Information flow and causality as rigorous notions *ab initio*

X. San Liang*
Nanjing Institute of Meteorology, Nanjing 210044, China
and China Institute for Advanced Study, Beijing 100081, China
(Received 23 June 2016; revised manuscript received 30 September 2016; published 1 November 2016)

Information flow or information transfer the widely applicable general physics notion can be rigorously derived from first principles, rather than axiomatically proposed as an ansatz. Its logical association with causality is firmly rooted in the dynamical system that lies beneath. The *principle of nil causality* that reads, an event is not causal to another if the evolution of the latter is independent of the former, which transfer entropy analysis and Granger causality test fail to verify in many situations, turns out to be a proven theorem here. Established in this study are the information flows among the components of time-discrete mappings and time-continuous dynamical systems, both deterministic and stochastic. They have been obtained explicitly in closed form, and put to applications with the benchmark systems such as the Kaplan-Yorke map, Rössler system, baker transformation, Hénon map, and stochastic potential flow. Besides unraveling the causal relations as expected from the respective systems, some of the applications show that the information flow structure underlying a complex trajectory pattern could be tractable. For linear systems, the resulting remarkably concise formula asserts analytically that causation implies correlation, while correlation does not imply causation, providing a mathematical basis for the long-standing philosophical debate over causation versus correlation.

DOI: 10.1103/PhysRevE.94.052201

I. INTRODUCTION

Information flow, or information transfer as it may be referred to in the literature, has been realized as a fundamental notion in general physics. Though literally one may associate it with communication, its importance lies far beyond in that it implies causation [1–5], uncertainty propagation [6], predictability transfer [7–9], etc. In fact, it is the recognition of its causality association that has attracted enormous interest from a wide variety of disciplines, e.g., neuroscience [10–16], finance [17,18], climate science [19,20], turbulence research [21,22], network dynamics [23–26], and dynamical systems particularly the field of synchronization [27–33].

This recognition has been further substantiated by the finding that transfer entropy [5] and Granger causality [34] are equivalent for Gaussian variables (up to a factor 2) [35].

Historically, many information theoretic quantities have been axiomatically or empirically proposed to measure information flow, including time-delayed mutual information [36], transfer entropy [5], momentary information transfer [20], and causation entropy [37]. Among these most notably is transfer entropy, which has spawned many varieties in its family, e.g., [13,17,38], and has been widely applied in different disciplines.

A fundamental question to ask is whether information flow needs to be axiomatically proposed as an ansatz (as the transfer entropy above), or whether it can be derived from first principles in information theory. Naturally, one would like to minimize or avoid the use of axioms in introducing new concepts in order to have the material more coherent within the field to which it belongs. In physics, "flow" or "transfer" does have definite meaning, albeit the meaning may differ depending on the context. One then naturally expects the concept to be rigorized. Indeed, as we will see soon, at least within the framework of dynamical systems, information flow/transfer can be rigorously derived from, rather than empirically or axiomatically proposed with, Shannon entropy.

Another impetus regards the inference of causality. As mentioned in the beginning, information flow arouses enormous interest in a wide range of fields not because of its original meaning in communication but because of its logical implication of causation. Whether the cause-effect relation underlying a system can be faithfully revealed is, therefore, the touchstone for a formalism of information flow. That is to say, information flow should be formulated with causality naturally embedded; it should, in particular, accurately reproduce a one-way causality (if existing), which is unambiguously equal to zero on one side. More specifically, a faithful formalism should verify the observational fact which we will henceforth refer to as *principle of nil causality*: an event is not causal to another event if the evolution of the latter does not depend on the former. In this light, the widely used formalism, namely, transfer entropy, is unfortunately not as satisfactory as one expects. This has even led to discussions on whether the two notions, namely, information flow and causality, should be differentiated (e.g., [39]). Since it is established that Granger causality and transfer entropy are equivalent, one may first look at the problems from the former. Now it is well known that spurious Granger causality may arise due to unobserved variables that influence the system dynamics (a problem identified by Granger himself) [40], due to low resolution in time [41,42], and due to observational noise [43]. Besides, Granger explicitly excludes deterministic systems in establishing the causality formalism, a case that is certainly important in realistic problems. For transfer entropy, the issue has just been systematically examined [45]. Aside from the failure in recovering the many preset one-way causalities, evidence has

*sanliang@courant.nyu.edu

shown that sometimes it may even give qualitatively wrong results; see [44] and [45] for such examples.

Realizing the limitation of transfer entropy, different alternatives have been proposed; the above momentary information transfer is one of these proposals. The purpose of this study is, instead of just remedying the deficiencies of the existing formalisms, to put information flow the fundamental physical notion on a rigorous footing so that it is universally applicable. The stringent one-way causality requirement will not be just verified with certain given examples, but rigorously proved as theorems.

With this faith, Liang and Kleeman [46] in 2005 took the initiative to study the problem with dynamical systems. In this framework, the information source and recipient are abstracted as the system components, and hence the problem is converted into the information flow or information transfer between dynamical system components. The basic idea can be best illustrated with a deterministic system of two components, say, x_1 and x_2:

$$\frac{dx_1}{dt} = F_1(x_1,x_2,t), \tag{1}$$

$$\frac{dx_2}{dt} = F_2(x_1,x_2,t), \tag{2}$$

where we follow the convention in physics and do not distinguish random and deterministic variables, which should be clear in the context. Now what we are to consider are the time evolutions of the marginal entropies of x_1 and x_2, denoted respectively as H_1 and H_2. Look at x_1; its marginal entropy evolution may be due to x_1 itself or subject to the influence of x_2. This partitions the set of mechanisms that cause H_1 to grow into two exclusive subsets. That is to say, if we write the contribution from the former mechanism as dH^*/dt and that from the latter as $T_{2\to1}$, then

$$\frac{dH_1}{dt} = \frac{dH_1^*}{dt} + T_{2\to1}. \tag{3}$$

This $T_{2\to1}$ is the very time rate of information flowing from x_2 to x_1. We remark that this setting is rather generic, except for the requirement of differentiability for the vector field $\mathbf{F} = (F_1,F_2)^T$. In particular, the input-output communication problem can be cast within the framework by letting, for example, $F_2 = F_2(x_1,t)$, $F_1 = F_1(x_1,t)$, where x_1 is the input/drive and x_2 the output/consequence, and the channel is represented by F_2.

From the above argument, the evaluation of the information flow $T_{2\to1}$ may be fulfilled through evaluating dH_1^*/dt. This is because that, when a dynamical system is given, the density evolution is known through the corresponding Liouville equation, and, accordingly, dH_1/dt can be obtained. In [46], Liang and Kleeman prove that the joint entropy of (x_1,x_2) follows a very concise law

$$\frac{dH}{dt} = E(\nabla\cdot\mathbf{F}), \tag{4}$$

where E is the operator of mathematical expectation. They then argue that

$$\frac{dH_1^*}{dt} = E\left(\frac{\partial F_1}{\partial x_1}\right), \tag{5}$$

and hence obtain the time rate of information flowing from x_2 to x_1

$$T_{2\to1} = \frac{dH_1}{dt} - \frac{dH_1^*}{dt} = -E\left(\frac{1}{\rho_1}\frac{\partial F_1\rho_1}{\partial x_1}\right), \tag{6}$$

where ρ_1 is the marginal probability density function of x_1. The thus-obtained information flow is asymmetric between x_1 and x_2; moreover, it possesses a *property of causality*, which reads, if the evolution of x_1 does not depend on x_2, then $T_{2\to1} = 0$. This is precisely the principle of nil causality.

The above result is later on rigorously proved [48,49]. It is remarkable in that the principle of nil causality can be stated as a proven theorem, rather than a fact for a formalism to verify; see [47] for a review. This result, however, is only for systems of dimension 2 (2D). For systems with many components, it does not work any more. We have endeavored to extend it to more general situations and have obtained results for deterministic systems of arbitrary dimensionality which possess the property of causality. But, as we will see in the following section, the extension relies on an assumption that is, again, axiomatically proposed. This makes the resulting formalism not one fully derived from first principles, and as we realize later on, it does not work for multidimensional stochastic systems. This line of work, though with a promising start, is stuck at this point.

In this study, we will show that the assumption can be completely removed. In a unified approach, the notion of information flow can be rigorously derived for both deterministic and stochastic systems of arbitrary dimensionality. In the following, we first briefly set up the framework, and show where the snag lies in the above approach. The solution is then presented, and applied to derive the information flows for deterministic mappings (Sec. III), continuous-time deterministic systems (Sec. IV), stochastic mappings (Sec. V), and continuous-time stochastic systems (Sec. VI). For the purpose of demonstration, each section contains one or more applications. As an important particular case, we specialize to do the derivation for linear systems, and the material is presented in Sec. VII. This study is summarized in Sec. VIII.

II. THE SNAG IN THE LIANG-KLEEMAN FORMALISM

The success of the Liang-Kleeman formalism is remarkable. It is, however, only for 2D dynamical systems. When the dimensionality exceeds 2, the resulting quantity, namely, (6), is not the information transfer from x_2 to x_1, but the cumulant transfer to x_1 from all other components $x_2,x_3,\ldots,x_n$. In this sense, the use of (6) is rather limited.

In order to extend the formalism to systems of higher dimensionality, Liang and Kleeman [48,49] reinterpret the term dH_1^*/dt in the above decomposition (3), for a 2D system, as the evolution of H_1 with the effect of x_2 excluded. More specifically, it is the evolution of H_1 with x_2 instantaneously frozen as a parameter at time t. To avoid confusing with dH_1^*/dt, denote it as $dH_{1\wr}/dt$, where the subscript $\wr$ signifies that x_2 is frozen, or that its effect is removed. With this, the disjoint decomposition (3) is restated as

$$\frac{dH_1}{dt} = \frac{dH_{1\wr}}{dt} + T_{2\to1}. \tag{7}$$

Note this decomposition, albeit seemingly with only a change of symbol, is actually fundamentally different from (3) in physical meaning; it now holds for systems of arbitrary dimensionality. The information flow is, therefore,

$$T_{2\to 1} = \frac{dH_1}{dt} - \frac{dH_{1\wr}}{dt}. \tag{8}$$

Of course, the key is how to find $\frac{dH_{1\wr}}{dt}$. In [48] and [49], Liang and Kleeman start with discrete mappings, and then take the limit as the time step size goes to zero. To illustrate, let $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping taking $\mathbf{x}(\tau)$ to $\mathbf{x}(\tau+1)$, as time moves on, from step τ to $\tau+1$. Correspondingly there is another mapping $\mathscr{P} : L^1(\mathbb{R}^n) \to L^1(\mathbb{R}^n)$ that steers its density ρ forward. This mapping is called a Frobenius-Perron operator; we will refer to it as the F-P operator henceforth. Loosely speaking, $\mathscr{P}$ is, for any $\omega \subset \mathbb{R}^n$, such that [50]

$$\int_\omega \mathscr{P}\rho(\mathbf{x})d\mathbf{x} = \int_{\Phi^{-1}(\omega)} \rho(\mathbf{x})d\mathbf{x}. \tag{9}$$

When the sample space is in a Cartesian product form, as it is in this case ($\mathbb{R}^n$), the operator can be evaluated. Let $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ be some constant point, and $\omega = [a_1, x_1] \times [a_2, x_2] \times \ldots \times [a_n, x_n]$. It has been established that (e.g., [50])

$$\mathscr{P}\rho(\mathbf{x}) = \frac{\partial^n}{\partial x_n \ldots \partial x_2 \partial x_1} \times \int_{\Phi^{-1}(\omega)} \rho(\xi_1, \xi_2, \ldots, \xi_n)d\xi_1 d\xi_2 \ldots d\xi_n. \tag{10}$$

For convenience, $\mathbf{a}$ is usually taken to be the origin. Furthermore, if Φ is nonsingular and invertible, then $\mathscr{P}$ can be explicitly written out

$$\mathscr{P}\rho(\mathbf{x}) = \rho[\Phi^{-1}(\mathbf{x})] \cdot |J^{-1}|, \tag{11}$$

where J is the Jacobian of Φ.

As the F-P operator carries ρ forth from time step τ to $\tau+1$, accordingly the entropies H, H_1, and H_2 are also steered forward. On $[\tau, \tau+1]$, let H_1 be incremented by ΔH_1. By the foregoing argument, the evolution of H_1 can be decomposed into two exclusive parts, namely, the information flow from x_2, $T_{2\to 1}$, and the evolution with the effect of x_2 excluded, $\Delta H_{1\wr}$. Hence, for discrete mappings, we have the following counterpart of (8):

$$T_{2\to 1} = \Delta H_1 - \Delta H_{1\wr}. \tag{12}$$

Liang and Kleeman derive the information flow for the continuous system from the discrete mapping. So the whole procedure relies on how

$$\Delta H_{1\wr} = H_{1\wr}(\tau+1) - H_1(\tau)$$

is evaluated, or, more specifically, how $H_{1\wr}(\tau+1)$ is evaluated [since $H_1(\tau)$ is known]. To see where lies its difficulty, first notice that

$$H_1(\tau+1) = -\int_{\mathbb{R}} (\mathscr{P}\rho)_1(x_1)\log(\mathscr{P}\rho)_1(x_1)dx_1,$$

which is the mean of $-\log(\mathscr{P}\rho)_1(x_1)$. Here the base of the log may be either 2 or e; it only affects the units of the results (resp. bits and nats). We hence do not specify it here and in the following formulas. But for consistency, let us use a base of e throughout. Given Φ, $\mathscr{P}$ can be found in the way as shown above, so $H_1(\tau+1)$ is known. For $H_{1\wr}$, however, things are much more difficult; $-\log(\mathscr{P}_\wr\rho)_1(x_1)$ involves not only the random variable $x_1(\tau+1)$, but also $x_2(\tau)$ (embedded in the subscript $\wr$). What is the joint density of $(x_2(\tau), x_1(\tau+1))$? We do not know. In [48] and [49], an approximation was proposed, which gives

$$H_{1\wr}(\tau+1) = -\int_\Omega (\mathscr{P}_\wr\rho)_1(y_1)\log(\mathscr{P}_\wr\rho)_1(y_1) \times \rho(x_2|x_1, x_3, \ldots, x_n)\rho_{3\ldots n}(x_3, \ldots, x_n) \times dy_1 dx_2 dx_3 \ldots dx_n,$$

where y_1 is employed to signify $x_1(\tau+1)$ and the symbol x_1 is reserved for $x_1(\tau)$. This is a natural extension of what the authors use in the their original study [46] for 2D discrete mappings. A central approximation is that they use

$$\rho(x_2|x_1, x_3, \ldots, x_n)(\mathscr{P}_\wr\rho)_1(y_1)\rho_{3\ldots n}(x_3, \ldots, x_n)$$

to represent the joint probability density function (pdf) of y_1 and x_2 (and $x_3, \ldots, x_n$) [think about $\rho(x_2|x_1)\rho(x_1) = \rho(x_1, x_2)$]. This is, however, only an approximation, since we really do not know what the joint pdf of (y_1, x_2) is. As we will see soon, though the resulting formalism verifies $dH_{1\wr}/dt = dH_1^*/dt$ for 2D systems and the principle of nil causality for one-way causal deterministic systems, when stochasticity gets in, the causality property cannot be recovered this way.

III. DETERMINISTIC MAPPING

A. Derivation

Fortunately, the issue that stuck the Liang-Kleeman formalism can be fixed; we actually can get the entropy without appealing to the joint probability density function of (y_1, x_2), i.e., that of $(x_1(\tau+1), x_2(\tau))$ as mentioned above. Consider a mapping

$$\Phi : \Omega \to \Omega, \quad \mathbf{x}(\tau) \mapsto \mathbf{x}(\tau+1) = (\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), \ldots, \Phi_n(\mathbf{x})),$$

where Ω is the sample space ($\mathbb{R}^n$ in particular). Let $\psi : \Omega \to \Omega$ be an arbitrary differentiable function of $\mathbf{x}$. We have the following theorem:

Theorem III.1.

$$E\psi(\mathbf{x}(\tau+1)) = E\psi(\Phi(\mathbf{x}(\tau))). \tag{13}$$

Remark 1. The expectation operator E on the right hand side applies to a function of $\mathbf{x}(\tau)$; it is thence with respect to $\rho(\tau)$. Differently, the left hand side E is with respect to $\rho(\tau+1) = \mathscr{P}\rho$, where $\mathscr{P}$ is the F-P operator as introduced in (9).

Remark 2. This equality is important in that one actually can obtain the expectation of $\psi(\mathbf{x}(\tau+1))$ without evaluating $\mathscr{P}\rho$.

Proof. The following proof is in the framework of Riemann-Stieltjes integration. A more general proof in terms of Lebesgue theory is also possible; in that case it may be used to introduce the Koopman operator (e.g., [50]). But here this is unnecessary, since the functions and vector fields we are dealing with in this study are assumed to differentiable.

Let $\{\omega_1, \omega_2, \ldots, \omega_n\}$ be a partitioning of the sample space Ω. The elements are mutually exclusive and $\Omega = \cup_{k=1}^n \omega_k$. To

make it simple, assume that these ω_k's have the same diameter (the maximal distance between any two points in ω_k). For clarity, write $\mathbf{x}(\tau+1)$ as $\mathbf{y}$, while $\mathbf{x}$ is reserved for $\mathbf{x}(\tau)$. Then

$$\begin{aligned} E\psi(\mathbf{x}(\tau+1)) &= \int_\Omega \mathscr{P}\rho(\mathbf{y})\psi(\mathbf{y})d\mathbf{y} \\ &= \lim_{n\to\infty}\sum_{k=1}^{n}\int_{\omega_k}\mathscr{P}\rho(\mathbf{y})\psi(\mathbf{y})d\mathbf{y} \\ &= \lim_{n\to\infty}\sum_{k=1}^{n}\psi(\mathbf{y}_k)\int_{\omega_k}\mathscr{P}\rho(\mathbf{y})d\mathbf{y}, \end{aligned}$$

where $\mathbf{y}_k \in \omega_k$ is some point in ω_k. The existence of the Riemann integral $\int_\Omega \mathscr{P}\rho(\mathbf{y})\psi(\mathbf{y})d\mathbf{y}$ assures that it can be any point in ω_k as n goes to infinity, while the resulting integral is the same. Now by (9),

$$\int_{\omega_k}\mathscr{P}\rho(\mathbf{y})d\mathbf{y} = \int_{\Phi^{-1}(\omega_k)}\rho(\mathbf{x})d\mathbf{x}.$$

So the above becomes

$$\begin{aligned} E\psi(\mathbf{x}(\tau+1)) &= \lim_{n\to\infty}\sum_{k=1}^{n}\psi(\mathbf{y}_k)\int_{\omega_k}\mathscr{P}\rho(\mathbf{y})d\mathbf{y} \\ &= \lim_{n\to\infty}\sum_{k=1}^{n}\psi(\mathbf{y}_k)\int_{\Phi^{-1}(\omega_k)}\rho(\mathbf{x})d\mathbf{x} \\ &= \lim_{n\to\infty}\sum_{k=1}^{n}\int_{\Phi^{-1}(\omega_k)}\rho(\mathbf{x})\psi(\Phi(\mathbf{x}))d\mathbf{x}. \end{aligned}$$

Notice, for $\Phi:\Omega\to\Omega$, $\Omega=\cup_k\omega_k$, it must be that $\cup_k\Phi^{-1}(\omega_k)=\Omega$. So the limit converges to $\int_\Omega\rho(\mathbf{x})\psi(\Phi(\mathbf{x}))d\mathbf{x}$. That is to say, $E\psi(\mathbf{x}(\tau+1)) = E\psi(\Phi(\mathbf{x}(\tau)))$. ■

The equality (13) actually can be utilized to derive the F-P operator. We look at the particular case when Φ is invertible. By definition, Eq. (13) means

$$\int_\Omega\psi(\mathbf{x})\rho(\tau+1,\mathbf{x})d\mathbf{x} = \int_\Omega\psi(\Phi(\mathbf{x}))\rho(\tau,\mathbf{x})d\mathbf{x}.$$

If Φ is invertible, the right hand side is $\int_\Omega\psi(\mathbf{y})\cdot\rho(\tau,\Phi^{-1}(\mathbf{y}))|J^{-1}|d\mathbf{y}$ by transformation of variables. Since ψ is arbitrary, we have

$$\mathscr{P}\rho = \rho(\tau+1,\mathbf{x}) = \rho(\tau,\Phi^{-1}(\mathbf{x}))\cdot|J^{-1}|,$$

which is precisely the Frobenius-Perron operator (11).

The above equality provides us a convenient and accurate way to evaluate $H_1(\tau+1)$ and $H_{1\mathfrak{Y}}(\tau+1)$. Picking ψ as $(\log\mathscr{P}\rho)_1$ and $\log(\mathscr{P}_\mathfrak{Y}\rho)_1$, we obtain, respectively, the following formulas:

Corollary III.1.

$$H_1(\tau+1) = -E\log(\mathscr{P}\rho)_1(\Phi_1(\mathbf{x})), \tag{14}$$

$$H_{1\mathfrak{Y}}(\tau+1) = -E\log(\mathscr{P}_\mathfrak{Y}\rho)_1(\Phi_1(\mathbf{x})). \tag{15}$$

In these formulas, both the expectations are taken with respect to $\rho(x_1,x_2,\ldots,x_n)$, i.e., the pdf at time step τ. In (15), we do not need to care about the joint pdf $\rho(y,x_2)$ any more. The information flow from x_2 to x_1 is, therefore,

Theorem III.2.

$$T_{2\to1} = E\log(\mathscr{P}_\mathfrak{Y}\rho)_1(\Phi_1(\mathbf{x})) - E\log(\mathscr{P}\rho)_1(\Phi_1(\mathbf{x})). \tag{16}$$

Proof.

$$\begin{aligned} T_{2\to1} &= \Delta H_1 - \Delta H_{1\mathfrak{Y}} = [H_1(\tau+1)-H_1(\tau)] \\ &\quad -[H_{1\mathfrak{Y}}(\tau+1)-H_1(\tau)] = H_1(\tau+1)-H_{1\mathfrak{Y}}(\tau+1). \end{aligned}$$

Substitute into the above formulas for H_1 and $H_{1\mathfrak{Y}}$ and (16) follows.

Note that the evaluation of $\mathscr{P}_\mathfrak{Y}\rho$ and $\mathscr{P}\rho$ generally depends on the system in question. But when Φ and $\Phi_\mathfrak{Y}$ are invertible, the information flow can be found explicitly in a closed form.

B. Properties

Theorem III.3. For 2D systems, if Φ_1 is invertible, then

$$\Delta H_{1\mathfrak{Y}} = H_{1\mathfrak{Y}}(\tau+1) - H_1(\tau) = E\log|J_1|.$$

Remark. This is the analog of (5) for discrete-time systems [46,48].

Proof. Let $\mathbf{x}(\tau+1)\equiv\mathbf{y}$. For a 2D system, and if Φ_1 is invertible, we have

$$(\mathscr{P}_\mathfrak{Y}\rho)_1(\mathbf{y}) = \rho_1\left(\Phi_1^{-1}(y_1)\right)\cdot\left|J_1^{-1}\right|,$$

which gives

$$\begin{aligned} H_{1\mathfrak{Y}} &= -E_x\log\left[\rho_1({\Phi^{-1}}_1(y_1))\cdot|J_1^{-1}|\right] \\ &= -E\log\left[\rho_1(x_1)\cdot|J_1^{-1}|\right] \\ &= -E\log\rho_1(x_1) + E\log|J_1|. \end{aligned}$$

To avoid confusion, we use E_x to indicate that the expectation is with respect to x when mixed variables x and y appear simultaneously. Note here $x_1 = {\Phi^{-1}}_1(y_1)$ since this is a 1D system after x_2 is frozen. Thus

$$\Delta H_{1\mathfrak{Y}} = E\log|J_1|.$$

■

Theorem III.4 (principle of nil causality). If Φ_1 is independent of x_2, then $T_{2\to1}=0$.

Proof. By the definition of the F-P operator,

$$\begin{aligned} &\int_{\omega_1}(\mathscr{P}_\mathfrak{Y}\rho)_1(x_1)dx_1 \\ &= \int_{\omega_1\times\mathbb{R}^{n-2}}\mathscr{P}_\mathfrak{Y}\rho(x_1,x_3,\ldots,x_n)dx_1dx_3\ldots dx_n \\ &= \int_{{\Phi^{-1}}_\mathfrak{Y}(\omega_1\times\mathbb{R}^{n-2})}\rho_\mathfrak{Y}(x_1,x_3,\ldots,x_n)dx_1dx_3\ldots dx_n \end{aligned}$$

for any $\omega_1\subset\mathbb{R}$. Note

$${\Phi^{-1}}_\mathfrak{Y}(\omega_1\times\mathbb{R}^{n-2}) = {\Phi^{-1}}_{1\mathfrak{Y}}(\omega_1\times\mathbb{R}^{n-2}).$$

That is to say,

$$\begin{aligned} &\int_{\omega_1}(\mathscr{P}_\mathfrak{Y}\rho)_1(x_1)dx_1 \\ &= \int_{{\Phi^{-1}}_{1\mathfrak{Y}}\omega_1}dx_1\int_{\mathbb{R}^{n-2}}\rho_\mathfrak{Y}(x_1,x_3,\ldots,x_n)dx_3\ldots dx_n \\ &= \int_{{\Phi^{-1}}_{1\mathfrak{Y}}\omega_1}\rho_1(x_1)dx_1 = \int_{{\Phi^{-1}}_1\omega_1}\rho_1(x_1)dx_1 \end{aligned}$$

since Φ_1 (hence ${\Phi^{-1}}_1$) is independent of x_2. On the other hand,

$$\begin{aligned}\int_{\omega_1}(\mathscr{P}\rho)_1(x_1)dx_1 &= \int_{\omega_1\times\mathbb{R}^{n-1}}\mathscr{P}\rho(\mathbf{x})d\mathbf{x}\\ &= \int_{\Phi^{-1}(\omega_1\times\mathbb{R}^{n-1})}\rho(\mathbf{x})d\mathbf{x}\\ &= \int_{{\Phi^{-1}}_1\omega_1\times\mathbb{R}^{n-1}}\rho(\mathbf{x})d\mathbf{x}\\ &= \int_{{\Phi^{-1}}_1\omega_1}dx_1\int_{\mathbb{R}^{n-1}}\rho(\mathbf{x})dx_2\ldots dx_n\\ &= \int_{{\Phi^{-1}}_1\omega_1}\rho_1(x_1)dx_1.\end{aligned}$$

So $\int_{\omega_1}(\mathscr{P}_{\not 2}\rho)_1(x_1)dx_1 = \int_{\omega_1}(\mathscr{P}\rho)_1(x_1)dx_1$, $\forall\omega_1\subset\mathbb{R}$, and hence $(\mathscr{P}_{\not 2}\rho)_1 \overset{a.e.}{=} (\mathscr{P}\rho)_1$. Therefore,

$$E\log(\mathscr{P}\rho)_1(x_1) = E\log(\mathscr{P}_{\not 2}\rho)_1(x_1)$$

and

$$T_{2\to 1} = H_1(\tau+1) - H_{1\not 2}(\tau+1) = 0.$$

■

C. Application: Kaplan-Yorke map

Once a dynamical system is specified, in principle the information flow can be obtained. This subsection presents an application with a discrete-time dynamical system, the Kaplan-Yorke map [51], that exhibits chaotic behavior.

The Kaplan-Yorke map is defined as a mapping $\Phi = (\Phi_1,\Phi_2):[0,1]\times\mathbb{R}\to[0,1]\times\mathbb{R}$, $(x_1,x_2)\mapsto(y_1,y_2)$, such that

$$y_1 = \Phi_1(x_1,x_2) = 2x_1 \mod 1, \tag{17}$$

$$y_2 = \Phi_2(x_1,x_2) = \alpha x_2 + \cos(4\pi x_1). \tag{18}$$

A typical trajectory for $\alpha = 0.2$ is plotted in Fig. 1. We now compute the information flows between the two components.

First we need to find the F-P operator $\mathscr{P}\rho(y_1,y_2)$. Pick a domain $\omega = [0,y_1]\times[0,y_2]$. By (10)

$$\mathscr{P}\rho(y_1,y_2) = \frac{\partial^2}{\partial y_2\partial y_1}\int_{\Phi^{-1}(\omega)}\rho(\xi_1,\xi_2)d\xi_1d\xi_2, \tag{19}$$

so the key is the finding of $\Phi^{-1}(\omega)$. Since

$$y_1 = \begin{cases}2x_1, & 0\leqslant x_1\leqslant\frac{1}{2},\\ 2x_1-1, & x_1>\frac{1}{2},\end{cases}$$

it is easy to obtain

$${\Phi_1}^{-1}([0,y_1]) = \left[0,\frac{y_1}{2}\right]\cup\left[\frac{1}{2},\frac{1+y_1}{2}\right]. \tag{20}$$

Given y_1, x_1 may be either $y_1/2$ or $(1+y_1)/2$, but either way, $\cos(4\pi x_1) = \cos(2\pi y_1)$. Thus

$${\Phi_2}^{-1}(\{y_1\}\times[0,y_2]) = \left[-\frac{\cos 2\pi y_1}{\alpha},\frac{y_2-\cos 2\pi y_1}{\alpha}\right]. \tag{21}$$

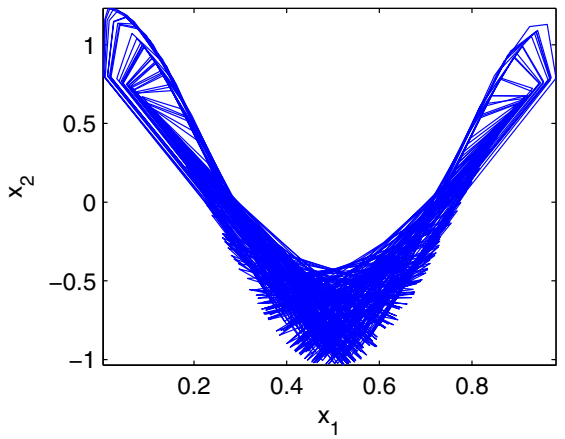

FIG. 1. The attractor of the Kaplan-Yorke map [Eqs. (17) and (18)] with $\alpha = 0{,}2$. To avoid the round-off error in the computation which will quickly lead to a zero x_1, we let $b = 9\,722\,377$, and instead compute $a_{n+1} = 2a_n \mod b$, $x_{1,n+1} = a_n/b$, $x_{2,n+1} = \alpha x_{2,n} + \cos(4\pi x_{1,n})$. The trajectory is initialized with $x_1 = 7\,722\,377/b$, $x_2 = 0$. (The initial points outside the attractor are not shown.)

Equation (19) is, therefore,

$$\begin{aligned}\mathscr{P}\rho(y_1,y_2) &= \frac{\partial^2}{\partial y_2\partial y_1}\int_0^{y_1/2}d\xi_1\int_{-\frac{\cos 2\pi y_1}{\alpha}}^{\frac{y_2-\cos 2\pi y_1}{\alpha}}\rho(\xi_1,\xi_2)d\xi_2\\ &\quad+\frac{\partial^2}{\partial y_2\partial y_1}\int_{1/2}^{(1+y_1)/2}d\xi_1\int_{-\frac{\cos 2\pi y_1}{\alpha}}^{\frac{y_2-\cos 2\pi y_1}{\alpha}}\rho(\xi_1,\xi_2)d\xi_2\\ &= \frac{1}{2\alpha}\left[\rho\left(\frac{y_1}{2},\frac{y_2-\cos 2\pi y_1}{\alpha}\right)\right.\\ &\quad\left.+\rho\left(\frac{1+y_1}{2},\frac{y_2-\cos 2\pi y_1}{\alpha}\right)\right]\\ &\quad+\frac{1}{\alpha}\left[\int_0^{y_1/2}\frac{\partial}{\partial y_1}\rho\left(\xi_1,\frac{y_2-\cos 2\pi y_1}{\alpha}\right)d\xi_1\right.\\ &\quad\left.+\int_{1/2}^{(1+y_1)/2}\frac{\partial}{\partial y_1}\rho\left(\xi_1,\frac{y_2-\cos 2\pi y_1}{\alpha}\right)d\xi_1\right].\end{aligned}$$

To compute $T_{2\to 1}$, freeze x_2. The resulting mapping $\Phi_{\not 2}$ is the dyadic mapping in the x_1 direction. As above,

$${\Phi_{\not 2}}^{-1}([0,y_1]) = \left[0,\frac{y_1}{2}\right]\cup\left[\frac{1}{2},\frac{1+y_1}{2}\right],$$

which gives

$$\begin{aligned}\mathscr{P}_{\not 2}\rho(y_1) &= \frac{\partial}{\partial y_1}\int_{{\Phi_{\not 2}}^{-1}([0,y_1])}\rho_1(\xi_1)d\xi_1\\ &= \frac{1}{2}\left[\rho_1\left(\frac{y_1}{2}\right)+\rho_1\left(\frac{1+y_1}{2}\right)\right].\end{aligned}$$

On the other hand,

$$\begin{aligned}(\mathscr{P})_1(y_1) &= \int_{\mathbb{R}}\mathscr{P}\rho(y_1,y_2)dy_2\\ &= \frac{1}{2}\rho_1\left(\frac{y_1}{2}\right)+\frac{1}{2}\rho_1\left(\frac{1+y_1}{2}\right)\end{aligned}$$

$$+\int_0^{y_1/2}\frac{\partial}{\partial y_1}\rho_1(\xi_1)d\xi_1+\int_{1/2}^{(1+y_1)/2}\frac{\partial}{\partial y_1}\rho_1(\xi_1)d\xi_1$$
$$=\frac{1}{2}\left[\rho_1\left(\frac{y_1}{2}\right)+\rho_1\left(\frac{1+y_1}{2}\right)\right].$$

So

$$T_{2\to1}=E\log(\mathscr{P}_{\xi}\rho)_1(y_1)-E\log(\mathscr{P}\rho)_1(y_1)=0, \quad (22)$$

just as one would expect based on the independence of Φ_1 on x_2. This serves as a validation of Theorem III.4.

To compute $T_{1\to2}$, notice

$$\Phi_{\xi}^{-1}([0,y_2])=\left[-\frac{\cos 4\pi x_1}{\alpha},\ \frac{y_2-\cos 4\pi x_1}{\alpha}\right].$$

The corresponding F-P operator is such that

$$(\mathscr{P}_{\xi}\rho)(y_2)=\frac{\partial}{\partial y_2}\int_{-\frac{\cos 4\pi x_1}{\alpha}}^{\frac{y_2-\cos 4\pi x_1}{\alpha}}\rho_2(\xi_2)d\xi_2$$
$$=\frac{1}{\alpha}\rho_2\left(\frac{y_2-\cos 4\pi x_1}{\alpha}\right)=\frac{1}{\alpha}\rho_2(x_2),$$

which makes sense, considering that, when x_1 is frozen, y_2 is just a translation followed by a rescaling of x_2. On the other hand, the marginal density

$$(\mathscr{P}\rho)_2(y_2)$$
$$=\int_0^1\mathscr{P}\rho(y_1,y_2)dy_1$$
$$=\frac{1}{2\alpha}\int_0^1\left[\rho\left(\frac{y_1}{2},\frac{y_2-\cos 2\pi y_1}{\alpha}\right)\right.$$
$$\left.+\rho\left(\frac{1+y_1}{2},\frac{y_2-\cos 2\pi y_1}{\alpha}\right)\right]dy_1$$
$$+\frac{1}{\alpha}\int_0^1 dy_1\left[\int_0^{y_1/2}\frac{\partial}{\partial y_1}\rho\left(\xi_1,\frac{y_2-\cos 2\pi y_1}{\alpha}\right)d\xi_1\right.$$
$$\left.+\int_{1/2}^{(1+y_1)/2}\frac{\partial}{\partial y_1}\rho\left(\xi_1,\frac{y_2-\cos 2\pi y_1}{\alpha}\right)d\xi_1\right].$$

Because of the intertwined y_1 and y_2, these integrals cannot be explicitly evaluated without specifications of ρ. But when ρ is given, it is a straightforward exercise to compute

$$-E\log(\mathscr{P}\rho)_2(y_2)$$
$$=-\int_0^1\int_{\mathbb{R}}\log(\mathscr{P}\rho)_2[\Phi_2(x_1,x_2)]\rho(x_1,x_2)dx_1dx_2.$$

Denote it by $\tilde{H}_2$. Then

$$T_{1\to2}=E\log(\mathscr{P}_{\xi}\rho)_2[\Phi_2(x_1,x_2)]-E\log(\mathscr{P}\rho)_2[\Phi_2(x_1,x_2)]$$
$$=\int_0^1\int_{\mathbb{R}}\frac{1}{\alpha}\rho_2(x_2)\rho(x_1,x_2)dx_1dx_2+\tilde{H}_2$$
$$=\tilde{H}_2-H_2/\alpha. \quad (23)$$

Generally this does not vanish. That is to say, within the Kaplan-Yorke map, there exists a one-way information flow from x_1 to x_2.

D. Applications: The baker transformation and Hénon map revisited

Since its establishment, the Liang-Kleeman formalism has been applied to a variety of dynamical system problems. Hereafter we will restudy some benchmark examples and see whether the results are different. In this subsection we look at the baker transformation and Hénon map.

1. Baker transformation

The baker transformation is an extensively studied prototype of area-conserving chaotic maps that has been used to model the diffusion process in the real physical world. It mimics the kneading of dough: first the dough is compressed, then cut in half; the two halves are stacked on one another, compressed, and so forth; see Fig. 2 for an illustration. In formal language, it is $\Phi:\Omega\to\Omega$, $\Omega=[0,1]\times[0,1]$ being a unit square,

$$\Phi(x_1,x_2)=\begin{cases}\left(2x_1,\ \frac{x_2}{2}\right), & 0\leqslant x_1\leqslant\frac{1}{2},\ 0\leqslant x_2\leqslant 1,\\ \left(2x_1-1,\ \frac{1}{2}x_2+\frac{1}{2}\right), & \frac{1}{2}<x_1\leqslant 1,\ 0\leqslant x_2\leqslant 1.\end{cases} \quad (24)$$

It is invertible, and the inverse is

$$\Phi^{-1}(x_1,x_2)=\begin{cases}\left(\frac{x_1}{2},\ 2x_2\right), & 0\leqslant x_2\leqslant\frac{1}{2},\ 0\leqslant x_1\leqslant 1,\\ \left(\frac{x_1+1}{2},\ 2x_2-1\right), & \frac{1}{2}\leqslant x_2\leqslant 1,\ 0\leqslant x_1\leqslant 1.\end{cases} \quad (25)$$

Thus the F-P operator $\mathscr{P}$ can be easily found:

$$\mathscr{P}\rho(x_1,x_2)=\rho[\Phi^{-1}(x_1,x_2)]\cdot|J^{-1}|$$
$$=\begin{cases}\rho\left(\frac{x_1}{2},2x_2\right), & 0\leqslant x_2<\frac{1}{2},\\ \rho\left(\frac{1+x_1}{2},2x_2-1\right), & \frac{1}{2}\leqslant x_2\leqslant 1.\end{cases} \quad (26)$$

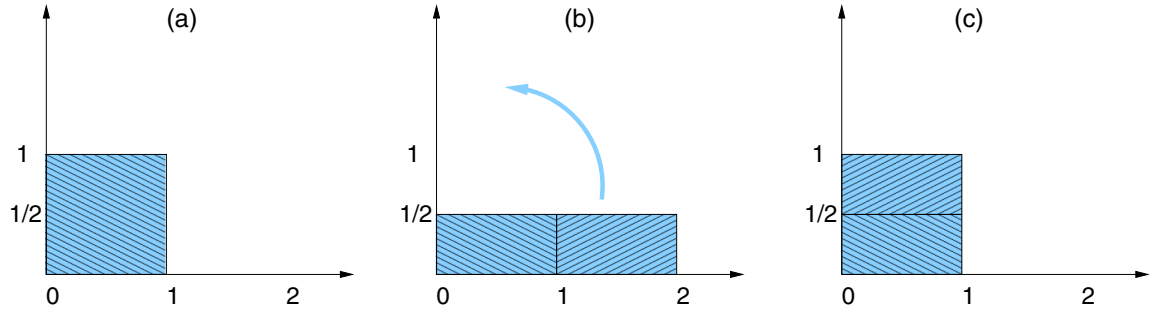

FIG. 2. A schematic of the unidirectional information flow from the abscissa to the ordinate upon applying the baker transformation.

We now use the above theorem to compute $T_{2\to 1}$. Integrating (26) with respect to x_2,

$$\begin{aligned}(\mathscr{P}\rho)_1(x_1) &= \int_0^{1/2} \rho\Big(\frac{x_1}{2},2x_2\Big)\,dx_2 + \int_{1/2}^1 \rho\Big(\frac{x_1+1}{2},2x_2-1\Big)dx_2\\ &= \frac{1}{2}\int_0^1 \Big[\rho\Big(\frac{x_1}{2},x_2\Big)+\rho\Big(\frac{x_1+1}{2},x_2\Big)\Big]dx_2\\ &= \frac{1}{2}\Big[\rho_1\Big(\frac{x_1}{2}\Big)+\rho_1\Big(\frac{x_1+1}{2}\Big)\Big]. \end{aligned} \tag{27}$$

When x_2 is frozen as a parameter, the baker transformation (24) becomes a dyadic mapping in the x_1 direction, i.e., a mapping $\Phi_1 : [0,1] \to [0,1]$,

$$\Phi_1(x_1) = 2x_1 \ (\mathrm{mod}\ 1).$$

For any $0 < x_1 < 1$, the counterimage of $[0,x_1]$ is

$$\Phi^{-1}([0,x_1]) = \Big[0,\frac{x_1}{2}\Big]\cup\Big[\frac{1}{2},\frac{1+x_1}{2}\Big].$$

So

$$\begin{aligned}(\mathscr{P}_{\wr}\rho)_1(x_1) &= \frac{\partial}{\partial x_1}\int_{\Phi^{-1}([0,x_1])}\rho(s)ds\\ &= \frac{\partial}{\partial x_1}\int_0^{x_1/2}\rho(s)ds + \frac{\partial}{\partial x_1}\int_{1/2}^{(1+x_1)/2}\rho(s)ds\\ &= \frac{1}{2}\Big[\rho\Big(\frac{x_1}{2}\Big)+\rho\Big(\frac{1+x_1}{2}\Big)\Big].\end{aligned}$$

Thus

$$(\mathscr{P}_{\wr}\rho)_1(x_1) = (\mathscr{P}\rho)_1(x_1).$$

By the above theorem,

$$T_{2\to 1} = E\log(\mathscr{P}_{\wr}\rho)_1[\Phi_1(\mathbf{x})] - E\log(\mathscr{P}\rho)_1[\Phi_1(\mathbf{x})] = 0.$$

To compute $T_{1\to 2}$, observe

$$(\mathscr{P}\rho)_2(x_2) = \int_0^1 \mathscr{P}\rho(x_1,x_2)dx_1 = \begin{cases}\int_0^1 \rho\big(\frac{x_1}{2},2x_2\big)dx_1, & 0\leqslant x_2<\frac{1}{2},\\ \int_0^1 \rho\big(\frac{x_1+1}{2},2x_2-1\big)dx_1, & \frac{1}{2}\leqslant x_2\leqslant 1.\end{cases} \tag{28}$$

By Corollary III.1,

$$\begin{aligned}H_2(\tau+1) &= -E_x\log(\mathscr{P}\rho)_2[\Phi_2(x_1,x_2)]\\ &= -\int_0^{1/2}\rho_2(x_2)\log\Big[\int_0^1\rho\Big(\frac{\lambda}{2},x_2\Big)d\lambda\Big]dx_2 - \int_{1/2}^1\rho_2(x_2)\log\Big[\int_0^1\rho\Big(\frac{\lambda+1}{2},x_2\Big)d\lambda\Big]dx_2\\ &= -\int_0^{1/2}\rho_2(x_2)\log\Big(2\int_0^{1/2}\rho(\xi,x_2)d\xi\Big)dx_2 - \int_{1/2}^1\rho_2(x_2)\log\Big(2\int_{1/2}^1\rho(\xi,x_2)d\xi\Big)dx_2\\ &= -\log 2 - \int_0^{1/2}\rho_2(x_2)\log\Big(\int_0^{1/2}\rho(\xi,x_2)d\xi\Big)dx_2 - \int_{1/2}^1\rho_2(x_2)\log\Big(\int_{1/2}^1\rho(\xi,x_2)d\xi\Big)dx_2,\end{aligned}$$

so

$$\begin{aligned}\Delta H_2 &= H_2(\tau+1)-H_2(\tau)\\ &= -\log 2 - \int_0^{1/2}\rho_2(x_2)\log\Big(\int_0^{1/2}\rho(\xi,x_2)d\xi\Big)dx_2 - \int_{1/2}^1\rho_2(x_2)\log\Big(\int_{1/2}^1\rho(\xi,x_2)d\xi\Big)dx_2\\ &\quad + \int_0^1\int_0^1\rho(x_1,x_2)\Big[\log\Big(\int_0^1\rho(\lambda,x_2)d\lambda\Big)\Big]dx_1dx_2\\ &= -\log 2 + (I+I\!I),\end{aligned}$$

where

$$I = \int_0^{1/2}\rho_2(x_2)\Big[\log\frac{\int_0^1\rho(\lambda,x_2)d\lambda}{\int_0^{1/2}\rho(\lambda,x_2)d\lambda}\Big]dx_2, \tag{29}$$

$$I\!I = \int_{1/2}^1\rho(x_2)\Big[\log\frac{\int_0^1\rho(\lambda,x_2)d\lambda}{\int_{1/2}^1\rho(\lambda,x_2)d\lambda}\Big]dx_2. \tag{30}$$

To compute $H_{2\wr}$, notice that, when x_1 is frozen, the transformation is invertible; moreover, the Jacobian $J_2 = 1/2$ is a constant. By Theorem III.3,

$$\Delta H_{2\wr} = \log\frac{1}{2} = -\log 2, \tag{31}$$

which gives

$$T_{1\to 2} = \Delta H_2 - \Delta H_{2\wr} = I + I\!I. \tag{32}$$

It is easy to show that $I + I\!I > 0$. In fact, obviously $I + I\!I$ is non-negative; besides, the two brackets cannot be zero simultaneously, so it cannot be zero. Hence $T_{1\to 2}$ is strictly positive; that is to say, there is always information flowing from the abscissa to the ordinate.

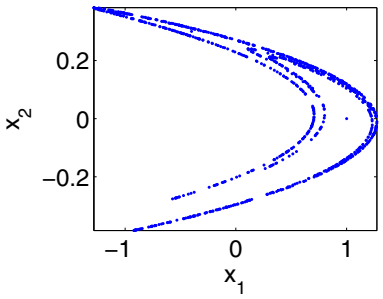

FIG. 3. A typical trajectory of the canonical Hénon map ($a = 1.4$, $b = 0.3$).

To summarize, $T_{2\to1} = 0$, $T_{1\to2} = I + II > 0$. These results are precisely the same as those obtained before in [46] and [48]. That is to say, for the baker transformation, the current formalism shows no difference from the previous one based on heuristic arguments and with approximations.

2. *Hénon map*

The Hénon map is a mapping $\Phi = (\Phi_1, \Phi_2) : \mathbb{R}^2 \mapsto \mathbb{R}^2$ defined such that

$$\begin{aligned} \Phi_1(x_1,x_2) &= 1 + x_2 - ax_1^2, \\ \Phi_2(x_1,x_2) &= bx_1, \end{aligned} \tag{33}$$

with $a > 0$, $b > 0$. The case with parameters $a = 1.4$ and $b = 0.3$ is called a "canonical Hénon map," whose attractor is shown in Fig. 3.

It is easy to see that the Hénon map is invertible; its inverse is

$$\Phi^{-1}(x_1,x_2) = \Big(\frac{x_2}{b},\, x_1 - 1 + \frac{a}{b^2}x_2^2\Big). \tag{34}$$

The F-P operator thus can be easily found from (11):

$$\begin{aligned} \mathscr{P}\rho(x_1,x_2) &= \rho(\Phi^{-1}(x_1,x_2))|J^{-1}| \\ &= \frac{1}{b}\rho\Big(\frac{x_2}{b},\, x_1 - 1 + \frac{a}{b^2}x_2^2\Big). \end{aligned} \tag{35}$$

In the following we compute the flows between the quadratic component x_1 and the linear component x_2.

Look at $T_{2\to1}$ first. By (16), we need to find the marginal density of x_1 at step $\tau + 1$ with and without the effect of x_2, i.e., $(\mathscr{P}\rho)_1$ and $(\mathscr{P}\rho)_{1\not2}$. From (35), $(\mathscr{P}\rho)_1$ is

$$\begin{aligned} (\mathscr{P}\rho)_{1(x_1)} &= \int_{\mathbb{R}} \mathscr{P}\rho(x_1,x_2)dx_2 \\ &= \int_{\mathbb{R}} \frac{1}{b}\rho\Big(\frac{x_2}{b}, x_1 - 1 + \frac{a}{b}x_2^2\Big)dx_2 \\ &= \int_{\mathbb{R}} \rho(\eta, x_1 - 1 + a\eta^2)d\eta \quad (x_2/b \equiv \eta). \end{aligned}$$

If $a = 0$, this would give $\rho_2(x_1 - 1)$, i.e., the marginal pdf of x_2 with argument $x_1 - 1$. But here $a > 0$, the integration is taken along a parabolic curve rather than a straight line. Still the final result will be related to the marginal density of x_2; for notational simplicity, write

$$(\mathscr{P}\rho)_{1(x_1)} = \tilde{\rho}_2(x_1). \tag{36}$$

To find $(\mathscr{P}_{\not2}\rho)_1$, use y_1 to denote

$$\Phi_1(x_1) = 1 + x_2 - ax_1^2,$$

following our convention to distinguish variables at different steps. Modify the system so that x_2 is now a parameter. As before, we need to find the counterimage of $(-\infty, y_1]$ under the transformation with x_2 frozen:

$$\begin{aligned} &\Phi_1^{-1}((-\infty,y_1]) \\ &\quad = (-\infty,\, -\sqrt{(1 + x_2 - y_1)/a}]\cup[\sqrt{(1 + x_2 - y_1)/a}, \infty). \end{aligned}$$

Therefore,

$$\begin{aligned} (\mathscr{P}_{\not2}\rho)_1(y_1) &= \frac{d}{dy_1}\int_{\Phi_1^{-1}((-\infty,y_1])} \rho_1(s)ds \\ &= \frac{d}{dy_1}\int_{-\infty}^{-\sqrt{(1+x_2-y_1)/a}} \rho_1(s)ds \\ &\quad + \frac{d}{dy_1}\int_{\sqrt{(1+x_2-y_1)/a}}^{\infty} \rho_1(s)ds \\ &= \frac{1}{2\sqrt{a(1 + x_2 - y_1)}}[\rho_1(-\sqrt{(1 + x_2 - y_1)/a}) \\ &\quad + \rho_1(\sqrt{(1 + x_2 - y_1)/a})] \quad (y_1 < 1 + x_2) \\ &= \frac{1}{2a|x_1|}[\rho_1(-x_1) + \rho_1(x_1)] \\ &\quad \text{(recall } y_1 = 1 + x_2 - ax_p 1^2\text{)}. \end{aligned}$$

Denote the average of $\rho_1(-x_1)$ and $\rho_1(x_1)$ as $\bar{\rho}_1(x_1)$ to make an even function of x_1. Then $(\mathscr{P}_{\not2}\rho)_1$ is simply

$$(\mathscr{P}_{\not2}\rho)_1(y_1) = \frac{\bar{\rho}_1(x_1)}{a|x_1|}. \tag{37}$$

Note that the parameter x_2 does not appear in the arguments. Substitute all the above into (16) to get

$$\begin{aligned} T_{2\to1} &= E\log(\mathscr{P}_{\not2}\rho)_1(y_1) - E\log(\mathscr{P}\rho)_1(y_1) \\ &= E\log\frac{\bar{\rho}_1(x_1)}{a|x_1|} - E\log\tilde{\rho}_2\big(1 + x_2 - ax_1^2\big) \\ &= E\log\bar{\rho}_1(x_1) - E\log|ax_1| - E\log\tilde{\rho}_2\big(1 + x_2 - ax_1^2\big). \end{aligned}$$

Comparing this to the result in [48], except for the term $-E\log|ax_1|$, all other terms are different.

Next consider $T_{1\to2}$. From (35), the marginal density of x_2 at $\tau + 1$ is

$$\begin{aligned} (\mathscr{P}\rho)_2(x_2) &= \int_{\mathbb{R}} \mathscr{P}\rho(x_1,x_2)dx_1 \\ &= \int_{\mathbb{R}} \frac{1}{b}\rho\Big(\frac{x_2}{b},\, x_1 - 1 + a\frac{x_2^2}{b^2}\Big)dx_1 \\ &= \frac{1}{b}\int_{\mathbb{R}} \rho(y,\xi)\,d\xi = \frac{1}{b}\rho_1\Big(\frac{x_2}{b}\Big). \end{aligned}$$

Thus

$$\begin{aligned} H_2 &= -E(\mathscr{P}\rho)_2(y_2) \\ &= -\int_{\mathbb{R}} \frac{1}{b}\rho_1\Big(\frac{x_2}{b}\Big)\log\Big[\frac{1}{b}\rho_1\Big(\frac{x_2}{b}\Big)\Big]dx_2 \\ &= H_1 + \log b. \end{aligned}$$

The evaluation of $H_{2\backslash}$ is much easier. As x_1 is frozen as a parameter, y_2 becomes definite. In this case, the 2D random variable is degenerated to a 1D variable. Correspondingly $\mathscr{P}_{\backslash}\rho$ becomes a pdf in x_1 only. So

$$(\mathscr{P}_{\backslash}\rho)_2 = \int_{\mathbb{R}} \mathscr{P}_{\backslash}\rho dx_1 = 1.$$

Thus $H_{2\backslash} = -E\log(\mathscr{P}_{\backslash}\rho)_2 = 0$. By (16), the information flow from x_1 to x_2 is, therefore,

$$T_{1\to 2} = H_2 - H_{2\backslash} = H_1 + \log b. \tag{38}$$

In other words, the flow from x_1 to x_2 is equal to the marginal entropy of x_1, modified by an amount related to the factor b. Particularly, when $b = 1$, $T_{1\to 2} = H_1$. This is precisely the same as what is obtained before in [48].

In summary, the information flows within the baker transformation are precisely the same as we have obtained before in [48]. For the Hénon map, the flow from x_1 to x_2 has recovered the benchmark result based on physical grounds, i.e., (38). But $T_{2\to 1}$ is generally different from that in [48].

IV. CONTINUOUS-TIME DETERMINISTIC SYSTEMS

A. Deriving the information flow

Now look at the information flow within continuous systems, the 2D version of which have motivated this line of work:

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t;\mathbf{x}). \tag{39}$$

Consider a time interval $[t, t+\Delta t]$. Following [49], we discretize the ordinary differential equation and construct a mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{x}(t) \mapsto \mathbf{x}(t+\Delta t) = \mathbf{x} + \mathbf{F}\Delta t$. Correspondingly there is a Frobenius-Perron operator $\mathscr{P} : L^1(\mathbb{R}^n) \to L^1(\mathbb{R}^n)$, $\rho(t) \mapsto \rho(t+\Delta t)$. Write $\mathbf{x}(t+\Delta t)$ as $\mathbf{y}$, a convention we have been using all the time to avoid confusion. Then the mapping $\Phi : \mathbf{x} \mapsto \mathbf{y}$ is such that

$$\begin{aligned} y_1 &= x_1 + F_1(x_1, x_2, \ldots, x_n)\Delta t, \\ y_2 &= x_2 + F_2(x_1, x_2, \ldots, x_n)\Delta t, \\ &\vdots \qquad\qquad \vdots \\ y_n &= x_n + F_n(x_1, x_2, \ldots, x_n)\Delta t. \end{aligned} \tag{40}$$

Its Jacobian is

$$\begin{aligned} J = \det\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right) &= \det\begin{bmatrix} 1 + \frac{\partial F_1}{\partial x_1}\Delta t & \cdots & \frac{\partial F_1}{\partial x_n}\Delta t \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_n}\Delta t & \cdots & 1 + \frac{\partial F_n}{\partial x_n}\Delta t \end{bmatrix} \\ &= 1 + \sum_i \frac{\partial F_i}{\partial x_i}\Delta t + o(\Delta t) \\ &= 1 + \nabla\cdot\mathbf{F}\Delta t + o(\Delta t). \end{aligned} \tag{41}$$

As $\Delta t \to 0$, $J \to 1 \neq 0$, so Φ thus constructed is always invertible for Δt small enough. Moreover, it is easy to obtain the inverse mapping

$$\Phi^{-1} : \mathbf{x} = \mathbf{y} - \mathbf{F}\Delta t + o(\Delta t) \tag{42}$$

and $J^{-1} = 1 - \nabla\cdot\mathbf{F}\Delta t + o(\Delta t)$. So

$$\begin{aligned} \mathscr{P}\rho(\mathbf{y}) &= \rho(\Phi^{-1}(\mathbf{y}))\cdot|J^{-1}| \\ &= \rho(\mathbf{y} - \mathbf{F}\Delta t)\cdot(1 - \nabla\cdot\mathbf{F}\Delta t) + o(\Delta t) \\ &= \rho(\mathbf{y}) - \nabla\rho\cdot\mathbf{F}\Delta t - \rho\nabla\cdot\mathbf{F}\Delta t + o(\Delta t) \\ &= \rho(\mathbf{y}) - \nabla\cdot(\rho\mathbf{F})\Delta t + o(\Delta t). \end{aligned}$$

As a verification, check

$$\frac{\partial\rho}{\partial t} = \lim_{\Delta t\to 0}\frac{\mathscr{P}\rho(\mathbf{x}) - \rho(\mathbf{x})}{\Delta t} = -\nabla\cdot(\rho\mathbf{F}).$$

This yields the Liouville equation $\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{F}) = 0$, as expected.

With $\mathscr{P}\rho$ we now can compute the marginal density

$$(\mathscr{P})_1(y_1) = \rho_1(y_1) - \Delta t\int_{R^{n-1}}\frac{\partial\rho F_1}{\partial y_1}dy_2\ldots dy_n + o(\Delta t)$$

which gives

$$\begin{aligned} -\log(\mathscr{P}\rho)_1(y_1) &= -\log\rho_1(y_1) - \left[\log\left(1 - \frac{\Delta t}{\rho_1}\int_{\mathbb{R}^{n-1}}\frac{\partial\rho F_1}{\partial y_1}dy_2\ldots dy_n + o(\Delta t)\right)\right] \\ &= -\log\rho_1(y_1) + \frac{\Delta t}{\rho_1(y_1)}\int_{\mathbb{R}^{n-1}}\frac{\partial\rho F_1}{\partial y_1}dy_2\ldots dy_n + o(\Delta t) \\ &= -\log\rho_1(x_1 + F_1\Delta t) + \frac{\Delta t}{\rho_1(x_1)}\int_{\mathbb{R}^{n-1}}\frac{\partial\rho F_1}{\partial x_1}dx_2\ldots dx_n + o(\Delta t). \end{aligned}$$

At the last step, the y's have been replaced by x's in the integral term. This is legitimate since the difference goes to higher order terms.

We now evaluate the marginal entropy increase at $t + \Delta t$. The following is the key step: Take expectation on both sides, the left hand side with respect to $(\mathscr{P}\rho)_1(y_1)$, while the right hand side with respect to $\rho_1(x_1)$. This yields

$$\begin{aligned} H_1(t+\Delta t) &= -E\log\rho_1(x_1 + F_1\Delta t) + \Delta t E\left(\frac{1}{\rho_1}\int_{\mathbb{R}^{n-1}}\frac{\partial\rho F_1}{\partial x_1}dx_2\ldots dx_n\right) + o(\Delta t) \\ &= H_1(t) - E\frac{\partial\log\rho_1}{\partial x_1}F_1\Delta t + \Delta t\int_{\mathbb{R}}\rho_1\frac{1}{\rho_1}dx_1\int_{\mathbb{R}^{n-1}}\frac{\partial\rho F_1}{\partial x_1}dx_2\ldots dx_n + o(\Delta t). \end{aligned}$$

Note the third term on the right hand side vanishes after integration with respect to x_1 due to the compactness of the functions. So

$$H_1(t+\Delta t) = H_1(t) - \Delta t E\left(F_1 \frac{\partial \log \rho_1}{\partial x_1}\right) + o(\Delta t),$$

and hence

$$\frac{dH_1}{dt} = \lim_{\Delta t\to 0} \frac{H_1(t+\Delta t)-H_1(t)}{\Delta t} = -E\left(F_1 \frac{\partial \log \rho_1}{\partial x_1}\right). \quad (43)$$

This is precisely the same as that either from the F-P operator [48] or directly from the Liouville equation [46], serving a validation of our approach in this study.

When x_2 is frozen as a parameter on $[t,t+\Delta t]$, we need to examine the modified mapping $\Phi_{\natural}: \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$:

$$\begin{aligned} y_1 &= x_1 + F_1(x_1,x_2,\ldots,x_n)\Delta t,\\ y_3 &= x_3 + F_3(x_1,x_2,\ldots,x_n)\Delta t,\\ &\vdots \qquad\qquad \vdots\\ y_n &= x_n + F_n(x_1,x_2,\ldots,x_n)\Delta t, \end{aligned} \quad (44)$$

i.e., the mapping Φ with the equation $y_2 = x_2 + F_2\Delta t$ removed, and x_2 frozen as a parameter. Again, here x_i stands for $x_i(t)$, and y_i for $x_i(t+\Delta t)$. For convenience, we further adopt the following notations:

$$\begin{aligned} \mathbf{y}_{\natural} &= (y_1,y_3,\ldots,y_n)^T,\\ \mathbf{x}_{\natural} &= (x_1,x_3,\ldots,x_n)^T,\\ \mathbf{F}_{\natural} &= (F_1,F_3,\ldots,F_n)^T. \end{aligned}$$

Besides, use $\rho_{\natural}$ to signify the joint density of $\mathbf{x}_{\natural}$, and $\rho_{1\natural}$ to denote the density of x_1 with x_2 frozen as a parameter on $[t,t+\Delta t]$. Notice the fact that $\rho_{1\natural} = \rho_1$ at time t.

It is easy to know that the Jacobian of $\Phi_{\natural}$

$$J_{\natural} = \det\left(\frac{\mathbf{y}_{\natural}}{\mathbf{x}_{\natural}}\right) = 1 + \Delta t \sum_{i\neq 2} \frac{\partial F_i}{\partial x_i} + o(\Delta t). \quad (45)$$

The corresponding F-P operator $\mathscr{P}_{\natural}: L^1(\mathbb{R}^{n-1}) \to L^1(\mathbb{R}^{n-1})$ is such that

$$\begin{aligned} \mathscr{P}_{\natural}\rho_{\natural}(\mathbf{y}) &= \rho_{\natural}(\Phi^{-1}{}_{\natural}(\mathbf{y}))\cdot\left|J_{\natural}^{-1}\right|\\ &= \rho_{\natural}(\mathbf{y}_{\natural} - \mathbf{F}_{\natural}\Delta t)\cdot\left(1 - \sum_{i\neq 2}\frac{\partial F_i}{\partial x_i}\Delta t\right) + o(\Delta t)\\ &= \rho_{\natural}(\mathbf{y}_{\natural}) - \nabla\cdot(\rho_{\natural}\mathbf{F}_{\natural})\Delta t + o(\Delta t). \end{aligned}$$

Integrate with respect to $(y_3,\ldots,y_n)$ (recall that x_2 is now a parameter) to get

$$\begin{aligned} &(\mathscr{P}_{\natural}\rho_{\natural})_1(y_1)\\ &\quad = \rho_{1\natural}(y_1) - \Delta t\int_{\mathbb{R}^{n-2}} \frac{\partial \rho_{\natural}F_1}{\partial y_1} dy_3\ldots dy_n + o(\Delta t), \end{aligned} \quad (46)$$

where other terms vanish due to the compactness assumed for the functions. Hence

$$\begin{aligned} &-\log(\mathscr{P}_{\natural}\rho_{\natural})_1(y_1)\\ &= -\log\rho_{1\natural}(y_1) - \log\left(1 - \frac{\Delta t}{\rho_{1\natural}(y_1)}\int \frac{\partial \rho_{\natural}F_1}{\partial y_1} dy_3\ldots dy_n\right)\\ &\quad + o(\Delta t)\\ &= -\log\rho_{1\natural}(x_1+F_1\Delta t) + \frac{\Delta t}{\rho_{1\natural}(x_1)}\int \frac{\partial \rho_{\natural}F_1}{\partial x_1} dx_3\ldots dx_n\\ &\quad + o(\Delta t). \end{aligned}$$

Note that in the integral term, the y's have been replaced by x's; this is legitimate as the difference goes to the higher order terms. Since $\rho_{1\natural}(x_1) = \rho_1(x_1)$ at t, so

$$\rho_{1\natural}(x_1+F_1\Delta t) = \rho_1(x_1) + \frac{\partial \rho_1}{\partial x_1}F_1\Delta t + o(\Delta t)$$

and hence

$$\begin{aligned} -\log(\mathscr{P}_{\natural}\rho_{\natural})_1(y_1) &= -\log\rho_1 - \frac{1}{\rho_1}\frac{\partial \rho_1}{\partial x_1}F_1\Delta t\\ &\quad + \frac{\Delta t}{\rho_1(x_1)}\int \frac{\partial \rho_{\natural}F_1}{\partial x_1} dx_3\ldots dx_n + o(\Delta t). \end{aligned}$$

Take expectation on both sides, the left hand side with respect to the joint probability density of (y_1,x_2), while the right hand side with respect to (x_1,x_2). This is the key step that makes the present study fundamentally different from [49] which relies on an approximation to fulfill the derivation. This yields

$$\begin{aligned} H_{1\natural}(t+\Delta t) &= H_t(t) - \Delta t E\left(F_1\frac{\partial \log\rho_1}{\partial x_1}\right)\\ &\quad + \Delta t\int_{\mathbb{R}^2}\frac{\rho_{12}(x_1,x_2)}{\rho_1(x_1)}dx_1dx_2\\ &\quad \times \int_{\mathbb{R}^{n-2}}\frac{\partial \rho_{\natural}F_1}{\partial x_1}dx_3\ldots dx_n + o(\Delta t)\\ &= H_t(t) - \Delta t E\left(F_1\frac{\partial \log\rho_1}{\partial x_1}\right)\\ &\quad + \Delta t\int_{\mathbb{R}}\rho_{2|1}\frac{\partial \rho_{\natural}F_1}{\partial x_1}d\mathbf{x} + o(\Delta t), \end{aligned}$$

where $\rho_{2|1}$ is the conditional density of x_2 on x_1. Thus

$$\frac{dH_{1\natural}}{dt} = -E\left(F_1\frac{\partial \log\rho_1}{\partial x_1}\right) + \int_{\mathbb{R}^n}\rho_{2|1}\frac{\partial \rho_{\natural}F_1}{\partial x_1}d\mathbf{x}. \quad (47)$$

We therefore arrive at the following theorem:

Theorem IV.1.

$$\begin{aligned} T_{2\to 1} &= \frac{dH_1}{dt} - \frac{dH_{1\natural}}{dt} = -\int_{\mathbb{R}^n}\rho_{2|1}\frac{\partial \rho_{\natural}F_1}{\partial x_1}d\mathbf{x}\\ &= -E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\natural}}{\partial x_1}dx_3\ldots dx_n\right]. \end{aligned} \quad (48)$$

An alternative derivation of the theorem is deferred to the Appendix.

B. Properties

Theorem IV.2. For a 2D system

$$\begin{aligned} \frac{dx_1}{dt} &= F_1(x_1,x_2,t),\\ \frac{dx_2}{dt} &= F_2(x_1,x_2,t), \end{aligned}$$

we have

$$\frac{dH_{1\mathcal{Y}}}{dt} = E\left(\frac{\partial F_1}{\partial x_1}\right). \tag{49}$$

Remark. This recovers Eq. (5), the key equation originally obtained by Liang and Kleeman [46] through heuristic argument. Here we rigorously prove it.

Proof. When $n = 2$, $\rho_{\mathcal{Y}} = \rho_1$, hence

$$\begin{aligned}\frac{dH_{1\mathcal{Y}}}{dt} &= E\left(\frac{1}{\rho_1}\frac{\partial F_1\rho_1}{\partial x_1}\right) - E\left(\frac{\partial \log\rho_1}{\partial x_1}F_1\right)\\ &= E\left[\frac{\partial F_1}{\partial x_1} + F_1\frac{\partial\rho_1}{\partial x_1}\frac{1}{\rho_1} - F_1\frac{\partial\log\rho_1}{\partial x_1}\right] = E\left(\frac{\partial F_1}{\partial x_1}\right).\end{aligned}$$

■

Theorem IV.3 (principle of nil causality). For the system (39) if F_1 is independent of x_2, then $T_{2\to1} = 0$.

Proof. If F_1 has no dependence on x_2, so is $F_1\rho_{\mathcal{Y}}$. Thus

$$\begin{aligned}T_{2\to1} &= -E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\mathcal{Y}}}{\partial x_1}dx_3\ldots dx_n\right]\\ &= -\int_{\mathbb{R}^n}\rho(x_2|x_1)\frac{\partial F_1\rho_{\mathcal{Y}}}{\partial x_1}d\mathbf{x}\\ &= -\int_{\mathbb{R}^{n-1}}\frac{\partial F_1\rho_{\mathcal{Y}}}{\partial x_1}dx_1dx_3\ldots dx_n = 0,\end{aligned}$$

where the fact $\int\rho(x_2|x_1)dx_2 = 1$ and the assumption of compact support have been used. ■

C. Application: Rössler system

In this subsection, we present an application study of the information flows within the Rössler system:

$$\frac{dx}{dt} = F_x = -y - z, \tag{50}$$

$$\frac{dy}{dt} = F_y = x + ay, \tag{51}$$

$$\frac{dz}{dt} = F_z = b + z(x - c), \tag{52}$$

where a, b, and c are parameters. Rössler finds a chaotic attractor for $a = 0.2$, $b = 0.2$, $c = 5.7$ [52], as shown in Fig. 4. From the figure the trajectories are limited within $[-12,12]\times[-14,10]\times[0,25]$.

To calculate the information flows, one needs to obtain the joint probability density function $\rho(x_1,x_2,x_3)$. It is, of course, obtainable through solving the Liouville equation

$$\frac{\partial\rho}{\partial t} + \frac{\partial F_x\rho}{\partial x} + \frac{\partial F_y\rho}{\partial y} + \frac{\partial F_z\rho}{\partial z} = 0$$

with some initial condition ρ_0. However, there is another way, namely, ensemble forecast, which is more efficient in terms of

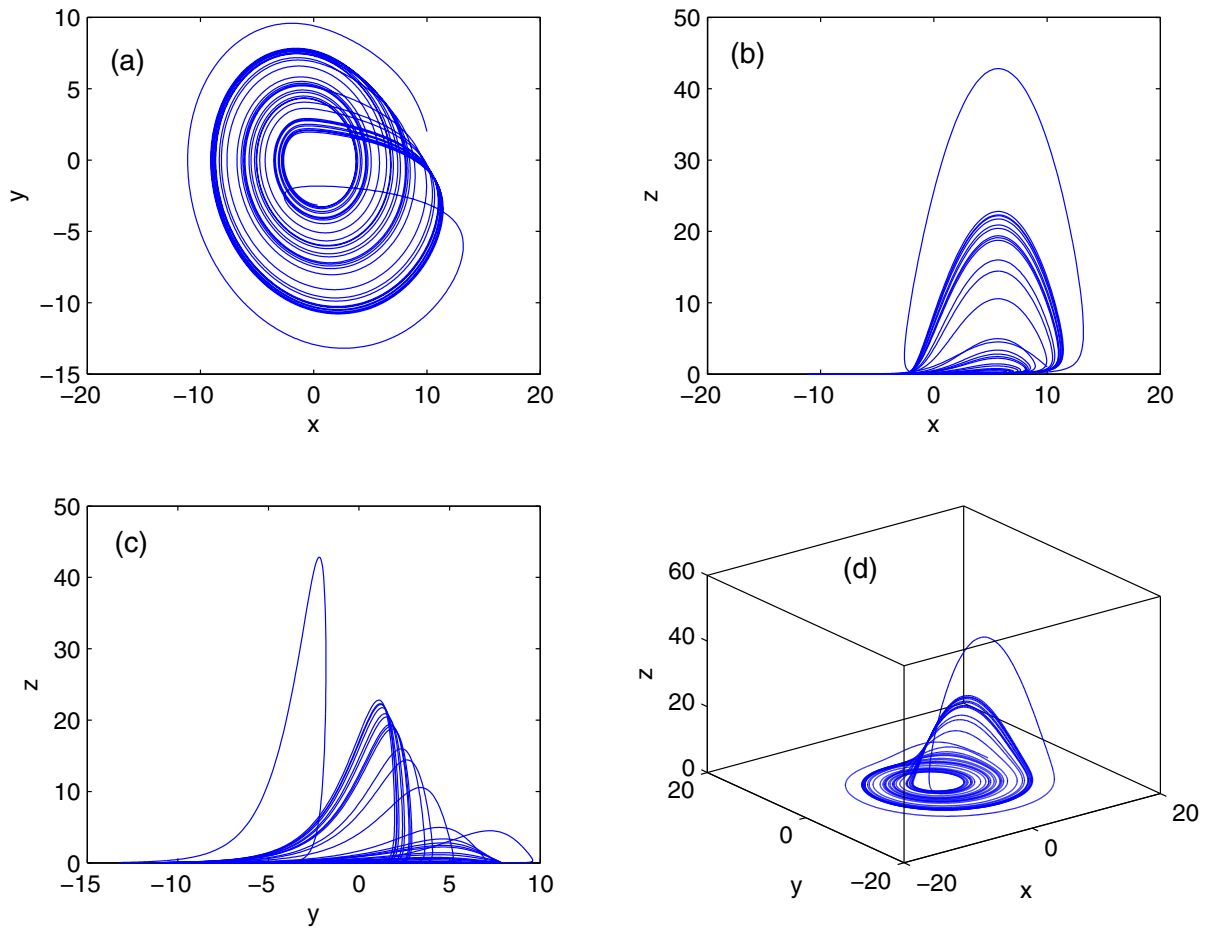

FIG. 4. The Rössler attractor.

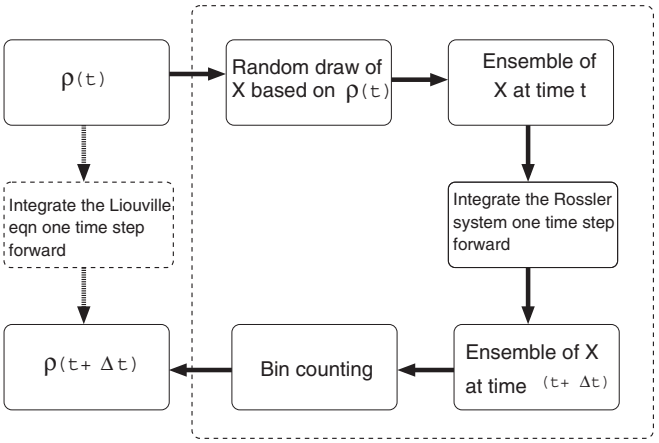

FIG. 5. A schematic of ensemble prediction. Instead of solving the Liouville equation for the density ρ, we make random draws according to the initial distribution $\rho(t_0)$ to form an ensemble, then let the Rössler system steer forth each member of the ensemble. At each time step, bins are counted and the probability density function is accordingly estimated.

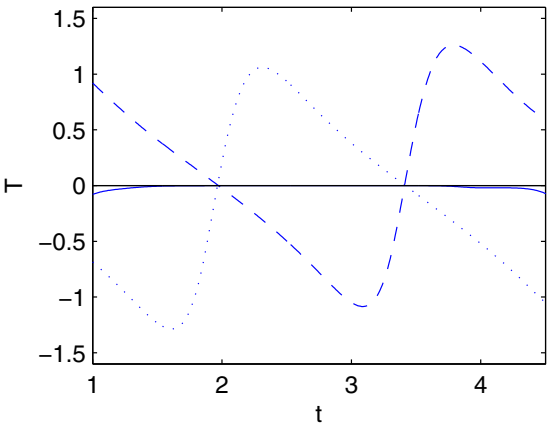

FIG. 6. The time series of the information flow rates within the Rössler system (in nats per unit time). Dashed: $T_{y\to x}$; dotted: $T_{x\to y}$; solid: $T_{z\to x}$. Other flows are essentially zero in this duration. The initial segments are not shown as some trajectories are still outside the attractor.

computational load. As illustrated in Fig. 5, instead of solving the Liouville equation, we solve the Rössler systems initialized with an ensemble of initial values of $\mathbf{x}$. This ensemble is formed with entries randomly drawn according to the initial pdf ρ_0. At each time step, we count the bins thus obtained and estimate the pdf. The resulting pdf is the desired ρ.

The Rössler system [Eqs. (50) and (51)] is solved using the second order Runge-Kutta method with a time step size $\Delta t = 0.01$. A typical computed trajectory is plotted in Fig. 4. The initial conditions are randomly drawn according to a Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the mean vector and covariance matrix being, respectively,

$$\boldsymbol{\mu} = \begin{bmatrix} 8 \\ 2 \\ 10 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

The initial mean values are chosen rather randomly (in reference to Fig. 4); μ_x is chosen large to make $\frac{dH}{dt} = E(\nabla \cdot \mathbf{F}) = \mu_x - 5.5$ positive.

Pick a computation domain $\Omega \equiv [-16,16] \times [-18,14] \times [-4,28]$, which clearly covers the attractor. We discretize it into $320 \times 320 \times 320 = 32\,768\,000$ bins with $\Delta x = \Delta y = \Delta z = 0.1$. To ensure one draw for each bin on average, in the beginning we make 32 768 000 random draws. As the ensemble scheme is carried forth, ρ and all other statistics can be estimated as a function of time. By Theorem IV.1 the information flow rates are computed accordingly.

For a system with three components (x,y,z), there are in total 6 flow pairs: $T_{x\to y}$, $T_{y\to z}$, $T_{z\to x}$, $T_{x\to z}$, $T_{y\to x}$, $T_{x\to z}$. A first examination of the system tells us that dy/dt does not depend on z and dz/dt does not depend on y. By the principle of nil causality (Theorem IV.3), $T_{z\to y}$ and $T_{y\to z}$ must vanish. The computational results reconfirm this. In Fig. 6, the two are essentially zero. What makes the results surprisingly interesting is that $T_{x\to z}$ is also insignificant (more than one order smaller in comparison to $T_{y\to x}$ and $T_{x\to y}$), while the dependence of dz/dt on x is explicitly specified. Besides, $T_{z\to x}$ is also small in certain periods, e.g., in the interval [1,4.5] as shown. In the figure are essentially the flows between x and y: $T_{y\to x}$ and $T_{x\to y}$.

The above information flow scenario motivates us to check the system, for the period as shown, with only x and y two components. This is an amplifying harmonic oscillator $\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}$ where $\mathbf{A} = \begin{bmatrix} 0 & -1 \\ 1 & a \end{bmatrix}$, a linear system allowing the information flow, say, from y to x, to be simply expressed as $T_{y\to x} = a_{12}\frac{\sigma_{12}}{\sigma_{11}}$ (see below in Sec. VII). That is to say, here the covariance matrix $\Sigma = (\sigma_{ij})$ completely determines the flow. The evolution of Σ follows

$$\frac{d\Sigma}{dt} = \mathbf{A}\Sigma + \Sigma\mathbf{A}^T.$$

Initialized by $\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$, σ_{ij} can be easily computed; the resulting $T_{y\to x}$ and $T_{x\to y}$ are shown in Fig. 7. Comparing to those in

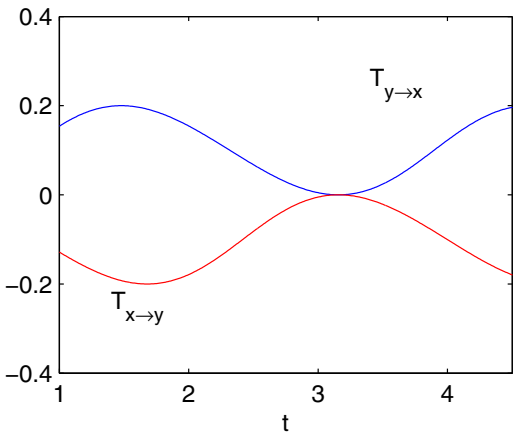

FIG. 7. The time series of the information flow rates within the amplifying harmonic system as shown in the text (in nats per unit time).

Fig. 6, the general trend, including the period, seems to be similar, though the geometry of the curves has been modified from harmonic into a seesaw one. Besides, the $T_{x\to y}$ ($T_{y\to x}$) is always negative (positive) for the harmonic oscillator, while for the Rössler system, they can be both negative and positive. Note the parameter a in $\mathbf{A}$ does not explicitly appear in the formula, but it does contributes to the generation of the information flow. One may easily check that, if it is zero, then $\frac{d\Sigma}{dt} = 0$, and hence the flow rates will stay zero if originally $\sigma_{12} = 0$.

The above example is just used for the demonstration of application and, in some cases, for the validation of the proven theorems such as the principle of nil causality. The seemingly vanishing $T_{x\to z}$ in spite of the dependence of dz/dx on x certainly deserves further investigation but is beyond the scope of this study. Here we just want to mention that this does conform to the observations with complex systems; emergence does not result from rules only (e.g., [53–55]). It has long been found that regular patterns may emerge out of irregular motions with some simple preset rules; a good example is the 2D turbulent flow in natural world (e.g., [56]). Clearly, these simple, rudimentary rules are not enough for explaining the causal efficacy and the bottom-up flow of information that leads to the emergence of the organized structure. As commented on by Corning [57], "Rules, or laws, have no causal efficacy; they do not in fact generate anything ... the underlying causal agencies must be separately specified." We shall see a more remarkable example in the following subsection.

D. Application: The truncated Burgers-Hopf system revisited

Here we reexamine the truncated Burgers-Hopf system (TBS hereafter), a chaotic system which seemingly has rather simple information flow structures as shown in the studies of Liang and Kleeman [49]. For a detailed description of the system itself, see [58]. In this section we only examine the following particular case:

$$\frac{dx_1}{dt} = F_1(\mathbf{x}) = x_1x_4 - x_3x_2, \tag{53}$$

$$\frac{dx_2}{dt} = F_2(\mathbf{x}) = -x_1x_3 - x_2x_4, \tag{54}$$

$$\frac{dx_3}{dt} = F_3(\mathbf{x}) = 2x_1x_2, \tag{55}$$

$$\frac{dx_4}{dt} = F_4(\mathbf{x}) = -x_1^2 + x_2^2. \tag{56}$$

As we have described before, the system is intrinsically chaotic, with a strange attractor embedded in

$$[-24.8,24.6] \times [-25.0,24.5] \times [-22.3,21.9] \times [-23.7,23.7].$$

The information flow within the TBS cannot be found analytically.

As before, we use the ensemble prediction technique to estimate the density evolution, and then evaluate the T's. The setting and procedure are made precisely the same as that in [49] in order to facilitate a comparison. Details are referred to the original paper and will not be presented here.

Figure 8 plots the results for the case with a Gaussian initial distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix},$$

with $\mu_i = 9$, $\sigma_i^2 = 9$, for $i = 1,2,3,4$.

Shown specifically are the time rates of the 12 information flows:

$$\begin{array}{lll} T_{2\to1}, & T_{3\to1}, & T_{4\to1}; \\ T_{1\to2}, & T_{3\to2}, & T_{4\to2}; \\ T_{1\to3}, & T_{2\to3}, & T_{4\to3}; \\ T_{1\to4}, & T_{2\to4}, & T_{3\to4}. \end{array}$$

The results are qualitatively the same as before in [49]. That is to say, except for $T_{3\to2}$, which is distinctly different from zero, all others are either negligible, or oscillatory around zero. But, of course, the present flows are much smaller in magnitude, in comparison to the one obtained before in [49] using the approximate formula.

V. STOCHASTIC MAPPING

A. Derivation

Consider the system

$$\mathbf{x}(\tau + 1) = \Phi(\mathbf{x}(\tau)) + \mathbf{B}(\mathbf{x})\mathbf{w}, \tag{57}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ is an n-dimensional mapping, $\mathbf{B}$ is an $n \times m$ matrix, and $\mathbf{w}$ an m-dimensional normally distributed random vector, representing an m-dimensional standard Wiener process. Without loss of generality, we assume the covariance matrix of $\mathbf{w}$, $\Sigma = \mathbf{I}$, since the perturbation amplitude can be put into $\mathbf{B}$.

In general, $\mathbf{B}$ may depend on $\mathbf{x}$. But this complicates the derivation a great deal. For simplicity, in this section we only consider the case when $\mathbf{B}$ is a constant $n \times m$ matrix. As $\mathbf{x}(\tau)$ is taken to $\mathbf{x}(\tau + 1)$, there exists an operator, written $\mathscr{P} : L^1(\mathbb{R}^n) \to L^1(\mathbb{R}^n)$, steering the pdf at time step τ, ρ, to the pdf at step $\tau + 1$, $\mathscr{P}\rho$. Since $\mathbf{x}(\tau)$ and $\mathbf{w}$ are independent, if $\mathbf{B}$ is a constant matrix, one may view $\mathbf{x}(\tau + 1)$ as the sum of two independent random variables, and then conjecture that $\mathscr{P}\rho$ be the convolution of $\mathscr{P}_\Phi\rho$ and some joint Gaussian distribution. Here $\mathscr{P}_\Phi$ stands for the F-P operator associated with the mapping Φ. This is indeed true, as is stated in the following theorem.

Theorem V.1.

$$\mathscr{P}\rho(\mathbf{y}) = \int_{\mathbb{R}^n} \mathscr{P}_\Phi\rho(\mathbf{y} - \mathbf{B}\mathbf{w}) \cdot \rho_w(\mathbf{w})d\mathbf{w}, \tag{58}$$

where

$$\rho_w(\mathbf{w}) = (2\pi)^{-m/2}(\det \Sigma)^{-1/2}e^{-\frac{1}{2}\mathbf{w}^T\Sigma^{-1}\mathbf{w}}.$$

Proof. We first assume that Φ is invertible to make the approach more transparent to the reader. As always, write $\mathbf{x}(\tau + 1)$ as $\mathbf{y}$ to avoid confusion. Make a transformation:

$$\Pi : \begin{cases} \mathbf{y} = \Phi(\mathbf{x}) + \mathbf{B}\mathbf{w}, \\ \mathbf{z} = \mathbf{w}. \end{cases} \tag{59}$$

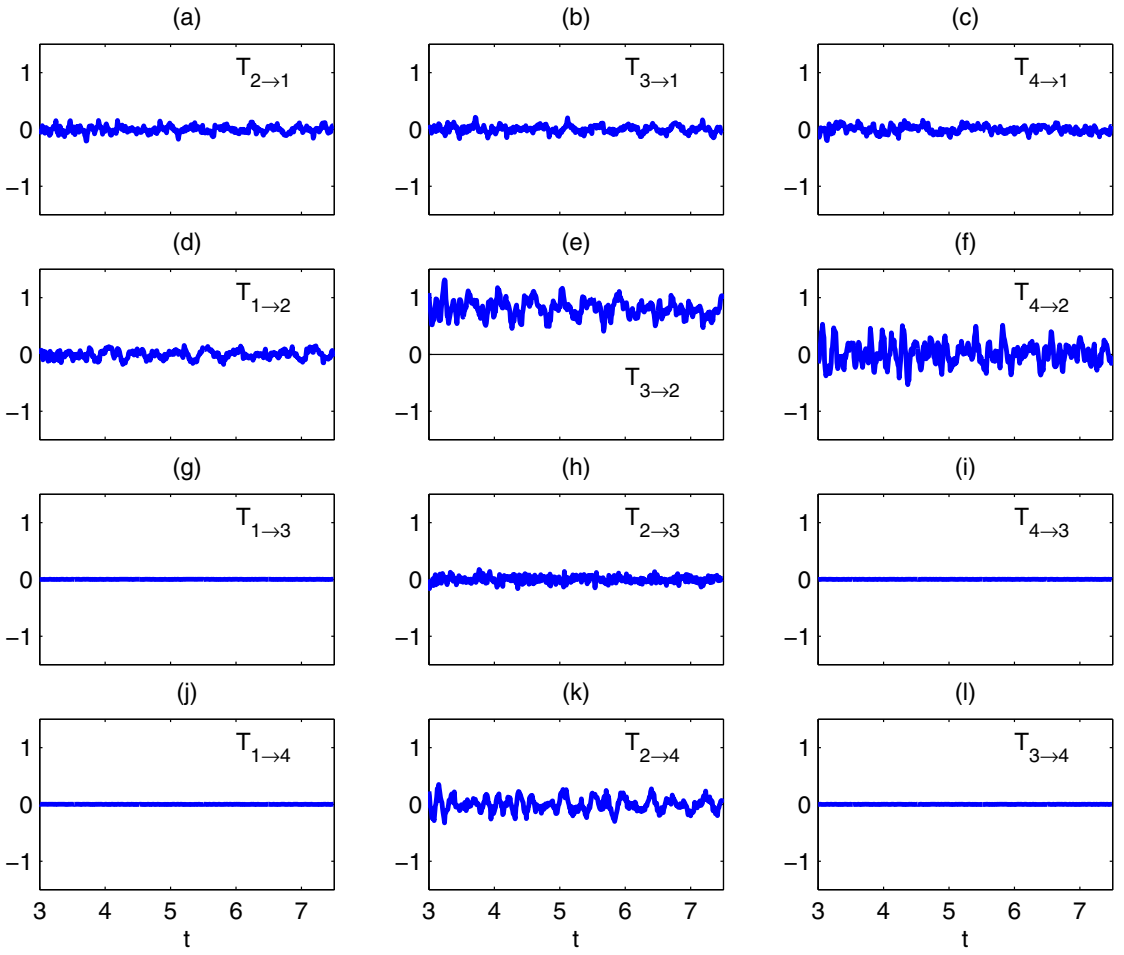

FIG. 8. Information flows between the components of the 4D truncated Burgers-Hopf system in the invariant chaotic attractor.

Its Jacobian

$$J_\pi = \det\left[\frac{\partial(\mathbf{y},\mathbf{z})}{\partial(\mathbf{x},\mathbf{w})}\right] = \det\begin{bmatrix}\frac{\partial\Phi}{\partial x} & \mathbf{B}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} = \det\left(\frac{\partial\Phi}{\partial\mathbf{x}}\right) = J_\Phi \equiv J. \tag{60}$$

The inverse mapping is

$$\Pi^{-1}: \begin{cases}\mathbf{x} = \Phi^{-1}(\mathbf{y} - \mathbf{B}\mathbf{z}),\\ \mathbf{w} = \mathbf{z}.\end{cases} \tag{61}$$

For any $S_y \in \mathbb{R}^n$, $S_z \in \mathbb{R}^m$,

$$\int_{S_y\times S_z}\rho_{yz}(\mathbf{y},\mathbf{z})d\mathbf{y}d\mathbf{z} = \int_{\Pi^{-1}(S_y\times S_z)}\rho_{xw}(\mathbf{x},\mathbf{w})d\mathbf{x}d\mathbf{w} = \int_{S_y\times S_z}\rho_{xw}(\Pi^{-1}(\mathbf{y},\mathbf{z}))\cdot\left|J_\pi^{-1}\right|d\mathbf{y}d\mathbf{z}.$$

So

$$\begin{aligned}\rho_{yz}(\mathbf{y},\mathbf{z}) &= \rho_{xw}(\Pi^{-1}(\mathbf{y},\mathbf{z}))\cdot\left|J_\pi^{-1}\right|\\ &= \rho_{xw}(\Phi^{-1}(\mathbf{y}-\mathbf{B}\mathbf{z}),\mathbf{z})\cdot|J^{-1}|\\ &= \rho(\Phi^{-1}(\mathbf{y}-\mathbf{B}\mathbf{z}))\cdot|J^{-1}|\cdot\rho_w(\mathbf{z}),\end{aligned}$$

where the independence between $\mathbf{x}$ and $\mathbf{w}$ has been used (hence $\rho_{xw} = \rho_x\rho_w$). $\mathscr{P}(\mathbf{y}) = \rho_y(\mathbf{y})$ is thence the marginal density by integrating out $\mathbf{z}$:

$$\mathscr{P}\rho(\mathbf{y}) = \int_{\mathbb{R}^n}\rho(\Phi^{-1}(\mathbf{y}-\mathbf{B}\mathbf{z}))\cdot|J^{-1}|\cdot\rho_w(\mathbf{z})d\mathbf{z}.$$

Since $\rho(\Phi^{-1}(\mathbf{y}))\cdot|J^{-1}| = \mathscr{P}_\Phi\rho(\mathbf{y})$, the theorem thus follows.

When Φ is singular or noninvertible, let its F-P operator be $\mathscr{P}_\Phi$, then $\forall S_y \in \mathbb{R}^n$, $S_z \in \mathbb{R}^m$,

$$\begin{aligned}\int_{S_y\times S_z}\rho_{yz}(\mathbf{y},\mathbf{z})d\mathbf{y}d\mathbf{z} &= \int_{\Pi^{-1}(S_y\times S_z)}\rho_{xw}(\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{z}\\ &= \int_{\Pi^{-1}(S_y\times S_z)}\rho(\mathbf{x})\cdot\rho_w(\mathbf{z})d\mathbf{x}d\mathbf{z}\\ &= \int_{S_z}\rho_w(\mathbf{z})d\mathbf{z}\int_{\Phi^{-1}S_{y-Bz}}\rho(\mathbf{x})d\mathbf{x}\\ &= \int_{S_z}\rho_w(\mathbf{z})d\mathbf{z}\int_{S_{y-Bz}}\mathscr{P}_\Phi\rho(\mathbf{x})d\mathbf{x}\\ &= \int_{S_z}\rho_w(\mathbf{z})d\mathbf{z}\int_{S_y}\mathscr{P}_\Phi(\mathbf{y}-\mathbf{B}\mathbf{z})d\mathbf{y}.\end{aligned}$$

The conclusion follows accordingly. ■

With the above theorem, the information flow can be easily computed. Note the theorem actually states that

$$\mathscr{P}\rho(\mathbf{y}) = E_w\mathscr{P}_\Phi\rho(\mathbf{y}-\mathbf{B}\mathbf{w}), \tag{62}$$

where E_w signifies the expectation taken with respect to $\mathbf{w}$. So

$$\begin{aligned} H_1(\tau+1) &= -E_y \log(\mathscr{P}\rho)_1(y_1) \\ &= -E_x\left[\log \int_{\mathbb{R}^{n-1}} E_w \mathscr{P}_\Phi \rho(\mathbf{y}-\mathbf{B}\mathbf{w}) dy_2 \dots dy_n\right] \\ &= -E_x[\log E_w(\mathscr{P}_\Phi\rho)_1(y_1-\mathbf{B}_1\mathbf{w})], \end{aligned}$$

where $\mathbf{B}_1 \equiv (b_{11}, b_{12}, \dots, b_{1m})$ is the row vector. Likewise,

$$\mathscr{P}_{\not 2}\rho(\mathbf{y}_{\not 2}) = E_w \mathscr{P}_{\Phi_{\not 2}}\rho(\mathbf{y}_{\not 2}-\mathbf{B}_{\not 2}\mathbf{w}). \tag{63}$$

In the equation, the subscript $\not 2$ in the vector(s) and matrix means the second row is removed from the corresponding entities. So

$$\begin{aligned} H_{1\not 2}(\tau+1) &= -E_y \log(\mathscr{P}_{\not 2}\rho)_1(y_1) \\ &= -E_x\left[\log \int_{\mathbb{R}^{n-2}} E_w \mathscr{P}_{\Phi_{\not 2}}\rho(\mathbf{y}_{\not 2}-\mathbf{B}_{\not 2}\mathbf{w}) dy_3 \dots dy_n\right] \\ &= -E_x[\log E_w(\mathscr{P}_{\Phi_{\not 2}}\rho)_1(y_1-\mathbf{B}_1\mathbf{w})]. \end{aligned}$$

Subtract $H_{1\not 2}(\tau+1)$ from $H_1(\tau+1)$, and the information flow $T_{2\to 1}$ follows:

Theorem V.2.

$$\begin{aligned} T_{2\to 1} = {} & E_x[\log E_w(\mathscr{P}_{\Phi_{\not 2}}\rho)_1(y_1-\mathbf{B}_1\mathbf{w})] \\ & -E_x[\log E_w(\mathscr{P}_\Phi\rho)_1(y_1-\mathbf{B}_1\mathbf{w})]. \end{aligned} \tag{64}$$

B. Properties

Theorem V.3 (principle of nil causality). For the system (57), if Φ_1 and $\mathbf{B}_1$ are independent of x_2, then $T_{2\to 1}=0$.

Proof. As we proved for the deterministic case, if Φ_1 is independent of x_2, then $(\mathscr{P}_\Phi\rho)_1 \overset{a.e.}{=} (\mathscr{P}_{\Phi_{\not 2}}\rho)_1$. If further $\mathbf{B}_1$ has no dependence on x_2, then the above $H_1(\tau+1)$ and $H_{1\not 2}(\tau+1)$ are equal, and hence $T_{2\to 1}=0$. ■

C. Application: A noisy Hénon map

We now reconsider the benchmark systems that have been examined before, but with Gaussian noise added. The baker transformation is not appropriate here, since addition of noise perturbation will take $\mathbf{x}$ outside the domain $[0,1]$. We hence only look at the Hénon map $\Phi:\mathbb{R}^2\to\mathbb{R}^2$:

$$\begin{aligned} \Phi_1(x_1,x_2) &= 1+x_2-\alpha x_1^2, \\ \Phi_2(x_1,x_2) &= \beta x_1, \end{aligned} \tag{65}$$

with parameters $\alpha,\beta>0$, and consider only the flow $T_{1\to 2}$ which has been shown as a benchmark case. Now perturb Φ to make a stochastic mapping:

$$\mathbf{x}(\tau+1) = \Phi(\mathbf{x}(\tau)) + \mathbf{B}\mathbf{w}, \tag{66}$$

where $\mathbf{B}=(b_{ij})$ is a constant matrix, $\mathbf{w}\sim N(0,\mathbf{I})$. Let $\mathbf{B}_i \equiv (b_{i1},b_{i2})$ denote a row vector. It is easy to see that Φ is invertible; in fact, $J=\left[\begin{smallmatrix} -2\alpha^2 & 1 \\ \beta & 0\end{smallmatrix}\right] = -\beta \neq 0$. The inverse is

$$\Phi^{-1}(x_1,x_2) = \left(\frac{x_2}{\beta}, x_1 - 1 + \frac{\alpha}{\beta^2}x_2^2\right). \tag{67}$$

Thus

$$\begin{aligned} \mathscr{P}_\Phi\rho(x_1,x_2) &= \rho(\Phi^{-1}(x_1,x_2))|J^{-1}| \\ &= \rho\left(\frac{x_2}{\beta}, x_1-1+\frac{\alpha}{\beta^2}x_2^2\right)\beta^{-1}. \end{aligned} \tag{68}$$

So $\mathscr{P}\rho(\mathbf{y}) = E_w\mathscr{P}_\Phi\rho(\mathbf{y}-\mathbf{B}\mathbf{w})$, and

$$\begin{aligned} (\mathscr{P}\rho)_2(y_2) &= \int_{\mathbb{R}} dy_1 E_w \frac{1}{\beta}\rho\left(\frac{y_2-\mathbf{B}_2\mathbf{w}}{\beta}, y_1-\mathbf{B}_1\mathbf{w}\right. \\ &\qquad \left. -1+\frac{\alpha}{\beta^2}(y_2-\mathbf{B}_2\mathbf{w})^2\right) \\ &= \frac{1}{\beta}E_w\rho_1\left(\frac{y_2-\mathbf{B}_2\mathbf{w}}{\beta}\right). \end{aligned}$$

If x_1 is frozen, $\Phi_2(x_1,x_2)=\beta x_1$ is a constant. Hence $H_{2\not 1}(\tau+1)=0$, and

$$\begin{aligned} T_{1\to 2} &= H_2(\tau+1) - H_{2\not 1}(\tau+1) \\ &= -E\left[\log\frac{1}{\beta}E_w\rho_1\left(\frac{y_2-\mathbf{B}_2\mathbf{w}}{\beta}\right)\right] - 0 \\ &= \log\beta - E_w E_x \log\rho_1\left(\frac{y_2-\mathbf{B}_2\mathbf{w}}{\beta}\right) \\ &= \log\beta - E_w E_x \log\rho_1\left(x_1-\frac{\mathbf{B}_2\mathbf{w}}{\beta}\right) \\ &= \log\beta + \mathscr{F}H_1. \end{aligned} \tag{69}$$

Here $\mathscr{F}H_1$ is the functional H_1 applied by a Gaussian filter. One may understand it as H_1 smeared out by a Gaussian filter. It is less than H_1, so the noise addition makes the system lose some information, compared to $T_{1\to 2}=\log\beta+H_1$ in the deterministic case.

VI. CONTINUOUS-TIME STOCHASTIC SYSTEMS

A. Derivation

Following what we have done in Sec. IV, we derive the information flow within a continuous-time stochastic system by taking the limit of the corresponding discrete stochastic mapping. In doing this, the results in the preceding section are ready for use. But, as noted, in the above derivation we have assumed a constant matrix $\mathbf{B}$, a simplified case allowing for a clear expression of information flow. (This case does have realistic relevance, though.) For a continuous-time system, this assumption actually can be completely relaxed. In the following we will see why.

Consider a system

$$d\mathbf{x} = \mathbf{F}(t;\mathbf{x})dt + \mathbf{B}(t;\mathbf{x})d\mathbf{w}, \tag{70}$$

where $\mathbf{x}$ and $\mathbf{F}$ are n-dimensional vectors, $\mathbf{B}$ is an $n\times m$ matrix, and $\mathbf{w}$ an m-vector of standard Wiener process. Note that $\mathbf{B}$ can be a function of both $\mathbf{x}$ and time t. This above equation may also be written as

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t;\mathbf{x}) + \mathbf{B}(t;\mathbf{x})\dot{\mathbf{w}}, \tag{71}$$

where $\dot{\mathbf{w}}$ a vector of white noise, or, in component form,

$$\frac{dx_1}{dt} = F_1(t;\mathbf{x}) + \mathbf{B}_1(t;\mathbf{x})\dot{\mathbf{w}}, \tag{72}$$

$$\frac{dx_2}{dt} = F_2(t;\mathbf{x}) + \mathbf{B}_2(t;\mathbf{x})\dot{\mathbf{w}}, \tag{73}$$

$$\vdots \qquad\qquad \vdots \tag{74}$$

$$\frac{dx_n}{dt} = F_n(t;\mathbf{x}) + \mathbf{B}_n(t;\mathbf{x})\dot{\mathbf{w}}. \tag{75}$$

In the equations we have used $\mathbf{B}_i$ to indicate the ith row vector of $\mathbf{B}$. Now consider (70) on a small interval $[t,t+\Delta t]$. Euler-Bernstein differencing,

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \mathbf{F}\Delta t + \mathbf{B}\Delta\mathbf{w}. \tag{76}$$

This motivates the introduction of a transformation

$$\Pi : \begin{cases} \mathbf{y} = \mathbf{x} + \mathbf{F}(\mathbf{x})\Delta t + \mathbf{B}(\mathbf{x})\Delta\mathbf{w}, \\ \mathbf{z} = \Delta\mathbf{w}. \end{cases} \tag{77}$$

As shown in the discrete mapping case, generally this transformation cannot be inverted. But for this special case where Δt and $\Delta\mathbf{w}$ are small, the inversion can be done asymptotically. In fact,

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + [\mathbf{F}(\mathbf{y}) + o(\Delta t)]\Delta t \\ &\quad + [\mathbf{B}(\mathbf{y})\Delta\mathbf{w} + \nabla(\mathbf{B}(\mathbf{y})\Delta\mathbf{w})(\mathbf{x}-\mathbf{y})] + o(\Delta w^2) \\ &= \mathbf{x} + \mathbf{F}(\mathbf{y})\Delta t + \mathbf{B}(\mathbf{y})\Delta\mathbf{w} \\ &\quad + [\nabla(\mathbf{B}\Delta\mathbf{w})](-\mathbf{F}\Delta t - \mathbf{B}\Delta\mathbf{w}) + o(\Delta t). \end{aligned}$$

Note, here the higher order terms mean terms with order higher than Δt or $(\Delta w)^2$; we will see soon that $E(\Delta w)^2 = \Delta t$. The above expansion helps invert Π to

$$\Pi^{-1} : \begin{cases} \mathbf{x} = \mathbf{y} - \mathbf{F}\Delta t - \mathbf{B}\mathbf{z} + \nabla(\mathbf{B}\mathbf{z})(\mathbf{B}\mathbf{z}) + o(\Delta t), \\ \Delta\mathbf{w} = \mathbf{z}. \end{cases} \tag{78}$$

The following proposition finds the Jacobian associated with the inverse transformation.

Proposition VI.1. Define the double dot of two dyadics $\mathbf{A}$ and $\mathbf{B}$ as $\mathbf{A} : \mathbf{B} = \sum_{i,j} a_{ij}b_{ji}$; then

$$J^{-1} = 1 - \nabla\cdot\mathbf{F}\Delta t - \nabla\cdot(\mathbf{B}\mathbf{z}) + \tfrac{1}{2}\nabla\nabla : (\mathbf{B}\mathbf{z}\mathbf{z}^T\mathbf{B}^T) + o(\Delta t). \tag{79}$$

Proof. By definition

$$J^{-1} = \left[\frac{\partial(\mathbf{x},\Delta\mathbf{w})}{\partial(\mathbf{y},\mathbf{z})}\right] = \det\begin{bmatrix} \frac{\partial\mathbf{x}}{\partial\mathbf{y}} & \frac{\partial\mathbf{x}}{\partial\mathbf{z}} \\ \frac{\partial\Delta\mathbf{w}}{\partial\mathbf{y}} & \frac{\partial\Delta\mathbf{w}}{\partial\mathbf{z}} \end{bmatrix} = \det\begin{bmatrix} \frac{\partial\mathbf{x}}{\partial\mathbf{y}} & -\mathbf{B}+\dots \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \det\left(\frac{\partial\mathbf{x}}{\partial\mathbf{y}}\right). \tag{80}$$

The key is the evaluation of $\det(\frac{\partial\mathbf{x}}{\partial\mathbf{y}})$. By (78), it is, up to $o(\Delta t)$, the determinant of

$$\begin{bmatrix} 1 - \frac{\partial F_1}{\partial y_1}\Delta t - \sum_k \frac{\partial b_{1k}}{\partial y_1}z_k + \frac{\partial}{\partial y_1}\left(\sum_{l,k,s}\frac{\partial b_{1k}}{\partial y_l}z_k b_{ls}z_s\right) & \dots & -\frac{\partial F_1}{\partial y_n}\Delta t - \sum_k \frac{\partial b_{1k}}{\partial y_n}z_k + \frac{\partial}{\partial y_n}\left(\sum_{l,k,s}\frac{\partial b_{1k}}{\partial y_l}z_k b_{ls}z_s\right) \\ \vdots & \ddots & \vdots \\ -\frac{\partial F_n}{\partial y_1}\Delta t - \sum_k \frac{\partial b_{nk}}{\partial y_1}z_k + \frac{\partial}{\partial y_1}\left(\sum_{l,k,s}\frac{\partial b_{nk}}{\partial y_l}z_k b_{ls}z_s\right) & \dots & 1 - \frac{\partial F_n}{\partial y_n}\Delta t - \sum_k \frac{\partial b_{nk}}{\partial y_n}z_k + \frac{\partial}{\partial y_n}\left(\sum_{l,k,s}\frac{\partial b_{nk}}{\partial y_l}z_k b_{ls}z_s\right) \end{bmatrix}. \tag{81}$$

Recall that, for an $n\times n$ matrix $\mathbf{A} = (a_{ij})$,

$$\det\mathbf{A} = \sum_{\sigma\in P_n}\operatorname{sgn}(\sigma)\prod_{i=1}^{n} a_{i,\sigma_i}, \tag{82}$$

where P_n is the totality of permutations of $\{1,2,\dots,n\}$. By this formula, the terms of order Δt and Δw are easy to find; they can only come from the diagonal entries. For terms of order $(\Delta w)^2$, there are three sources:

(1) the last term at each diagonal entry, together with $n-1$ 1's;
(2) multiplication of two entries at (i,i) and (j,j), $i\neq j$, together with $n-2$ 1's on the diagonal;
(3) similar to (2), but with entries at (i,j) and (j,i), $i\neq j$.

In (3) the order between i and j switches and it has $\operatorname{sgn} = -1$ for its permutation. Except for (3) involving off-diagonal entries, all others are from the diagonal. So

$$\begin{aligned} \det\left(\frac{\partial\mathbf{x}}{\partial\mathbf{y}}\right) &= \prod_{i=1}^{n}\left[1 - \frac{\partial F_i}{\partial y_i}\Delta t - \sum_k\frac{\partial b_{ik}}{\partial y_i}z_k + \sum_{l=1}^{n}\sum_{k,s=1}^{m}\frac{\partial}{\partial y_i}\left(b_{ls}\frac{\partial b_{ik}}{\partial y_l}\right)z_k z_s\right] \\ &\quad + \frac{1}{2}\sum_{\substack{i,j \\ i\neq j}}(-1)^1\left(-\sum_k\frac{\partial b_{ik}}{\partial y_j}z_k\right)\left(-\sum_s\frac{\partial b_{js}}{\partial y_i}z_s\right) + o(\Delta t). \end{aligned}$$

Notice the factor $\frac{1}{2}$ in the last term. Because of the symmetry between i and j they repeat once when summed over $i,j = 1,n$. Thus

$$\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right) = 1 - \sum_i \frac{\partial F_i}{\partial y_i}\Delta t - \sum_i \sum_k \frac{\partial b_{ik}}{\partial y_i} z_k + \frac{1}{2}\sum_{i\neq j}\sum_k \frac{\partial b_{ik}}{\partial y_i} z_k \sum_s \frac{\partial b_{js}}{\partial y_j} z_s + \sum_{i,j}\sum_{k,s}\frac{\partial}{\partial y_i}\left(b_{js}\frac{\partial b_{ik}}{\partial y_j}\right) z_k z_s$$
$$-\frac{1}{2}\sum_{i\neq j}\sum_k \frac{\partial b_{ik}}{\partial y_j} z_k \sum_s \frac{\partial b_{js}}{\partial y_i} z_s + o(\Delta t).$$

Notice

$$\frac{\partial^2 b_{ik} b_{js}}{\partial y_i \partial y_j} = \frac{\partial b_{ik}}{\partial y_i}\frac{\partial b_{js}}{\partial y_j} + b_{ik}\frac{\partial^2 b_{js}}{\partial y_i \partial y_j} + \frac{\partial b_{js}}{\partial y_i}\frac{\partial b_{ik}}{\partial y_j} + b_{js}\frac{\partial^2 b_{ik}}{\partial y_i \partial y_j},$$

and

$$\sum_{k,s}\sum_{i,j}\frac{\partial b}{\partial y_i}_{js}\frac{\partial b_{ik}}{\partial y_j} z_k z_s = \sum_{k,s}\sum_{i,j}\frac{\partial b_{js}}{\partial y_i}\frac{\partial b_{ik}}{\partial y_j} z_k z_s + \sum_{k,s}\sum_{i,j} b_{js}\frac{\partial^2 b_{ik}}{\partial y_i \partial y_j} z_k z_s$$
$$= \frac{1}{2}\sum_{k,s}\sum_{i\neq j}\frac{\partial b_{[}js}{\partial y_i}\frac{\partial b_{ik}}{\partial y_j} z_k z_s + \frac{1}{2}\sum_{k,s}\sum_i \frac{\partial b_{is}}{\partial y_i}\frac{\partial b_{ik}}{\partial y_i} z_k z_s + \frac{1}{2}\sum_{k,s}\sum_{i,j}\frac{\partial b_{js}}{\partial y_i}\frac{\partial b_{ik}}{\partial y_j} z_k z_s$$
$$+\frac{1}{2}\sum_{k,s}\sum_{i,j}\left(b_{js}\frac{\partial^2 b_{ik}}{\partial y_i \partial y_j} + b_{ik}\frac{\partial^2 b_{js}}{\partial y_i \partial y_j}\right) z_k z_s.$$

The last parenthesis holds because the two pairs (i,k) and (j,s) may be switched under the summation without changing the result. Thence

$$\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right) = 1 - \sum_i \frac{\partial F_i}{\partial y_i}\Delta t - \sum_i\sum_k \frac{\partial b_{ik}}{\partial y_i} z_k + \frac{1}{2}\sum_{i,j}\sum_{k,s}\frac{\partial b_{ik}}{\partial y_i}\frac{\partial b_{js}}{\partial y_j} z_k z_s$$
$$+\frac{1}{2}\sum_{i,j}\sum_{k,s}\frac{\partial b_{js}}{\partial y_i}\frac{\partial b_{ik}}{\partial y_j} z_k z_s + \frac{1}{2}\sum_{i,j}\sum_{k,s}\left(b_{js}\frac{\partial^2 b_{ik}}{\partial y_i \partial y_j} + b_{ik}\frac{\partial^2 b_{js}}{\partial y_i \partial y_j}\right) z_k z_s + o(\Delta t)$$
$$= 1 - \sum_i \frac{\partial F_i}{\partial y_i}\Delta t - \sum_i\sum_k \frac{\partial b_{ik}}{\partial y_i} z_k + \frac{1}{2}\sum_{i,j}\frac{\partial^2 \sum_{k,s} b_{ik} z_k z_s b_{js}}{\partial y_i \partial y_j} + o(\Delta t)$$
$$= 1 - \nabla\cdot\mathbf{F}\Delta t - \nabla\cdot(\mathbf{B}\mathbf{z}) + \frac{1}{2}\nabla\nabla : (\mathbf{B}\mathbf{z}\mathbf{z}^T\mathbf{B}^T) + o(\Delta t),$$

which is J^{-1} by (80). ■

With J^{-1}, we can then evaluate the operator $\mathscr{P}$ and hence arrive at $\frac{dH_1}{dt}$ and $\frac{dH_{12}}{dt}$.

Proposition VI.2. Let $\mathbf{B}\mathbf{B}^T \equiv \mathbf{G} = (g_{ij})$. The time rate of change of H_1 is

$$\frac{dH_1}{dt} = -E\left[F_1\frac{\partial \log\rho_1}{\partial x_1}\right] - \frac{1}{2}E\left[g_{11}\frac{\partial^2 \log\rho_1}{\partial x_1^2}\right]. \tag{83}$$

Proof. For any subset $S_y \in \mathbb{R}^n$, $S_z \in \mathbb{R}^m$,

$$\int_{S_y\times S_z}\rho_{yz}(\mathbf{y},\mathbf{z})d\mathbf{y}d\mathbf{z} = \int_{\Pi^{-1}(S_y\times S_z)}\rho_{xw}(\mathbf{x},\Delta\mathbf{w})d\mathbf{x}ddelw = \int_{S_y\times S_z}\rho_{xw}(\mathbf{y}-\mathbf{F}\Delta t - \mathbf{B}\mathbf{z} + \nabla(\mathbf{B}\mathbf{z})\cdot(\mathbf{B}\mathbf{z}),\,\mathbf{z})\cdot|J^{-1}|d\mathbf{y}d\mathbf{z}$$
$$= \int_{S_y\times S_z}\rho(\mathbf{y}-\mathbf{F}\Delta t - \mathbf{B}\mathbf{z} + \nabla(\mathbf{B}\mathbf{z})\cdot(\mathbf{B}\mathbf{z}))|J^{-1}|\cdot\rho_w(\mathbf{z})$$

because $\rho_{xw}(\mathbf{a},\mathbf{b}) = \rho_x(\mathbf{a})\cdot\rho_w(\mathbf{b}) = \rho(\mathbf{a})\cdot\rho_w(\mathbf{b})$ due to the independence between $\mathbf{x}$ and $\Delta\mathbf{w}$. Since S_y and S_z are arbitrarily chosen, the integrand is the very joint pdf $\rho_{yz}(\mathbf{y},\mathbf{z})$. Thus

$$\mathscr{P}\rho(\mathbf{y}) = \rho_y(\mathbf{y}) = \int_{\mathbb{R}^m}\rho_{yz}(\mathbf{y},\mathbf{z})d\mathbf{z} = \int_{\mathbb{R}^m}[\rho(\mathbf{y}-\mathbf{F}\Delta t - \mathbf{B}\mathbf{z} + \nabla(\mathbf{B}\mathbf{z})\cdot(\mathbf{B}\mathbf{z}))\cdot|J^{-1}|]\cdot\rho_w(\mathbf{z})d\mathbf{z}$$
$$= E_w\{\rho(\mathbf{y}-\mathbf{F}\Delta t - \mathbf{B}\Delta\mathbf{w} + \nabla(\mathbf{B}\Delta\mathbf{w})\cdot(\mathbf{B}\Delta\mathbf{w}))\cdot|J^{-1}|\}$$
$$= E_w\left[\rho(\mathbf{y}) - \nabla\rho\cdot[\mathbf{F}\Delta t + \mathbf{B}\Delta\mathbf{w} + \nabla(\mathbf{B}\Delta\mathbf{w})\cdot(\mathbf{B}\Delta\mathbf{w})] + \frac{1}{2}(\mathbf{B}\Delta\mathbf{w})(\mathbf{B}\Delta\mathbf{w})^T : \nabla\nabla\rho\right]$$

$$\cdot\left[1-\nabla\cdot\mathbf{F}\Delta t-\nabla\cdot(\mathbf{B}\Delta\mathbf{w})+\frac{1}{2}\nabla\nabla:(\mathbf{B}\Delta\mathbf{w}\Delta\mathbf{w}^T\mathbf{B}^T)\right]+o(\Delta t)$$

$$=\rho(\mathbf{y})-(\mathbf{F}\cdot\nabla\rho+\rho\nabla\cdot\mathbf{F})\Delta t+\frac{1}{2}\left\{\rho\nabla\nabla:(\mathbf{B}\mathbf{B}^T)+2\nabla\rho\cdot[\nabla\cdot(\mathbf{B}\mathbf{B}^T)]+(\mathbf{B}\mathbf{B}^T):(\nabla\nabla\rho)\right\}\Delta t+o(\Delta t)$$

$$=\rho(\mathbf{y})-\nabla\cdot(\mathbf{F}\rho)\Delta t+\frac{1}{2}\nabla\nabla:(\mathbf{B}\mathbf{B}^T\rho)\Delta t+o(\Delta t). \tag{84}$$

Note here the fact

$$E\Delta\mathbf{w}=\mathbf{0},\qquad E\Delta\mathbf{w}\Delta\mathbf{w}^T=\Delta t\mathbf{I} \tag{85}$$

about Wiener process has been used. As a verification, one may obtain from this step

$$\frac{\partial\rho}{\partial t}=\lim_{\Delta t\to 0}\frac{\mathscr{P}\rho(\mathbf{y})-\rho(\mathbf{y})}{\Delta t}=-\nabla\cdot(\mathbf{F}\rho)+\frac{1}{2}\nabla\nabla:(\mathbf{B}\mathbf{B}^T\rho), \tag{86}$$

which is precisely the Fokker-Planck equation.

Denote $\mathbf{B}\mathbf{B}^T$ by $\mathbf{G}$. Integrate both sides of the above equation with respect to $(y_2,y_3,\ldots,y_n)$ to obtain

$$(\mathscr{P}\rho)_1(y_1)=\rho_1(y_1)-\Delta t\int_{\mathbb{R}^{n-1}}\frac{\partial F_1\rho}{\partial y_1}dy_2\ldots dy_n+\frac{\Delta t}{2}\int_{\mathbb{R}^{n-1}}\frac{\partial^2 g_{11}\rho}{\partial y_1^2}dy_2\ldots dy_n+o(\Delta t),$$

and hence

$$\log(\mathscr{P}\rho)_1(y_1)=\log\rho_1(y_1)-\frac{\Delta t}{\rho_1}\int_{\mathbb{R}^{n-1}}\frac{\partial F_1\rho}{\partial y_1}dy_2\ldots dy_n+\frac{\Delta t}{2\rho_1}\int_{\mathbb{R}^{n-1}}\frac{\partial^2 g_{11}\rho}{\partial y_1^2}dy_2\ldots dy_n+o(\Delta t). \tag{87}$$

So

$$H_1(t+\Delta t)=-E\log(\mathscr{P}\rho)_1(y_1)=-E\log\rho_1(y_1)$$

as the rest two terms vanish after applying the operator $E(\cdot)=\int_{\mathbb{R}}\rho_1(\cdot)dy_1$. Expanding y_1 around x_1, and denoting $\mathbf{B}_1\equiv(b_{11},b_{12},\ldots,b_{1n})$, we have

$$H_1(t+\Delta t)=-E\log\rho_1(x_1+F_1\Delta t+\mathbf{B}_1\Delta\mathbf{w})$$

$$=-E\left[\log\rho_1(x_1)+\frac{\partial\log\rho_1}{\partial x_1}(F_1\Delta t+\mathbf{B}_1\Delta\mathbf{w})+\frac{1}{2}\frac{\partial^2\log\rho_1}{\partial x_1^2}\mathbf{B}_1\Delta\mathbf{w}\Delta\mathbf{w}^T\mathbf{B}_1^T\right]+o(\Delta t)$$

$$=H_1(t)-E\left[F_1\frac{\partial\log\rho_1}{\partial x_1}\right]\Delta t-\frac{1}{2}E\left[g_{11}\frac{\partial^2\log\rho_1}{\partial x_1^2}\right]\Delta t+o(\Delta t).$$

Let $\Delta t\to 0$ and we finally arrive at

$$\frac{dH_1}{dt}=-E\left[F_1\frac{\partial\log\rho_1}{\partial x_1}\right]-\frac{1}{2}E\left[g_{11}\frac{\partial^2\log\rho_1}{\partial x_1^2}\right].$$

■

Now consider during the time interval $[t,t+\Delta t]$ to freeze x_2 as a parameter, and examine how the marginal entropy of x_1 evolves. In this case we are actually considering a density $\rho_{1\not 2}$, with $\rho_{1\not 2}(t)=\rho_1(t)$ under an $(n-1)$-dimensional transformation: $\mathbb{R}^{n-1}\to\mathbb{R}^{n-1}$, $\mathbf{x}_{1\not 2}\to\mathbf{y}_{1\not 2}$:

$$\begin{aligned}
y_1&=x_1(t+\Delta t)=x_1(t)+F_1\Delta t+\mathbf{B}_1\Delta\mathbf{w},\\
y_3&=x_3(t+\Delta t)=x_3(t)+F_3\Delta t+\mathbf{B}_3\Delta\mathbf{w},\\
&\ldots\\
y_n&=x_n(t+\Delta t)=x_n(t)+F_n\Delta t+\mathbf{B}_n\Delta\mathbf{w}.
\end{aligned}$$

With this system we have the following proposition.

Proposition VI.3. Let $\rho_{\not 2}$ be $\int_{\mathbb{R}}\rho(\mathbf{x})dx_2$; then

$$\frac{dH_{1\not 2}}{dt}=-E\left[F_1\frac{\partial\log\rho_1}{\partial x_1}\right]-\frac{1}{2}E\left[g_{11}\frac{\partial^2\log\rho_1}{\partial x_1^2}\right]+E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\not 2}}{\partial x_1}dx_3\ldots dx_n\right]-\frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\not 2}}{\partial x_1^2}dx_3\ldots dx_n\right]. \tag{88}$$

Proof. Following the same procedure as above, we arrive at an equation for $\log(\mathscr{P}_{\not 2})_1(y_1)$ similar to (87):

$$\log(\mathscr{P}_{\not 2}\rho)_1(y_1)=\log\rho_{1\not 2}(y_1)-\frac{\Delta t}{\rho_{1\not 2}}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\not 2}}{\partial y_1}dy_3\ldots dy_n+\frac{\Delta t}{2\rho_{1\not 2}}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\not 2}}{\partial y_1^2}dy_3\ldots dy_n+o(\Delta t).$$

So

$$H_{1\mathfrak{Y}}(t+\Delta t) = -E\log(\mathscr{P}_{\mathfrak{Y}}\rho)_1(y_1) = -E\log\rho_{1\mathfrak{Y}}(y_1) + E\left[\frac{1}{\rho_{1\mathfrak{Y}}}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\mathfrak{Y}}}{\partial y_1}dy_3\dots dy_n\right]\Delta t$$
$$-\frac{1}{2}E\left[\frac{1}{\rho_{1\mathfrak{Y}}}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial y_1^2}dy_3\dots dy_n\right]\Delta t + o(\Delta t).$$

Note at time t, $\rho_{1\mathfrak{Y}} = \rho_1$, and in the last two terms $\mathbf{y}$ can be replaced by $\mathbf{x}$ with error going to higher order terms. Thus

$$H_{1\mathfrak{Y}}(t+\Delta t) = -E\left[\log\rho_1(x_1) + \frac{\partial\log\rho_1}{\partial x_1}(F_1\Delta t + \mathbf{B}_1\Delta\mathbf{w}) + \frac{1}{2}\frac{\partial^2\log\rho_1}{\partial x_1^2}\mathbf{B}_1\Delta\mathbf{w}\Delta\mathbf{w}^T\mathbf{B}_1^T\right]$$
$$+E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\mathfrak{Y}}}{\partial x_1}dx_3\dots dx_n\right]\Delta t - \frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial x_1^2}dx_3\dots dx_n\right]\Delta t + o(\Delta t)$$
$$= H_1(t) - E\left[F_1\frac{\partial\log\rho_1}{\partial x_1}\right]\Delta t - \frac{1}{2}E\left[g_{11}\frac{\partial^2\log\rho_1}{\partial x_1^2}\right]\Delta t + E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\mathfrak{Y}}}{\partial x_1}dx_3\dots dx_n\right]\Delta t$$
$$-\frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial x_1^2}dx_3\dots dx_n\right]\Delta t + o(\Delta t).$$

Take the limit

$$\frac{dH_{1\mathfrak{Y}}}{dt} = \lim_{\Delta t\to 0}\frac{H_{1\mathfrak{Y}}(t+\Delta t) - H_1(t)}{\Delta t}$$

and we arrive at the conclusion. ■

Theorem VI.1.

$$T_{2\to 1} = -E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\mathfrak{Y}}}{\partial x_1}dx_3\dots dx_n\right]$$
$$+\frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial x_1^2}dx_3\dots dx_n\right] \quad (89)$$
$$= -\int_{\mathbb{R}^n}\rho_{2|1}(x_2|x_1)\frac{\partial F_1\rho_{\mathfrak{Y}}}{\partial x_1}d\mathbf{x}$$
$$+\frac{1}{2}\int_{\mathbb{R}^n}\rho_{2|1}(x_2|x_1)\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial x_1^2}d\mathbf{x}. \quad (90)$$

Proof. Subtract (88) from (83) and the conclusion follows. ■

B. Properties

Theorem VI.2. For a 2D system

$$\frac{dH_{1\mathfrak{Y}}}{dt} = E\left(\frac{\partial F_1}{\partial x_1}\right) \quad (91)$$

in the absence of stochasticity.

Remark. This recovers the heuristic argument by Liang and Kleeman in [46]; see Eq. (5).

Proof. In this case $g_{11} = 0$, $\rho_{\mathfrak{Y}} = \rho_1$, so

$$\frac{dH_{1\mathfrak{Y}}}{dt} = -E\left[F_1\frac{\partial\log\rho_1}{\partial x_1}\right] + E\left[\frac{1}{\rho_1}\frac{\partial F_1\rho_1}{\partial x_1}\right]$$
$$= E\left[\frac{\rho_1}{\rho_1}\frac{\partial F_1}{\partial x_1} + F_1\frac{\partial\log\rho_1}{\partial x_1} - F_1\frac{\partial\log\rho_1}{\partial x_1}\right]$$
$$= E\left(\frac{\partial F_1}{\partial x_1}\right).$$

■

Theorem VI.3. If $g_{11} = \sum_{k=1}^m b_{1k}b_{1k}$ is independent of x_2, the resulting $T_{2\to 1}$ has a form the same as its deterministic counterpart.

Proof. If g_{11} is independent of x_2, so is $\int\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial x_1^2}dx_3\dots dx_n$. Hence the integration can be simplified:

$$\int_{\mathbb{R}^n}\rho_{2|1}\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial x_1^2}d\mathbf{x}$$
$$= \int_{\mathbb{R}^{n-1}}\left(\int_{\mathbb{R}}\frac{\rho_{12}}{\rho_1}dx_2\right)\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial x_1^2}dx_1dx_3\dots dx_n$$
$$= \int_{\mathbb{R}^{n-1}}\frac{\partial^2 g_{11}\rho_{\mathfrak{Y}}}{\partial x_1^2}dx_1dx_3\dots dx_n = 0.$$

■

Theorem VI.4 (principle of nil causality). If both F_1 and g_{11} are independent of x_2, then $T_{2\to 1} = 0$.

Proof. As proved above, when g_{11} has no dependence on x_2, the last term of $T_{2\to 1}$ becomes zero. If, moreover, F_1 does not depend on x_2, then $\frac{\partial F_1\rho_{\mathfrak{Y}}}{\partial x_1}$ does not, either. So the integration with respect to x_2 can be taken inside directly to $\rho_{12}/\rho_1 = \rho_{2|1}(x_2|x_1)$:

$$T_{2\to 1} = -\int_{\mathbb{R}}dx_1\int_{\mathbb{R}}\rho_{2|1}(x_2|x_1)dx_2\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\mathfrak{Y}}}{\partial x_1}dx_3\dots dx_n$$
$$= \int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\mathfrak{Y}}}{\partial x_1}dx_3\dots dx_n = 0.$$

■

C. Application: A stochastic gradient system

We are about to study the information flow within a system which has a drift function in the gradient form. We particularly want to understand how stochastic perturbation may exert

influence on the flow. The gradient systems are chosen because their corresponding Fokker-Planck equations admit explicit equilibrium solutions, i.e., solutions of the Boltzmann type. To see this, let

$$\mathbf{F} = -\nabla V, \tag{92}$$

where $V = V(\mathbf{x})$ is the potential function. For simplicity, suppose that the stochastic perturbation amplitude $\mathbf{B} = b\mathbf{I}$ where I is the identity matrix and $b =$ constant. Hence $\mathbf{G} = \mathbf{B}\mathbf{B}^T = g\mathbf{I}$, and $g = b^2$ is a constant. It is trivial to verify that

$$\rho = \frac{1}{Z} e^{-2V/g}, \tag{93}$$

where Z is the normalizer (or partition function as is called in statistical physics), solves

$$\nabla \cdot (\rho \mathbf{F}) = \tfrac{1}{2} g \nabla^2 \rho,$$

the equilibrium density equation for the system

$$d\mathbf{x} = -\nabla V dt + b\mathbf{I} d\mathbf{w}. \tag{94}$$

As an example, consider the potential function

$$V = \tfrac{1}{2}\left(x_1^2 x_2^2 + x_2^2 x_3^2 + x_1^2 + x_2^2 + x_3^2\right). \tag{95}$$

This system, though simple, results in a compactly supported density function, while allowing for asymmetric nonlinear interactions among x_1, x_2, and x_3. The resulting vector field is

$$\begin{aligned} F_1 &= -x_1 x_2^2 - x_1, \\ F_2 &= -x_2 x_3^2 - x_2 x_1^2 - x_2, \\ F_3 &= -x_3 x_2^2 - x_3. \end{aligned}$$

Obviously, $T_{3\to 1} = T_{1\to 3} = 0$ by the principle of nil causality. The general flow from x_j to x_i is

$$\begin{aligned} T_{j\to i} &= -\int_{\mathbb{R}^3} \rho_{j|i}(x_j|x_i) \frac{\partial F_i \rho_{\backslash j}}{\partial x_i}\, d\mathbf{x} \\ &= -\int_{\mathbb{R}^3} \rho_{j|i} \left(F_i \frac{\partial \rho_{\backslash j}}{\partial x_i} + \rho_{\backslash j} \frac{\partial F_i}{\partial x_i} \right) \\ &= -\int_{\mathbb{R}^3} \rho_{j|i} \left(F_i \int_{\mathbb{R}} \frac{2}{g} \rho F_i dx_j + \rho_{\backslash j} \frac{\partial F_i}{\partial x_i} \right) d\mathbf{x}. \end{aligned} \tag{96}$$

The computation seems to be easy, but by no means trivial. The difficulty comes from the evaluation of the conditional density $\rho_{j|i}(x_j|x_i)$. Theoretically this is not a problem, but in realizing the computation we have to consider the problem on a limited domain, which may not effectively cover the support of the density function. Here we choose a domain

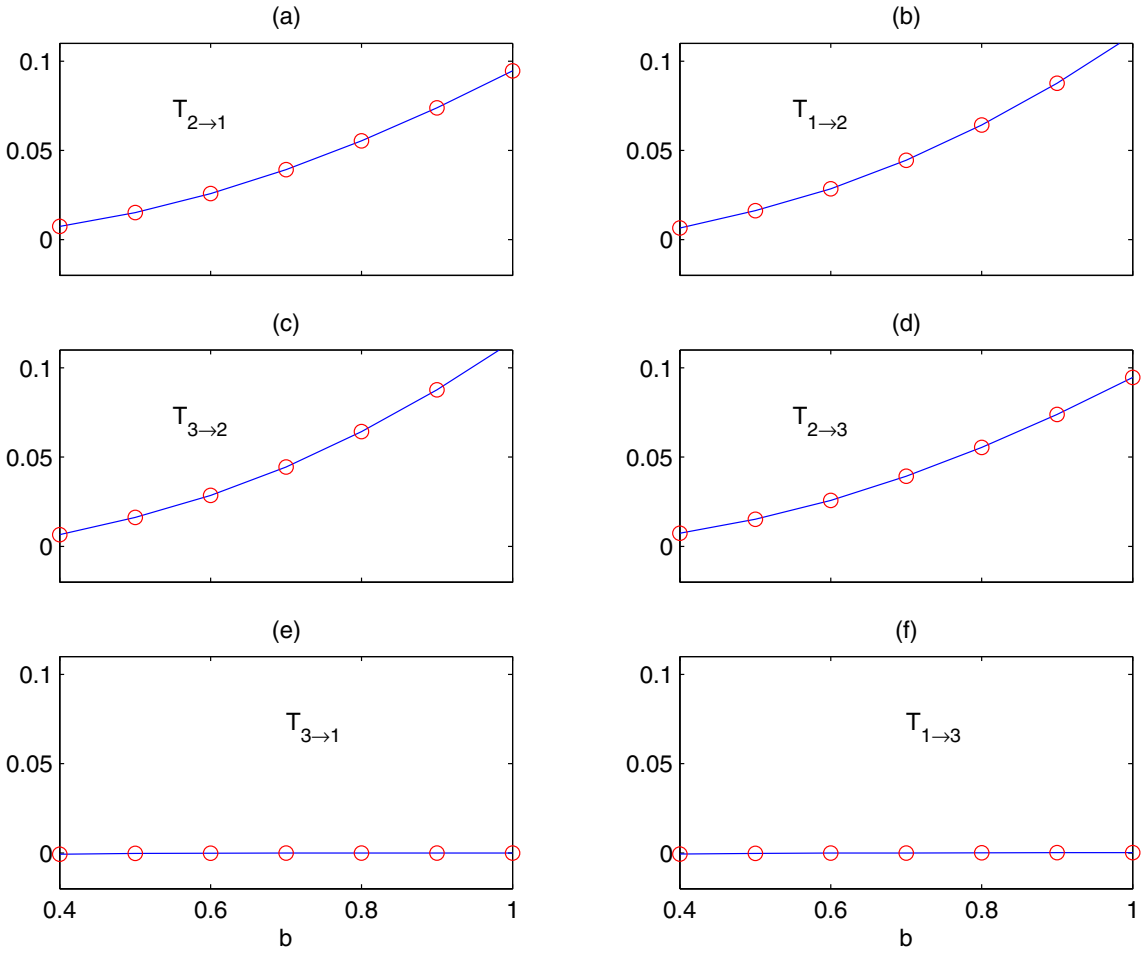

FIG. 9. Information flow within a gradient system with the potential function (95).

$[-5,5]\times[-5,5]\times[-5,5]$, and a spacing size $\Delta x = 0.05$. The computation is implemented henceforth.

To test how the stochastic perturbation may affect the information flow, tune b to see the response. The tuning range is rather limited, though, with the present computational domain. Shown in Fig. 9 are the results. As expected, $T_{3\to1}$ and $T_{1\to3}$ are identically zero. For others, the flow rates generally increase with b. That is to say, they tend to increase the uncertainty of their corresponding target components. This makes sense, since g functions like temperature in thermodynamics, and increase in T surely will lead to increase in uncertainty. If examining more carefully, one finds that the increase is actually not symmetric. Those going to x_2 ($T_{3\to2}$ and $T_{1\to2}$) are faster than those leaving x_2 ($T_{2\to1}$ and $T_{2\to3}$), reflecting the property of asymmetry of information flow.

Since ρ can be accurately obtained, this example can be utilized to validate our numerical computations for more general cases.

VII. LINEAR STOCHASTIC SYSTEMS

As always, it would be of interest to look at the particular case, namely, the case with linear systems:

$$d\mathbf{x} = \mathbf{A}\mathbf{x}dt + \mathbf{B}d\mathbf{w}, \tag{97}$$

where $\mathbf{A}$ and $\mathbf{B}$ are constant matrices. If originally $\mathbf{x}$ is normally distributed, then it is normal/Gaussian forever. Let its mean vector be $\boldsymbol{\mu}$ and its covariance matrix be $\boldsymbol{\Sigma}$. Then

$$\frac{d\boldsymbol{\mu}}{dt} = \mathbf{A}\boldsymbol{\mu}, \tag{98}$$

$$\frac{d\boldsymbol{\Sigma}}{dt} = \mathbf{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T. \tag{99}$$

In component form $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^T$, $\boldsymbol{\Sigma} = (\sigma_{ij})_{n\times n}$, and $\mathbf{B}\mathbf{B}^T$ has been denoted by $\mathbf{G}$ in the above. The distribution is, therefore,

$$\rho = \frac{1}{\sqrt{(2\pi)^n \det\boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

We need to find

$$\rho_1, \rho_{12}, \rho_{\not 2},$$

and the following facts will help.

Fact 1: $\rho_{\not 2}$ is a multivariate Gaussian $N(\boldsymbol{\mu}_{\not 2}, \boldsymbol{\Sigma}_{\not 2})$ where $\boldsymbol{\mu}_{\not 2} = (\mu_1,\mu_3,\mu_4,\ldots,\mu_n)^n$, and $\boldsymbol{\Sigma}_{\not 2}$ is the covariance matrix of $(x_1,x_3,x_4,\ldots,x_n)^n$.

Fact 2: The conditional probability density function $\rho_{2|1}$ is

$$\rho_{2|1}(x_2|x_1) \propto e^{-\frac{\sigma_{11}}{2\Delta_{12}}[x_2-\mu_2-\frac{\sigma_{12}}{\sigma_{11}}(x_1-\mu_1)]^2}, \tag{100}$$

in other words,

$$x_2|x_1 \sim N\left(\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(x_1-\mu_1),\ \frac{\Delta_{12}}{\sigma_{11}}\right). \tag{101}$$

In the above equations, we have used, and will be using, Δ_{ij} to shorten $\det\begin{bmatrix}\sigma_{ii} & \sigma_{ij}\\ \sigma_{ij} & \sigma_{jj}\end{bmatrix}$.

We now compute the information flow $T_{2\to1}$. Since $\mathbf{B}$ is constant (hence independent of x_1), the stochastic term vanishes by Theorem VI.3. So we need only consider its deterministic part:

$$T_{2\to1} = -E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\not 2}}{\partial x_1}\right] = -\int_{\mathbb{R}^n}\rho_{2|1}(x_2|x_1)\frac{\partial F_1\rho_{\not 2}}{\partial x_1}d\mathbf{x}.$$

As a starting point, let us consider the case $n=3$. By the proposition above,

$$\begin{aligned}\rho_{\not 2} &= \rho_{13}\\ &= \frac{1}{\sqrt{(2\pi)^2\Delta_{13}}}e^{-\frac{1}{\Delta_{13}}[\sigma_{33}(x_1-\mu_1)^2+\sigma_{11}(x_3-\mu_3)^2-2\sigma_{13}(x_1-\mu_1)(x_3-\mu_3)]}.\end{aligned}$$

So

$$\begin{aligned}&\int_{\mathbb{R}}\frac{\partial F_1\rho_{\not 2}}{\partial x_1}dx_3\\ &\quad= \int_{\mathbb{R}}\rho_{13}\{a_{11} + [\sigma_{13}(x_3-\mu_3)\\ &\qquad -\sigma_{33}(x_1-\mu_1)](a_{11}x_1+a_{12}x_2+a_{13}x_3)/\Delta_{13}\}dx_3\\ &\quad= a_{11}\rho_1 - \frac{\sigma_{13}\mu_3+\sigma_{33}(x_1-\mu_1)(a_{11}x_1+a_{12}x_2)}{\Delta_{13}}\rho_1\\ &\qquad+\frac{1}{\Delta_{13}}\int_{\mathbb{R}}\rho_{13}\{a_{13}\sigma_{13}x_3^2 + (a_{11}x_1+a_{12}x_2)\sigma_{13}x_3\\ &\qquad -[\sigma_{13}\mu_3+\sigma_{33}(x_1-\mu_1)]a_{13}x_3\}dx_3.\end{aligned}$$

We need to find $\int_{\mathbb{R}}x_3\rho_{13}dx_3$ and $\int_{\mathbb{R}}x_3^2\rho_{13}dx_3$. Since (x_1,x_3) is a bivariate Gaussian,

$$x_3|x_1 \sim N\left(\mu_3 + \frac{\sigma_{13}}{\sigma_{11}}(x_1-\mu_1),\ \frac{\Delta_{13}}{\sigma_{11}}\right),$$

we thence have

$$\int_{\mathbb{R}}\rho_{13}x_3dx_3 = \rho_1\int_{\mathbb{R}}\rho_{3|1}x_3dx_3 = \rho_1\left(\mu_3+\frac{\sigma_{13}}{\sigma_{11}}(x_1-\mu_1)\right),$$

$$\begin{aligned}\int_{\mathbb{R}}\rho_{13}x_3^2dx_3 &= \rho_1\int_{\mathbb{R}}\rho_{3|1}x_3^2dx_3\\ &= \rho_1\left[\frac{\Delta_{13}}{\sigma_{11}} + \left(\mu_3+\frac{\sigma_{13}}{\sigma_{11}}(x_1-\mu_1)\right)^2\right].\end{aligned}$$

Substituting back, we obtain

$$\begin{aligned}&\int_{\mathbb{R}}\frac{\partial F_1\rho_{\not 2}}{\partial x_1}dx_3\\ &\quad= a_{11}\rho_1 - \frac{\sigma_{13}\mu_3+\sigma_{33}(x_1-\mu_1)(a_{11}x_1+a_{12}x_2)}{\Delta_{13}}\rho_1\\ &\qquad+a_{13}\sigma_{13}\left(\frac{\Delta_{13}}{\sigma_{11}} + \left[\mu_3+\frac{\sigma_{13}}{\sigma_{11}}(x_1-\mu_1)\right]^2\right)\frac{\rho_1}{\Delta_{13}}\\ &\qquad+\{(a_{11}x_1+a_{12}x_2)\sigma_{13} - [\sigma_{13}\mu_3+\sigma_{33}(x_1-\mu_1)]a_{13}\}\\ &\qquad\times\left[\mu_3+\frac{\sigma_{13}}{\sigma_{11}}(x_1-\mu_1)\right]\frac{\rho_1}{\Delta_{13}}.\end{aligned}$$

Thus

$$\begin{aligned}T_{2\to1} &= -E\frac{1}{\rho_1}\frac{\partial F_1\rho_{13}}{\partial x_1}dx_3\\ &= -a_{11} - \frac{1}{\Delta_{13}}\Big\{-\sigma_{13}\mu_3a_{11}\mu_1 - \sigma_{13}\mu_3a_{12}\mu_2\\ &\qquad -a_{11}\sigma_{33}\sigma_{11} - a_{12}\sigma_{33}\sigma_{12}\end{aligned}$$

$$
\begin{aligned}
&+a_{13}\sigma_{13}\frac{\Delta_{13}}{\sigma_{11}}+a_{13}\sigma_{13}\left[\mu_3^2+\left(\sigma_{13}^2/\sigma_{11}^2\right)\sigma_{11}\right]\\
&+a_{11}\sigma_{13}\mu_3\mu_1\\
&+a_{12}\mu_3\sigma_{13}\mu_2-a_{13}\sigma_{13}\mu_3^2-0+\left(a_{11}\sigma_{13}^2/\sigma_{11}\right)\sigma_{11}\\
&+\left(a_{12}\sigma_{13}^2/\sigma_{11}\right)\sigma_{12}-0-\left(a_{13}\sigma_{33}\sigma_{13}/\sigma_{11}\right)\sigma_{11}\Bigg\}\\
=\;&a_{12}\frac{\sigma_{12}}{\sigma_{11}}.
\end{aligned}
$$

Here so many terms are canceled out, and the result turns out to be precisely the same as that for the 2D case we have derived before ever since Liang and Kleeman [46] in 2005.

The above remarkably concise formula actually holds for systems of arbitrary dimensionality. This makes the following theorem:

Theorem VII.1. If an n-dimensional ($n \geqslant 2$) vector of random variables $(x_1,\ldots,x_n)^T$ evolves subject to the linear system

$$d\mathbf{x} = \mathbf{A}\mathbf{x}dt + \mathbf{B}d\mathbf{w},$$

where $\mathbf{A} = (a_{ij})$ and $\mathbf{B}$ are constant matrices, and if its covariance matrix is (σ_{ij}), then the information flow from x_j to x_i is

$$T_{j\to i} = a_{ij}\frac{\sigma_{ij}}{\sigma_{ii}}, \tag{102}$$

for any $i,j = 1,\ldots,n$, $i \neq j$.

Proof. It suffices to prove the case $(i,j) = (1,2)$; if not, we may always reorder the components to make them so. We prove by induction. The 3D case has just been shown above. Now suppose (102) holds for n-dimensional systems. Consider an n+1-dimensional system

$$
\begin{aligned}
\frac{dx_1}{dt} &= \sum_{j=1}^{n} a_{1j}x_j + a_{1,n+1}x_{n+1}\\
&\vdots \qquad\qquad\qquad \vdots\\
\frac{dx_2}{dt} &= \sum_{j=1}^{n} a_{nj}x_j + a_{n,n+1}x_{n+1}\\
\frac{dx_{n+1}}{dt} &= \sum_{j=1}^{n} a_{n+1,j}x_j + a_{n+1,n+1}x_{n+1}.
\end{aligned}
$$

To distinguish, we now use ρ^n to denote the joint density for the n-dimensional system. The information flow from x_2 to x_1 is

$$
\begin{aligned}
T_{2\to 1} &= -\int_{\mathbb{R}^{n+1}} \rho_{2|1}(x_2|x_1)\frac{\partial F_1\rho_{\S}}{\partial x_1}d\mathbf{x}\\
&= \int_{\mathbb{R}^{n+1}} \rho_{2|1}\frac{\partial}{\partial x_1}\left[\left(\sum_{j=1}^{n} a_{1j}x_j\right)\rho_{\S} + (a_{1,n+1}x_{n+1})\rho_{\S}\right]d\mathbf{x}\\
&= \int_{\mathbb{R}^{n}} \rho_{2|1}\frac{\partial}{\partial x_1}\left(\sum_{j=1}^{n} a_{1j}x_j\rho_{\S}^n\right)\\
&\quad + \int_{\mathbb{R}^{n+1}} \rho_{2|1}\frac{\partial}{\partial x_1}(a_{1,n+1}x_{n+1}\rho_{\S})d\mathbf{x}.
\end{aligned}
$$

Note the first term results from integration with respect x_{n+1}, since all the variables except $\rho_{\S}$ are independent of x_{n+1}. This is precisely the information flow from x_2 to x_1 for an n-dimensional system; by our assumption it is $a_{12}\sigma_{12}/\sigma_{11}$. For the second term, note that all variables, except $\rho_{2|1}$, are independent of x_2, so we may take the integral with respect to x_2 directly inside with $\rho_{2|1}$. But $\int_{\mathbb{R}} \rho_{2|1}dx_2 = 1$, so the second term results in the integral of $\frac{\partial}{\partial x_1}(a_{1,n+1}x_{n+1}\rho_{\S})$ which vanishes by the compactness of ρ. Therefore (102) holds for n+1-dimensional systems. By induction, it holds for systems of arbitrary dimensionality. ■

Let us see an example: $\mathbf{A} = \begin{bmatrix} 1 & -2 & 0\\ 1 & 0 & -5\\ -1 & 2 & -1 \end{bmatrix}$, and $\mathbf{B} = \begin{bmatrix} 1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 3 \end{bmatrix}$. In component form, the equation is

$$\frac{dx_1}{dt} = x_1 - 2x_2 + 0x_3 + \dot{w}_1, \tag{103}$$

$$\frac{dx_2}{dt} = x_1 + 0x_2 - 5x_3 + 2\dot{w}_2, \tag{104}$$

$$\frac{dx_3}{dt} = -x_1 + 2x_2 - x_3 + 3\dot{w}_3. \tag{105}$$

The evolution of the covariance matrix $\mathbf{C}$ is governed by

$$\frac{d\mathbf{C}}{dt} = \mathbf{A}\mathbf{C} + \mathbf{C}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T. \tag{106}$$

Let it be initialized by $\begin{bmatrix} 1 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 9 \end{bmatrix}$. The solution is shown in Fig. 10.

The rates of information flow are subsequently obtained and plotted in Fig. 11. Among them, $T_{3\to 1} = 0$, just as expected by the principle of nil causality. $T_{3\to 2}$ and $T_{2\to 3}$ oscillate around a value near zero, and $T_{2\to 1}$ oscillates around -0.9. The remaining transfers, $T_{1\to 2}$ and $T_{1\to 3}$, albeit still oscillatory, approximately approach two constant values. The former approaches 0.16, while the latter approaches 1.

In applied sciences, it has been a common practice to infer causality via analyzing correlation, coherence, etc. However, there has also been a long-time debate about whether this makes sense. Though it has been generally agreed, based on philosophical and/or physical arguments, that correlation is not equivalent to causation, a clear-cut statement about their relation is yet to be found. Here, Theorem VII.1 says that, in the linear sense, the relation can be analytically expressed. Specifically, $\sigma_{ij} = 0$ implies $T_{j\to i} = 0$, but the converse is not true due to the existence of another term a_{ij}. That is to say, two uncorrelated events do not have a causality in between, but not vice versa. Contrapositively, this means that causation implies correlation, but correlation does not imply causation. Although this holds only in the linear limit, the implication is far-reaching, considering that when "correlation" is mentioned, we usually talk about "linear correlation" characterized by Pearson's population correlation coefficient $r = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$.

VIII. CONCLUSIONS AND DISCUSSION

Information flow, or information transfer as it may appear in the literature, is a fundamental notion in general physics

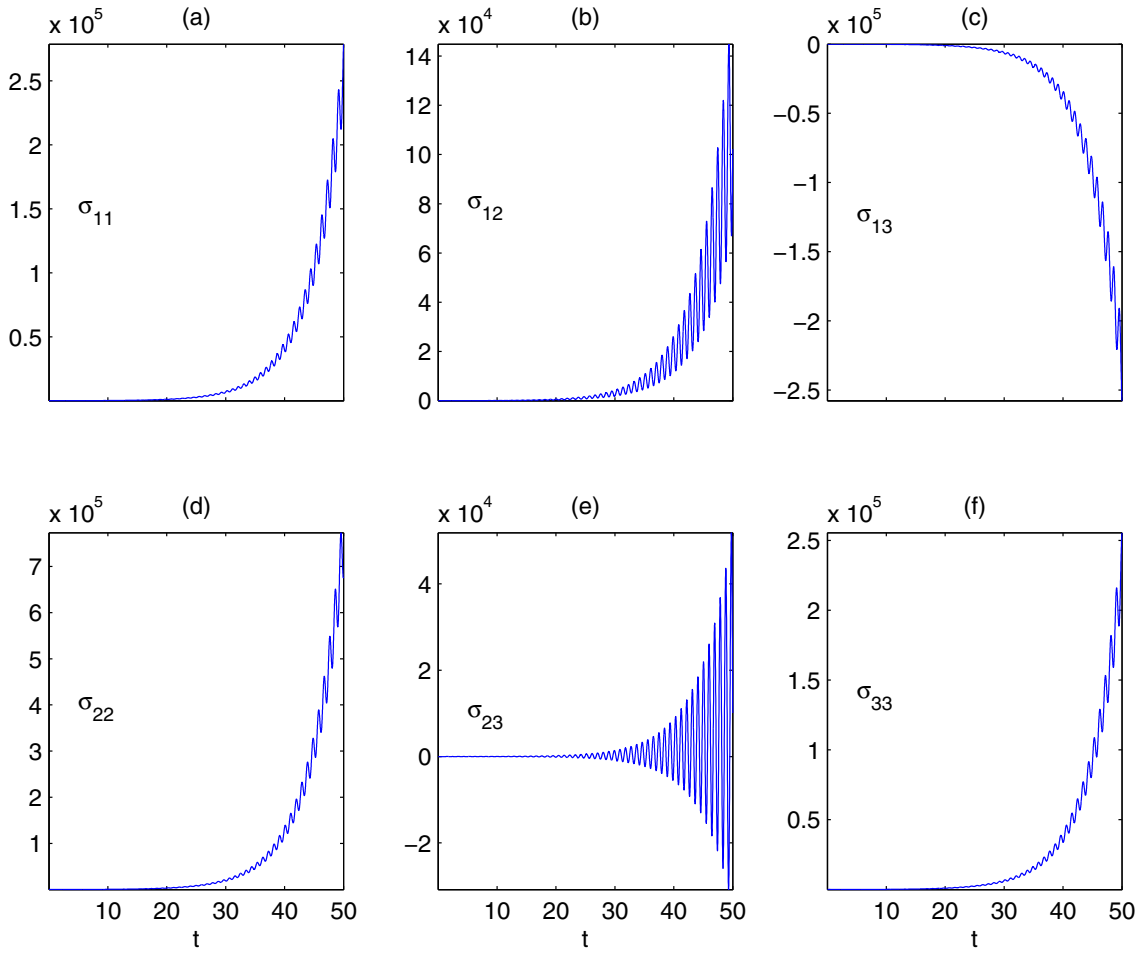

FIG. 10. Covariance evolution with the linear system (103)–(105).

which has wide applications in different disciplines. In this study we have shown that, within the framework of dynamical systems, it can be rigorously derived from first principles. That is to say, it is a notion *ab initio*, quite different from the existing axiomatic postulates or empirical proposals. In this light we have studied the information flow for both time-discrete and time-continuous differentiable vector fields in both deterministic and stochastic settings. In a nutshell, the results can be summarized as follows.

Consider an n-dimensional state variable $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the corresponding probability density function (pdf) being $\rho(x_1, x_2, \ldots, x_n)$, and the marginal pdf of x_i being ρ_i. For a deterministic mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^n$,

$$\mathbf{x}(\tau) \mapsto \mathbf{x}(\tau + 1) = (\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), \ldots, \Phi_n(\mathbf{x})),$$

the rate of information flowing from x_2 to x_1 proves to be

$$T_{2\to 1} = E \log(\mathscr{P}_{\hat{2}}\rho)_1(\Phi_1(\mathbf{x})) - E \log(\mathscr{P}\rho)_1(\Phi_1(\mathbf{x})),$$

where E is the mathematical expectation with respect to $\mathbf{x}$, $\mathscr{P}$ the Frobenius-Perron operator of Φ, and $\mathscr{P}_{\hat{2}}$ the same operator of Φ but with x_2 frozen as a parameter [so $(\mathscr{P}_{\hat{2}})_{1(x_1)}$ has no dependence on x_2]. The units are in nats per unit time; the same below. If the system is continuous in time, i.e.,

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}, t),$$

then

$$\begin{aligned} T_{2\to 1} &= -\int_{\mathbb{R}^n} \rho_{2|1} \frac{\partial \rho_{\hat{2}} F_1}{\partial x_1} d\mathbf{x} \\ &= -E\left[\frac{1}{\rho_1} \int_{\mathbb{R}^{n-2}} \frac{\partial \rho_{\hat{2}} F_1}{\partial x_1} dx_3 \ldots dx_n\right], \end{aligned}$$

where $\rho_{\hat{2}} = \int_{\mathbb{R}} \rho(x_1, x_2, \ldots, x_n) dx_2$, and $\rho_{2|1}$ is the conditional pdf of x_2 on x_1. When stochasticity comes in, in the discrete mapping case,

$$\mathbf{x}(\tau + 1) = \Phi(\mathbf{x}(\tau)) + \mathbf{B}(\mathbf{x})\mathbf{w},$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ is an n-dimensional mapping, $\mathbf{B}$ an $n \times m$ constant matrix, and $\mathbf{w}$ an m-dimensional standard Wiener process, then

$$\begin{aligned} T_{2\to 1} = {} & E_x[\log E_w(\mathscr{P}_{\Phi_{\hat{2}}}\rho)_1(y_1 - \mathbf{B}_1\mathbf{w})] \\ & - E_x[\log E_w(\mathscr{P}_{\Phi}\rho)_1(y_1 - \mathbf{B}_1\mathbf{w})], \end{aligned}$$

with $\mathbf{B}_1 = (b_{11}, b_{12}, \ldots, b_{1m})$ a row vector of the matrix $\mathbf{B}$. Here we use E_x and E_w to indicate that the expectation is taken

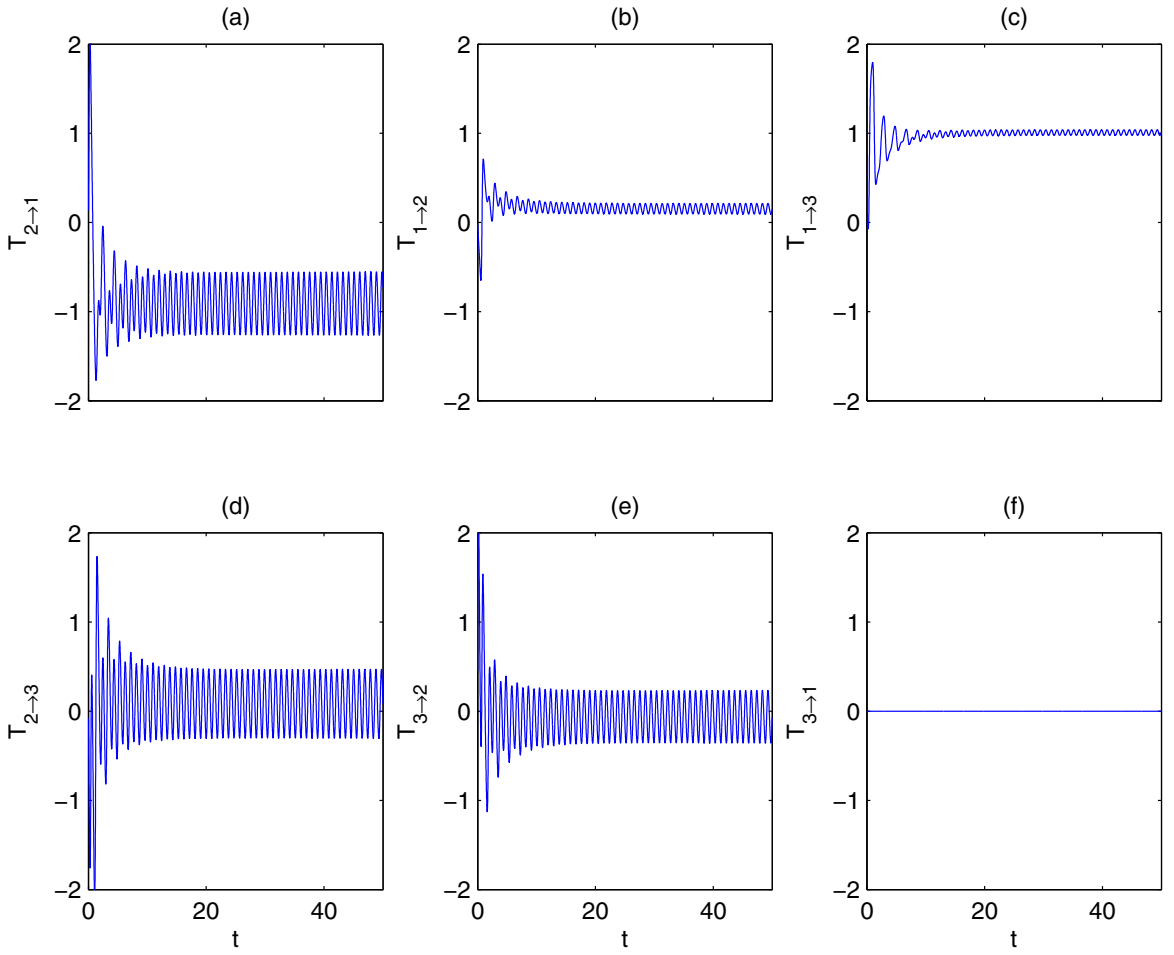

FIG. 11. As Fig. 10, but for rates of information flow.

with respect to x and w, respectively. If what we consider is a continuous-time stochastic system, i.e., a system as

$$d\mathbf{x} = \mathbf{F}(\mathbf{x},t)dt + \mathbf{B}d\mathbf{w},$$

or alternatively written as

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x},t) + \mathbf{B}\dot{\mathbf{w}},$$

where $\dot{\mathbf{w}}$ is the white noise ($\mathbf{B}$ need not be constant), then the result can be explicitly evaluated:

$$\begin{aligned} T_{2\to 1} &= -E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\mathbb{x}}}{\partial x_1}dx_3\ldots dx_n\right] \\ &\quad +\frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\mathbb{x}}}{\partial x_1^2}dx_3\ldots dx_n\right] \qquad (107) \\ &= -\int_{\mathbb{R}^n}\rho_{2|1}(x_2|x_1)\frac{\partial F_1\rho_{\mathbb{x}}}{\partial x_1}d\mathbf{x} \\ &\quad +\frac{1}{2}\int_{\mathbb{R}^n}\rho_{2|1}(x_2|x_1)\frac{\partial^2 g_{11}\rho_{\mathbb{x}}}{\partial x_1^2}d\mathbf{x}, \qquad (108) \end{aligned}$$

where $g_{11} = \sum_{j=1}^{m} b_{1j}b_{1j}$. Note the first term is just from the deterministic vector field, while the second the contribution from the noise. It has been proved that, if b_{1j} has no dependence on x_2, then the stochastic contribution vanishes, making the information flow the same in form as that from its deterministic counterpart. We have particularly examined the case $\mathbf{F} = \mathbf{A}\mathbf{x}$, i.e., the case when the system is linear and autonomous,

$$d\mathbf{x} = \mathbf{A}\mathbf{x}dt + \mathbf{B}d\mathbf{w}$$

with $\mathbf{A} = (a_{ij})_{n\times n}$ and $\mathbf{B} = (b_{ij})_{n\times m}$ being constant matrices; then the information flow from x_j to x_i is remarkably simple:

$$T_{j\to i} = a_{ij}\frac{\sigma_{ij}}{\sigma_{ii}},$$

for any (i,j), $1 \leqslant i,j \leqslant n$, $i \neq j$. This result is precisely the same in form as what Liang and Kleeman obtained in 2005 for 2D deterministic systems based on intuitive arguments [46].

The above results have been put to applications with a variety of benchmark systems. Particularly we have reexamined the baker transformation, Hénon map, and truncated Burgers-Hopf system. The results are qualitatively similar to what we have obtained before using an approximate formalism, but with magnitudes significantly smaller. Also shown are the information flows within a Kaplan-Yorke map, a noisy Hénon map, a Rössler system, and a stochastic gradient flow. We look forward to more applications in the near future.

Historically it has been a long-time endeavor to relate information flow to causality. We want specifically to have that, if $T_{j\to i} \neq 0$, then x_j causes x_i; otherwise x_j is not causal. With

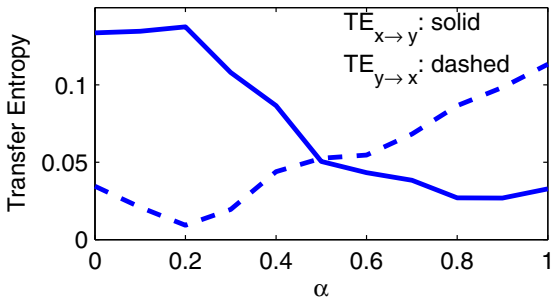

FIG. 12. The transfer entropy (TE) (in nats) between the components of the anticipatory system (109) and (110) as a function of α (a reproduction of the Fig. 1a of [44]). Ideally $\mathrm{TE}_{y\to x}$ (dashed line) should be zero, but it is not. With our formalism, $T_{y\to x}$ is always zero, as guaranteed by the *principle of nil causality*, i.e., Theorem III.4.

the existing empirical/half-empirical measures for information flow, such as the widely used transfer entropy (denoted as TE), the endeavor has been fruitful for some problems but unsuccessful for others (e.g., [45]), and the inconsistency has even led people to doubt about the association between information flow and causality (e.g., [39]). An excellent example is the anticipatory system introduced by Hahs and Pethel [44]:

$$x_{n+1} = f(x_n), \tag{109}$$

$$y_{n+1} = 0.7f(y_n) + 0.3[(1-\alpha)f(x_n) + \alpha f(f(x_n))], \tag{110}$$

where $f(x) = 4x(1-x)$ is the chaotic logistic map. Obviously x_n is the drive and y_n the response. If transfer entropy were a faithful measure of information flow/transfer, then ideally we would have $\mathrm{TE}_{y\to x} = 0$. However, as shown in Fig. 12, this is not the case. Moreover, as the parameter α (called "anticipation" in [44]) increases from 0 to 1, $\mathrm{TE}_{y\to x}$ (dashed line) even exceeds $\mathrm{TE}_{x\to y}$ (solid line).

The above example, among others [45], is a disaster to transfer entropy analysis and the like such as Granger causality test. In contrast, this is not at all a problem with our rigorous formalism. By Theorem III.4, $T_{y\to x} \equiv 0$, which gives the precise result just as expected. This reflects how our formalism differs fundamentally from transfer entropy and other formalisms of the like: For any dynamical system, the implied causality, the touchstone one-way causality in particular, is a proved fact as stated in various theorems, rather than something to be verified in applications. More specifically, when the evolution of x_i does not depend on x_j, then $T_{j\to i} = 0$. This is particularly clear in the above linear case; the dependence of x_i on x_j is from the entry a_{ij} of $\mathbf{A}$, so when it is zero, then x_j is not causal to x_i. This result also quantitatively, and unambiguously, tells us that causation implies correlation, but not vice versa, resolving the long-standing debate over correlation versus causation.

The derivation of information flow is based on the time rate of change of entropy. One naturally wonders what it would be if the system is stationary, since in that case the entropy change vanishes. This is actually not a problem, considering that an information flow rate $T_{2\to 1}$ can be negative as well as positive; a negative $T_{2\to 1}$ means that x_2 acts to reduce the marginal entropy of x_1, i.e., H_1. In the decomposition (3), for example, it is very possible that dH_1^*/dt and $T_{2\to 1}$ cancel out to make a zero dH_1/dt. But this does make an issue in normalizing the obtained information flow; a detailed study can be found in [59]. In some sense, that an information flow rate can be both positive and negative makes a point in which our formalism differs from transfer entropy, though in inferring causality only the magnitude (absolute value) is needed.

In our previous formalism for 2D systems, we have shown that the information flow rates actually can be estimated from time series through maximum likelihood estimation. Specifically, it has been established that [60], for two time series x_1 and x_2, under the assumption of a linear model, the maximum likelihood estimator of the information flow rate from x_2 to x_1 is

$$\hat{T}_{2\to 1} = \frac{C_{11}C_{12}C_{2,d1} - C_{12}^2 C_{1,d1}}{C_{11}^2 C_{22} - C_{11}C_{12}^2}, \tag{111}$$

where $\mathbf{C} = (C_{ij})$ is the sample covariance matrix between time series x_1 and x_2, and $C_{i,dj}$ the sample covariance between x_i and a series derived from x_j using the Euler forward differencing scheme: $\dot{x}_{j,n} = (x_{j,n+1} - x_{j,n})/\Delta t$ [or $\dot{x}_{j,n} = (x_{j,n+2} - x_{j,n})/2\Delta t$ in some special cases, where Δt is the time step size]. This remarkably concise formula has been successfully applied to many real world problems, such as the relation between CO_2 and global warming [61], financial economics [59], etc. And, even though it is under an assumption of linearity, it proves to be very successful [60] in the highly chaotic anticipatory system problem as shown above in Fig. 12. We are therefore working on the estimation of the information flows from time series for systems of arbitrary dimensionality.

ACKNOWLEDGMENTS

This study was partially supported by the 2015 Jiangsu Program for Innovation Research and Entrepreneurship Groups, by the Jiangsu Chair Professorship Program, and by the National Program on Global Change and Air-Sea Interaction (GASI-IPOVAI-06).

APPENDIX: AN ALTERNATIVE PROOF OF THEOREM IV.1

For the continuous system

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x},t), \tag{A1}$$

consider an interval $[t, t+\Delta t]$, and a mapping

$$\Phi : \mathbb{R}^n \to \mathbb{R}^n, \qquad \mathbf{x}(t) \mapsto \mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \mathbf{F}\Delta t.$$

Recall that by definition $E\psi(\mathbf{x}(t+\Delta t)) = \int \psi(\mathbf{x})\rho(\mathbf{x}, t+\Delta t)d\mathbf{x}$, for any test function ψ, and

$$\begin{aligned} E\psi(\mathbf{x}(t+\Delta t)) &= E\psi(\mathbf{x}(t) + \mathbf{F}\Delta t + o(\Delta t)) \\ &= E[\psi(\mathbf{x}(t)) + \nabla\psi \cdot \mathbf{F}\Delta t + o(\Delta t)]. \end{aligned}$$

Note the expectation on the left hand side is with respect to $\rho(t+\Delta t)$, and that on the right is with respect to $\rho(t)$. This way we obtain the Liouville equation.

Now let ψ be the functional $\log(\mathscr{P}_{\hat{\rho}}\rho)_1$. When x_2 is frozen, on interval $[t,t+\Delta t]$, there is a Liouville equation

$$\frac{\partial \rho_{\hat{\rho}}}{\partial t}+\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}+\frac{\partial F_3\rho_{\hat{\rho}}}{\partial x_3}+\ldots+\frac{\partial F_n\rho_{\hat{\rho}}}{\partial x_n}=0 \qquad \text{(A2)}$$

for $\rho_{\hat{\rho}}$ the joint density of $(x_1,x_3,\ldots,x_n)$. The equation for its marginal density $\rho_{1\hat{\rho}}=\int_{\mathbb{R}^{n-2}}\rho_{\hat{\rho}}dx_3\ldots dx_n$ is, after integration with respect to $(x_3,x_4,\ldots,x_n)$ and with the consideration of the compact support assumption,

$$\frac{\partial \rho_{1\hat{\rho}}}{\partial t}+\frac{\partial}{\partial x_1}\int_{\mathbb{R}^{n-2}}F_1\rho_{\hat{\rho}}dx_3\ldots dx_n=0.$$

Divided by $\rho_{1\hat{\rho}}$, this yields

$$\frac{\partial \log\rho_{1\hat{\rho}}}{\partial t}+\int_{\mathbb{R}^{n-2}}\frac{1}{\rho_{1\hat{\rho}}}\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}dx_3\ldots dx_n=0.$$

Discretizing, and noticing the fact $\rho_{1\hat{\rho}}(t)=\rho_1(t)$,

$$\begin{aligned}&\log\rho_{1\hat{\rho}}(t+\Delta t;x_1)\\&=\log\rho_1(t;x_1)-\Delta t\int_{R^{n-2}}\frac{1}{\rho_{1\hat{\rho}}}\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}dx_3\ldots dx_n+o(\Delta t),\end{aligned}$$

which is $\log(\mathscr{P}_{\hat{\rho}}\rho)_{1(x_1)}$. As conventional, let $\mathbf{x}(t+\Delta t)\equiv\mathbf{y}$ and leave $\mathbf{x}$ for $\mathbf{x}(t)$ to avoid confusion. We actually need to find

$$\begin{aligned}\log(\mathscr{P}_{\hat{\rho}}\rho)_1(y_1)&=\log\rho_{1\hat{\rho}}(t+\Delta t;y_1)\\&=\log\rho_1(t;x(t+\Delta t))\\&\quad-\Delta t\int_{\mathbb{R}^{n-2}}\frac{1}{\rho_{1\hat{\rho}}}\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}dx_3\ldots dx_n+o(\Delta t)\\&=\log\rho_1(t;x)+\frac{\partial\log\rho_1}{\partial x_1}F_1\Delta t\\&\quad-\Delta t\int_{\mathbb{R}^{n-2}}\frac{1}{\rho_{1\hat{\rho}}}\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}dx_3\ldots dx_n+o(\Delta t).\end{aligned}$$

Taking expectation and multiplying by (-1) on both sides, we obtain

$$\begin{aligned}H_{1\hat{\rho}}(t+\Delta t)&=H_1(t)-\Delta tE\left(F_1\frac{\partial\log\rho_1}{\partial x_1}\right)\\&\quad+\Delta tE\int_{\mathbb{R}^{n-2}}\frac{1}{\rho_1}\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}dx_3\ldots dx_n.\end{aligned}$$

So

$$\begin{aligned}\frac{dH_{1\hat{\rho}}}{dt}&=\lim_{\Delta t\to0}\frac{H_{1\hat{\rho}}(t+\Delta t)-H_1(\Delta t)}{\Delta t}\\&=E\int_{\mathbb{R}^{n-2}}\frac{1}{\rho_1}\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}dx_3\ldots dx_n-E\left(F_1\frac{\partial\log\rho_1}{\partial x_1}\right).\end{aligned}$$

On the other hand, from the Liouville equation it is easy to obtain

$$\frac{dH_1}{dt}=\int_{\mathbb{R}^n}\log\rho_1\frac{\partial F_1\rho}{\partial x_1}d\mathbf{x}. \qquad \text{(A3)}$$

Hence

$$\begin{aligned}T_{2\to1}&=\frac{dH_1}{dt}-\frac{dH_{1\hat{\rho}}}{dt}\\&=\int_{\mathbb{R}^n}\log\rho_1\frac{\partial F_1\rho}{\partial x_1}d\mathbf{x}-E\int_{\mathbb{R}^{n-2}}\frac{1}{\rho_1}\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}dx_3\ldots dx_n\\&\quad+E\left(F_1\frac{\partial\log\rho_1}{\partial x_1}\right)\\&=-E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\hat{\rho}}}{\partial x_1}dx_3\ldots dx_n\right], \qquad \text{(A4)}\end{aligned}$$

which is the same as (48) in Theorem IV.1.

[1] K. Hlavackova-Schindler, M. Palus, M. Vejmelka, and J. Bhattacharya, Causality detection based on information-theoretic approaches in time series analysis, Phys. Rep. **441**, 1 (2007).

[2] J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, 2000).

[3] P. Sprites, C. Glymour, and R. Scheines, *Causation, Prediction, and Search* (MIT Press, Cambridge, 2001).

[4] B. P. Bezruchko and D. A. Smirnov, *Extracting Knowledge from Time Series: An Introduction to Nonlinear Empirical Modeling* (Springer, Berlin, 2010).

[5] T. Schreiber, Measuring Information Transfer, Phys. Rev. Lett. **85**, 461 (2000).

[6] X. S. Liang, Uncertainty generation in deterministic fluid flows, Dyn. Atmos. Oceans **52**, 51 (2011).

[7] R. Kleeman, Information flow in ensemble weather prediction, J. Atmos. Sci. **64**, 1005 (2007).

[8] T. Schneider, and S. M. Griffes, A conceptual framework for predictability studies, J. Clim. **12**, 3133 (1999).

[9] X. S. Liang, Local predictability and information flow in complex dynamical systems, Physica D **248**, 1 (2013).

[10] E. Pereda, R. Quian Quiroga, and J. Bhattacharya, Nonlinear multivariate analysis of neurophysiological signals, Progr. Neurobiol. **77**, 1 (2005).

[11] K. J. Friston, L. Harrison, and W. Penny, Dynamic causal modeling, NeuroImage **19**, 1273 (2003).

[12] B. Schelter, M. Winterhalder, M. Eichler, M. Peifer, B. Hellwig, B. Guschlbauer, C. Leucking, R. Dahlhaus, and J. Timmer, Testing for directed influences among neural signals using partial directed coherence, J. Neurosci. Methods **152**, 210 (2006).

[13] M. Staniek and K. Lehnertz, Symbolic Transfer Entropy, Phys. Rev. Lett. **100**, 158101 (2008).

[14] R. G. Andrzejak and T. Kreuz, Characterizing unidirectional couplings between point processes and flows, Europhys. Lett. **96**, 50012 (2011).

[15] S. Stramaglia, G. R. Wu, M. Pellicoro, and D. Marinazzo, Expanding the transfer entropy to identify information circuits in complex systems, Phys. Rev. E **86**, 066211 (2012).

[16] J. Wu, X. Liu, and J. Feng, Detecting causality between different frequencies, J. Neurosci. Methods **167**, 367 (2008).
[17] R. Marschinski and H. Kantz, Analysing the information flow between financial time series: An improved estimator for transfer entropy, Eur. Phys. J. B **30**, 275 (2002).
[18] S. S. Lee, Jumps and information flow in financial markets, Rev. Financ. Stud. **25**, 439 (2012).
[19] W. Wang, B. T. Anderson, R. K. Kaufmann, and R. B. Myneni, The relation between the North Atlantic Oscillation and SSTs in the North Atlantic basin, J. Climate **17**, 4752 (2004).
[20] J. Runge, J. Heitzig, N. Marwan, and J. Kurths, Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy, Phys. Rev. E **86**, 061121 (2012).
[21] G. Tissot, A. Lozano-Durán, L. Cordier, J. Jiménez, and B. R. Noack, Granger causality in wall-bounded turbulence, J. Phys.: Conf. Ser. **506**, 1 (2014).
[22] X. S. Liang and A. Lozano-Durán, A preliminary study of the causal structure in fully developed near-wall turbulence, in *Proceedings of the Summer Program 2016*, (Center for Turbulence Research, Stanford University, CA, in press).
[23] R. Sun, A neural network model of causality, IEEE Transactions on Neural Networks **5**, 604 (1994).
[24] N. Ay and D. Polani, Information flows in causal networks, Adv. Complex Syst. **11**, 17 (2008).
[25] L. Sommerlade, M. Eichler, M. Jachan, K. Henschel, J. Timmer, and B. Schelter, Estimating causal dependencies in networks of nonlinear stochastic dynamical systems, Phys. Rev. E **80**, 051128 (2009).
[26] M. Timme and J. Casadiego, Revealing networks from dynamics: An introduction, J. Phys. A: Math. Theor. **47**, 343001 (2014).
[27] A. S. Pikovsky, M. G. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences* (Cambridge University Press, Cambridge, 2001).
[28] K. Lehnertz, S. Bialonski, M.-T. Horstmann, D. Krug, A. Rothkegel, M. Staniek, and T. Wagner, Synchronization phenomena in human epileptic brain networks, J. Neurosci. Methods **183**, 42 (2009).
[29] S. Boccaletti, J. Kurths, G. Osipov, D. Valladares, and C. Zhou, The synchronization of chaotic systems, Phys. Rep. **366**, 1 (2002).
[30] E. Mosekilde, Yu. Maistrenko, and D. Postnov, *Chaotic Synchronization: Applications to Living Systems* (World Scientific, Singapore, 2002).
[31] G. V. Osipov, J. Kurths, and C. Zhou, *Synchronization in Oscillatory Networks* (Springer, Berlin, 2007).
[32] A. Balanov, N. Janson, D. Postnov, and O. Sosnovtseva, *Syncrhonization: From Simple to Complex* (Springer, Berlin, 2008).
[33] N. J. Corron and S. D. Pethel, Information flow in synchronization, in *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 4 (IEEE, 2005), p. 3546.
[34] C. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica **37**, 424 (1969).
[35] L. Barnett, A. B. Barrett, and A. K. Seth, Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables, Phys. Rev. Lett. **103**, 238701 (2009).
[36] J. A. Vastano and H. L. Swinney, Information Transport in Spatiotemporal Systems, Phys. Rev. Lett. **60**, 1773 (1988).
[37] J. Sun and E. Bolt, Causation entropy identifies indirect influences, dominance of neighbors, and anticipatory couplings, Physica D **267**, 49 (2014).
[38] P. Duan, F. Yang, T. Chen, and S. L. Shah, Direct causality detection via the transfer entropy approach, IEEE Trans. Control Systems Tech. **21**, 2052 (2013).
[39] J. T. Lizier and M. Prokopenko, Differentiating information transfer and causal effect, Eur. Phys. J. B **73**, 605 (2010).
[40] C. W. J. Granger, Testing for causality: A personal viewpoint, J. Econ. Dyn. Control **2**, 329 (1980).
[41] C. A. Sims, Discrete approximations to continuous time distributed lags in econometrics, Econometrica **39**, 545 (1971).
[42] D. A. Smirnov and B. P. Bezruchko, Spurious causalities due to low temporal resolution: Towards detection of directional coupling from time series, Europhys. Lett. **100**, 10005 (2012).
[43] H. Nalatore, M. Ding, and G. Rangarajan, Mitigating the effects of measurement noise on Granger causality, Phys. Rev. E **75**, 031123 (2007).
[44] D. W. Hahs and S. D. Pethel, Distinguishing Anticipation from Causality: Anticipatory Bias in the Estimation of Information Flow, Phys. Rev. Lett. **107**, 128701 (2011).
[45] D. A. Smirnov, Spurious causalities with transfer entropy, Phys. Rev. E **87**, 042917 (2013).
[46] X. S. Liang and R. Kleeman, Information Transfer between Dynamical System Components, Phys. Rev. Lett. **95**, 244101 (2005).
[47] X. S. Liang, The Liang-Kleeman information flow: Theory and application, Entropy **15**, 327 (2013).
[48] X. S. Liang and R. Kleeman, A rigorous formalism of information transfer between dynamical system components. I. Discrete mapping, Physica D **231**, 1 (2007).
[49] X. S. Liang and R. Kleeman, A rigorous formalism of information transfer between dynamical system components. II. Continuous flow, Physica D **227**, 173 (2007).
[50] A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics* (Springer, New York, 1994).
[51] J. L. Kaplan and J. A. Yorke, *Functional Differential Equations and Approximations of Fixed Points*, Lecture Notes in Mathematics, Vol. 730 (Springer-Verlag, 1979).
[52] O. E. Rössler, An equation for continuous chaos, Phys. Lett. A **57**, 397 (1976).
[53] Y. Bar-Yam, *Dynamics of Complex Systems* (Addison-Wesley Press, Reading, MA, 1997).
[54] J. P. Crutchfield, The calculi of emergence: Computation, dynamics, and induction induction, Physica D **75**, 11 (1994).
[55] J. Goldstein, Emergence as a construct: History and issues, Emerg. Complex. Org. **1**, 49 (1999).
[56] J. C. McWilliams, The emergence of isolated, coherent vortices in turbulence flows, J. Fluid Mech. **146**, 21 (1984).
[57] P. A. Corning, The re-emergence of emergence: A venerable concept in search of a theory, Complexity **7**, 18 (2002).

[58] R. V. Abramov, G. Kovacic, and A. J. Majda, Hamiltonian structure and statistically relevant conservative quantities for the truncated Burgers-Hopf equation, Commun. Pure Appl. Math. **56**, 1 (2003).

[59] X. S. Liang, Normalizing the causality between time series, Phys. Rev. E **92**, 022126 (2015).

[60] X. S. Liang, Unraveling the cause-effect relation between time series, Phys. Rev. E **90**, 052150 (2014).

[61] A. Stips, D. Macias, C. Coughlan, E. Garcia-Gorriz and X. S. Liang, On the causal structure between CO_2 and global temperature, Sci. Rep. **6**, 21691 (2016).