Entropy **2013**, *15*, 524-543; doi:10.3390/e15020524

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

On Thermodynamic Interpretation of Transfer Entropy

Mikhail Prokopenko [1,2,]*, Joseph T. Lizier [1] and Don C. Price [3]

[1] CSIRO Information and Communications Technologies Centre, PO Box 76, Epping, NSW 1710, Australia; E-Mail: joseph.lizier@csiro.au

[2] School of Physics A28, University of Sydney, NSW 2006, Australia

[3] CSIRO Materials Science and Engineering, Bradfield Road, West Lindfield, NSW 2070, Australia; E-Mail: don.price@csiro.au

* Author to whom correspondence should be addressed; E-Mail: mikhail.prokopenko@csiro.au; Tel.: +61-29372-4716; Fax: +61-29372-4161.

Received: 16 November 2012; in revised form: 16 January 2013 / Accepted: 28 January 2013 / Published: 1 February 2013

Abstract: We propose a thermodynamic interpretation of transfer entropy near equilibrium, using a specialised Boltzmann's principle. The approach relates conditional probabilities to the probabilities of the corresponding state transitions. This in turn characterises transfer entropy as a difference of two entropy rates: the rate for a resultant transition and another rate for a possibly irreversible transition within the system affected by an additional source. We then show that this difference, the local transfer entropy, is proportional to the external entropy production, possibly due to irreversibility. Near equilibrium, transfer entropy is also interpreted as the difference in equilibrium stabilities with respect to two scenarios: a default case and the case with an additional source. Finally, we demonstrated that such a thermodynamic treatment is not applicable to information flow, a measure of causal effect.

Keywords: transfer entropy; information transfer; entropy production; irreversibility; Kullback–Leibler divergence; thermodynamic equilibrium; Boltzmann's principle; causal effect

1. Introduction

Transfer entropy has been introduced as an information-theoretic measure that quantifies the statistical coherence between systems evolving in time [1]. Moreover, it was designed to detect asymmetry in the interaction of subsystems by distinguishing between "driving" and "responding" elements. In constructing the measure, Schreiber considered several candidates as measures of directional information transfer, including symmetric mutual information, time-delayed mutual information, as well as asymmetric conditional information. All these alternatives were argued to be inadequate for determining the direction of information transfer between two, possibly coupled, processes.

In particular, defining information transfer simply as the dependence of the next state of the receiver on the previous state of the source [2] is incomplete according to Schreiber's criteria requiring the definition to be both *directional* and *dynamic*. Instead, the (predictive) information transfer is defined as the average information contained in the source about the next state of the destination in the context of what was already contained in the destination's past.

Following the seminal work of Schreiber [1] numerous applications of transfer entropy have been successfully developed, by capturing information transfer within complex systems, e.g., the stock market [3], food webs [4], EEG signals [5], biochemicals [6], cellular automata and distributed computation in general [7–10], modular robotics [11], random and small-world Boolean networks [12,13], inter-regional interactions within a brain [14], swarm dynamics [15], cascading failures in power grids [16], *etc*. Also, several studies further capitalised on transition probabilities used in the measure, highlighting fundamental connections of the measure to entropy rate and Kullback–Leibler divergence noted by Kaiser and Schreiber [17], as well as causal flows [18]. At the same time there are several recent studies investigating ties between information theory and thermodynamics [19–23]. This is primarily through Landauer's principle [24], which states that irreversible destruction of one bit of information results in dissipation of at least $kT \ln 2$ J of energy (T is the absolute temperature and k is Boltzmann's constant.) into the environment (*i.e.*, an entropy increase in the environment by this amount). (Maroney [25] argues that while a logically irreversible transformation of information does generate this amount of heat, it can in fact be accomplished by a *thermodynamically reversible* mechanism.)

Nevertheless, transfer entropy *per se* has not been precisely interpreted thermodynamically. Of course, as a measure of directed information transfer, it does not need to have an explicit thermodynamic meaning. Yet, one may still put forward several questions attempting to cast the measure in terms more familiar to a physicist rather than an information theorist or a computer scientist: Is transfer entropy a measure of some entropy transferred between subsystems or coupled processes? Is it instead an entropy of some transfer happening within the system under consideration (and what is then the nature of such transfer)? If it is simply a difference between some entropy rates, as can be seen from the definition itself, one may still inquire about the thermodynamic nature of the underlying processes.

Obviously, once the subject relating entropy definitions from information theory and thermodynamics is touched, one may expect vigorous debates that have been ongoing since Shannon introduced the term entropy itself. While this paper will attempt to produce a thermodynamic interpretation of transfer entropy, it is out of scope to comment here on rich connections between Boltzmann entropy and Shannon entropy, or provide a review of quite involved discussions on the topic. It suffices to point out prominent works of Jaynes [26,27] who convincingly demonstrated that information theory can be applied to the problem of justification of statistical mechanics, producing predictions of equilibrium thermodynamic properties. The statistical definition of entropy is widely considered more general and fundamental than the original thermodynamic definition, sometimes allowing for extensions to the situations where the system is not in thermal equilibrium [23,28]. In this study, however, we treat the problem of finding a thermodynamic interpretation of transfer entropy somewhat separately from the body of work relating Boltzmann and Shannon entropies—and the reason for this is mainly that, even staying within Jaynes' framework, one still needs to provide a possible thermodynamic treatment for transfer entropy *per se*. As will become clear, this task is not trivial, and needs to be approached carefully.

Another contribution of this paper is a clarification that similar thermodynamic treatment is not applicable to information flow—a measure introduced by Ay and Polani [18] in order to capture causal effect. That correlation is not causation is well-understood. Yet while authors increasingly consider the notions of information transfer and information flow and how they fit with our understanding of correlation and causality [1,18,29–34], several questions nag. Is information transfer, captured by transfer entropy, akin to causal effect? If not, what is the distinction between them? When examining the "effect" of one variable on another (e.g., between brain regions), should one seek to measure information transfer or causal effect?

Unfortunately, these concepts have become somewhat tangled in discussions of information transfer. Measures for both predictive transfer [1] and causal effect [18] have been inferred to capture information transfer in general, and measures of predictive transfer have been used to infer causality [33,35–37] with the two sometimes (problematically) directly equated (e.g., [29,32,34,38–40]). The study of Lizier and Prokopenko [41] clarified the relationship between these concepts and described the manner in which they should be considered separately. Here, in addition, we demonstrate that a thermodynamic interpretation of transfer entropy is not applicable to causal effect (information flow), and clarify the reasons behind this.

This paper is organised as follows. We begin with Section 2 that introduces relevant information-theoretic measures both in average and local terms. Section 3 defines the system and the range of applicability of our approach. In providing a thermodynamic interpretation for transfer entropy in Section 4 we relate conditional probabilities to the probabilities of the corresponding state transitions, and use a specialised Boltzmann's principle. This allows us to define components of transfer entropy with the entropy rate of (i) the resultant transition and (ii) the internal entropy production. Sub-section 4.3 presents an interpretation of transfer entropy near equilibrium. The following Section 5 discusses the challenges for supplying a similar interpretation to causal effect (information flow). A brief discussion in Section 6 concludes the paper.

2. Definitions

In the following sections we describe relevant background on transfer entropy and causal effect (information flow), along some technical preliminaries.

2.1. Transfer Entropy

Mutual information $I_{Y;X}$ has been something of a de facto measure for information transfer between Y and X in complex systems science in the past (e.g., [42–44]). A major problem however is that mutual information contains no inherent *directionality*. Attempts to address this include using the previous state of the "source" variable Y and the next state of the "destination" variable X' (known as *time-lagged mutual information* $I_{Y;X'}$). However, Schreiber [1] points out that this ignores the more fundamental problem that mutual information measures the *statically* shared information between the two elements. (The same criticism applies to equivalent non-information-theoretic definitions such as that in [2].)

To address these inadequacies Schreiber introduced *transfer entropy* [1] (TE), the deviation from independence (in bits) of the state transition (from the previous state to the next state) of an information destination X from the previous state of an information source Y:

$$T_{Y \to X}(k,l) = \sum_{x_{n+1}, x_n^{(k)}, y^{(l)}} p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log_2 \frac{p(x_{n+1} \mid x_n^{(k)}, y_n^{(l)})}{p(x_{n+1} \mid x_n^{(k)})} \tag{1}$$

Here n is a time index, $x_n^{(k)}$ and $y_n^{(l)}$ represent past states of X and Y (*i.e.*, the k and l past values of X and Y up to and including time n). Schreiber points out that this formulation is a truly *directional*, *dynamic* measure of information transfer, and is a generalisation of the entropy rate to more than one element to form a mutual information *rate*. That is, transfer entropy may be seen as the difference between two entropy rates:

$$T_{Y \to X}(k,l) = h_X - h_{X,Y} \tag{2}$$

where h_X is the entropy rate:

$$h_X = -\sum p(x_{n+1}, x_n^{(k)}) \log_2 p(x_{n+1} \mid x_n^{(k)}) \tag{3}$$

and $h_{X,Y}$ is a generalised entropy rate conditioning on the source state as well:

$$h_{X,Y} = -\sum p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log_2 p(x_{n+1} \mid x_n^{(k)}, y_n^{(l)}) \tag{4}$$

The entropy rate h_X accounts for the average number of bits needed to encode one additional state of the system if all previous states are known [1], while the entropy rate $h_{X,Y}$ is the entropy rate capturing the average number of bits required to represent the value of the next destination's state if source states are included in addition. Since one can always write

$$h_X = -\sum p(x_{n+1}, x_n^{(k)}) \log_2 p(x_{n+1} \mid x_n^{(k)}) = -\sum p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log_2 p(x_{n+1} \mid x_n^{(k)}) \tag{5}$$

it is easy to see that the entropy rate h_X is equivalent to the rate $h_{X,Y}$ when the next state of destination is independent of the source [1]:

$$p(x_{n+1} \mid x_n^{(k)}) = p(x_{n+1} \mid x_n^{(k)}, y_n^{(l)}) \tag{6}$$

Thus, in this case the transfer entropy reduces to zero.

Similarly, the TE can be viewed as a *conditional* mutual information $I(Y^{(l)}; X' \mid X^{(k)})$ [17], that is as the average information contained in the source about the next state X' of the destination that was not already contained in the destination's past $X^{(k)}$:

$$T_{Y \to X}(k,l) = I_{Y^{(l)};X'|X^{(k)}} = H_{X'|X^{(k)}} - H_{X'|X^{(k)},Y^{(l)}} \tag{7}$$

This could be interpreted (following [44,45]) as the diversity of state transitions in the destination minus assortative noise between those state transitions and the state of the source.

Furthermore, we note that Schreiber's original description can be rephrased as the information provided by the source about the state transition in the destination. That $x_n^{(k)} \rightarrow x_{n+1}$ (or including redundant information $x_n^{(k)} \rightarrow x_{n+1}^{(k)}$) is a *state transition* is underlined in that the $x_n^{(k)}$ are *embedding vectors* [46], which capture the underlying *state* of the process. Indeed, since all of the above mathematics for the transfer entropy is equivalent if we consider the next source *state* $x_{n+1}^{(k)}$ instead of the next source value x_{n+1}, we shall adjust our notation from here onwards to consider the next source state $x_{n+1}^{(k)}$, so that we are always speaking about interactions between source states $\mathbf{y_n}$ and destination state transitions $\mathbf{x_n} \rightarrow \mathbf{x_{n+1}}$ (with embedding lengths l and k implied).

Importantly, the TE remains a measure of observed (conditional) *correlation* rather than direct effect. In fact, the TE is a non-linear extension of a concept known as the "Granger causality" [47], the nomenclature for which may have added to the confusion associating information transfer and causal effect. Importantly, as an information-theoretic measure based on observational probabilities, the TE is applicable to both deterministic and stochastic systems.

2.2. Local Transfer Entropy

Information-theoretic variables are generally defined and used as an *average* uncertainty or information. We are interested in considering *local* information-theoretic values, *i.e.*, the uncertainty or information associated with a *particular observation* of the variables rather than the average over all observations. *Local* information-theoretic measures are sometimes called *point-wise* measures [48,49]. Local measures within a global average are known to provide important insights into the *dynamics* of non-linear systems [50].

Using the technique originally described in [7], we observe that the TE is an average (or *expectation value*) of a *local transfer entropy* at each observation n, *i.e.*,:

$$T_{Y\rightarrow X} = \langle t_{Y\rightarrow X}(n+1)\rangle \tag{8}$$

$$t_{Y\rightarrow X}(n+1) = \log_2 \frac{p(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n})}{p(\mathbf{x_{n+1}} \mid \mathbf{x_n})} \tag{9}$$

with embedding lengths l and k implied as described above. The local transfer entropy quantifies the information contained in the source state $\mathbf{y_n}$ about the next state of the destination $\mathbf{x_{n+1}}$ at time step $n+1$, in the context of what was already contained in the past state of the destination $\mathbf{x_n}$. The measure is *local* in that it is defined at each time n for each destination X in the system and each causal information source Y of the destination.

The local TE may also be expressed as a local conditional mutual information, or a difference between local conditional entropies:

$$t_{Y\rightarrow X}(n+1) = i(\mathbf{y_n}; \mathbf{x_{n+1}} \mid \mathbf{x_n}) = h(\mathbf{x_{n+1}} \mid \mathbf{x_n}) - h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) \tag{10}$$

where local conditional mutual information is given by

$$i(\mathbf{y_n}; \mathbf{x_{n+1}} \mid \mathbf{x_n}) = \log_2 \frac{p(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n})}{p(\mathbf{x_{n+1}} \mid \mathbf{x_n})} \tag{11}$$

and local conditional entropies are defined analogously:

$$h(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = -\log_2 p(\mathbf{x_{n+1}} \mid \mathbf{x_n}) \tag{12}$$

$$h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) = -\log_2 p(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) \tag{13}$$

The average transfer entropy $T_{Y \to X}(k)$ is always positive but is bounded above by the information capacity of a single observation of the destination. For a discrete system with b possible observations this is $\log_2 b$ bits. As a conditional mutual information, it can be either larger *or* smaller than the corresponding mutual information [51]. The *local* TE however is not constrained so long as it averages into this range: it can be greater than $\log_2 b$ for a large local information transfer, and can also in fact be measured to be negative. Local transfer entropy is negative where (in the context of the history of the destination) the probability of observing the actual next state of the destination given the source state $p(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n})$, is lower than that of observing that actual next state independently of the source $p(\mathbf{x_{n+1}} \mid \mathbf{x_n})$. In this case, the source variable is actually *misinformative* or misleading about the state transition of the destination. It is possible for the source to be misleading where other causal information sources influence the destination, or in a stochastic system. Full examples are described by Lizier *et al*. [7].

2.3. Causal Effect as Information Flow

As noted earlier, predictive information transfer refers to the amount of information that a source variable adds to the next state of a destination variable; *i.e.*, "if I know the state of the source, how much does that help to predict the state of the destination?". Causal effect, on the contrary, refers to the extent to which the source variable has a direct influence or drive on the next state of a destination variable, *i.e.*, "if I change the state of the source, to what extent does that alter the state of the destination?". Information from causal effect can be seen to *flow* through the system, like injecting dye into a river [18].

It is well-recognised that measurement of causal effect necessitates some type of *perturbation* or *intervention* of the source so as to detect the effect of the intervention on the destination (e.g., see [52]). Attempting to infer causality without doing so leaves one measuring correlations of observations, regardless of how directional they may be [18]. In this section, we adopt the measure information flow for this purpose, and describe a method introduced by Lizier and Prokopenko [41] for applying it on a local scale.

Following Pearl's probabilistic formulation of causal Bayesian networks [52], Ay and Polani [18] consider how to measure causal information flow via *interventional conditional probability distribution functions*. For instance, an interventional conditional PDF $p(y \mid \hat{s})$ considers the distribution of y resulting from *imposing* the value of $\hat{s}$. *Imposing* means intervening in the system to *set* the value of the imposed variable, and is at the essence of the definition of causal information flow. As an illustration of the difference between interventional and standard conditional PDFs, consider two correlated variables S and Y: their correlation alters $p(y \mid s)$ in general from $p(y)$. If both variables are solely caused by another variable G however, then even where they remain correlated we have $p(y \mid \hat{s}) = p(y)$ because imposing a value $\hat{s}$ has no effect on the value of y.

In a similar fashion to the definition of transfer entropy as the deviation of a destination from *stochastic* independence on the source in the content of the destination's past, Ay and Polani propose the measure *information flow* as the deviation of the destination X from *causal* independence on the source Y *imposing* another set of nodes S. Mathematically, this is written as:

$$I_p(Y \to X \mid \hat{\mathrm{S}}) = \sum_{\mathrm{s}} p(\mathrm{s}) \sum_{y} p(y \mid \hat{\mathrm{s}}) \sum_{x} p(x \mid \hat{y}, \hat{\mathrm{s}}) \log_2 \frac{p(x \mid \hat{y}, \hat{\mathrm{s}})}{\sum_{y'} p(y' \mid \hat{\mathrm{s}}) p(x \mid \hat{y}', \hat{\mathrm{s}})} \tag{14}$$

The value of the measure is dependent on the choice of the set of nodes S. It is possible to obtain a measure of apparent causal information flow $I_p(Y \to X)$ from Y to X without any S (*i.e.*, $\mathrm{S} = \oslash$), yet this can be misleading. In particular, it ignores causal information flow arising from interactions of the source with another source variable. For example, if $x = y \,\mathrm{XOR}\, s$ and $p(y, s) = 0.25$ for each combination of binary y and s, then $I_p(Y \to X) = 0$ despite the clear causal effect of Y, while $I_p(Y \to X \mid \hat{S}) = 1$ bit. Also, we may have $I_p(Y \to X) > 0$ only because Y effects S which in turn effects X; where we are interested in *direct* causal information flow from Y to X only $I_p(Y \to X \mid \hat{\mathrm{S}})$ validly infers no direct causal effect.

Here we are interested in measuring the *direct* causal information flow from Y to X, so we must either include all possible other sources in S or at least include enough sources to "block" (A set of nodes U *blocks* a path of causal links where there is a node v on the path such that either:

- $v \in U$ and the causal links through v on the path are not both into v, or
- the causal links through v on the path are both into v, and v and all its causal descendants are not in U.)

all non-immediate directed paths from Y to X [18]. The minimum to satisfy this is the set of all direct causal sources of X excluding Y, including any past states of X that are direct causal sources. That is, in alignment with transfer entropy S would include $X^{(k)}$.

The major task in computing $I_p(Y \to X \mid \hat{\mathrm{S}})$ is the determination of the underlying interventional conditional PDFs in Equation (14). By definition these may be gleaned by observing the results of intervening in the system, however this is not possible in many cases.

One alternative is to use detailed knowledge of the dynamics, in particular the structure of the causal links and possibly the underlying rules of the causal interactions. This also is often not available in many cases, and indeed is often the very goal for which one turned to such analysis in the first place. Regardless, where such knowledge is available it may allow one to make direct inferences.

Under certain constrained circumstances, one can construct these values from observational probabilities only [18], e.g., with the "back-door adjustment" [52]. A particularly important constraint on using the back-door adjustment here is that *all* $\{\mathrm{s}, y\}$ combinations must be observed.

2.4. Local Information Flow

A *local information flow* can be defined following the argument that was used to define local information transfer:

$$f(y \to x \mid \hat{\mathrm{s}}) = \log_2 \frac{p(x \mid \hat{y}, \hat{\mathrm{s}})}{\sum_{y'} p(y' \mid \hat{\mathrm{s}}) p(x \mid \hat{y}', \hat{\mathrm{s}})} \tag{15}$$

The meaning of the local information flow is slightly different however. Certainly, it is an *attribution* of local causal effect of y on x were $\hat{\mathrm{s}}$ imposed at the given observation (y, x, s). However, one must be aware that $I_p(Y \to X \mid \hat{\mathrm{S}})$ is not the *average* of the local values $f(y \to x \mid \hat{\mathrm{s}})$ in exactly the same manner as the local values derived for information transfer. Unlike standard information-theoretical measures, the information flow is averaged over a product of *interventional* conditional probabilities ($p(\mathrm{s})p(y \mid \hat{\mathrm{s}})p(x \mid \hat{y}, \hat{\mathrm{s}})$, see Equation (14) which in general does not reduce down to the probability of the given observation $p(\mathrm{s}, y, x) = p(\mathrm{s})p(y \mid \mathrm{s})p(x \mid y, \mathrm{s})$. For instance, it is possible that not all of the tuples $\{y, x, \mathrm{s}\}$ will actually be observed, so averaging over observations would ignore the important contribution that any unobserved tuples provide to the determination of information flow. Again, the local information flow is specifically tied not to the *given observation* at time step n but to the *general configuration* (y, x, s), and only *attributed* to the associated observation of this configuration at time n.

3. Preliminaries

3.1. System Definition

Let us consider the non-equilibrium thermodynamics of a physical system close to equilibrium. At any given moment in time, n, the thermodynamic state of the physical system X is given by a vector $\mathbf{x} \in R^d$, comprising d variables, for instance the (local) pressure, temperature, chemical concentrations and so on. A state vector completely describes the physical macrostate as far as predictions of the outcomes of all possible measurements performed on the system are concerned [53]. The state space of the system is the set of all possible states of the system.

The thermodynamic state is generally considered as a fluctuating entity so that transition probabilities like $p\left(\mathbf{x_{n+1}}|\mathbf{x_n}\right)$ are clearly defined and can be related to a sampling procedure. Each macrostate can be realised by a number of different microstates consistent with the given thermodynamic variables. Importantly, in the theory of non-equilibrium thermodynamics close to equilibrium, the microstates belonging to one macrostate $\mathbf{x}$ are equally probable.

3.2. Entropy Definitions

The thermodynamic entropy was originally defined by Clausius as a state function S that satisfies

$$S_B - S_A = \int_A^B \mathrm{d}q_{rev}/T \tag{16}$$

where q_{rev} is the heat transferred to an equilibrium thermodynamic system during a reversible process from state A to state B. Note that this *path integral* is the same for all reversible paths between the past and next states.

It was shown by Jaynes that thermodynamic entropy could be interpreted, from the perspective of statistical mechanics, as a measure of the amount of information about the microstate of a system that an observer lacks if they know only the macrostate of the system [53].

This is encapsulated in the famous Boltzmann's equation $S = k \log W$, where k is Boltzmann's constant and W is the number of microstates corresponding to a given macrostate (an integer greater than or equal to one). While it is not a mathematical probability between zero and one, it is sometimes called "thermodynamic probability", noting that W can be normalized to a probability $p = W/N$, where N is the number of possible microstates for all macrostates.

The Shannon entropy that corresponds to the Boltzmann entropy $S = k \log W$ is the uncertainty in the microstate that has produced the given macrostate. That is, given the number W of microscopic configurations that correspond to the given macrostate, we have $p_i = 1/W$ for each equiprobable microstate i. As such, we can compute the local entropy for each of these W microstates as $-\log_2 1/W = \log_2 W$ bits. Note that the average entropy across all of these equiprobable microstates is $\log_2 W$ bits also. This is equivalent to the Boltzmann entropy up to Boltzmann's constant k and the base of the logarithms (see [54,55] for more details).

3.3. Transition Probabilities

A specialisation of Boltzmann's principle by Einstein [56], for two states with entropies S and S_0 and "relative probability" W_r (the ratio of numbers W and W_0 that account for the numbers of microstates in the macrostates with S and S_0 respectively), is given by:

$$S - S_0 = k \log W_r \tag{17}$$

The expression in these relative terms is important, as pointed out by Norton [57], because the probability *W_r is the probability of the transition between the two states under the system's normal time evolution.*

In the example considered by Einstein [56,57], S_0 is the entropy of an (equilibrium) state, e.g., "a volume V_0 of space containing n non-interacting, moving points, whose dynamics are such as to favor no portion of the space over any other", while S is the entropy of the (non-equilibrium) state with the "same system of points, but now confined to a sub-volume V of V_0". Specifically, Einstein defined the transition probability $W_r = (V/V_0)^n$, yielding

$$S - S_0 = kn \log(V/V_0) \tag{18}$$

Since dynamics favour no portion of the space over any other, all the microstates are equiprobable.

3.4. Entropy Production

In general, the variation of entropy of a system ΔS is equal to the sum of the internal entropy production σ inside the system and the entropy change due to the interactions with the surroundings ΔS_{ext}:

$$\Delta S = \sigma + \Delta S_{ext} \tag{19}$$

In the case of a closed system, ΔS_{ext} is given by the expression

$$\Delta S_{ext} = \int \mathrm{d}q/T \tag{20}$$

where q represents the heat flow received by the system from the exterior and T is the temperature of the system. This expression is often written as

$$\sigma = \Delta S - \Delta S_{ext} = (S - S_0) - \Delta S_{ext} \tag{21}$$

so that when the transition from the initial state S_0 to the final state S is irreversible, the entropy production $\sigma > 0$, while for reversible processes $\sigma = 0$, that is

$$S - S_0 = \int \mathrm{d}q_{rev}/T \tag{22}$$

We shall consider another state vector, $\mathbf{y}$, describing a state of a part Y of the exterior possibly coupled to the system represented by X. In other words, X and Y may or may not be dependent. In general, we shall say that σ_y is the internal entropy production *in the context of* some source Y, while ΔS_{ext} is the entropy production *attributed to* Y.

Alternatively, one may consider two scenarios for such a general physical system. In the first scenario, the entropy changes only due to reversible transitions, amounting to $S - S_0$. In the second scenario, the entropy changes partly irreversibly due to the interactions with the external environment affected by $\mathbf{y}$, but still achieves the same total change $S - S_0$.

3.5. Range of Applicability

In an attempt to provide a thermodynamic interpretation of transfer entropy we make two important assumptions, defining the range of applicability for such an interpretation. The first one relates the transition probability W_{r_1} of the system's reversible state change to the conditional probability $p(\mathbf{x_{n+1}} \mid \mathbf{x_n})$, obtained by sampling the process X:

$$p(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = \frac{1}{Z_1} W_{r_1} \tag{23}$$

where Z_1 is a normalisation factor that depends on $\mathbf{x_n}$. According to the expression for transition probability (17), under this assumption the conditional probability of the system's transition from state $\mathbf{x_n}$ to state $\mathbf{x_{n+1}}$ corresponds to some number W_{r_1}, such that $S(\mathbf{x_{n+1}}) - S(\mathbf{x_n}) = k \log W_{r_1}$, and hence

$$p(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = \frac{1}{Z_1} e^{(S(\mathbf{x_{n+1}}) - S(\mathbf{x_n}))/k} \tag{24}$$

The second assumption relates the transition probability W_{r_2} of the system's possibly irreversible internal state change, due to the interactions with the external surroundings represented in the state vector $\mathbf{y}$, to the conditional probability $p(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n})$, obtained by sampling the systems X and Y:

$$p(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) = \frac{1}{Z_2} W_{r_2} \tag{25}$$

Under this assumption the conditional probability of the system's (irreversible) transition from state $\mathbf{x_n}$ to state $\mathbf{x_{n+1}}$ in the context of $\mathbf{y_n}$, corresponds to some number W_{r_2}, such that $\sigma_y = k \log W_{r_2}$, where σ_y is the system's internal entropy production in the context of $\mathbf{y}$, and thus

$$p(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) = \frac{1}{Z_2} e^{\sigma_y/k} \tag{26}$$

where Z_2 is a normalisation factor that depends on $\mathbf{x_n}$.

3.6. An Example: Random Fluctuation Near Equilibrium

Let us consider the above-defined stochastic process X for a small random fluctuation around equilibrium:

$$\mathbf{x_{n+1}} = \Lambda \mathbf{x_n} + \xi \tag{27}$$

where ξ is a multi-variate Gaussian noise process, with covariance matrix Σ_ξ, uncorrelated in time. Starting at time n with state $\mathbf{x_n}$ having entropy $S\left(\mathbf{x_n}\right)$, the state develops into $\mathbf{x_{n+1}}$, with entropy $S\left(\mathbf{x_{n+1}}\right)$.

From the probability distribution function of the above multi-variate Gaussian process, we obtain

$$p\left(\mathbf{x_{n+1}}|\mathbf{x_n}\right) = \frac{1}{Z} e^{-\frac{1}{2}\left(\mathbf{x_{n+1}}-\Lambda\mathbf{x_n}\right)^T \Sigma_\xi^{-1}\left(\mathbf{x_{n+1}}-\Lambda\mathbf{x_n}\right)} \tag{28}$$

We now demonstrate that this expression concurs with the corresponding expression obtained under assumption (24). To do so we expand the entropies around $\mathbf{x} = 0$ with entropy $S(0)$:

$$S\left(\mathbf{x_n}\right) = S(0) - k\frac{1}{2}\mathbf{x_n}^T \Sigma_x^{-1}\mathbf{x_n} \tag{29}$$

where Σ_x is the covariance matrix of the process X.

Then, according to the assumption (24)

$$p(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = \frac{1}{Z_1} e^{(S(\mathbf{x_{n+1}})-S(\mathbf{x_n}))/k} = \frac{1}{Z_1} e^{-\frac{1}{2}\left(\mathbf{x_{n+1}}^T \Sigma_x^{-1}\mathbf{x_{n+1}} - \mathbf{x_n}^T \Sigma_x^{-1}\mathbf{x_n}\right)} = \frac{1}{\tilde{Z}_1} e^{-\frac{1}{2}\mathbf{x_{n+1}}^T \Sigma_x^{-1}\mathbf{x_{n+1}}} \tag{30}$$

where the term $e^{\frac{1}{2}\mathbf{x_n}^T \Sigma^{-1}\mathbf{x_n}}$ is absorbed into the normalisation factor being only dependent on $\mathbf{x_n}$. In general [58,59], we have

$$\Sigma_x = \sum_{j=0}^{\infty} \Lambda^j \, \Sigma_\xi \, \Lambda^{j^T} \tag{31}$$

Given the quasistationarity of the relaxation process, assumed near an equilibrium, $\Lambda \to 0$, and hence $\Sigma_x \to \Sigma_\xi$. Then the Equation (30) reduces to

$$p(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = \frac{1}{\tilde{Z}_1} e^{-\frac{1}{2}\left(\mathbf{x_{n+1}}^T \Sigma_\xi^{-1}\mathbf{x_{n+1}}\right)} \tag{32}$$

The last expression concurs with Equation (28) when $\Lambda \to 0$.

4. Transfer Entropy: Thermodynamic Interpretation

4.1. Transitions Near Equilibrium

Supported by this background, we proceed to interpret transfer entropy via transitions between states. In doing so, we shall operate with local information theoretic measures (such as the local transfer entropy (9)), as we are dealing with (transitions between) *specific* states $\mathbf{y_n}$, $\mathbf{x_n}$, $\mathbf{x_{n+1}}$, *etc.* and not with all possible state-spaces X, Y, *etc.* containing all realizations of specific states.

Transfer entropy is a difference not between entropies, but rather between entropy rates or conditional entropies, specified on average by Equations (2) or (7), or for local values by Equation (10):

$$t_{Y\to X}(n+1) = h(\mathbf{x_{n+1}} \mid \mathbf{x_n}) - h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) \tag{33}$$

As mentioned above, the first assumption (23), taken to define the range of applicability for our interpretation, entails (24). It then follows that the first component of Equation (33), $h(\mathbf{x_{n+1}} \mid \mathbf{x_n})$, accounts for $S(\mathbf{x_{n+1}}) - S(\mathbf{x_n})$:

$$h(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = -\log_2 p(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = -\log_2 \frac{1}{Z_1} e^{(S(\mathbf{x_{n+1}})-S(\mathbf{x_n}))/k} \tag{34}$$

$$= \log_2 Z_1 - \frac{1}{k\log 2}\left(S(\mathbf{x_{n+1}}) - S(\mathbf{x_n})\right) \tag{35}$$

That is, the local conditional entropy $h(\mathbf{x_{n+1}} \mid \mathbf{x_n})$ corresponds to resultant entropy change of the transition from the past state $\mathbf{x_n}$ to the next state $\mathbf{x_{n+1}}$.

Now we need to interpret the second component of Equation (33): the local conditional entropy $h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n})$ in presence of some other factor or extra source, $\mathbf{y_n}$. Importantly, we must keep both the past state $\mathbf{x_n}$ and the next state $\mathbf{x_{n+1}}$ the same—only then we can characterise the internal entropy change, offset by some contribution of the source $\mathbf{y_n}$.

Our second constraint on the system (25) entails (26), and so

$$h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) = -\log_2 p(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) = -\log_2 \frac{1}{Z_2} e^{\sigma_y/k} = \log_2 Z_2 - \frac{1}{k \log 2}(\sigma_y) \tag{36}$$

4.2. Transfer Entropy as Entropy Production

At this stage we can bring two right-hand side components of transfer entropy (33), represented by Equations (35) and (36), together:

$$t_{Y \to X}(n+1) = \log_2 \frac{Z_1}{Z_2} + \frac{1}{k \log 2}\left(-\left(S(\mathbf{x_{n+1}}) - S(\mathbf{x_n})\right) + \sigma_y\right) \tag{37}$$

When one considers a small fluctuation near an equilibrium, $Z_1 \approx Z_2$, as the number of microstates does not change much in the relevant macrostates. This removes the additive constant. Then, using the expression for entropy production (21), we obtain

$$t_{Y \to X}(n+1) = -\frac{\Delta S_{ext}}{k \log 2} \tag{38}$$

If $Z_1 \neq Z_2$, the relationship includes some additive constant $\log_2 \frac{Z_1}{Z_2}$.

That is, the transfer entropy is proportional to the external entropy production, brought about by the source of irreversibility Y. It captures the difference between the entropy rates that correspond to two scenarios: the reversible process and the irreversible process affected by another source Y. It is neither a transfer of entropy, nor an entropy of some transfer—it is formally a difference between two entropy rates. The opposite sign reflects the different direction of entropy production attributed to the source Y: when $\Delta S_{ext} > 0$, *i.e.*, the entropy increased during the transition in X more than the entropy produced internally, then the local transfer entropy is negative, and the source misinforms about the macroscopic state transition. When, on the other hand, $\Delta S_{ext} < 0$, *i.e.*, some of the internal entropy produced during the transition in X dissipated to the exterior, then the local transfer entropy is positive, and better predictions can be made about the macroscopic state transitions in X if source Y is measured.

As mentioned earlier, while transfer entropy is non-negative on average, some local transfer entropies can be negative when (in the context of the history of the destination) the source variable is misinformative or misleading about the state transition. This, obviously, concurs with the fact that, while a statistical ensemble average of time averages of the entropy change is always non-negative, at certain times entropy change can be negative. This follows from the fluctuation theorem [60], the Second law inequality [61], and can be illustrated with other examples of backward transformations and local violations of the second law [62,63].

Another observation follows from our assumptions (24) and (26) and the representation (37) when $Z_1 \approx Z_2$. If the local conditional entropy $h(\mathbf{x_{n+1}} \mid \mathbf{x_n})$, corresponding to the resultant entropy change of the transition, is different from the local conditional entropy $h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n})$ capturing the internal entropy production in context of the external source Y, then X and Y are dependent. Conversely, whenever these two conditional entropies are equal to each other, X and Y are independent.

4.3. Transfer Entropy as a Measure of Equilibrium's Stability

There is another possible interpretation that considers a fluctuation near the equilibrium. Using Kullback–Leibler divergence between discrete probability distributions p and q:

$$D_{\mathrm{KL}}(p \| q) = \sum_i p(i) \log \frac{p(i)}{q(i)} \tag{39}$$

and its local counterpart:

$$d_{\mathrm{KL}}(p \| q) = \log \frac{p(i)}{q(i)} \tag{40}$$

we may also express the local conditional entropy as follows:

$$h(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = h(\mathbf{x_{n+1}}, \mathbf{x_n}) - h(\mathbf{x_n}) = d_{\mathrm{KL}}\left(p(\mathbf{x_{n+1}}, \mathbf{x_n}) \| p(\mathbf{x_n})\right) \tag{41}$$

It is known in macroscopic thermodynamics that stability of an equilibrium can be measured with Kullback–Leibler divergence between the initial (past) state, e.g., $\mathbf{x_n}$, and the state brought about by some fluctuation (a new observation), e.g., $\mathbf{x_{n+1}}$ [64]. That is, we can also interpret the local conditional entropy $h(\mathbf{x_{n+1}} \mid \mathbf{x_n})$ as the entropy change (or entropy rate) of the fluctuation near the equilibrium.

Analogously, the entropy change in another scenario, where an additional source $\mathbf{y}$ contributes to the fluctuation around the equilibrium, corresponds now to Kullback–Leibler divergence

$$h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) = h(\mathbf{x_{n+1}}, \mathbf{x_n}, \mathbf{y_n}) - h(\mathbf{x_n}, \mathbf{y_n}) = d_{\mathrm{KL}}\left(p(\mathbf{x_{n+1}}, \mathbf{x_n}, \mathbf{y_n}) \| p(\mathbf{x_n}, \mathbf{y_n})\right) \tag{42}$$

and can be seen as a measure of stability with respect to the fluctuation that is now affected by the extra source $\mathbf{y}$.

Contrasting both these fluctuations around the same equilibrium, we obtain in terms of Kullback–Leibler divergences:

$$t_{Y \to X}(n+1) = d_{\mathrm{KL}}\left(p(\mathbf{x_{n+1}}, \mathbf{x_n}) \| p(\mathbf{x_n})\right) - d_{\mathrm{KL}}\left(p(\mathbf{x_{n+1}}, \mathbf{x_n}, \mathbf{y_n}) \| p(\mathbf{x_n}, \mathbf{y_n})\right) \tag{43}$$

In these terms, transfer entropy contrasts stability of the equilibrium between two scenarios: the first one corresponds to the original system, and the second one disturbs the system by the source Y. If, for instance, the source Y is such that the system X is independent of it, then there is no difference in the extents of disturbances to the equilibrium, and the transfer entropy is zero.

4.4. Heat Transfer

It is possible to provide a similar thermodynamic interpretation relating directly to the Clausius definition of entropy. However, in this case we need to make assumptions stronger than Equations (23) and (25). Specifically, we assume Equations (24) and (26) which do not necessarily entail Equations (23) and (25) respectively. For example, setting the conditional probability $p(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = \frac{1}{Z_1} e^{(S-S_0)/k}$ does not mean that $W_1 = e^{(S-S_0)/k}$ is the transition probability.

Under the new stronger assumptions, the conditional entropies can be related to the heat transferred in the transition, per temperature. Specifically, assumption (24) entails

$$h(\mathbf{x_{n+1}} \mid \mathbf{x_n}) = \log_2 Z_1 - \frac{1}{k \log 2}\left(S(\mathbf{x_{n+1}}) - S(\mathbf{x_n})\right) = \log_2 Z_1 - \frac{1}{k \log 2} \int_{\mathbf{x_n}}^{\mathbf{x_{n+1}}} \mathrm{d}q_{rev}/T \tag{44}$$

where the last step used the definition of Clausius entropy (16). As per (16), this quantity is the same for all reversible paths between the past and next states. An example illustrating the transition $(\mathbf{x_n} \rightarrow \mathbf{x_{n+1}})$ can be given by a simple thermal system $\mathbf{x_n}$ that is connected to a heat bath—that is, to a system in contact with a source of energy at temperature T. When the system X reaches a (new) equilibrium, e.g., the state $\mathbf{x_{n+1}}$, due to its connection to the heat bath, the local conditional entropy $h(\mathbf{x_{n+1}} \mid \mathbf{x_n})$ of the transition undergone by system X represents the heat transferred in the transition, per temperature.

Similarly, assumption (26) leads to

$$h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n}) = \log_2 Z_2 - \frac{1}{k \log 2}\left(\sigma_y\right) = \log_2 Z_2 - \frac{1}{k \log 2} \int_{\mathbf{x_n} \xrightarrow{\mathbf{y_n}} \mathbf{x_{n+1}}} \mathrm{d}q/T \tag{45}$$

where $\mathbf{x_n} \xrightarrow{\mathbf{y_n}} \mathbf{x_{n+1}}$ is the new path between $\mathbf{x_n}$ and $\mathbf{x_{n+1}}$ brought about by $\mathbf{y_n}$, and the entropy produced along this path is σ_y. That is, the first and the last points of the path over which we integrate heat transfers per temperature are unchanged but the path is affected by the source $\mathbf{y}$. This can be illustrated by a modified thermal system, still at temperature T but with heat flowing through some thermal resistance Y, while the system X repeats its transition from $\mathbf{x_n}$ to $\mathbf{x_{n+1}}$.

Transfer entropy captures the difference between expressions (44) and (45), *i.e.*, between the relevant amounts of heat transferred to the system X, per temperature.

$$t_{Y \rightarrow X}(n+1) = \log_2 \frac{Z_1}{Z_2} + \frac{1}{k \log 2}\left(\int_{\mathbf{x_n} \xrightarrow{\mathbf{y_n}} \mathbf{x_{n+1}}} \mathrm{d}q/T - \int_{\mathbf{x_n}}^{\mathbf{x_{n+1}}} \mathrm{d}q_{rev}/T \right) \tag{46}$$

Assuming that $Z_1 \approx Z_2$ is realistic, e.g., for quasistatic processes, then the additive constant disappears as well.

It is clear that if the new path is still reversible (e.g., when the thermal resistance is zero) then the source $\mathbf{y}$ has not affected the resultant entropy change and we must have

$$\int_{\mathbf{x_n}}^{\mathbf{x_{n+1}}} \mathrm{d}q_{rev}/T = \int_{\mathbf{x_n} \xrightarrow{\mathbf{y_n}} \mathbf{x_{n+1}}} \mathrm{d}q/T \tag{47}$$

and $t_{Y \rightarrow X}(n+1) = 0$. This obviously occurs if and only if the source Y satisfies the independence condition (6), making the transfer entropy (46) equal to zero. In other words, we may again observe that if the local conditional entropy $h(\mathbf{x_{n+1}} \mid \mathbf{x_n})$ corresponds to the resultant entropy change of the transition, then X and Y are dependent only when the external source Y, captured in the local conditional entropy $h(\mathbf{x_{n+1}} \mid \mathbf{x_n}, \mathbf{y_n})$, brings about an irreversible internal change. If, however, the source Y changed the path in such a way that the process became irreversible, then $t_{Y \rightarrow X}(n+1) \neq 0$.

Finally, according to Equations (19) and (20), the difference between the relevant heats transferred is $\int \mathrm{d}q/T$, where q represents the heat flow received by the system from the exterior via the source Y, and hence

$$t_{Y \rightarrow X}(n+1) = \log_2 \frac{Z_1}{Z_2} - \frac{1}{k \log 2} \int \mathrm{d}q/T \tag{48}$$

In other words, local transfer entropy is proportional to the heat received or dissipated by the system from/to the exterior.

5. Causal Effect: Thermodynamic Interpretation?

In this section we shall demonstrate that a similar treatment is not possible in general for causal effect. Again, we begin by considering local causal effect (15) of the source $\mathbf{y_n}$ on destination $\mathbf{x_{n+1}}$, while selecting s as the destination's past state $\mathbf{x_n}$:

$$f(\mathbf{y_n} \to \mathbf{x_{n+1}} \mid \hat{\mathbf{x}}_\mathbf{n}) = \log_2 \frac{p(\mathbf{x_{n+1}} \mid \hat{\mathbf{y}}_\mathbf{n}, \hat{\mathbf{x}}_\mathbf{n})}{\sum_{\mathbf{y'_n}} p(\mathbf{y'_n} \mid \hat{\mathbf{x}}_\mathbf{n}) p(\mathbf{x_{n+1}} \mid \hat{\mathbf{y}}'_\mathbf{n}, \hat{\mathbf{x}}_\mathbf{n})} \tag{49}$$

Let us first consider conditions under which this representation reduces to the local transfer entropy. As pointed out by Lizier and Prokopenko [41], there are several conditions for such a reduction.

Firstly, $\mathbf{y_n}$ and $\mathbf{x_n}$ must be the only causal contributors to $\mathbf{x_{n+1}}$. In a thermodynamic setting, this means that there are no other sources affecting the transition from $\mathbf{x_n}$ to $\mathbf{x_{n+1}}$, apart from $\mathbf{y_n}$.

Whenever this condition is met, and in addition, the combination $(\mathbf{y_n}, \mathbf{x_n})$ is observed, it follows that

$$p(\mathbf{x_{n+1}} \mid \hat{\mathbf{y}}_\mathbf{n}, \hat{\mathbf{x}}_\mathbf{n}) = p(\mathbf{x_{n+1}} \mid \mathbf{y_n}, \mathbf{x_n}) \tag{50}$$

simplifying the numerator of Equation (49).

Furthermore, there is another condition:

$$p(\mathbf{y_n} \mid \hat{\mathbf{x}}_\mathbf{n}) \equiv p(\mathbf{y_n} \mid \mathbf{x_n}) \tag{51}$$

For example, it is met when the source $\mathbf{y_n}$ is both causally and conditionally independent of the destination's past $\mathbf{x_n}$. Specifically, causal independence means $p(\mathbf{y_n}) \equiv p(\mathbf{y_n} \mid \hat{\mathbf{x}}_\mathbf{n})$, while conditional independence is simply $p(\mathbf{y_n}) \equiv p(\mathbf{y_n} \mid \mathbf{x_n})$. Intuitively, the situation of causal and conditional independence means that inner workings of the system X under consideration do not interfere with the source Y. Alternatively, if X is the only causal influence on Y, the condition (51) also holds, as Y is perfectly "explained" by X, whether X is observed or imposed on. In general, though, the condition (51) means that the probability of $\mathbf{y_n}$ if we impose a value $\hat{\mathbf{x}}_\mathbf{n}$ is the same as if we had simply observed the value $\mathbf{x_n} = \hat{\mathbf{x}}_\mathbf{n}$ without imposing in the system X.

Under the conditions (50) and (51), the denominator of Equation (49) reduces to $p(\mathbf{x_{n+1}} \mid \mathbf{x_n})$, yielding the equivalence between local causal effect and local transfer entropy

$$f(\mathbf{y_n} \to \mathbf{x_{n+1}} \mid \hat{\mathbf{x}}_\mathbf{n}) = t_{Y \to X}(n+1) \tag{52}$$

In this case, the thermodynamic interpretation of transfer entropy would be applicable to causal effect as well.

Whenever one of these conditions is not met, however, the reduction fails. Consider, for instance, the case when the condition (51) is satisfied, but the condition (50) is violated. For example, we may assume that there is some hidden source affecting the transition to $\mathbf{x_{n+1}}$. In this case, the denominator of Equation (49) does not simplify much, and the component that may have corresponded to the entropy rate of the transition between $\mathbf{x_n}$ and $\mathbf{x_{n+1}}$ becomes

$$\log_2 \sum_{\mathbf{y'_n}} p(\mathbf{y'_n} \mid \mathbf{x_n}) p(\mathbf{x_{n+1}} \mid \hat{\mathbf{y}}'_\mathbf{n}, \hat{\mathbf{x}}_\mathbf{n}) \tag{53}$$

The interpretation of this irreducible component is important: the presence of the imposed term $\hat{\mathbf{y}'_{\mathbf{n}}}$ means that one should estimate individual contributions of all possible states $\mathbf{y}$ of the source Y, while varying (*i.e.*, imposing on) the state $\mathbf{x_n}$. This procedure becomes necessary because, in order to estimate the causal effect of source $\mathbf{y}$, *in presence of some other hidden source*, one needs to check all possible impositions on the source state $\mathbf{y}$. The terms of the sum under the logarithm in Equation (53) inevitably vary in their specific contribution, and so the sum cannot be analytically expressed as a single product under the logarithm. This means that we cannot construct a direct thermodynamic interpretation of causal effect in the same way that we did for the transfer entropy.

6. Discussion and Conclusions

In this paper we proposed a thermodynamic interpretation of transfer entropy: an information-theoretic measure introduced by Schreiber [1] as the average information contained in the source about the next state of the destination in the context of what was already contained in the destination's past. In doing so we used a specialised Boltzmann's principle. This in turn produced a representation of transfer entropy $t_{Y \to X}(n+1)$ as a difference of two entropy rates: one rate for a resultant transition within the system of interest X and another rate for a possibly irreversible transition within the system affected by an addition source Y. This difference was further shown to be proportional to the external entropy production, Δ_{ext}, attributed to the source of irreversibility Y.

At this stage we would like to point out a difference between our main result, $t_{Y \to X}(n+1) \propto -\Delta_{ext}$, and a representation for entropy production discussed by Parrondo *et al.* [22]. The latter work characterised the entropy production in the total device, in terms of relative entropy, the Kullback–Leibler divergence between the probability density ρ in phase space of some forward process and the probability density $\tilde{\rho}$ of the corresponding and suitably defined time-reversed process. The consideration of Parrondo *et al.* [22] does not involve any additional sources Y, and so transfer entropy is outside of the scope of their study. Their main result characterised entropy production as $k \; d_{\mathrm{KL}} \left(\rho \| \tilde{\rho} \right)$, which is equal to the total entropy change in the total device. In contrast, in our study we consider the system of interest X specifically, and characterise various entropy rates of X, but in doing so compare how these entropy rates are affected by some source of irreversibility Y. In short, transfer entropy is shown to concur with the entropy produced/dissipated by the system attributed to the external source Y.

We also briefly considered a case of fluctuations in the system X near an equilibrium, relating transfer entropy to the difference in stabilities of the equilibrium, with respect to two scenarios: a default case and the case with an additional source Y. This comparison was carried out with Kullback–Leibler divergences of the corresponding transition probabilities.

Finally, we demonstrated that such a thermodynamic treatment is not applicable to information flow, a measure introduced by Ay and Polani [18] in order to capture a causal effect. We argue that the main reason is the interventional approach adopted in the definition of causal effect. We identified several conditions ensuring certain dependencies between the involved variables, and showed that the causal effect may also be interpreted thermodynamically—but in this case it reduces to transfer entropy anyway. The highlighted difference once more shows a fundamental difference between transfer entropy and causal effect: the former has a thermodynamic interpretation relating to the source of irreversibility

Y, while the latter is a construct that in general assumes an observer intervening in the system in a particular way.

We hope that the proposed interpretation will further advance studies relating information theory and thermodynamics, both in equilibrium and non-equilibrium settings, reversible and irreversible scenarios, average and local scopes, *etc*.

Acknowledgements

The Authors are thankful to Ralf Der (Max Planck Institute for Mathematics in the Sciences, Leipzig) who suggested and co-developed the example in subsection 3.6, and anonymous reviewers whose suggestions significantly improved the paper.

References

1. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.
2. Jakubowski, M.H.; Steiglitz, K.; Squier, R. Information transfer between solitary waves in the saturable Schrödinger equation. *Phys. Rev. E* **1997**, *56*, 7267.
3. Baek, S.K.; Jung, W.S.; Kwon, O.; Moon, H.T. Transfer Entropy Analysis of the Stock Market, **2005**, arXiv:physics/0509014v2.
4. Moniz, L.J.; Cooch, E.G.; Ellner, S.P.; Nichols, J.D.; Nichols, J.M. Application of information theory methods to food web reconstruction. *Ecol. Model.* **2007**, *208*, 145–158.
5. Chávez, M.; Martinerie, J.; Le Van Quyen, M. Statistical assessment of nonlinear causality: Application to epileptic EEG signals. *J. Neurosci. Methods* **2003**, *124*, 113–128.
6. Pahle, J.; Green, A.K.; Dixon, C.J.; Kummer, U. Information transfer in signaling pathways: A study using coupled simulated and experimental data. *BMC Bioinforma.* **2008**, *9*, 139.
7. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Local information transfer as a spatiotemporal filter for complex systems. *Phys. Rev. E* **2008**, *77*, 026110.
8. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Information modification and particle collisions in distributed computation. *Chaos* **2010**, *20*, 037109.
9. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Coherent information structure in complex computation. *Theory Biosci.* **2012**, *131*, 193–203.
10. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Local measures of information storage in complex distributed computation. *Inf. Sci.* **2012**, *208*, 39–54.
11. Lizier, J.T.; Prokopenko, M.; Tanev, I.; Zomaya, A.Y. Emergence of Glider-like Structures in a Modular Robotic System. In Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems (ALife XI), Winchester, UK, 5–8 August 2008; Bullock, S., Noble, J., Watson, R., Bedau, M.A., Eds.; MIT Press: Cambridge, MA, USA, 2008; pp. 366–373.
12. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. The Information Dynamics of Phase Transitions in Random Boolean Networks. In Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems (ALife XI), Winchester, UK, 5–8 August 2008;

Bullock, S., Noble, J., Watson, R., Bedau, M.A., Eds.; MIT Press: Cambridge, MA, USA, 2008; pp. 374–381.

13. Lizier, J.T.; Pritam, S.; Prokopenko, M. Information dynamics in small-world Boolean networks. *Artif. Life* **2011**, *17*, 293–314.
14. Lizier, J.T.; Heinzle, J.; Horstmann, A.; Haynes, J.D.; Prokopenko, M. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *J. Comput. Neurosci.* **2011**, *30*, 85–107.
15. Wang, X.R.; Miller, J.M.; Lizier, J.T.; Prokopenko, M.; Rossi, L.F. Quantifying and tracing information cascades in swarms. *PLoS One* **2012**, *7*, e40084.
16. Lizier, J.T.; Prokopenko, M.; Cornforth, D.J. The Information Dynamics of Cascading Failures in Energy Networks. In Proceedings of the European Conference on Complex Systems (ECCS), Warwick, UK, 21–25 October 2009; p. 54. ISBN: 978-0-9554123-1-8.
17. Kaiser, A.; Schreiber, T. Information transfer in continuous processes. *Physica D* **2002**, *166*, 43–62.
18. Ay, N.; Polani, D. Information flows in causal networks. *Adv. Complex Syst.* **2008**, *11*, 17–41.
19. Bennett, C.H. Notes on Landauer's principle, reversible computation, and Maxwell's Demon. *Stud. History Philos. Sci. Part B* **2003**, *34*, 501–510.
20. Piechocinska, B. Information erasure. *Phys. Rev. A* **2000**, *61*, 062314.
21. Lloyd, S. *Programming the Universe*; Vintage Books: New York, NY, USA, 2006.
22. Parrondo, J.M.R.; den Broeck, C.V.; Kawai, R. Entropy production and the arrow of time. *New J. Phys.* **2009**, *11*, 073008.
23. Prokopenko, M.; Lizier, J.T.; Obst, O.; Wang, X.R. Relating Fisher information to order parameters. *Phys. Rev. E* **2011**, *84*, 041116.
24. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **1961**, *5*, 183–191.
25. Maroney, O.J.E. Generalizing Landauer's principle. *Phys. Rev. E* **2009**, *79*, 031105.
26. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
27. Jaynes, E.T. Information theory and statistical mechanics. II. *Phys. Rev.* **1957**, *108*, 171–190.
28. Crooks, G. Measuring thermodynamic length. *Phys. Rev. Lett.* **2007**, *99*, 100602+.
29. Liang, X.S. Information flow within stochastic dynamical systems. *Phys. Rev. E* **2008**, *78*, 031113.
30. Lüdtke, N.; Panzeri, S.; Brown, M.; Broomhead, D.S.; Knowles, J.; Montemurro, M.A.; Kell, D.B. Information-theoretic sensitivity analysis: A general method for credit assignment in complex networks. *J. R. Soc. Interface* **2008**, *5*, 223–235.
31. Auletta, G.; Ellis, G.F.R.; Jaeger, L. Top-down causation by information control: From a philosophical problem to a scientific research programme. *J. R. Soc. Interface* **2008**, *5*, 1159–1172.
32. Hlaváčková-Schindler, K.; Paluš, M.; Vejmelka, M.; Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* **2007**, *441*, 1–46.
33. Lungarella, M.; Ishiguro, K.; Kuniyoshi, Y.; Otsu, N. Methods for quantifying the causal structure of bivariate time series. *Int. J. Bifurc. Chaos* **2007**, *17*, 903–921.
34. Ishiguro, K.; Otsu, N.; Lungarella, M.; Kuniyoshi, Y. Detecting direction of causal interactions between dynamically coupled signals. *Phys. Rev. E* **2008**, *77*, 026216.

35. Sumioka, H.; Yoshikawa, Y.; Asada, M. Causality Detected by Transfer Entropy Leads Acquisition of Joint Attention. In Proceedings of the 6th IEEE International Conference on Development and Learning (ICDL 2007), London, UK, 11–13 July 2007; pp. 264–269.
36. Vejmelka, M.; Palus, M. Inferring the directionality of coupling with conditional mutual information. *Phys. Rev. E* **2008**, *77*, 026214.
37. Verdes, P.F. Assessing causality from multivariate time series. *Phys. Rev. E* **2005**, *72*, 026222:1–026222:9.
38. Tung, T.Q.; Ryu, T.; Lee, K.H.; Lee, D. Inferring Gene Regulatory Networks from Microarray Time Series Data Using Transfer Entropy. In Proceedings of the Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS '07), Maribor, Slovenia, 20–22 June 2007; Kokol, P., Podgorelec, V., Mičetič-Turk, D., Zorman, M., Verlič, M., Eds.; IEEE: Los Alamitos, CA, USA, 2007; pp. 383–388.
39. Van Dijck, G.; van Vaerenbergh, J.; van Hulle, M.M. Information Theoretic Derivations for Causality Detection: Application to Human Gait. In Proceedings of the International Conference on Artificial Neural Networks (ICANN 2007), Porto, Portugal, 9–13 September 2007; Sá, J.M.d., Alexandre, L.A., Duch, W., Mandic, D., Eds.; Springer-Verlag: Berlin/Heidelberg, Germany, 2007; Volume 4669, *Lecture Notes in Computer Science*, pp. 159–168.
40. Hung, Y.C.; Hu, C.K. Chaotic communication via temporal transfer entropy. *Phys. Rev. Lett.* **2008**, *101*, 244102.
41. Lizier, J.T.; Prokopenko, M. Differentiating information transfer and causal effect. *Eur. Phys. J. B* **2010**, *73*, 605–615.
42. Wuensche, A. Classifying cellular automata automatically: Finding gliders, filtering, and relating space-time patterns, attractor basins, and the Z parameter. *Complexity* **1999**, *4*, 47–66.
43. Solé, R.V.; Valverde, S. Information transfer and phase transitions in a model of internet traffic. *Physica A* **2001**, *289*, 595–605.
44. Solé, R.V.; Valverde, S. Information Theory of Complex Networks: On Evolution and Architectural Constraints. In *Complex Networks*; Ben-Naim, E., Frauenfelder, H., Toroczkai, Z., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; Volume 650, *Lecture Notes in Physics*, pp. 189–207.
45. Prokopenko, M.; Boschietti, F.; Ryan, A.J. An information-theoretic primer on complexity, self-organization, and emergence. *Complexity* **2009**, *15*, 11–28.
46. Takens, F. Detecting Strange Attractors in Turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*; Rand, D., Young, L.S., Eds.; Springer: Berlin/Heidelberg, Germany, 1981; pp. 366–381.
47. Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438.
48. Fano, R. *Transmission of Information: A Statistical Theory of Communications*; The MIT Press: Cambridge, MA, USA, 1961.
49. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; The MIT Press: Cambridge, MA, USA, 1999.
50. Dasan, J.; Ramamohan, T.R.; Singh, A.; Nott, P.R. Stress fluctuations in sheared Stokesian suspensions. *Phys. Rev. E* **2002**, *66*, 021409.

51. MacKay, D.J. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, MA, USA, 2003.
52. Pearl, J. *Causality: Models, Reasoning, and Inference*; Cambridge University Press: Cambridge, MA, USA, 2000.
53. Goyal, P. Information physics–towards a new conception of physical reality. *Information* **2012**, *3*, 567–594.
54. Sethna, J.P. *Statistical Mechanics: Entropy, Order Parameters, and Complexity*; Oxford University Press: Oxford, UK, 2006.
55. Seife, C. *Decoding the Universe*; Penguin Group: New York, NY, USA, 2006.
56. Einstein, A. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Ann. Phys.* **1905**, *322*, 132–148.
57. Norton, J.D. Atoms, entropy, quanta: Einstein's miraculous argument of 1905. *Stud. History Philos. Mod. Phys.* **2006**, *37*, 71–100.
58. Barnett, L.; Buckley, C.L.; Bullock, S. Neural complexity and structural connectivity. *Phys. Rev. E* **2009**, *79*, 051914.
59. Ay, N.; Bernigau, H.; Der, R.; Prokopenko, M. Information-driven self-organization: The dynamical system approach to autonomous robot behavior. *Theory Biosci.* **2012**, *131*, 161–179.
60. Evans, D.J.; Cohen, E.G.D.; Morriss, G.P. Probability of second law violations in shearing steady states. *Phys. Rev. Lett.* **1993**, *71*, 2401–2404.
61. Searles, D.J.; Evans, D.J. Fluctuations relations for nonequilibrium systems. *Aus. J. Chem.* **2004**, *57*, 1129–1123.
62. Crooks, G.E. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E* **1999**, *60*, 2721–2726.
63. Jarzynski, C. Nonequilibrium work relations: Foundations and applications. *Eur. Phys. J. B-Condens. Matter Complex Syst.* **2008**, *64*, 331–340.
64. Schlögl, F. Information Measures and Thermodynamic Criteria for Motion. In Structural Stability in Physics: Proceedings of Two International Symposia on Applications of Catastrophe Theory and Topological Concepts in Physics, Tübingen, Germany, 2–6 May and 11–14 December 1978; Güttinger, W., Eikemeier, H., Eds.; Springer: Berlin/Heidelberg, Germany, 1979; Volume 4, *Springer series in synergetics*, pp. 199–209.