J Stat Phys (2010) 138: 767–779
DOI 10.1007/s10955-009-9897-8

On the Entropy Production of Time Series with Unidirectional Linearity

Dominik Janzing

Received: 5 August 2009 / Accepted: 25 November 2009 / Published online: 10 December 2009

Abstract There are non-Gaussian time series that admit a causal linear autoregressive moving average (ARMA) model when regressing the future on the past, but not when regressing the past on the future. The reason is that, in the latter case, the regression residuals are not statistically independent of the regressor. In previous work, we have experimentally verified that many empirical time series indeed show such a time inversion asymmetry.

For various physical systems, it is known that time-inversion asymmetries are linked to the thermodynamic entropy production in non-equilibrium states. Here we argue that unidirectional linearity is also accompanied by entropy generation.

To this end, we study the dynamical evolution of a physical toy system with linear coupling to an infinite environment and show that the linearity of the dynamics is inherited by the forward-time conditional probabilities, but not by the backward-time conditionals. The reason is that the environment permanently provides particles that are in a product state before they interact with the system, but show statistical dependence afterwards. From a coarse-grained perspective, the interaction thus generates entropy. We quantitatively relate the strength of the non-linearity of the backward process to the minimal amount of entropy generation.

The paper thus shows that unidirectional linearity is an indirect implication of the thermodynamic arrow of time, given that the joint dynamics of the system and its environment is linear.

Keywords Arrow of time · Entropy production · Irreversible processes · Time series · ARMA models

1 Unidirectional Linearity in Time Series

To study the implications and the different versions of the thermodynamic arrow of time has attracted interest of theoretical physicists and philosophers since a long time [1–7]. More

D. Janzing (✉)
Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany
e-mail: dominik.janzing@tuebingen.mpg.de

D. Janzing (✉)
e-mail: janzing@ira.uka.de

specifically, it is the question how the difference between time reversibility of microscopic physical dynamics is consistent with the existence of irreversible processes on the macroscopic level. The most prominent examples of irreversibilities (e.g. heat always flows from the hot to the cold reservoir, never vice versa, every kind of energy can be converted into heat, but not vice versa) can directly be explained by the fact that the processes generate entropy and their inverted counterpart is therefore forbidden by the second law.

Here we describe an asymmetry between past and future whose connection to the second law is more subtle. An extensive analysis of more than 1000 time series [8] showed that there are many cases where the statistics could be better explained by a linear autoregressive model from the past to the future and fewer cases where regressing the past on the future yields a better model [8, 9]. The goal of this paper is to describe how this asymmetry is connected to non-equilibrium thermodynamics. It has been shown for various physical models (e.g. [10–12], and also in a more abstract setting [13]) that statistical asymmetries between past and future can be related to thermodynamic entropy production.[1]

This paper does not focus on general time-asymmetries between past and future, but only on the unidirectional linearity observed in our experiments. To link this phenomenon to the entropy production we will try to use only those assumptions about the underlying physical system that are necessary to make the case and try to simplify the argument as much as possible. The ingredients are (1) a system interacting with an environment consisting of infinitely many copies of the same system where the joint system is initially in a product state, each copy having an abstract vector space as phase space, (2) linear volume preserving dynamical equations for the joint system. We will not refer to any other ingredients from physics, like energy levels, thermal Gibbs states, etc. Of course, this raises the question of how to define entropy production. Here, we interpret the generation of statistical dependence among an increasing number of particles as a *phenomenological* entropy production because the sum of the Shannon entropies of the particles increases.

To describe the model more precisely, we start with preliminary remarks on statistical dependence. First we introduce the following terminology.

Definition 1 (Linear Models) The joint distribution $P_{X,Y}$ of two real-valued random variables X and Y is said to admit a linear model $X \to Y$ with additive noise (linear model, for short) if Y can be written as

$$Y := \alpha X + \epsilon$$

with a structure coefficient $\alpha \in \mathbb{R}$ and a noise term ϵ that is statistically independent of X ($X \perp\!\!\!\perp \epsilon$, for short).

It should be emphasized that statistical independence between two random variables Z, W is defined by factorizing probabilities

$$P_{Z,W} = P_Z \otimes P_W,$$

[1] It should be mentioned, however, that some authors (e.g. [14]) *define* "entropy production" by a relative entropy distance between a forward and a backward stochastic process. This is motivated by the fact that the relative entropy distance has been shown to coincide with an increase of thermodynamic entropy for various physical models. Since we are not aware of any model studied in the literature that is sufficiently general to include the system described in this paper, we will not, a priori, assume such a connection between asymmetry and entropy generation.

instead of the weaker condition of uncorrelatedness, which is defined by factorizing expectations:

$$\mathbb{E}(ZW) = \mathbb{E}(Z)\mathbb{E}(W). \tag{1}$$

Uncorrelatedness between X and ϵ is automatically satisfied if α is chosen to minimize the least square error. Whether or not $P_{Y,X}$ admits a linear model from X to Y is actually a property of the conditional $P_{Y|X}$ alone (provided that the conditional distribution $P(Y|X = x)$ is defined for all $x \in \mathbb{R}$). We will therefore also say that the *conditional* $P_{Y|X}$ admits a linear model.

Except for the trivial cases of independence or deterministic dependence, $P_{X,Y}$ can only admit linear models in both directions at the same time if it is bivariate Gaussian. This can be shown using the theorem of Darmois and Skitovich [15, 16], which we rephrase now because it will also be used later.

Lemma 1 (Theorem of Darmois & Skitovich) *Let $Y_1, Y_2, \ldots, Y_k$ be statistically independent random variables and the two linear combinations*

$$W_1 := \sum_{j=1}^{k} \beta_j^{(1)} Y_j,$$

$$W_2 := \sum_{j=1}^{k} \beta_j^{(2)} Y_j$$

be independent. Then all Y_j with $\beta_j^{(1)}\beta_j^{(2)} \neq 0$ are Gaussian.

In the context of inferring causal directions from statistical data, it has been proposed to consider the direction of the linear model as the *causal* direction [17, 18]. In [8] we have shown that the same idea can be used to solve the following binary classification problem: Given numbers $X_1, X_2, X_3, \ldots$ that are known to be the values of an empirical time series in their correct or in their time reversed order. Decide whether $X_1, X_2, X_3, \ldots$ or $\ldots, X_3, X_2, X_1$ is the correct order. Certainly, this problem is less relevant than the problem of inferring causal directions since our experiment required to artificially blur the true direction even though it was actually known. The motivation for our study was to test causal inference principles by applying them to this artificial problem.

To explain the "time direction inference rule" proposed in [8] we first introduce an important class of stochastic processes [19]:

Definition 2 (ARMA Models) We call a time series $(X_t)_{t\in\mathbb{Z}}$ an autoregressive moving average process of order (p, q) if it is weakly stationary and there is an iid noise ϵ_t with mean zero such that

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \quad \forall t \in \mathbb{Z},$$

where ϕ_i and θ_j are real-valued coefficients and ϵ_t is independent of all $X_{t'}$ with $t' < t$. For $q = 0$ the process reduces to an autoregressive process and for $p = 0$ to a moving average process. The short-hand notations are $ARMA(p, q)$, $AR(p)$, and $MA(q)$. The first and the second sums are called the AR-part and the MA-part, respectively.

The process is called causal[2] if

$$\epsilon_t \perp\!\!\!\perp X_{t-i} \quad \forall i > 0. \tag{2}$$

Note that a process is called weakly stationary if the mean $\mathbb{E}(X_t)$ and second order moments $\mathbb{E}(X_t X_{t+h})$ are constant in time [19]. In [8] we have shown the following theorem:

Theorem 1 (Non-invertibility of Non-Gaussian Processes) *If $(X_t)_{t\in\mathbb{Z}}$ is a causal ARMA process with non-vanishing AR-part, then $(X_{-t})_{t\in\mathbb{Z}}$ is a causal ARMA process if and only if (X_t) is a Gaussian process.*

In particular, a process with long-tailed distributions like e.g. Cauchy can only be causal in one direction (provided that it has an AR-part). In [8] we have postulated that whenever a time series has a causal ARMA model in one direction but not the other the former is likely to be the true one. Our experiments in [8] support this hypothesis, but we want to give the reader the opportunity to judge the strength of the evidence by himself. Therefore, we need to add some comments on the practical implementation.

Testing condition (2) yields p-values for the hypothesis of independence. The performance of our inference method depends heavily on how these p-values are used to decide whether a linear model is accepted for one and only one of the directions. Our rule depends on two parameters α and δ, the significance level and the gap, respectively. We say that an ARMA model is accepted for one direction but not the other if the p-value for the direction under consideration is above α and it is below α for the converse direction and, moreover, the gap is at least δ. By choosing a small value α and a large value δ one gets fewer decisions but also the fraction of wrong classifications decreases. On 1180 empirical time series from EEGs [8] we where able to classify around 82% correctly when the parameters are set to yields decisions for about 4% of the time series. When decisions were made for a larger fraction of time series, the number of correct answers still significantly exceeded chance level. Qualitatively similar results were obtained for 200 time series from different areas, like finance, physics, transportation, crime, production of goods, demography, economy, neuroscience, and agriculture [9]. It thus seems that nature more often generates linearity in forward than in backward time.

2 Physical Toy Model

Here we describe a physical model that suggests that the observed asymmetry is an implication of generally accepted asymmetries between past and future. We assume that the values X_t are observables of a classical physical system.[3] For our toy model, we use only two properties of physical models that we consider decisive for the argument:

(1) The state of a system is a point in some phase-space $\mathcal{S}$ that is a sub-manifold of $\mathbb{R}^n$.
(2) The dynamical evolution of an isolated system is given by a family M_t of volume-preserving bijections on $\mathcal{S}$.

[2] Reference [19] chooses a different definition, but we have argued in [8] that it is equivalent to ours.

[3] Of course, such an embedding is hard to imagine for time series from stock markets, for instance. However, other time series, e.g., EEG-data, are closer related to *physical* observables.

Due to Liouville's Theorem, the latter condition holds for the dynamics of all Hamiltonian systems, i.e., those that are energetically closed. Non-Hamiltonian dynamics occur in physics only as the time-evolution of *open* system. Since we want to include all relevant parts of the environment of the physical system under consideration we consider the joint system as closed.

For simplicity, we restrict the attention to an $AR(1)$ process:

$$X_t = \phi X_{t-1} + \epsilon_t. \tag{3}$$

We will now interpret X_t as a physical observable of a system $S^{(0)}$, whose state is changed by interacting with its environment. The latter consists of an infinite collection of subsystems $S^{(j)}$ with $j \in \mathbb{Z} \setminus \{0\}$. Each subsystem is described by the real-valued observable $Z^{(j)}$. Its value at time t is denoted by $Z_t^{(j)}$, hence $X_t = Z_t^{(0)}$, but we will keep the notation X_t whenever its special status among the variables should be emphasized.

Then we define a joint time evolution by

$$Z_{t+1}^{(0)} = \gamma_{11} Z_t^{(-1)} + \gamma_{12} Z_t^{(0)}, \tag{4}$$

$$Z_{t+1}^{(1)} = \gamma_{21} Z_t^{(-1)} + \gamma_{22} Z_t^{(0)}, \tag{5}$$

$$Z_{t+1}^{(j)} = Z_t^{(j-1)} \quad \text{for } j \neq 0, 1. \tag{6}$$

The dynamics thus is a concatenation of the map Γ on the variable pair $(Z_t^{(-1)}, Z_t^{(0)})$, given by the entries γ_{kl}, with a shift propagating the state of subsystem $S^{(j)}$ to $S^{(j+1)}$.

The environment may be thought of as a beam of particles that approaches site $S^{(0)}$, interacts with it, and disappears to infinity; we have discretized the propagation only to make it compatible with the discrete stochastic process. The interaction is given by Γ. The phase space of the systems $S^{(j)}$ may be larger than one-dimensional, but we assume that the variables $Z_t^{(j)}$ define the observables that are relevant for the interaction. To ensure conservation of volume in the entire phase space, Γ needs to be volume-preserving, i.e. $|\det(\Gamma)| = 1$. Since our model should be interpreted as the discretization of a *continuous* time process we assume $\Gamma \in SL(2)$.

One checks easily that the above dynamical system generates for $t > 0$ the causal $AR(1)$-process

$$X_t = \gamma_{12} X_{t-1} + \epsilon_t \quad \text{with } \epsilon_t := \gamma_{11} Z_{t-1}^{(-1)},$$

if we impose the initial conditions

$$Z_0^{(j)} \quad \text{i.i.d. with some distribution } Q. \tag{7}$$

Actually, it would be sufficient to impose independence only for the non-positive j, but later it will be convenient to include also positive values j and assume that the whole ARMA process has a starting time $t = 0$. This will make it easier to track the increase of statistical dependence over time.

We will now show that, under generic conditions, the dynamics creates statistical dependence between the subsystems. We will later see that this is the reason why the time-inverted version of the above scenario would not be a reasonable physical model for the process (X_{-t}). We need the following Lemma:

Lemma 2 (Dependence Created by Sequences of Adjacent Operations) *Let* $\Gamma \in SL(2)$ *have non-diagonal and diagonal entries. Denote by* $\Gamma_{l,l+1}^{(n)}$ *the embedding into the two-dimensional subspaces of* $\mathbb{R}^n$ *that correspond to consecutive components* $l, l+1$ *with*

$l = 0, \ldots, n-1$, *i.e.*,

$$\Gamma^{(n)}_{l,l+1} := \mathbf{1}_{l-1} \oplus \Gamma \oplus \mathbf{1}_{n-l-1},$$

where $\mathbf{1}_m$ *denotes the identity matrix in* m *dimensions. Let* P *be a non-Gaussian distribution on* $\mathbb{R}$. *Then the application of*

$$\Gamma^{(n)}_{0,1} \circ \Gamma^{(n)}_{2,3} \circ \cdots \circ \Gamma^{(n)}_{n-2,n-1}$$

to $\mathbb{R}^n$ *turns the product distribution* $P^{\otimes n}$ *into a non-product distribution.*

Proof Due to Lemma 1, $\Gamma^{(n)}_{n-2,n-1}$ generates dependence between the last and the second last component. Since none of the other operations acts on the last component, the dependence between the last component and the joint system given by the remaining $n-2$ components, is preserved. □

To apply Lemma 2 to our system, it is sufficient to focus on the region of the chain on which the statistical dependence has been generated after the time t under consideration. It is given by

$$S^{0,\ldots,t} := S^{(0)} \times S^{(1)} \times \cdots \times S^{(t)}. \tag{8}$$

Its state at time t can be found by using the variable transformation

$$(Z^{(0)}_t, Z^{(1)}_t, \ldots, Z^{(t)}_t) = (\Gamma^{(t+1)}_{0,1} \circ \Gamma^{(t+1)}_{1,2} \circ \cdots \circ \Gamma^{(t+1)}_{t-1,t})(Z^{(-j)}_0, \ldots, Z^{(0)}_0), \tag{9}$$

and all the other sites are still jointly independent and independent of region (8). If the relation between X_t and X_{t+1} is non-trivial (i.e., neither deterministic nor independent) Γ must have diagonal and non-diagonal entries, which implies that the state given by the left side of (9) is not a product state.

The following argument shows that the dependence between the outgoing particles is closely linked to the irreversibility of the scenario: The fact that the time evolution generates a *causal* $AR(1)$-process is ensured by independence of $Z^{(0)}_t, Z^{(-1)}_t, Z^{(-2)}_t, \ldots$ describing the incoming particles. If the variables $Z^{(1)}_t, Z^{(2)}_t, \ldots$ are also independent we can run the process backwards to induce the *causal* $AR(1)$-process (X_{-t}). However, by virtue of Theorem 1, this is only possible for (X_t) Gaussian.

Summarizing the essential part of the argument, the joint distribution $P_{X_t, X_{t+1}}$ has a linear model from X_t to X_{t+1} but not vice versa because the incoming particles are jointly independent but the outgoing particles are dependent. Now we show a quantitative relation between the non-linearity in backward time direction and the generated dependence. To this end, we measure the strength of the statistical dependence of the joint system as follows. If a system consists of finitely many subsystems its multi-information is defined by

$$I(Y_1, \ldots, Y_k) := \sum_{j=1}^{k} H(Y_j) - H(Y_1, \ldots, Y_k).$$

Here, $H(.)$ is the differential Shannon entropy [20]

$$H(Y_1, \ldots, Y_n) := -\int p(y_1, \ldots, y_n) \log p(y_1, \ldots, y_n) dy_1 \cdots dy_n,$$

where $p(y_1, \ldots, y_n)$ denotes the joint probability density of the random variables $Y_1, \ldots, Y_n$. For $k = 2$, the multi-information coincides with the usual mutual information $I(Y_1 : Y_2) = H(Y_1) + H(Y_2) - H(Y_1, Y_2)$. Multi-information is always non-negative and zero if and only if the variables are jointly independent. For our infinite system we define multi-information as follows:

Definition 3 (Multi-information) The multi-information of the joint system of all $S^{(j)}$ at time t is defined by

$$I(t) := \lim_{m\to\infty} I(Z_t^{(-m)}, Z_t^{(-m+1)}, \ldots, Z_t^{(m)}),$$

whenever the limit exists.

If the sum of the entropies of all subsystems is interpreted as a coarse-grained entropy, multi-information measures the difference between coarse-grained and non-coarse-grained entropy. The increase of multi-information in time can thus be interpreted as a phenomenological increase of entropy. Here, it can easily be computed:

Lemma 3 (Multi-information as Pairwise Information) *Let the initial state of $S^{-\infty\ldots\infty}$ satisfy the conditions* (7) *and the time evolution be given by* (4) *to* (6) *with $\Gamma \in SL(2)$. Then the generation of multi-information per time step is given by*

$$I(t) - I(t-1) = I(Z_t^{(0)} : Z_t^{(1)}) \quad \forall t \geq 0.$$

Proof We consider the state of the system $S^{0,\ldots,t}$ at time t that we would have obtained if the interaction had been inactive (i.e., $\Gamma = \mathbf{1}$) during the last time step. This state is described by the transformed variables

$$(\tilde{Z}_t^{(0)}, \ldots, \tilde{Z}_t^{(t)}) := (\Gamma_{1,2}^{(t+1)} \circ \Gamma_{2,3}^{(t+1)} \circ \cdots \circ \Gamma_{t-1,t}^{(t+1)})(Z_0^{(-t)}, \ldots, Z_0^{(0)}). \tag{10}$$

We have

$$I(\tilde{Z}_t^{(0)}, \ldots, \tilde{Z}_t^{(t)}) = I(t-1),$$

because the shift part of the dynamics is irrelevant. Hence,

$$I(t) - I(t-1) = I(Z_t^{(0)}, \ldots, Z_t^{(t)}) - I(\tilde{Z}_t^{(0)}, \ldots, \tilde{Z}_t^{(t)}), \tag{11}$$

since the systems $S^{(j)}$ with j outside the interval $[0, \ldots, t]$ do not contribute to the overall multi-information. The true state of system $S^{0,\ldots,t}$ at time t is, by assumption, given by additionally applying $\Gamma_{0,1}^{(t)}$ to (10), i.e.,

$$(Z_t^{(0)}, \ldots, Z_t^{(t)}) = \Gamma_{0,1}^{(t)}(\tilde{Z}_t^{(0)}, \ldots, \tilde{Z}_t^{(t)}).$$

We thus obtain

$$I(t) - I(t-1) = \sum_{j=0}^{t} H(Z_t^{(j)}) - H(Z_1, \ldots, Z_t^{(t)}) - \left(\sum_{j=0}^{t} H(\tilde{Z}_t^{(j)}) + H(\tilde{Z}_1, \ldots, \tilde{Z}_t^{(t)}) \right)$$

$$
\begin{aligned}
&= \sum_{j=0}^{t} H(Z_t^{(j)}) - \sum_{j=0}^{t} H(\tilde{Z}_t^{(j)}) \\
&= \sum_{j=0}^{2} H(Z_t^{(j)}) - H(Z_t^{(0)}, Z_t^{(1)}) \\
&\quad - \left(\sum_{j=0}^{2} H(\tilde{Z}_t^{(j)}) - H(\tilde{Z}_t^{(0)}, \tilde{Z}_t^{(1)}) \right) \\
&= I(Z_t^{(0)} : Z_t^{(1)}) - I(\tilde{Z}_t^{(0)} : \tilde{Z}_t^{(1)}) \\
&= I(Z_t^{(0)} : Z_t^{(1)}),
\end{aligned}
$$

where the second equality holds because $\Gamma_{0,1}^{(t)}$ preserves the joint entropy of $S^{0,\dots,t}$ and the third one holds because it also preserves all the marginal entropies for $j \neq 0, 1$ and the joint entropy of $S^{0,1}$. The last equality is due to the independence of $\tilde{Z}_t^{(0)}$ and $\tilde{Z}_t^{(1)}$. □

To show the link between the amount of generated multi-information and the non-linearity of the backward process, we measure the latter as follows.

Definition 4 (Measuring Non-linearity of Joint Distributions) Let L be the set of joint distributions $R_{X,Y}$ that admit a linear model from X to Y. Set

$$D(P_{X,Y}||L) := \inf_{R_{X,Y} \in L} D(P_{X,Y}||R_{X,Y}),$$

where the infimum is taken over all distributions in L. Here, D denotes the relative entropy distance [20], defined as follows. If P, Q are arbitrary distributions with strictly positive densities p and q it is given by

$$D(P||Q) := \int p(x) \log \frac{p(x)}{q(x)} dx.$$

Then we have:

Theorem 2 (Non-linearity of Backwards Model and Multi-inf.) *Let (X_t) be a causal AR(1)-process and $I(t)$ the multi-information of all the "particles" in the toy model given by (4) to (6). Then,*

$$I(t) - I(t-1) \geq D(P_{X_t, X_{t-1}}||L).$$

Proof Assume X_t and X_{t-1} are neither linear dependent nor statistically independent because otherwise the bound becomes trivial since we had $P_{X_t, X_{t-1}} \in L$. The idea of the proof is the following: we figure out how much the joint distribution of X_t and X_{t-1} has to be modified to admit a linear model from X_t to X_{t-1}. We have already argued that the entire stochastic process would admit a linear model in backward direction if all the outgoing particles were statistically independent. To obtain a linear model only from X_t to X_{t-1} by reversing the physical toy model it is sufficient to replace the systems with statistically independent ones. More precisely, we replace P, the joint distribution of

$$\dots Z_t^{-1}, Z_t^0, Z_t^1, Z_t^2, \dots$$

by the product distribution with the same marginals, which we shall denote by $\tilde{P}$. Then we check how this changes the joint distribution of X_t and X_{t-1}. The inverse dynamics $t \mapsto t-1$ is given by

$$Z_{t-1}^{(-1)} = \tilde{\gamma}_{11} Z_t^{(0)} + \tilde{\gamma}_{12} Z_t^{(1)}, \tag{12}$$

$$Z_{t-1}^{(0)} = \tilde{\gamma}_{21} Z_t^{(0)} + \tilde{\gamma}_{22} Z_t^{(1)}, \tag{13}$$

$$Z_{t-1}^{(j)} = Z_t^{(j-1)} \quad \text{for } j \neq 0, -1, \tag{14}$$

where $\tilde{\gamma}_{kl}$ denote the entries of Γ^{-1}.

Since $X_t = Z_t^{(0)}$ and

$$X_{t-1} = \tilde{\gamma}_{21} Z_t^{(0)} + \tilde{\gamma}_{22} Z_t^{(1)}, \tag{15}$$

which is implied by (12), the pair $(Z_t^{(0)}, Z_t^{(1)})$ and (X_t, X_{t-1}) span the same probability space (note that both coefficients in (15) are non-zero because we have excluded the cases of linear dependency and statistical independence). Hence $\tilde{P}_{Z_t^{(0)}, Z_t^{(1)}}$ induces by variable transformation a distribution $\tilde{P}_{X_t, X_{t-1}}$ satisfying

$$D(P_{X_t, X_{t-1}} || \tilde{P}_{X_t, X_{t-1}}) = D(P_{Z_t^{(0)}, Z_t^{(1)}} || \tilde{P}_{Z_t^{(0)}, Z_t^{(1)}}).$$

The left hand side is an upper bound for the distance of P_{X_{t-1}, X_t} to a linear model from X_t to X_{t-1} because $\tilde{P}_{X_t, X_{t-1}}$ admits such a model, namely

$$X_{t-1} = \tilde{\gamma}_{21} X_t + \tilde{\gamma}_{22} Z_t^{(1)}.$$

Thus, we have

$$D(P_{X_t, X_{t-1}} || L) \leq D(P_{Z_t^{(0)}, Z_t^{(1)}} || \tilde{P}_{Z_t^{(0)}, Z_t^{(1)}}). \tag{16}$$

The right-hand side of (16) is equal to the mutual information $I(Z_t^{(1)} : Z_t^{(0)})$ since this mutual information is known [20] to equal the relative entropy distance from $P_{Z_t^{(0)}, Z_t^{(1)}}$ to the product distribution with the same marginals. It follows, by Lemma 3, that the right-hand side of (16) is equal to $I(t) - I(t-1)$, and so (16) is equivalent to the formula to be proved. □

If X_t is Gaussian, the stochastic process need not generate multi-information: The easiest case is where all $Z_0^{(j)}$ are independent identically distributed Gaussians with zero mean and Γ rotates the space $\mathbb{R}^2$ by some angle α. Since Γ preserves isotropic Gaussians in $\mathbb{R}^2$, the dynamics induced on $S^{-\infty \cdots \infty}$ leaves the entire joint state invariant. This model can induce any stationary Gaussian $AR(1)$-process, because then $|\phi|^2 \leq 1$ in (3) and we can thus write

$$X_{t+1} = \sin\alpha X_t + \epsilon_t$$

with $\epsilon_t := \cos\alpha Z_t^{(-1)}$.

Note that Gaussian processes can also be realized by a system that *does* generate multi-information. For instance,

$$\Gamma := \begin{pmatrix} \cos\alpha & \sin\alpha \\ 0 & (\cos\alpha)^{-1} \end{pmatrix}.$$

induces the same process (X_t) as a rotation by the angle α, but induces dependent outgoing particles because $\Gamma^T\Gamma$ is non-diagonal. This way, one can easily generate a *stationary* Gaussian process (which is automatically time-symmetric) that generates entropy. This shows that the quantitative correspondence between entropy production and time-inversion asymmetry of (X_t) can only consist of *lower* bounds.

3 Interpretation

We first discuss the interpretation of the Gaussian case. To show an even closer link to thermodynamics, we recall that Gaussian distributions often occur in the context of thermal equilibrium states. For instance, the variable position and momentum of a harmonic oscillator are Gaussian distributed in thermal equilibrium. The case where all $Z_0^{(0)}$ are identically independently distributed Gaussians and Γ is a rotation (which preserves the joint state) can therefore be interpreted as thermal equilibrium dynamics. The fact that the joint distribution $P_{X_t,X_{t+1}}$ coincides with $P_{X_t,X_{t-1}}$ is exactly the symmetry imposed by the well-known *detailed-balance* condition [21] that holds for every Gibbs state.

In order to interpret the scenario in the non-Gaussian case as entropy production, we note that the sum of the marginal entropies of the subsystems increase linearly in time. The fact that the joint Shannon entropy remains constant loses more and more its practical relevance since it requires complex joint operations to undo all the dependence. From a coarse-grained point of view, the entropy increases at every step if we interpret the sum of marginal entropies as coarse-grained entropy. Since the joint entropy remains constant, the increase of multi-information then coincides with the coarse-grained entropy production.

In our experiments we found several examples of time series that could better be fitted with a causal ARMA model from the future to the past than vice versa, even though this was only a minority of those for which a decision was made. Of course, we do not want to suggest that the physical systems corresponding to those negative examples violate the second law. To avoid such misconceptions we discuss which assumptions could be violated to generate time series that admit non-Gaussian ARMA models in the *wrong* direction.

To this end, we list the requirements that jointly make the time-inverted scenario of the above dynamics extremely unlikely:

1. The "incoming particles" (which correspond to the outgoing ones in the original scenario) and $S^{(0)}$ must be statistically dependent.[4]
2. The coupling between $S^{(0)}$ and the incoming particles must be chosen such that it exactly removes the statistical dependence of the incoming particles. There is nothing wrong with *dependent* particles approaching $S^{(0)}$, and a coupling that destroys dependences between the particles and $S^{(0)}$ while creating additional dependence with a third party. However, removing statistical dependence in a *closed* system requires transformations that are specifically adapted to the kind of dependence that is present. In other words, the coupling between $S^{(0)}$ and the incoming particles had to be one of the "few" linear maps $\tilde{\Gamma} \in SL(2)$ needed for undoing the operation that created the statistical dependence of the incoming particles.

[4]This indicates that they have already been interacting earlier, cf. Reichenbach's principle of the common cause [1], which is meanwhile one of the cornerstones of causal inference.

We want to be more explicit about the last item and recall that the joint state of $S^{0,\dots,t}$ at time t is given by

$$(\Gamma_{0,1}^{(t+1)} \circ \Gamma_{1,2}^{(t+1)} \circ \cdots \circ \Gamma_{t-1,t}^{(t+1)}) Q^{\otimes(t+1)}.$$

We now run the time inverted dynamics (12)–(13) (starting from t and ending at 0) to this input using some arbitrary $\tilde{\Gamma} \in SL(2)$. The state of $S^{-t,\dots,0}$ then reads

$$(\hat{\Gamma}_{0,1}^{(t+1)} \circ \hat{\Gamma}_{1,2}^{(t+1)} \circ \cdots \circ \hat{\Gamma}_{t-1,t}^{(t+1)}) Q^{\otimes(t+1)},$$

where we have defined

$$\hat{\Gamma} := \tilde{\Gamma} \circ \Gamma.$$

Due to Lemma 2, this can only be a product state if $\hat{\Gamma}$ has only diagonal or only off-diagonal entries (or if Q is Gaussian). This shows that the statistical dependence of the incoming particles can be removed by $\tilde{\Gamma}$ only if $\tilde{\Gamma}$ is adjusted to the specific form of this dependence (whose characteristic feature is the map Γ that has created the dependence from independent states in the remote past).

This kind of mutual adjustment between mechanism and incoming state is unlikely. Similar arguments have been used in causal inference recently [22, 23]. According to the language used there, the incoming state and the coupling share *algorithmic* information, which indicates that the incoming state and the coupling have not been chosen independently.[5]

To generate a process $(X_t)_{t\in\mathbb{Z}}$ that admits a linear model in backward direction thus requires a different class of dynamical models. For instance, the joint dynamics could be non-linear.

4 Conclusions and Discussion

We have discussed time series that admit a causal ARMA model in forward direction but require non-linear transitions in backward directions to remain causal. Previous experiments verified that some empirical time series indeed show this asymmetry. Here we have related this asymmetry to the thermodynamic arrow of time.

To this end, we have presented a toy model of a physical system coupled to an infinite environment where the asymmetry is due to the production of entropy (if the joint system is considered from a coarse-grained perspective).

The essential feature of the irreversible process studied here is that the linearity of the joint dynamics is passed down to the forward but not to the backward conditionals. Of course, not every physical dynamics is linear, but the result suggests a more general statement for irreversible processes: there seems to be a sense in which the simplicity of the joint dynamics of system and environment is passed down to the forward conditionals of the system but not the backward conditionals. To study for which notions of simplicity, other than linearity, this holds has to be left to the future. Results of this kind would provide a better understanding of more subtle time-asymmetries in physics than the obvious implications of second law. This would be particularly relevant for stochastic processes because they usually describe the state of a system that strongly interacts with its environment and there is thus no simple entropy criterion to distinguish between the true and the wrong time direction.

[5] Note that the thermodynamic relevance of *algorithmic* information has also been pointed out in [24].

Understanding asymmetries between past of future also helps in understanding asymmetries between cause and effect, which is relevant for the field of causal inference: [25–27] developed algorithms that inferred whether X causes Y or Y causes X for just two observed random variables X and Y. Their approaches where based on the observation that the conditional $P(\texttt{effect}|\texttt{cause})$ is usually simpler then $P(\texttt{cause}|\texttt{effect})$, where different notions of simplicity were used. It should be emphasized that this kind of reasoning cannot be justified by referring to Occam's Razor only, i.e., the principle to prefer simple models if possible. The point that deserves our attention is not why science should look for simple laws. Instead, we are asking why we should expect that *causal* conditionals $P(\texttt{effect}|\texttt{cause})$ are simple instead of expecting non-causal conditionals $P(\texttt{cause}|\texttt{effect})$ to be simple. Some first explanations were provided by observations in [28, 29]. These papers discuss two interacting physical systems described by random variables X and Y where the causal influence was mainly from X to Y and the backaction from Y to X was negligible. Then they show that $P(Y|X)$ is simple and $P(X|Y)$ complex for the models under consideration. References [28, 29] as well as the present paper, therefore explore the thermodynamic foundation of a novel type of causal inference rules.

Acknowledgements This work has been inspired by discussions with Armen Allahverdyan in a meeting that was part of the VW-project "Quantum thermodynamics: energy and information flow at the nanoscale". Thanks to Jonas Peters for comments on an early draft.

References

1. Reichenbach, H.: The Direction of Time. University of California Press, Berkley (1956). Dover, New York (1999)
2. Penrose, O., Percival, I.: The direction of time. Proc. Phys. Soc. **79**, 605–616 (1962)
3. Balian, R.: From Microphysics to Macrophysics. Springer, Berlin (1992)
4. Gibbs, J.W.: Elementary Principles in Statistical Mechanics. Ox Bow Press, Woodbridge (1902)
5. Jaynes, E.T.: Gibbs vs. Boltzmann entropies. Am. J. Phys. **33**, 391 (1965)
6. Lebowitz, J.: Macroscopic dynamics, time's arrow and Boltzmann entropy. Physica A **194**, 1–27 (1993)
7. Wallace, C.: Statistical and Inductive Inference by Minimum Message Length. Springer, Berlin (2005)
8. Peters, J., Janzing, D., Gretton, A., Schölkopf, B.: Detecting the direction of causal time series. In: Proceedings of the International Conference on Machine Learning, Montreal. ACM International Conference Proceeding Series, vol. 382, pp. 801–808. ACM, New York (2009). http://www.cs.mcgill.ca/~icml2009/papers/503.pdf and http://portal.acm.org/citation.cfm?doid=1553374.1553477
9. Peters, J., Janzing, D., Gretton, A., Schölkopf, B.: Kernel methods for detecting the direction of time series. In: Proceedings of the 32nd Annual Conference of the German Classification Society (GfCKI 2008), pp. 1–10. Springer, Berlin (2009)
10. Maes, C., Redig, F., Van Moffaert, A.: On the definition of entropy production via examples. J. Math. Phys. **41**, 1528–1554 (2000)
11. Gallavotti, G., Cohen, E.: Dynamical ensembles and nonequilibrium statistical mechanics. Phys. Rev. Lett. **74**, 2694–2697 (1995)
12. Horowitz, E., Sahni, S.: Fundamentals of Data Structures. Computer Science Press, New York (1976)
13. Maes, C., Netocný, K.: Time reversal and entropy. J. Stat. Phys. **110**(1–2), 269–309 (2003)
14. Chazottes, J.-R., Redig, F.: Testing the irreversibility of a Gibbsian process via hitting and return times. Nonlinearity **18**(18), 2477–2489 (2005)
15. Darmois, G.: Analyse générale des liaisons stochastiques. Rev. Inst. Int. Stat. **21**, 2–8 (1953)
16. Skitovic, V.: Linear combinations of independent random variables and the normal distribution law. Sel. Transl. Math. Stat. Probab. **2**, 211–228 (1962)

17. Kano, Y., Shimizu, S.: Causal inference using nonnormality. In: Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion, Tokyo, Japan, pp. 261–270 (2003)
18. Shimizu, S., Hyvärinen, A., Kano, Y., Hoyer, P.O.: Discovery of non-Gaussian linear causal models using ICA. In: Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, Edinburgh, UK, pp. 526–533 (2005)
19. Brockwell, P., Davis, R.: Time Series: Theory and Methods. Springer, Berlin (1991)
20. Cover, T., Thomas, J.: Elements of Information Theory. Wileys Series in Telecommunications. Wiley, New York (1991)
21. Tolman, R.: The Principles of Statistical Mechanics. Oxford University Press, Oxford (1938)
22. Lemeire, J., Dirkx, E.: Causal models as minimal descriptions of multivariate systems. http://parallel.vub.ac.be/~jan/ (2006)
23. Janzing, D., Schölkopf, B.: Causal inference using the algorithmic Markov condition. http://arxiv.org/abs/0804.3678 (2008)
24. Zurek, W.: Algorithmic randomness and physical entropy. Phys Rev A **40**(8), 4731–4751 (1989)
25. Hoyer, P., Janzing, D., Mooij, J., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Proceedings of the Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, 2008. MIT Press, Cambridge (2009). http://books.nips.cc/papers/files/nips21/NIPS2008_0266.pdf
26. Mooij, J., Janzing, D., Peters, J., Schölkopf, B.: Regression by dependence minimization and its application to causal inference. In: Proceedings of the International Conference on Machine Learning, Montreal. ACM International Conference Proceeding Series, vol. 382, pp. 745–752. ACM, New York (2009) http://www.cs.mcgill.ca/~icml2009/papers/279.pdf and http://portal.acm.org/citation.cfm?id=1553374.1553470
27. Zhang, K., Hyvärinen, A.: On the identifiability of the post-nonlinear causal model. In: 25th Conference on Uncertainty in Artificial Intelligence, Montreal, Canada (2009)
28. Janzing, D.: On causally asymmetric versions of Occam's Razor and their relation to thermodynamics. http://arxiv.org/abs/0708.3411v2 (2008)
29. Allahverdyan, A., Janzing, D.: Relating the thermodynamic arrow of time to the causal arrow. J. Stat. Mech. P04001 (2008)