse. We can therefore simulate the circuit in backward direction. It is remarkable that in such a setup not only $P(Y|X)$ but also $P(X|Y)$ is efficiently computable provided that Y is the *complete* output. This is because we can compute the complete input $(x', z) = F^{-1}(y)$ from y . If $x' = x$ we know already that $P(x|y) = 0$. Otherwise we compute the probability for z which is exactly the desired value for $P(x|y)$.

It should be emphasized that *local* reversibility of the network is essential, i.e., it is not sufficient that the computed function F is *bijective*. In order to compute F^{-1} efficiently we have inverted each single gate.³

³An example for a bijective function whose inverse is believed to be not efficiently

We conclude: If y is the complete output of a locally reversible circuit we can compute $P(X|Y)$ efficiently. This means: Using the complete effect of the cause we would be able to compute $P(\text{cause}|\text{effect})$ efficiently, no matter whether we have probabilistic causality where y is additionally influenced by a latent variable.

In order to understand the reason for the asymmetry between cause and effect we should recall that it has here actually been put in *by assumption*: We have assumed that all input bits were uncorrelated, in particular that noise and cause variables are stochastically independent. We could think of the time-inverted scenario where x, z is distributed according to some probability measure P having the property that the distribution of y, w is a product measure. Then we could efficiently compute $P(x|y)$ applying the method in Observation 2 and obtain an efficient simulation of the time-reversed circuit. Obviously, such a scenario is unlikely unless *we have calculated* how to randomize the input such that a product distribution of the output is obtained.

The question is why uncorrelated inputs generating correlated outputs are natural to assume but correlated inputs and uncorrelated outputs are artificial. The first possible answer is that the joint distribution of inputs has been created by the acting subject. This is, however, not really satisfactory because we would like to also make statements about processes in *nature*. A more interesting answer is based on Reichenbach's principle of the common cause: Observed stochastic dependences between physical systems can only be explained by interactions that happened in the past. Interactions that happen in the future do not *explain* the dependences even though they may also occur. Recalling that this makes exactly the difference between a causal fork and a collider (see Fig 1), we have again described a link between the asymmetry of the complexity of the conditionals and the constrained-based causal inference principles. Since Reichenbach's principle has proven to be useful as a basis for causal reasoning, our world seems to contain an abundance of stochastically independent systems. If stochastic independence would be a rare exception as an initial condition the inference principle would be almost void. Given that our world contains many objects that have never been interacting, it is a priori likely to obtain stochastically independent sys-

computable (because it is not *locally* invertible) is $f(n) = (a^n) \bmod b$ where a and b are chosen appropriately and $n \in \{0, \dots, b-1\}$. The security of the crypto-system RSA relies on the assumption that the inverse of this function is hard to find.

tems when choosing randomly some objects that are not located too close together.

If we think of the bits of the devices discussed above as toy models for states of physical systems that have never been interacting before some time t_0 . After they interact, a collective dynamics (represented by the circuit) creates stochastic dependences between initially independent systems which makes on the one hand the statistical asymmetries discussed above and we have, again, the same asymmetry that we already considered in the context of the asymmetries between fork and collider.

This argument could, in a similar way, also be phrased in terms of von-Neumann entropy⁴.

9 Justifying independence by symmetry: the de Finetti theorem

In the preceding section we have emphasized that the input and output bits of the devices may thought of representing the states of physical systems that are initially stochastically independent because they have never been interacting in the past. In this section, we want to mention that the assumption of no interaction in the past is stronger than necessary. Even though there may have been very effective interactions in the past, we often have to assign product states just for *symmetry* reasons. To explain the idea we introduce the concept of *exchangeable* probability assignments. First we introduce the notion of symmetric probability assignments [41]. Let X_1, \dots, X_n be random variables, each having the value set \mathcal{X} . Then a joint distribution on these n variables is called *symmetric* if $P(x_{\pi(1)}, \dots, x_{\pi(n)}) = P(x_1, \dots, x_n)$ for every permutation π . A symmetric joint measure is called *exchangeable* if there exists for every $m \in \mathbb{N}$ a joint measure on $n + m$ random variables which is symmetric and whose restriction to X_1, \dots, X_n coincides with P . The de Finetti theorem states that every exchangeable measure is a mixture of

⁴Even though the relation between thermodynamic entropy and statistical entropy is subject of controversial discussions I assume (in agreement with Landauer's principle) that statistical entropy has indeed thermodynamical relevance. It has been argued that the Kolmogorov complexity of physical states contributes also to the physical entropy, in addition to the statistical entropy [39, 40]. But from this point of view, the application of the circuit would also increase the physical entropy.

identical product measures, i.e., P can be written as

$$P(x_1, \dots, x_n) = \int Q(R)R(x_1) \dots R(x_n) dR.$$

Here R runs over the set \mathcal{R} of all possible distribution for one variable X_j and $Q(R)$ is a probability distribution⁵ on \mathcal{R} .

Assume now we have a box containing a large number of identical physical particles or objects. For symmetry reasons we should assume that the joint state is exchangeable. After drawing as many samples of these objects we may determine R up to a small error. Hence the conditional joint state of the remaining system, given the statistics of the observations, is *one specific* product measure

$$R(x_1) \dots R(x_n).$$

We could, for instance, assume that the bath contains two-level systems like spin-1/2 systems. Assuming that their quantum state is diagonal with respect to some preferred basis (due to decoherence effects) we have just a joint distribution on classical bits. After drawing a large number of these spin systems and counting the number of systems found in their upper level we have a product state on the *remaining* bits. If we feed the input wires of our logical circuits in Section 8 with the values of these bits we obtain indeed stochastically independent inputs.

The fact that nature provides us with an abundance of mutually stochastically independent systems (that represent, for instance, the noise variables entering into every node in a Markovian causal network) is therefore not *necessarily* due to the fact that all these systems have never been interacting.

This raises the question whether we should modify Reichenbach's principle of the common cause in the following way:

Complete the statement "two systems that have never been interacted in the past (i.e., no common cause exists) should be stochastically independent" with the less concise formulation: "Given a huge ensemble of identical systems. After drawing sufficiently many samples we should assign a product distribution to every pair of system unless we have some specific knowledge that refers to the history of interactions that refers to this particular pair."

⁵For a generalization to quantum states see [42].

10 Conclusions

We have described several physical settings where the conditional probability for an effect given its cause is less complex than the probability for the cause given its effect. Here we have considered complexity with respect to a hierarchy of exponential families as well as with respect to computational complexity.

To link this kind of “asymmetric Occam’s Razor principles” with the thermodynamic arrow of time we have constructed models where the statistical asymmetries between cause and effect are implications of the irreversibility of mixing processes. The common root between all the known asymmetries is therefore the tendency of *specific* initial conditions to evolve into *typical* final states. Specific initial conditions can, for instance, be product probability distributions of joint systems that evolve typically to distributions with strong stochastic dependences. The fact that specific initial conditions occur often rather than specific final conditions is at least linked with the second law if it is not considered as its essential content.

Already Reichenbach discussed the statistical asymmetry between causal forks and colliders (on three random variables) in the context of the direction of time. The asymmetry of the causal Markov condition with respect to inverting causal directions formalizes this asymmetry in a more general setting. However, we believe that this asymmetry is only the tip of an iceberg and further asymmetries have yet to be discovered. Since it is impossible to draw reliable causal conclusions from statistical observations that do not involve interventions, we have to restrict ourselves to finding causal inference rules which are *often* valid. These have to be based upon observing which transition probabilities $P(\text{effect}|\text{cause})$ are likely to occur in nature and which one are likely to correspond to non-causal conditionals. To explore this asymmetry in a systematic way is an important challenge for both machine learning and theoretical physics.

References

- [1] H. Reichenbach. *The direction of time*. Dover, 1999.
- [2] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Lecture Notes in Statistics. Springer, New York, 1993.

- [3] J. Pearl. *Causality*. Cambridge University Press, 2000.
- [4] J. Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings, Cognitive Science Society*, pages 329–334, Greenwich, Albex, 1985.
- [5] N. Cartwright. *How the laws of physics lie*. Oxford, U.K., Clarendon, 1983.
- [6] D. Heckerman. A Bayesian approach to causal discovery. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 141–165. 1999.
- [7] C. Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 411–418. Morgan Kaufmann, Montreal, QU, 1995.
- [8] C. Rasmussen and Z. Ghahramani. *Advances in Neural Information Processing Systems 13*, chapter Occam’s Razor. MIT Press, 2001.
- [9] Y. Kano and S. Shimizu. Causal inference using nonnormality. In *In Proc. International Symposium on Science of modeling-The 30th Anniversary of the Information Criterion (AIC)*, pages 261–270, Tokyo, Japan, 2003.
- [10] X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceeding of the 9th Int. Symp. Art. Int. and Math.*, Fort Lauderdale, Florida, 2006.
- [11] X. Sun, D. Janzing, and B. Schölkopf. Exploring the causal order of binary variables via plausible markov kernels. To be presented at the 11th European Symposium on Artificial Neural Networks, April 2007.
- [12] N. Friedman and I. Nachman. Gaussian process networks. In *Uncertainty in Artificial Intelligence*, 2000.
- [13] D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgement under uncertainty: heuristics and biases*. Cambridge University Press, 1982.

- [14] O. Stern. The method of molecular rays. Nolel Price Lecture, Imprimerie Royale Norstedt and Soner, Stockholm, 1948. <http://www.nobel.se/physics/laureates/1943/stern-lecture.html>.
- [15] D. Walls and G. Milburn. *Quantum optics*. Springer, Heidelberg, 1994.
- [16] V. Vapnik. *Statistical learning theory*. John Wileys & Sons, New York, 1998.
- [17] R. Christensen. *Log-linear models*. Springer Verlag, 1990.
- [18] L. Goodman. *Analyzing Qualitative/Categorical Data*. Addison Wesley, 1978.
- [19] D. Dowson and A. Wragg. Maximum-entropy distributions having prescribed first and second moments. *IEEE Trans. Inf. Th.*, 19(5):689–693, 1973.
- [20] S. Amari. Information geometry on hierrachy of probability distributions. *IEEE Trans. Inf. Th.*, 47(5):434–438, 2001.
- [21] H. Callen. *Thermodynamics*. J. Wiley and Sons, New York, 1960.
- [22] B. McCoy and T. Wu. *The Two-Dimensional Ising Model*. Haward University Press, Cambridge Massachusetts, 1973.
- [23] E. Jaynes. Information theory and statistical mechanics. In K. Ford, editor, *Statistical Physics*, page 181. Benjamin, 1963.
- [24] D. Poulin. Macroscopic observables. *Phys. Rev. A*, 71:022102, 2005.
- [25] X. Sun, D. Janzing, and B. Schölkopf. Distinguishing between cause and effect via kernel-based complexity measures for conditional distributions. In *Proceedings of the 15th European Symposium on Artificial Neural Networks*,, pages 441–446, 2007.
- [26] A. Allahverdyan and D. Janzing. Relating the thermodynamic arrow of time to the causal arrow. arXiv:0708.1175.
- [27] A. Allahverdyan and T. Nieuwenhuizen. Steady adiabatic state: Its thermodynamics, entropy production, energy dissipation, and violation of Onsager relations. *Phys. Rev. E*, 62(1):845–850, 2000.

- [28] T. Cover and J. Thomas. *Elements of Information Theory*. Wileys Series in Telecommunications, New York, 1991.
- [29] E. T. Jaynes. Gibbs vs. Boltzmann entropies. *Am. J. Phys.*, 33:391, 1965.
- [30] O. Penrose and I. Percival. The direction of time, 1962.
- [31] W. J. Gibbs. *Elementary Principles in Statistical Mechanics*. Ox Bow Press, 1902.
- [32] J. Lebowitz. Macroscopic dynamics, time’s arrow and Boltzmann entropy. *Physica A*, 194, pp 1–27, 1993.
- [33] Sandu Popescu, Anthony J. Short, and Andreas Winter. The foundations of statistical mechanics from entanglement: Individual states vs. averages, arXiv.org:quant-ph/0511225.
- [34] D. Janzing, X. Sun, and B. Schölkopf. Causal inference based on properties of conditional distributions. *submitted to JMLR*.
- [35] Ch. Papadimitriou. *Computational Complexity*. Addison Wesley, Reading, Massachusetts, 1994.
- [36] R. Landauer. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop.*, 5:183–191, 1961.
- [37] C. H. Bennett. Logical reversibility of computation. *IBM J. Res. Develop.*, 17:525–532, 1973.
- [38] T. Toffoli. Reversible computing. *MIT Report MIT/LCS/TM-151*, 1980.
- [39] W. Zurek. Algorithmic randomness and physical entropy. *Phys Rev A*, 40(8):4731–4751, 1989.
- [40] C. Mora, B. Kraus, and H. Briegel. Quantum Kolmogorov complexity and its applications. arXiv.org:quant-ph/0610109.
- [41] *Probabilità e Induzione – Induction and Probability*. Bibliotheca di Statistica, CLUEB, Bologna, 1993.

- [42] C. Caves, C. Fuchs, and Schack. R. Unknown quantum states: The quantum de Finetti representation. *J. Math. Phys.*, 43(9):4537–4559, 2002.