Transfer entropy computation using the Perron-Frobenius operator

David Diego, Kristian Agasøster Haaga, and Bjørte Hannisdal
Department of Earth Science, University of Bergen
PO Box 7803, NO-5020 Bergen, Norway
(Dated: April 3, 2019)

We propose a method for computing the transfer entropy between time series using Ulam’s approximation of the Perron-Frobenius (transfer) operator associated with the map generating the dynamics. Our method differs from standard transfer entropy estimators in that the invariant measure is estimated not directly from the data points but from the invariant distribution of the transfer operator approximated from the data points. For sparse time series and low embedding dimension, the transfer operator is approximated using a triangulation of the attractor, whereas for data-rich time series or higher embedding dimension we use a faster grid approach. We compare the performance of our methods with existing estimators such as the k nearest neighbors method and kernel density estimation method, using coupled instances of well known chaotic systems: coupled logistic maps and a coupled Rössler-Lorenz system. We find that our estimators are robust against moderate levels of noise. For sparse time series with less than a hundred observations and low embedding dimension, our triangulation estimator shows improved ability to detect coupling directionality, relative to standard transfer entropy estimators.

I. INTRODUCTION

Time series analysis is used to study the dynamics of complex systems across many disciplines, including macro-scale activity of the brain [1] and interactions in the global climate system [2]. A long-standing problem in time series analysis is the detection of causal connections between different components of a system from observed time series. Several theoretical frameworks have been proposed to address this problem [3], including information-theoretic approaches [4, 5]. A popular information-theoretic method is the so-called transfer entropy (TE) [6], or conditional mutual information [7], which quantifies whether knowledge of changes in one variable reduces uncertainty about changes in another variable. For deterministic systems, the concept of information entropy relies on the existence of invariant densities associated to attractors [8]. Standard methods for computing mutual information (and from it, TE), estimate the invariant distribution directly from the embedding of the data. For instance, the k nearest neighbors method (kNN) [9], which uses counting of nearest neighbors, or approaches using visitation frequency, either directly [6] or through kernel density estimation (KDE) [10]. Other TE estimators use the concept of permutation entropy [11]. In our work, we propose to compute TE (or any standard information theoretic measure [12]) between time series, based on a numerical approximation of the Perron-Frobenius (transfer) operator of the underlying dynamics. The transfer operator [13] dictates how densities in phase space are transformed under the action of maps and its approximation has been used to identify long-term emergent behavior in dynamical systems. Applications include computation of the stretching rate of chaotic maps, and identification of attracting regions in ocean circulation [14, 15]. The transfer operator enables estimation of invariant distributions from which information entropies may be computed [16–18]. Transfer operator approximation may also be used to generate dynamically informed surrogates for null-hypothesis testing, and to interpolate and forecast time series. These possibilities are explored in a forthcoming study [19].

In the following, we describe our approach starting with the general notion of TE, and a brief overview of entropies in the context of deterministic systems (for a more in-depth review, see [3, 8]).

II. TRANSFER ENTROPY

Suppose that for two variables, $X_1$ and $X_2$, one is given the probability density $P(X_1 = x_1, X_2 = x_2)$ with support $S$. From this density one may compute their mutual information [12, 20] as

$$I(X_1, X_2) = \int_S P(x_1, x_2) \log \frac{P(x_1, x_2)}{P(x_1) P(x_2)}. \quad (1)$$

For the case of three variables, $X_1, X_2$ and $X_3$, a related quantity is the TE

$$TE_{X_1, X_2|X_3} = \int_S P(x_1, x_2, x_3) \log \frac{P(x_1|x_2, x_3)}{P(x_1|x_3)}. \quad (2)$$

* david.castro@uib.no; diegocastro.david@gmail.com
† https://www.earthsystemevolution.com;
https://github.com/kahaaga/
One can easily check the identity
\[ TE_{X_1, X_2|X_3} = I(X_1, X_{23}) - I(X_1, X_2), \]
with \( X_{23} = (X_2, X_3) \). \( TE_{X_1, X_2|X_3} \) thus quantifies the
amount of information shared between \( X_1 \) and \( (X_2, X_3) \) beyond the
information already shared between \( X_1 \) and \( X_2 \). The TE was originally introduced in the context of
time series analysis [6] as a way of estimating the
information transfer from one time series to another.
Given two time series \( X \) and \( Y \), the TE measures how
much information is lost by assuming that the the variables \( X \) and \( Y \) are independent, i.e., by assuming that
\( P(x(t + \tau)|x(t), y(t)) = P(x(t + \tau)|x(t)) \). There is, however, no absolute scale associated with information
entropy [12, 20], hence the TE only determines whether
the information transfer from \( Y \) to \( X \) is greater than in the
opposite direction.

III. MEASURE THEORETIC ENTROPIES FOR
DETERMINISTIC SYSTEMS

In the context of dynamical systems, mutual information
(or any information theoretic measure) between variables may be computed from the density distribution of
invariant measures associated with the attractor of the
dynamical system. In the following, we briefly review
these concepts.

A. Attracting sets and attractors

Assume the dynamics is generated by a diffeomorphism
[21] \( \psi : \mathbb{R}^n \to \mathbb{R}^n \). A set \( A \subset \mathbb{R}^n \) is said to be an
attracting set for \( \psi \) if the following conditions are met
[22]:

1. There is an open set \( U \supset A \) and a natural number
   \( N \) such that for any open set \( V \supset A \), \( \psi^m(U) \subset V \)
   for all \( m \geq N \).
2. \( \psi(A) \subset A \)

Here \( \psi^m \) denotes the \( m \)-fold iterate of \( \psi \). The open set
\( U \) is called a fundamental neighborhood of \( A \). This definition implies [22] that \( A = \cap_{m \geq 1} \psi^m(U) \)
and \( \psi(A) = A \). Moreover, if there is an open set \( U \subset \mathbb{R}^n \) such that for all
\( m \) big enough \( \psi^m(U) \) has compact adherence contained in \( U \), then \( A = \cap_{r \geq 1} \psi^r(U) \) is a compact attracting set
with fundamental neighborhood \( U \). Because the open set \( V \) can be arbitrarily small around \( A \), all the trajectorys entering \( U \) asymptotically approach \( A \). In addition, \( B = \cup_{m \geq 0} \psi^{-m}(U) \), where \( \psi^{-m} \) denotes the pre-image
of the \( m \)-fold iterate of \( \psi \), is such that for any \( p \in B \)
there is \( m \) with \( \psi^m(p) \in U \). Thus, the corresponding orbit approaches \( A \) asymptotically. \( B \) is called the basin
of attraction of \( A \) and if \( B = \mathbb{R}^n \), \( A \) is called a global
attracting set. An attractor is, however, a somewhat
more restrictive concept than an attracting set. Roughly
speaking, an attractor is what is left of an attracting set
after removing the wandering points (see [22] for a precise
definition). Trivial examples of attractors are asymptotically
stable fixed points while less trivial ones are stable
limit cycles or quasi periodic limiting orbits [23]. More
complicated ('strange') attractors contain unstable orbits
(i.e. sensitivity to initial conditions) and usually fractal geometries. Most dynamical systems associated
with natural processes, even simple processes involving very
few variables, give rise to highly complex dynamics in
the form of strange attractors [24–26].

B. Invariant measures and ergodicity

The trajectory of a typical orbit of a dynamical system having an attractor generates a distribution of
points in the phase space with a certain density which
seems to be intrinsic to the system. Different portions
of the attractor are visited by the orbit with different
frequency, and this frequency of visitations naturally
defines a density on the attractor, clearly invariant
under the dynamics. The notion of invariance leads to
the notion of ergodicity. Intuitively, a dynamical system
is said to be ergodic if a generic trajectory fills in the
attractor (according to the above notion of invariant
density). A crucial result pertaining to ergodic systems
is the celebrated Birkhoff’s ergodic theorem:

Given a space \( M \) and a map \( h : M \to M \), let \( \mu \)
be measure on \( M \) such that \( \mu \) is invariant under \( h \).
Then for any \( \phi : M \to \mathbb{R} \), measurable, it holds that
\[ \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(h^k(x)) = \int_M \phi \, d\mu, \]
for \( \mu \)-almost every \( x \in M \).

Applied to the characteristic function [27] of any
measurable set, \( K \), the theorem implies that the
measure of \( K \), \( \mu(K) \), equals the frequency of visits to
\( K \), in concordance with the above notion of invariant
measure. In appendix B, we show the equality between
time and spatial averages obtained from the estimates
of invariant measures considered in this work.

Assuming that a density, \( \delta \), of \( \mu \) is well defined for
\( \mu \)-almost all points [28] \( (\mu(K) = \int_K \delta \, dm) \), with \( dm \) the
Euclidean (Riemannian) volume element induced on the
attractor), any standard information theoretic entropy [12] can be computed from $\delta$.

Axiom-A systems are known to possess an (unique) invariant measure of physical significance [29] (compatible with the volume measure on the attractor). The property of being axiom-A refers to the existence of a continuous and invariant splitting of the tangent space, at each point of the attractor, into stable and unstable directions [30]. This property is difficult (if not impossible) to check from an observed times series. However, a system with a sufficiently large number of degrees of freedom, and in a stationary state, can be regarded, for the purpose of computing macroscopic properties, as a smooth dynamical system with a transitive axiom-A global attractor [31].

**C. Estimation of the transfer operator and invariant measures**

Let $\psi : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable invertible and $\mathcal{A} \subset \mathbb{R}^n$ a compact attractor with $m$ denoting the induced volume measure on $\mathcal{A}$. Suppose that $\mu$ is a measure compatible with $m$, with support contained in $\mathcal{A}$ and having density $\delta$ with respect to $m$. The map $\psi$ acts on the measure as $(\psi_*\mu)(K) := \mu(\psi^{-1}(K))$, for any measurable set $K \subset \mathcal{A}$. Thus its density is modified as $(\psi_*\delta)(x) := |d_x\psi^{-1}| \cdot \delta \circ \psi^{-1}(x)$, $|\cdot|$ denoting the absolute value of the determinant. The linear map between functions

$$P(f)(x) := |d_x\psi^{-1}| \cdot f \circ \psi^{-1}(x),$$

is known as the Perron-Frobenius (or transfer) operator associated to the map $\psi$. If $\mu$ is a $\psi$-invariant measure, then

$$\delta(x) = |d_x\psi^{-1}| \cdot \delta \circ \psi^{-1}(x),$$

and thus $\psi$-invariant densities correspond to fixed points of $P$. There is a rich literature on the approximation of the transfer operator and the estimation of invariant measures [14, 15]. Ulam’s method [32] approximates the transfer operator by a row-stochastic Markov matrix acting on distributions defined over a given partition of $\mathcal{A}$. More specifically, let $\{B_1, \cdots, B_N\}$, be a partition of $\mathcal{A}$ into measurable sets and for each $1 \leq a \leq N$, let $\chi_a : \mathbb{R}^n \to \mathbb{R}$ be defined as $\chi_a(x) = 1$ if $x \in B_a$ and 0, otherwise. Any (measurable) function $\rho : \mathbb{R}^n \to \mathbb{R}_+$ can be approximated as constant on each partition element, that is

$$\rho = \sum_{i=1}^N \frac{\int_{B_i} \rho \, dm}{m(B_i)} \chi_i,$$

where $m(B_i)$ denotes the volume of the partition element $B_i$ and $dm$ is the (Lebesgue) volume element. From this piecewise constant approximation and from equation (5) it follows that

$$P(\chi_i) = \sum_{j=1}^N \frac{1}{m(B_j)} \int_{B_j} |d_x\psi^{-1}| \chi_i \circ \psi^{-1} \chi_j$$

$$= \sum_{j=1}^N \frac{\int_{\psi^{-1}(B_j)} \rho \, dm}{m(B_j)} \chi_j$$

$$= \sum_{j=1}^N \frac{m(B_i \cap \psi^{-1}(B_j))}{m(B_i)} \chi_jymj.$$

The linearity of $P$ implies

$$\hat{\rho} := P(\rho) = \sum_{i,j} \frac{1}{m(B_i)} \rho_i \frac{m(B_i \cap \psi^{-1}(B_j))}{m(B_i)} \chi_j,$$

where $\rho_i := \int_{B_i} \rho \, dm$ is the measure of $B_i$ according to the density $\rho$. Taking $\hat{\rho}_i := \int_{B_i} \rho_i \, dm$ as the updated measure of $B_i$, one finds

$$\hat{\rho}_i = \sum_{j} \rho_i \frac{m(B_i \cap \psi^{-1}(B_j))}{m(B_i)}.$$

The row stochastic matrix with entries

$$P^{(N)}_{ij} = \frac{m(B_i \cap \psi^{-1}(B_j))}{m(B_i)},$$

constitutes the Ulam’s approximation to the transfer operator and it approximates how distributions defined over a given partition do change under the map generating the dynamics. Accordingly, its left invariant distribution $(\rho^{(N)} \cdot P^{(N)})$ corresponds to the approximation of the invariant density of the system subject to the partition. From the left invariant distribution of $P^{(N)}$, a measure on $\mathcal{A}$ can be defined as

$$\mu_N(E) = \sum_{i=1}^N \rho_i^{(N)} \frac{m(E \cap B_i)}{m(B_i)}.$$

In [15] it is shown that using a piecewise linear approximation of the map $\psi$, the above sequence of measures, $\{\mu_N\}$, approaches a $\psi$-invariant measure as the partition gets infinitely refined (the maximum size of the sets in the partition approaches 0 as $N \to \infty$). In the following sections we give a detailed description of the implementation of these approximations.
D. Computation of TE

Suppose $X$ and $Y$ are time series of two variables of a dynamical system with attractor $A$. The attractor may be reconstructed using a generalized delay embedding from both time series as $(x^{(i)}(t), y^{(k)}(t))$ [33]. Where $x^{(i)}(t) = (x(t), \ldots, x(t - (l - 1)\tau))$ and $y^{(k)}(t) = (y(t), \ldots, y(t - (k - 1)\tau))$, for appropriate delay $\tau$ and embedding dimension $l + k$. Denote the resulting embedded attractor by $\hat{A}$. The transfer operator and the invariant measure may be approximated using equations (11) and (12) in the $(x^{(i)}, y^{(k)})$ embedding space. Suppose $\hat{\mu}$ is the invariant measure on $\hat{A}$ and $P(x^{(i)}, y^{(k)})$ is the corresponding density of $\hat{\mu}$. The TE from $Y \rightarrow X$, is then

$$TE_{Y \rightarrow X} = \int \text{d}m \frac{P(x^{(j)}|x^{(l-j)}, y^{(k)})}{P(x^{(j)}|x^{(l-j)})} \log \frac{P(x^{(j)}|x^{(l-j)}, y^{(k)})}{P(x^{(j)}|x^{(l-j)})}.$$  

(13)

Notice that this procedure can be easily extended to compute any of the standard information theoretic measures by using the appropriate embedding. For instance, for the conditional TE, $TE_{Y \rightarrow X|Z}$, one may use a generalized embedding of the form $(x^{(i)}(t), y^{(m)}(t), z^{(k)}(t))$ [34].

During the revision of this manuscript we were made aware of the work by Bollt [17], who proposed to use Ulam’s approximation to the transfer operator to estimate the transfer entropy between coupled systems in order to identify synchronization. In his work, Bollt interprets the transfer matrix, that constitutes the Ulam’s approximation, as a conditional probability between states in the phase space and computes the TE using Bayes’s rule. In our work, we follow a different strategy: we use a generalized embedding to approximate the transfer operator and the invariant distribution of this transfer operator is then interpreted as a joint probability on the phase space, from which TE is computed. Our method also differs from Bollt’s in the use of a triangulation estimator for sparse time series, as described in the following section.

E. Numerical implementation

Suppose that $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_N\}$ are time series of two variables of some dynamical system generated by the map $\psi$, and that $TE_{Y \rightarrow X}$ is to be computed. The collection of points $E = \{(x^{(n)}, y^{(n)}(\tau))\}_{n=1}^N$, for $x^{(n)} = (x_n, \ldots, x_{n-l+1})$ (analogously for $y^{(n)}(\tau)$) is a reconstruction of the attractor for suitable time delay $k$ and embedding dimension $j + l + \tau$ [33]. Several methods for estimating both parameters can be found in the literature [3, 35]. Suppose for the moment that the transfer operator has already been approximated using equation (11) and an estimate for an invariant measure for $E$, $\mu$, has been obtained from equation (12). For convenience we relabel the axes corresponding to $x^{(j)}_{n+k}, x^{(l)}_n, y^{(r)}_n$ as $1, 2, 3$, respectively. We use a regular grid into (hyper) rectangular bins, say $\{C_i\}_{i=1,\ldots,J}$, and uniquely decompose each bin index, $i$, into the triplet $(i_1, i_2, i_3)$ (appendix A for details). The integral expression for the TE, equation (2), can then be approximated as

$$TE_{Y \rightarrow X} \simeq \sum_{i_1, i_2, i_3} m(C_i) P(i_1, i_2, i_3) \log \frac{P(i_1|i_2, i_3)}{P(i_1|i_2)},$$

(14)

where $m(C_i)$ is the Euclidean volume of the $i$-th bin and $P(i_1, i_2, i_3) = \frac{\mu(C_i)}{m(C_i)}$ i.e. the density of $\mu$ over the bin $C_i$. Defining $\mu_{i_1,i_2,i_3} = \mu(C_i)$, one easily checks that

$$\sum_{i_1, i_2, i_3} m(C_i) P(i_1, i_2, i_3) \log \frac{P(i_1|i_2, i_3)}{P(i_1|i_2)}$$

$$= -H(i_1, i_2, i_3) - H(i_2) + H(i_1, i_2) + H(i_2, i_3),$$

(15)

where $H(i_1, i_2, i_3)$ denotes the Shannon entropy of the distribution $\mu_{i_1,i_2,i_3}$, $H(i_1, i_2)$ is the entropy of the marginal distribution $\sum_k \mu_{i_1,k,i_3}$, $H(i_2)$ corresponds to the Shannon entropy of the marginal distribution $\sum_i \mu_{i_2,i,i_3}$ and so on.

To estimate the transfer operator and the invariant measure, we propose two different approaches depending on the length of the time series (the number of observations). Denote the set of points in the reconstructed attractor $E$ by $\{p_n\}_{n=1}^N$.

1. Grid estimator.

If the time series $X$ and $Y$ contain a sufficient number of observations [36], the transition matrix in equation (11) can be approximated by a coarse-grained estimation [37] as

$$P_{ij} \simeq \frac{\sharp \{p_n \mid \psi(p_n) \in C_j \cap p_n \in C_i\}}{\sharp \{p_m \mid p_m \in C_i\}},$$

(16)

with $\sharp$ denoting the cardinal. In this case the left invariant distribution $\rho$ from equation (12) coincides with the measure of the bins, that is: $\rho_i = \mu(C_i) = \mu_{i_1,i_2,i_3}$. We
clarify that the name grid estimator refers to the fact that the transfer operator is approximated using a partition into rectangular bins, as opposed to using a partition consisting of simplices, which we consider in the next section. The word grid does not imply that we use a visitation frequency estimator. For time series with a sufficient number of observations, however, the visitation frequency and the grid estimators converge to the same invariant distribution (appendix C). The motivation for obtaining an estimate of the transfer operator is that it provides an approximation to the underlying map that has applications beyond the computation of TE. For the purpose of this study, and as we show in the following sections, computing TE from the transfer operator is advantageous for sparse and noisy time series.

2. Triangulation estimator.

For time series with fewer observations, the estimation of the transfer operator using equation (16) might become inaccurate. Actually, the transfer matrix obtained with the grid method for time series with few observations might fail to be Markovian. This is because the bin containing the last point in the embedding might not contain any other point. In that case such a bin is a sink of information. We thus adopt the method developed in [15], by which the reconstructed attractor $E$ is triangulated into simplices [38] (the vertices of each simplex being points from the embedding $E$). Suppose \{$S_1, \ldots, S_N$\} is such a triangulation. The map $\psi$ is then approximated by a linear map, $\tilde{\psi}$, on each simplex such that if \{$p_{a0}, \ldots, p_{ad}$\} are the vertices of the simplex $S_a$, then \{$p_{a0+1}, \ldots, p_{a(d+1)}$\} are the vertices of its image under the map, $\tilde{\psi}(S_a)$. The transfer matrix is obtained as

$$P_{ab} = \frac{m(S_b \cap \tilde{\psi}(S_a))}{m(\tilde{\psi}(S_a))},$$  \hspace{1cm} (17)

and according to equation (12), the measure of the simplex $a$ is $\rho_a$. To compute the simplex intersection volume we follow a direct approach outlined in appendix D, although several methods for polytope volume computation can be found in the literature [39]. Once we have obtained the invariant distribution over the simplices, finding $\mu(C_i)$ exactly is computationally rather demanding. Instead, we estimate the measure of each bin by evenly sampling each simplex of the triangulation with $M_s$ points, and then assuming that each sampling point carries a fraction $1/M_s$ of the measure of the simplex they belong to. Thus, if the set of sampling points belonging to the bin $C_i$ is formed by $N_a$ points from simplex $a$, for $a = 1, \ldots, N$ (possibly with some $N_a$ being zero), its measure is estimated as

$$\mu(C_i) \approx \sum_{a=1}^{N} \frac{N_a}{M_s} \rho_a. \hspace{1cm} (18)$$

Using this sampling enables a virtually unlimited number of points to estimate the density of the measure. Notice that we do not introduce any bias by doing so (provided the sampling is even over each simplex) because the sampling points do not contain any information beyond that encoded in $\rho$. Using embedding dimension 3 and time series with a few hundred observations, the number of simplices with positive measure out of the triangulation is on the order of hundreds (appendix F). For the examples we study here, TE becomes independent of the (total) number of sampling points beyond $\sim 5000$ [40]. Therefore, $M_s$ will be on the order of tens. We subsample the simplices using a shape preserving simplex splitting routine developed in [41]. This algebraic procedure uses an edge wise splitting factor $r$ and splits a simplex in dimension $d$ into $r^d$ subsimplices, all with the same volume. We use the centroids of the subsimplices resulting from the splitting as the sampling points.

IV. EXAMPLE DYNAMICAL SYSTEMS

We apply our TE estimators to time series generated by coupled instances of well known dynamical systems. Several realizations of the time series are generated from randomly chosen initial values. In all cases, TE in each direction is computed using a bin size adapted to the number of points available [42] and to the size of the reconstructed attractor (appendix A). We will generically denote the TE computed from time series $X$ and $Y$, as $TE_{X \rightarrow Y}$ and $TE_{Y \rightarrow X}$, where $X \rightarrow Y$ corresponds to the direction of the coupling (in the case of unidirectional coupling) or to the direction of the strongest coupling (in the case of bidirectional coupling). In both cases, one expects $TE_{X \rightarrow Y} - TE_{Y \rightarrow X} > 0$. To check the ability of our methods to detect the direction of the coupling between time series, we study the dependence of the average values of $TE_{X \rightarrow Y}$ and $TE_{Y \rightarrow X}$ across realizations, on the number of observations in the time series. We also study the response of our estimators to the strength of the coupling and to observational and dynamical noise. We compare our results with those obtained with the kNN [9] and the KDE [10] estimators. Although these methods, strictly speaking, estimate mutual information, TE can be computed from the identity in equation (3).
Because we are primarily interested in the sensitivity of our new estimators to noise and time series length, and not the absolute value of the TE, we do not apply any bias correction to the estimators [43–45]. In appendix A we detail the embedding used for computing TE for each dynamical system example.

A. Dynamical and measurement noise

Measurement noise is independently added to each time series as follows: given a time series $X$, measurement noise of intensity $\epsilon$ is simulated by adding independently to each observation of $X$ a random value drawn from a Gaussian distribution with zero mean and standard deviation $\sigma$, where $\sigma$ is the standard deviation corresponding to the values in $X$, and $0 \leq \epsilon \leq 1$. For instance, a measurement noise of intensity $\epsilon = 0.1$ will be referred to as 10% measurement noise. Dynamical noise is simulated for each system as explained in the following sections (equations (21), (24) and (31)).

B. Unidirectionally coupled logistic maps (UCLM)

The logistic map is one of the hallmarks of chaotic behavior in dynamical systems [25]. It was originally proposed by R. May as a model for population growth. Here we consider two logistic maps unidirectionally coupled, given by

\[
x(n + 1) = 3.78 \, x(n) [1 - x(n)] ,
\]

\[
y(n + 1) = 3.66 \, f_n [1 - f_n] ,
\]

\[
f_n = \frac{y(n) + c \, x(n)}{1 + \epsilon} .
\]

We also allow for the presence of a moderate level of dynamical noise by modifying the sequence $f_n$ as

\[
\tilde{f}_n = \frac{y(n) + c \, x(n) + \epsilon \, \xi}{1 + c(1 + \epsilon)} ,
\]

where $\xi$ is a random number drawn from $[0, 1]$ with a flat distribution and $0 \leq \epsilon \leq 0.5$.

C. Bidirectionally coupled logistic maps (BCLM)

In this case we consider the system generated by the map

\[
x(n + 1) = \frac{3.78 \, x(n) [1 - x(n)] + 0.03 \, y(n)^2}{1.03} ,
\]

\[
y(n + 1) = \frac{3.66 \, g_n [1 - g_n] + c \, x(n)^2}{1 + c} ,
\]

\[
g_n = \frac{y(n) + 0.06 \, x(n)}{1.06} .
\]

As in the UCLM case, we also allow for the presence of dynamical noise by modifying the sequence $g_n$ as

\[
\tilde{g}_n = \frac{y(n) + 0.06 \, x(n) + \epsilon \, \xi}{1 + 0.06(1 + \epsilon)} ,
\]

where $\xi$ is a random number drawn from $[0, 1]$ with a flat distribution and $0 \leq \epsilon \leq 0.5$.

D. Coupled Rössler-Lorenz system

Historically relevant in the study of chaos are also the Lorenz system [24] and the Rössler system [46]. The former was developed by E. Lorenz in 1963 as a simplified model of viscous fluid flow. O. Rössler proposed his system in 1976 as a simpler version of the Lorenz attractor in order to more easily study its chaotic properties. Here we study a coupled version of both systems also studied in [42, 47], generated by the vector field

\[
\dot{x}_1 = -6 \, (x_2 + x_3) ,
\]

\[
\dot{x}_2 = 6 \, (x_1 + 0.2 \, x_2) ,
\]

\[
\dot{x}_3 = 6 \, (0.2 + x_3 (x_1 - 5.7)) ,
\]

\[
\dot{y}_1 = 10 \, (y_2 - y_1) ,
\]

\[
\dot{y}_2 = y_1 \, (28 - y_3) - y_2 + c \, (x_2)^2 ,
\]

\[
\dot{y}_3 = y_1 \, y_2 - (8/3) \, y_3 .
\]

In this case, the dynamical noise is introduced by modifying the coupling term in the equation (29) above as

\[
c \, (x_2)^2 \, (1 + \epsilon \, \xi)^2 ,
\]

where $\xi$ is a random number drawn from $[-1, 1]$ with a flat distribution and $0 \leq \epsilon \leq 0.5$. 

\[
x(n + 1) = \frac{3.78 \, x(n) [1 - x(n)] + 0.03 \, y(n)^2}{1.03} ,
\]

\[
y(n + 1) = \frac{3.66 \, g_n [1 - g_n] + c \, x(n)^2}{1 + c} ,
\]

\[
g_n = \frac{y(n) + 0.06 \, x(n)}{1.06} .
\]
E. Synchronization

The synchronization threshold for the coupled instances of the logistic maps can be easily estimated from the shape of the attractor itself. It turns out that for both cases the synchronization seems to take place around $c \sim 1$. In the UCLM case, the attractor clearly shrinks to the diagonal for $c \sim 1$ (figure 1). For the BCLM, however, one has a generalized synchronization and hence the attractor does not collapse to the diagonal when synchronization sets in (figure 2). In fact, by taking the limit when $c \to \infty$, the BCLM system reduces to the new dynamical system

$$
\begin{align*}
x(n+1) &= 3.67 x(n) [1 - x(n)] + 0.029 y(n)^2, \quad (32) \\
y(n+1) &= x(n)^2. \quad (33)
\end{align*}
$$

As $c$ increases, the attractor generated by the BCLM approaches the attractor obtained with the system (32)-(33) (figure 3).

For the case of the Rössler-Lorenz system, the generalized synchronization seems to take place around $c \gtrsim 2$ [47]. Indeed, one can observe a great distortion of the usual butterfly shape Lorenz attractor for $c > 2.5$ (figure 4).

V. SENSITIVITY OF TE TO TIME SERIES LENGTH

In this section we study the dependence of the TE computed using our estimators on the number of observations in the time series.
FIG. 4. 3d sections of the Rössler-Lorenz phase space generated from $10^4$-point long orbits without noise and for different values of the coupling: $c = 0$ (a), $c = 1$ (b), $c = 2.5$ (c) and $c = 3.5$ (d).

A. Coupled logistic maps

1. Data-rich time series

In both UCLM and BCLM systems, we compute the TE as a function of the time series length in the range 1000 to 5000 observations and with a low level (10%) of measurement noise. For the UCLM case the coupling constant is set to $c = 0.4$, while for the BCLM instance we set $c = 0.2$.

In both cases of coupled logistic maps, the grid and the kNN estimators seem to be the least sensitive to the number of observations in the time series (figures 5a,c and 6a,c), while the KDE estimator shows a mildly stronger dependence on the time series length. For the grid estimator, and using the adapted bin size described in appendix A, TE for both UCLM and BCLM saturates to a fixed value for time series with more than $\sim 5000$ observations.

2. Sparse time series

In this section we check the ability of our estimators to yield directional asymmetry in the TE for sparse data in the range 50 to 400 observations and adding 10% measurement noise. We also test the stability of the TE against the time series length. For both UCLM and BCLM systems, our estimators yield the correct TE asymmetry ($T_{E_{x\rightarrow y}} > T_{E_{y\rightarrow x}}$) even for time series sparsely sampled with 50 values (figures 7a,b and 8a,b).

FIG. 5. Sensitivity of $TE$ to time series length for the UCLM with coupling constant $c = 0.4$ and 10% measurement noise, using data rich time series. Values are the mean and standard deviation of $T_{E_{x\rightarrow y}}$ (red line) and $T_{E_{y\rightarrow x}}$ (dashed blue line) over 50 realizations, computed with the grid estimator (a), the KDE estimator (b) and the kNN estimator (c).

FIG. 6. Sensitivity of $TE$ to time series length for the BCLM with coupling constant $c = 0.2$ and 10% measurement noise, using data rich time series. Values are the mean and standard deviation of of $T_{E_{x\rightarrow y}}$ (red line) and $T_{E_{y\rightarrow x}}$ (dashed blue line) over 50 realizations, computed with the grid estimator (a), the KDE estimator (b) and the kNN estimator (c).

FIG. 7. Sensitivity of $TE$ to time series length for the UCLM with coupling constant $c = 0.4$ and 10% measurement noise, using sparse time series. Values are the mean and standard deviation of $T_{E_{x\rightarrow y}}$ (red line) and $T_{E_{y\rightarrow x}}$ (dashed blue line) over 50 realizations, computed with the grid estimator (a), the KDE estimator (b) and the kNN estimator (c).

In the case of the BCLM system, the triangulation estimator outcompetes the rest of the estimators at detecting asymmetry in the TE (in the expected direction) for very sparse time series, with less than 100 observations (figure 8). As for the sensitivity of the TE on the number of...
observations, our estimators yield relatively stable TE in the range $100 - 300$ observations (figures 7a,b and 8a,b). The KDE estimator seems to be the least sensitive to the number of observations while the kNN estimator shows the highest sensitivity (figures 7c,d and 8c,d).
We note that the computational requirements of the triangulation estimator become prohibitive for embedding dimensions beyond $\sim 5$. The current implementation of this estimator is based on computing exact simplex volume intersections (appendix D) which suffers from the curse of dimensionality. There is room for optimization, however, through more efficient (approximate) polytope intersection algorithms. Alternatively, a coarse grained sampling of the simplices might also be used to estimate the volume intersection. We have implemented both exact and approximate volume intersection routines in our CausalityTools.jl [48] Julia [49] package, which also provides an implementation of the grid estimator. We also suggest a more efficient alternative in appendix E. We leave a comprehensive investigation of these optimizations for future work.

VI. DEPENDENCE OF TE ON THE COUPLING CONSTANT AND ITS RESPONSE TO NOISE

In this section we study the dependence of the TE computed using our estimators on the coupling constant and its response to observational and dynamical noise. We compare the results with the above standard estimators. Because our main interest in this study is the estimation of TE from sparse time series, we fix the time series length to 100 observations for the coupled logistic maps (both UCLM and BCLM) and to 1000 observations for the the Rössler-Lorenz system.

A. Coupled logistic maps

For the coupled logistic maps we compute the TE for values of the coupling constant in the range 0 to 1.6 in steps of 0.2. We also include dynamical noise as well as measurement noise with intensities ranging from 0 to 0.5 in steps of 0.1. For the case of UCLM without noise, the asymmetry $TE_{x\rightarrow y} - TE_{y\rightarrow x}$, computed with our estimators, starts at zero (or a very small value) for $c = 0$, it then increases up to a maximum value around $c \sim 0.2 - 0.4$ and decreases back to zero as the synchronization triggers for $c \gtrsim 1$ (figures 10a,b). In the UCLM system, synchronization causes the evolution of $y$ to closely follow the evolution of $x$ (figure 1). Accordingly, one expects $TE_{x\rightarrow y} \rightarrow TE_{y\rightarrow x}$ as $c$ increases beyond 1. However, when dynamical noise is added, the effect of the synchronization is attenuated (figures 12a,b). Dynamical noise may be interpreted as a hidden process affecting the system. It is then expected that the synchronization effect breaks down for sufficiently intense dynamical noise. On the other hand, when measurement noise is added, the asymmetry $TE_{x\rightarrow y} - TE_{y\rightarrow x}$ decreases with increasing noise intensity (figure 13), as the effect of the coupling is masked by the noise. Remarkably, the ability of the triangulation estimator to detect the correct directionality of the coupling enhances for low to moderate levels of measurement noise and weak coupling ($c \sim 0.2 - 0.4$) (figure 13b). This finding is congruent with the concept of random perturbation approximation to the map, on which the triangulation estimator is based [15]. In the case of BCLM, synchronization does not decrease the asymmetry between $TE_{x\rightarrow y}$ and $TE_{y\rightarrow x}$. Synchronization reduces the BCLM system to the map in equations (32)-(33). In that limiting case, the coupling in the $x \rightarrow y$ direction is much stronger than in the opposite direction. Both the grid and the triangulation estimators do capture this saturation of the asymmetry in the TE for high values of the coupling constant (figures 11a,b). In contrast, the KDE estimator yields a monotonically decreasing asymmetry $c \gtrsim 1$ (figure 11c) whereas the kNN estimator shows a less obvious decrease (figure 11d). When dynamical noise is included into the BCLM system, the asymmetry $TE_{x\rightarrow y} - TE_{y\rightarrow x}$ responds similarly to that of the UCLM system (figure 14). Observational noise, however, causes the TE asymmetry to decreases with increasing levels of noise, as expected. (figure 15). Although less obvious than for the UCLM system, the triangulation estimator also shows a local maximum in TE asymmetry for weak coupling and moderate observational noise (figure 15b).
FIG. 11. Dependence of $TE_{x\to y} - TE_{y\to x}$ on the coupling constant for the noise-free BCLM. Values are the mean and standard deviation of $TE_{x\to y} - TE_{y\to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c) and the kNN estimator (d).

FIG. 12. Dependence of $TE_{x\to y} - TE_{y\to x}$ on the coupling constant and dynamical noise level for the UCLM. Values are the mean and standard deviation of $TE_{x\to y} - TE_{y\to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c) and the kNN estimator (d).

FIG. 13. Dependence of $TE_{x\to y} - TE_{y\to x}$ on the coupling constant and measurement noise level for the UCLM. Values are the mean and standard deviation of $TE_{x\to y} - TE_{y\to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c) and the kNN estimator (d).

FIG. 14. Dependence of $TE_{x\to y} - TE_{y\to x}$ on the coupling constant and dynamical noise level for the BCLM. Values are the mean and standard deviation of $TE_{x\to y} - TE_{y\to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c) and the kNN estimator (d).

B. Rössler-Lorenz system

For the Rössler-Lorenz system, we compute the TE for values of the coupling constant in the range 0 to 4 in steps of 0.2 also including observational and dynamical noise with intensities ranging from 0 to 0.5 in steps of 0.1. The asymmetry $TE_{x\to y} - TE_{y\to x}$ computed with the grid estimator saturates for high values of dynamical noise and coupling constant (figure 16d), a trait also seen for the kNN estimator (figure 16c). The TE asymmetry computed using the grid estimator increases for strong coupling and high levels of dynamical noise (figure...
FIG. 15. Dependence of $TE_{x \rightarrow y} - TE_{y \rightarrow x}$ on the coupling constant and measurement noise level for the BCLM. Values are the mean and standard deviation of $TE_{x \rightarrow y} - TE_{y \rightarrow x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c) and the kNN estimator (d).

However, the asymmetry saturates for higher levels of dynamical noise (figure 16d). With increasing levels of measurement noise, all estimators yield a decreasing value for the TE asymmetry (figure 17). Remarkably, nonetheless, the TE asymmetry computed with the grid estimator only starts to decrease substantially once the level of measurement noise goes beyond 50% (figure 17d). These findings suggest that our grid estimator is robust to both observational and dynamical noise.

VII. DIRECT VS INDIRECT COUPLING

At the end of section III D, we mentioned that our approach to estimate TE can be easily extended to compute conditional transfer entropy between 3 time series. As an example of this, we apply the grid estimator to a chain of three coupled Lorenz systems, studied in [50]. The flow is generated by the vector field

\[
\begin{align*}
\dot{x}_1 &= 10(y_1 - x_1), \\
\dot{y}_1 &= x_1(28 - z_1) - y_1, \\
\dot{z}_1 &= x_1y_1 - 8/3z_1,
\end{align*}
\]

\[
\begin{align*}
\dot{x}_i &= 10(y_i - x_i) + c(x_{i-1} - x_i), \\
\dot{y}_i &= x_i(28 - z_i) - y_i, \\
\dot{z}_i &= x_iy_i - 8/3z_i,
\end{align*}
\]

with $i = 2, 3$. The direct coupling chain is $x_1 \rightarrow x_2 \rightarrow x_3$. We use coupling constant values in the range 0 to 8 in steps of 0.4 (according to [50], the full synchronization takes place for $c > 8$). For each instance of the coupling constant, we generate 50 orbits starting at randomly cho-
The data generation for this system and the delay embeddings used to compute TE are specified in Appendix A. Our grid estimator detects the direct coupling $x_1 \rightarrow x_2$ for $c \gtrsim 2$ (figure 18a) and the coupling $x_2 \rightarrow x_3$ is detected for $c \gtrsim 4$ (figure 18b). In addition, the grid estimator detects the indirect coupling $x_1 \rightarrow x_3$ for $c \gtrsim 4$ (figure 18c). On the other hand, when the transfer entropy is conditioned on the mediating variable $x_2$, the TE for the indirect coupling $x_1 \rightarrow x_3$ vanishes (figure 18d), which indicates that our method holds some promise for detecting indirect coupling.

FIG. 18. Distinguishing direct from indirect coupling for the chain of coupled Lorenz systems. Values are the mean and standard deviation of TE asymmetry over 50 realizations of the chain of coupled Lorenz systems (equations (34)-(39)). TE asymmetry for the direct coupling $x_1 \rightarrow x_2$ (a), direct coupling $x_2 \rightarrow x_3$ (b), indirect coupling $x_1 \rightarrow x_3$ (c) and conditional TE for the indirect coupling given the mediating variable (d).

VIII. CONCLUSIONS

In this work we propose the computation of transfer entropy (TE) between time series corresponding to variables of some dynamical system, based on a numerical approximation of the Perron-Frobenius operator (transfer operator) associated to the map (or vector field) giving rise to the dynamics. More specifically, the TE is computed using the invariant distribution of the transfer operator. Depending on the number of observations in the time series and the embedding dimension, we propose two methods to estimate the transfer operator. For sparse time series (less than a few hundred points) and low embedding dimension, we use a triangulation of the delay reconstructed attractor to estimate the transfer operator, whereas for data-rich time series (thousands of points) or high embedding dimension we estimate TE using a faster rectangular grid approach. The TE computed using our estimators shows robustness to both observational and dynamical noise, even for a high dimensional system such as the Rössler-Lorenz system. Our results suggest that our estimators of TE are relevant for the detection of causal directionality between sparse and noisy time series, which are commonly encountered in many disciplines.

Computer code required for reproducing the numerical results presented in this work is available in our CausalityTools.jl Julia package, for which the source code is found at https://github.com/kahaaga/CausalityTools.jl.

ACKNOWLEDGEMENTS

This work has been funded by the Bergen Research Foundation and by the Norwegian Research Council grant no. 231259.

Appendix A: Numerical implementation details.

1. Generating time series and embedding

   a. Logistic maps

   To generate the time series for the UCLM and BCLM systems, the variables $x$ and $y$ are sampled every second iterate, after a lapse of $10^3$ iterations. For the computation of $TE_{x \rightarrow y}$, we used the embedding $(y(i+1), y(i), x(i))$ while the computation of $TE_{y \rightarrow x}$ was done with the embedding $(x(i+1), x(i), y(i))$.

   b. Rössler-Lorenz system

   The system of equations (25)-(30) is solved using a 4th order Runge-Kutta routine with time step $dt = 0.005$. The time series are generated by recording the variables every 6 time steps of integration and after an initial lapse of 500 steps, to avoid transients. The embeddings we used in this case were $(y_2(i+3), y_2(i+2), y_2(i+1), x_2(i+2), x_2(i+1), x_2(i))$ for $TE_{x_2 \rightarrow y_2}$ and $(x_2(i+3), x_2(i+2), x_2(i+1), y_2(i+2), y_2(i+1), y_2(i))$ for $TE_{y_2 \rightarrow x_2}$. 
c. Chain of coupled Lorenz systems

The system of equations (34-39) is also solved using a 4th order Runge-Kutta method with the same integration step as for the Rössler-Lorenz system. The variables $x_1$, $x_2$ and $x_3$ are sampled using the same sampling time and initial lapse. To compute $TE_{x_1 \rightarrow x_2}$, with $i \neq j \in \{1, 2, 3\}$, we use the embedding $(x_j(t+3), x_j(t+2), x_j(t+1), x_j(t))$ while for the estimation of the conditional transfer entropy $TE_{x_1 \rightarrow x_3 \mid x_2}$ we use the embedding $(x_3(t+3), x_3(t+2), x_1(t+2), x_2(t+2), x_2(t+1), x_2(t))$. For the case $TE_{x_3 \rightarrow x_1 \mid x_2}$, the embedding is obtained as in the case of $TE_{x_1 \rightarrow x_3 \mid x_2}$ by simply interchanging the roles of $x_3$ and $x_1$.

2. Bin sizes

The choice of the size of the intervals along each axis is adapted to the size of the attractor and the number of points available. Following [42], if $N$ is the number of points furnishing the (embedded) attractor and $d$ is the embedding dimension, the number of intervals along each axis is taken as $N_{\text{int}} = \min \left\{ \text{cell}(N^{1/(d+1)}), n_{\text{max}} \right\}$, where cell($\cdot$) denotes the ceiling and $n_{\text{max}}$ is taken to be 9 for $d = 3$ and 4 for $d = 6$. If $\{p_k\}$ is the set of points furnishing the attractor (it could either be the actual set of points in the embedding or the result of the sampling of the simplices in the triangulation), the size of the intervals along the $a$-th axis is chosen as follows:

Let $O_a = \left(1 - \frac{1}{10N_{\text{int}}}\right) \min \{ |p_i| | 1 \leq i \leq N \}$ and $T_a = \left(1 + \frac{1}{10N_{\text{int}}}\right) \max \{ |p_i| | 1 \leq i \leq N \}$. $O$ will be referred to as the origin of the attractor. The interval size along the $a$-th axis is determined as $\epsilon_a = (T - O)_a / N_{\text{int}}$.

The results for the KDE estimator are obtained using the minimum interval size for each case, that is min $\{\epsilon_a\}$.

3. Constructing the binning

Suppose $E = \{p_1, \cdots, p_N\} \subset \mathbb{R}^d$ is the set of points furnishing the reconstructed attractor (in the case of the grid estimator) or the set of final sampling points (in the case of the triangulation estimator). Let $\epsilon = \{\epsilon_a\}$ and $O = \{O_a\}$ be the bin size and the origin of the attractor (see the previous section). Call $(x_1, \cdots, x_d)$ the coordinate axes on the embedding space, and generically denote as $A_{n+1} := (x_1, \cdots, x_{n_1})$, $A_{n} := (x_{n_1+1}, \cdots, x_{n_2})$ and $B_n := (x_{n_2+1}, \cdots, x_{d})$, the variables on which the transfer entropy $TE_{B \rightarrow A}$ is computed. Each point $p_i \in E$ is assigned a unique triplet of integer tuples $I_i = (i_1, i_2, i_3)$, with $i_1 = (j_{i_1}^1, \cdots, j_{i_1}^d)$, $i_2 = (k_{i_2}^{n_1+1}, \cdots, k_{i_2}^{n_2})$ and $i_3 = (m_{i_3}^{n_2+1}, \cdots, m_{i_3}^{d})$, and such that $O_a + j_{i_1}^a \epsilon_a < \langle p_i \rangle_a \leq O_a + j_{i_1}^a \epsilon_a$, for all $1 \leq a \leq n_1$ (analogously for $i_2$ and $i_3$). The unique elements in the set $\{I_1, \cdots, I_N\}$, say $I_1, I_2, \cdots, I_M$, identify the bins that contain at least one point from the set $E$ and constitute the binning used to compute the transfer operator and the TE (in the case of the grid estimator) and just the TE in the case of the triangulation estimator.

4. Grouping of variables for TE computation

For the case of the coupled logistic maps (both UCLM and BCLM), the TE corresponding to $x \rightarrow y$ uses the grouping of variables $A_{n+1} = (x(n+1))$, $B_n = (y(n))$, while the TE corresponding to $y \rightarrow x$ is computed with the variable grouping $A_{n+1} = (y(n+1))$, $A_n = (x(n))$ and $B_n = (x(n))$.

For the case of the Rössler-Lorenz system, the gathering of variables used to compute $TE_{x_2 \rightarrow x_3}$ is $A_{n+1} = \{y_2(i+3)\}$, $A_n = \{y_2(i+2), y_2(i+1)\}$ and $B_n = \{x_2(i+2), x_2(i+1), x_2(i)\}$. For computing $TE_{x_3 \rightarrow x_2}$ we use the same gathering of variables but interchanging the symbols $x_2$ and $y_2$.

The TE corresponding to the coupling $x_i \rightarrow x_j$ (both direct and indirect) in section VII is computed using the gathering of variables $A_{n+1} = (x_j(t+3))$, $A_n = (x_j(t+2))$ and $B_n = (x_i(t+2), x_i(t+1), x_i(t))$.

5. Computation of conditional TE

Given the variables $A_{n+1}$, $A_n$, $B_n$, and $C_n$, the conditional transfer entropy $TE_{B \rightarrow A|C}$ is computed as

$$\int P(A_{n+1}, A_n, B_n, C_n) \log \frac{P(A_{n+1}|A_n, B_n, C_n)}{P(A_{n+1}|A_n, C_n)} \, dA_n \, dA_{n+1} \, dB_n \, dC_n.$$ 

The conditional $TE_{x_1 \rightarrow x_3 \mid x_2}$ in section VII is computed using the gathering of variables $A_{n+1} = (x_3(t+3))$, $A_n = (x_3(t+2))$, $B_n = (x_2(t+2), x_2(t+1), x_2(t))$. The conditional $TE_{x_3 \rightarrow x_1 \mid x_2}$ is computed using the same gathering of variables but interchanging $x_1$ and $x_3$.

6. $k$ nearest neighbors counting

For the case of kNN estimator of mutual information for dimension 3 or less, we used 5 nearest neighbors to compute $I(a, b, c)$ while 10 nearest neighbors were used to compute $I(a, b)$, where $a$, $b$ and $c$ denote generic variables. For higher dimensions ($\geq 4$), the same number of
nearest neighbors may be taken for both mutual infor-

Appendix B: Ergodicity cross check

As a way of testing the ergodicity of the invariant measure estimated with our method, we compare the temporal and spatial averages of several functions for the UCLM and for the Rössler-Lorenz system. In particular, we consider the functions: \( h_a := \text{sech} \left( \sqrt{x^2 + y^2} \right) \), \( h_b := \beta (1 + x^2, 1 + y^2) \) and \( h_c = \psi \left( \sqrt{x^2 + y^2} \right) \), for the coupled logistic maps, and \( h_d := \psi \left( \sqrt{(x_2)^2 + (y_2)^2} \right) \), \( h_e := \beta (1 + (x_2)^2, 1 + (y_2)^2) \) and \( h_f := \psi \left( \sqrt{(x_2)^2 + (y_2)^2} \right) \), for the Rössler-Lorenz system, where \( \beta(x, y) \) is the Euler \( \beta \) function and \( \psi(x) \) is the digamma function. There is no particular reason behind the choice of these functions, other than being complicated functions having no obvious connection with the systems.

The spatial averages rapidly converge to the temporal averages as the bin size decreases (figure 19). Also, we point out that the rate of convergence seems to be fairly independent of the functions chosen to be averaged. This is maybe not so evident in the case of the Rössler-Lorenz system (figures 19d,e and f) but all the spatial averages seem to saturate beyond \( \sim 15 \) intervals per axis. This rate of convergence is rather dependent on the system and, likely more strongly, on the embedding dimension, suggesting that such a saturation could be used as a criterion for choosing a suitable bin size.

Appendix C: Grid estimator vs visitation frequency estimator

Here we compare the invariant density obtained using the grid estimator with the density that a direct visitation frequency estimation yields, as a function of the number of observations in the time series. For each instance of time series length, say \( n \), we set a bin size (as explained in appendix A) and consider a partition into rectangular bins. We then apply the grid estimator to 50 realizations of time series with \( n \) observations and generated from randomly chosen initial values. Hence, we obtain 50 estimates for the invariant distribution, say \( \rho_{\text{grid}}(r, n) \), for \( 1 \leq r \leq 50 \). Using the same time series, we also compute the visitation frequency to each bin, obtaining thus \( \rho_{vf}(r, n) \). We consider the discrepancy measure

\[
\delta(n) = \frac{1}{50} \sum_{r=1}^{50} \max \left\{ \| \rho_{\text{grid}}(r, n) - \rho_{vf}(r, n) \|, \| \rho_{vf}(r, n) \| \right\}
\]

with \( \| v \| := \max \{|v_k|\} \).

We apply this procedure to our example systems and find that both methods for estimating invariant densities produce the same outcomes (within very small discrepancies) for long enough time series (figure 20). By virtue of the ergodic theorem, the invariant density of the transfer operator and the invariant density yielded by the frequency of visitations must coincide. Hence, the convergence of the density estimates is expected for long enough time series.

Appendix D: Computing the intersecting volume between simplices in dimension \( d \).

We outline the method used in this paper for computing the volume of the intersection between two simplices. Let \( V = \{ p_0, \ldots, p_d \} \) be \( d+1 \) affinely independent points in \( \mathbb{R}^d \). The simplex \( S \) with vertices \( V \) is the convex hull of them, usually denoted as \( S = CH(V) \), and defined as: all the points in \( R^d \) constructed as \( x = \alpha_0 p_0 + \cdots + \alpha_d p_d \), with \( \alpha_i \geq 0 \) and \( \alpha_0 + \cdots + \alpha_d = 1 \). In addition, \( x \) lies in the interior of \( S \), denoted as \( x \in S \), if and only if all \( \alpha_i > 0 \). A generic boundary of \( S \) is the simplex with vertices \( \{ p_{\sigma_0}, \ldots, p_{\sigma_k} \} \), where \( \{ \sigma_0 < \cdots < \sigma_k \} \) is a (non-empty) selection of \( \{0, \cdots, d\} \), for \( k = 0, \cdots, d \). The proper
faces of the simplex correspond to $k = d - 1$. Suppose $S_1 = CH(V_1)$ and $S_2 = CH(V_2)$, with $V_1 = \{p_0, \ldots, p_d\}$ and $V_2 = \{q_0, \ldots, q_d\}$. The method for computing the volume of the intersection $S_1 \cap S_2$ used in this work is based on the following result (the proof of which is given at the end of this appendix):

**Theorem 1.** Let $S_1, S_2 \subset \mathbb{R}^n$ be two simplices of dimensions $n \geq m \geq 1$, respectively, and with $S_1 \cap S_2 \neq \emptyset$. Let $\mathcal{I}$ be the set of points in $\mathbb{R}^n$ constructed as follows: $p \in \mathcal{I}$ if $p \in B_1 \cap B_2$, where $B_1$ and $B_2$ are boundaries of $S_1$ and $S_2$, respectively, and not supporting any common direction. Then it holds that $S_1 \cap S_2 = CH(\mathcal{I})$.

In other words, if $\mathcal{I} = \{x_1, \ldots, x_N\}$ is such a set, then $S_1 \cap S_2$ consists of all the points of the form $\beta_1 x_1 + \cdots + \beta_N x_N$, for $\beta_n \geq 0$ and $\beta_1 + \cdots + \beta_N = 1$.

The set $\mathcal{I}$ may be found as follows: let $B_1 = CH(\{p_{r_1}, \ldots, p_{r_s}\})$ and $B_2 = CH(\{q_{s_1}, \ldots, q_{s_r}\})$ and, with no loss of generality, assume that $r \geq s$. Reorder the vertices of $S_1$ as $\{p_{i_1}, \ldots, p_{i_r}, p_{r+1}, \ldots, p_{r+s}\}$. Every point in $\mathbb{R}^d$ can be expressed as a unique affine linear combination of these vertices, possibly with negative coefficients. In particular, $q_{s_i} = \sum_{j=1}^{d} c_{ji} p_{r+j}$, where the coefficients on the rest of the vertices of $S_1$ are omitted. Denote the least dimensional affine space containing the boundary $B_1$ (respectively $B_2$) as $\Pi_1$ (respectively $\Pi_2$). By definition, the affine spaces generated by the sets of (affinely independent) points $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ intersect uniquely if the equations

$$\sum_{i=1}^{n} \alpha_i x_i = \sum_{j=1}^{m} \beta_j y_j,$$  \hspace{1cm} \text{(D1)}

$$\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{m} \beta_j = 1,$$  \hspace{1cm} \text{(D2)}

have unique solution. In our case, this translates into the conditions

$$\text{rank } \gamma = s - 1$$  \hspace{1cm} \text{(D3)}

$$\text{rank } \begin{pmatrix} \gamma \\ 1 \end{pmatrix} = s$$  \hspace{1cm} \text{(D4)}

where $\gamma$ is the $(d-r+1) \times s$ matrix with entries $\gamma_{ab}$ and 1 denotes a row of ones. If $x$ is the unique intersecting point, then

$$x = \sum_{i=1}^{r} \alpha_i p_{ri} = \sum_{j=1}^{s} \beta_j q_{sj},$$  \hspace{1cm} \text{(D5)}

with $\sum_i \alpha_i = \sum_j \beta_j = 1$, and only if $\alpha_i, \beta_j > 0$, the point $x$ is in the interior of both boundaries and thus an element of $\mathcal{I}$.

Once the set $\mathcal{I}$ has been found it can be further decomposed as $\mathcal{I} = \cup_{a=1,2,\ldots,d} \mathcal{I}_a^{(a)}$ where $\mathcal{I}_a^{(a)} := \mathcal{I} \cap F_a^{(a)}$, and $F_a^{(a)}$ denotes the face of the simplex $S_a$ that lies opposite to the $j$-th vertex of that simplex. Some of these subsets might be empty and there might be repetitions. Call $\mathcal{F}$ the set of the unique elements in the collection $\{\mathcal{I}_a^{(a)}\}_{a=1,2,\ldots,d}$ that appear at most once for $a = 1$ or once for $a = 2$. If the element $\mathcal{I}_i^{(1)}$ appears only for $a = 1$ then it is easy to check that the intersection $F_i^{(1)} \cap S_2$ produces a convex set of dimension $d - 1$, i.e. a proper face of $S_1 \cap S_2$. On the other hand, theorem 1 applied to the simplices $F_i^{(1)}$ and $S_2$ precisely yields the set of vertices $\mathcal{I}_i^{(1)}$. The same holds if the set of vertices $\mathcal{I}_i^{(2)}$ appears only for $a = 2$. In the case where any such set, $\mathcal{I}_i^{(a)}$, appears for both $a = 1$ and $a = 2$, its convex hull is a proper face of $S_1 \cap S_2$ only when the corresponding faces containing $\mathcal{I}_i^{(a)}$ are parallel. Each face of $S_1 \cap S_2$ may be triangulated (some of them might be already simplices) and the triangulation of these faces, together with any point in the interior of $S_1 \cap S_2$, for instance its centroid, form a triangulation of the intersection between the simplices. Its volume is then computed as the sum of the volumes of the simplices in this last triangulation.

### 1. Proof of theorem 1

The statement in theorem 1 is actually a corollary of a well known result on convex geometry:
Theorem 2 (Minkowski). Every convex and compact set in $\mathbb{R}^n$ is the convex hull of its extreme points.

The proof of theorem 2 can be found in [51]. A point $x$ in a convex set $P$ is said to be extreme if the equality $x = \lambda y + (1 - \lambda) z$, for $y, z \in P$ and $0 < \lambda < 1$, requires $x = y = z$. In other words, $x$ is not found in the interior of any segment contained in $P$. For our purposes, it is convenient to use an (easily shown to be) equivalent definition: $x$ is extreme if for every unit vector $u$ and for every $\epsilon > 0$, there is $|\lambda| < \epsilon$, such that $x + \lambda u \notin P$. The set of extreme points of $P$ is denoted as $\text{ext}(P)$.

\textit{(Proof of theorem 1).} We claim that the set $\mathcal{I}$, as defined in theorem 1, equals $\text{ext} (S_1 \cap S_2)$. To see this, let \{p_0, \cdots, p_n\} be the vertices of $S_1$ and \{q_0, \cdots, q_m\} be the vertices of $S_2$ and let $x \in \hat{B}_1 \cap \hat{B}_2$ with $B_1$ and $B_2$, boundaries verifying the properties required in theorem 1. Next, let $u$ be an arbitrary unit vector and assume, without loss of generality (w.l.o.g.), that the direction $u$ is not supported by $\Pi_1$ (the least dimensional affine space supporting $B_1$). Further assume, w.l.o.g., that \{p_0, \cdots, p_r\} are the vertices of the boundary $B_1$. Then $u = \sum_{i=1}^{r} \gamma_i (p_i - p_0) + \sum_{j=r+1}^{n} \mu_j (p_j - p_0)$ with not all $\mu_j$ vanishing. Assume, w.l.o.g., that $\mu_{r+1} \neq 0$. Therefore, $x + \lambda u = x + \lambda \mu_{r+1} p_{r+1} + \sum_{i=1}^{r} \alpha_i (\lambda) p_i + \sum_{j=r+1}^{n} \beta_j (\lambda) p_j$.

Given $\epsilon > 0$ arbitrary, take $\lambda = -\text{sign} (\mu_{r+1}) \frac{\epsilon}{2}$, it then holds that $x + \lambda u = -|\mu_{r+1}| \frac{\epsilon}{2} p_{r+1} + \cdots$ and therefore $x + \lambda u \notin S_1$. This shows that $x$ is an extreme point of $S_1 \cap S_2$ and since $x \in \mathcal{I}$ was arbitrary, it follows that $\mathcal{I} \subset \text{ext} (S_1 \cap S_2)$.

To see the reverse inclusion, let $x \in \text{ext} (S_1 \cap S_2)$ arbitrary. W.l.o.g., assume that $x = \sum_{i=0}^{r} \alpha_i p_i = \sum_{j=0}^{s} \beta_j q_j$ with all $\alpha_i > 0$ and all $\beta_j > 0$. Call $\Pi_r$ the least dimensional affine space containing the vertices \{p_0, \cdots, p_r\} and $\Pi_s$ the affine space with the same property with respect to the vertices \{q_0, \cdots, q_s\}. It holds that $\Pi_r$ and $\Pi_s$ do not support any common direction. Indeed, suppose $u$ is a unit vector along a direction supported by both $\Pi_r$ and $\Pi_s$ then, having that all the coefficients $\alpha_i$ and $\beta_j$ are strictly positive, it follows that for some $\epsilon > 0$ small enough, $x + \lambda u \in S_1 \cap S_2$, for all $|\lambda| < \epsilon$, contradicting that $x$ is an extreme point. This shows that $\mathcal{I} = \text{ext} (S_1 \cap S_2)$. Given that both $S_1$ and $S_2$ are compact and convex sets, so it is $S_1 \cap S_2$ and the proof is completed by using theorem 2.

\textit{Appendix E: Sample-and-filter approach.}

Here we provide a tentative modification of the triangulation estimator in order to reduce its high computational demands. As described in III E 2, the triangulation estimator is based on an initial partition of the embedded attractor into simplices and the map generating the dynamics is approximated linearly onto each simplex [15], say $\tilde{\psi}$. Given the initial triangulation of the attractor, we make use of the piecewise linear approximation of the map to generate sampling points as an input to the grid estimator (equation (16)):

1. Given an embedded attractor in a $d$ dimensional space (we assume it consists of few hundred points), let \{S_1, \cdots, S_N\} be its triangulation into $d$-simplices and \{B_1, \cdots, B_M\} be the set of bins in a regular grid that are visited by the embedded points (figure 21(a)). The size of the bins in the regular grid is adapted to the number of points in the reconstructed attractor (appendix A).

2. Each simplex is sampled with $N_s$ points using a predefined matrix of convex coefficients. More specifically, let $C$ be a $N_s \times (d+1)$ matrix such that $C_{ai} \geq 0$, no two rows are equal and $\sum_{j=1}^{d+1} C_{aj} = 1$, for all $1 \leq a \leq N_s$ and $1 \leq i \leq d + 1$. Let \{v_1, \cdots, v_{d+1}\} be the vertices of the simplex $S_n$ in the triangulation of the attractor and \{\tilde{\psi}(v_1), \cdots, \tilde{\psi}(v_{d+1})\} be the vertices of the simplex being the image of $S_n$ under the map (see III E 2 for details). The $a$-th sampling point of the simplex $S_n$ is given by $p_{n,a} := \sum_{j=1}^{d+1} C_{aj} v_j$ and its image under $\tilde{\psi}$ is given by $\tilde{\psi}(p_{n,a}) = \sum_{j=1}^{d+1} C_{aj} \tilde{\psi}(v_j)$ (the map $\tilde{\psi}$ is linear on each simplex). Call $P := \{p_{n,a}\}_{1 \leq n \leq N_s, 1 \leq a \leq N_s}$, the set of all the sampling points of the simplices in the triangulation and $\tilde{\psi}(P) = \{\tilde{\psi}(p_{n,a})\}$, the set of image points.

3. From $P$, discard all those points not lying in $\bigcup_{i=1}^{M} B_i$. The resulting set, $\hat{P}$, contains (possibly thousands of) points that are distributed more tightly to the volume occupied by the reconstructed attractor (figure 21b).

4. Finally, a new bin size is adapted to the number of points in $\hat{P}$. Their images are found in the set $\tilde{\psi}(P) = \{\tilde{\psi}(p_{n,a})\}_{p_{n,a} \in \bigcup_{i=1}^{M} B_i}$. Therefore, the transfer operator may be approximated using equation (16), which using the above notation reads.
\[
P_{ij} \approx \frac{\sharp \{ p_{n,a} | \tilde{\psi}(p_{n,a}) \in B_j \cap p_{n,a} \in B_i \}}{\sharp \{ p_{m,b} | p_{m,b} \in B_i \}}.
\]

Note that generating sampling points in this manner only assumes piecewise linearity of the map and does not introduce any further bias. We applied this approach to both UCLM and BCLM, using time series with 100 observations and 50 realizations from randomly chosen initial conditions. The dependence of \( \Delta TE := TE_{x \rightarrow y} - TE_{y \rightarrow x} \) on the coupling constant and noise is comparable to that obtained from the grid estimator (figures 22a,b). Interestingly, the values for the \( \Delta TE \) obtained with the sample-and-filter approach are, in general, higher than those obtained with the grid estimator (figure 23). This result suggests that from a sparse time series one may generate thousands of points from a piecewise linear approximation and obtain reliable results for the TE. We also apply the sample-and-filter approach to the Rössler-Lorenz system, in which case the results are less impressive (figure 24). Note however, that these analyses used time series with only 200 observations in embedding dimension 6 (appendix A).

**FIG. 21.** Example of the sampling and filtration of 100 embedding points for the UCLM system with \( c = 0.4, \) no noise and using the embedding \((x(n+1), x(n), y(n))\). Solid black points represent the original embedding points and the grid is shown in green. (a) initial triangulation of the reconstructed attractor. (b) each simplex is sampled with 80 points (small blue points), where those not lying in the grid have been filtered out. The number of points after filtering (small blue points) is about 23000.

**FIG. 22.** Mean value of \( TE_{x \rightarrow y} - TE_{y \rightarrow x} \) over 50 realizations as a function of coupling constant and noise level obtained with the sample-and-filter approach. (a) UCLM with dynamical noise; (b) UCLM with measurement noise; (c) BCLM with dynamical noise; (d) BCLM with measurement noise.

**FIG. 23.** \( \Delta TE(SF) - \Delta TE(grid) \) (SF indicates sample-and-filter method) as a function of the coupling constant and noise level for: (a) UCLM with dynamical noise; (b) UCLM with measurement noise; (c) BCLM with dynamical noise; (d) BCLM with measurement noise.

**Appendix F: Computational times for the triangulation estimator**

In this appendix we provide CPU times required to obtain the transfer operator using the triangulation estimator (figure 25b) as well as the CPU time required for ob-
FIG. 24. Mean value of $TE_{x_2 \rightarrow y_2} - TE_{y_2 \rightarrow x_2}$ over 50 realizations as a function of the coupling constant and noise level for the Rössler-Lorenz system and: (a) with dynamical noise obtained with the sample-and-filter approach; (b) the same as in (a) but with measurement noise instead; (c) obtained with the grid estimator and adding dynamical noise; (d) the same as in (c) but with measurement noise instead.

FIG. 25. (a): mean and standard deviation over 50 realizations of the CPU time needed to obtain the volume of a non-trivial simplex intersection; (b) mean and standard deviation over 10 realizations of the CPU time required to obtain an estimate of the transfer operator for to the UCLM system with $c = 0.4$ and no noise, in embedding dimension 3. The timing values were obtained using the cputime function in Matlab; (c) mean and standard deviation over 50 realizations of the number of simplices in a triangulation in 3d; (d) the same as in (c), fixing the number of observations in the time series to 500.

Intersecting the volume of a (non-trivial) simplex intersection (figure 25a). The simulations were run in Matlab using a MacBook Pro with a 2.8 GHz Intel Core i7 processor. Computation times increase for higher dimension because the CPU time to obtain the volume for a simplex intersection and the number of simplices in the triangulation both scale exponentially (figures 25a,d). In particular, for dimension 5, the CPU time required to obtain the volume of a non-trivial simplex intersection is $t_s \sim 5 \cdot 10^{-2}$ s and the typical number of simplices in a triangulation generated from 500 points is $n_s \sim 5 \cdot 10^4$. If each simplex intersects non-trivially with just 10 of the simplices in the triangulation (which is a quite optimistic estimate) we are left with a computation time for obtaining the transfer operator in the order $t_{TO} \sim 10 n_s t_s \sim 2.5 \cdot 10^3 s \sim 7 h$. For comparison, our grid estimator applied to 500-point long time series in dimension 5 requires a CPU time of $\sim 5 \cdot 10^{-3}$ s to estimate the transfer operator.
view.” (2013), arXiv:1211.3169.
[6] T. Schreiber, Physical Review Letters 85, 461 (2000), arXiv:0001042v1 [nlin].
[7] M. Pulš, V. Komárek, Z. Hrnčí, and K. Štěrbová, Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 63, 046211 (2001).
[8] J.-P. Eckmann and D. Ruelle, Reviews of Modern Physics 57, 1115 (1985).
[9] A. Kraskov, H. Stögbauer, and P. Grassberger, Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 69, 16 (2004), arXiv:0305641 [cond-mat].
[10] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, Bioinformatics 18, 5231 (2002).
[11] M. Staniek and K. Lehnertz, Physical Review Letters 100, 158101 (2008); H. Dickten and K. Lehnertz, Physical Review E 90, 062706 (2014); C. Bandt and B. Pompe, Physical Review Letters 88, 174102 (2002).
[12] T. M. Cover and J. A. Thomas, Elements of information theory, 2nd ed. (Wiley, New Jersey, 2006) p. 748.
[13] A. Lasota and M. C. Mackey, Applied mathematical sciences, 2nd ed., 97 (Springer-Verlag, New York, 1994) p. 472 p., arXiv:arXiv:1011.1669v3; A. Berman and R. J. Plemmons, Nonnegative matrices in the mathematical sciences (Academic Press, New York, 1979) p. 337; C. Beck and F. Schlogl, Thermodynamics of chaotic systems: an introduction (Cambridge University Press, Cambridge, UK, 1993) p. 286.
[14] G. Froyland and K. Padberg-Gehle, Physica D 241, 1612 (2012); P. Miron, F. J. Beron-Vera, M. J. Olascoaga, J. Sheinbaum, P. Pérez-Brunius, and G. Froyland, Scientific Reports 7, 1 (2017), arXiv:1703.10684; G. Froyland, C. Horenkamp, V. Rossi, and E. van Sebille, Chaos: An Interdisciplinary Journal of Nonlinear Science 25, 083119 (2015); G. Froyland, R. M. Stuart, and E. van Sebille, ibid. 24, 033126 (2014); E. Ser-Giacomi, V. Rodríguez-Méndez, C. López, and E. Hernández-García, The European Physical Journal Special Topics 226, 2057 (2017); C. Maes, N. Grima, B. Blanke, E. Martinez, T. Paviet-Salomon, and T. Huck, Geophysical Research Letters 45, 1915 (2018); R. McAdam and E. van Sebille, Journal of Geophysical Research: Oceans 123, 514 (2018).
[15] G. Froyland, Bulletin of the Australian Mathematical Society 56, 157 (1997).
[16] G. Froyland, Nonlinearity 12, 79 (1999).
[17] E. M. Bollt, International Journal of Bifurcation and Chaos 22, 1250261 (2012).
[18] X. S. Liang and R. Kleeman, Physical Review Letters 95, 244101 (2005).
[19] K.A. Haaga, D. Diego and B. Hannisdal. doi:10.17605/OSF.IO/M57EX.
[20] C. E. Shannon, The Bell System Technical Journal 27, 379 (1948).
[21] The same would be true for a dynamical system generated by a vector field.
[22] D. Ruelle, Communications in Mathematical Physics 82, 137 (1981).
[23] A typical example of a quasi periodic orbit is given by the map on the torus defined as $T(x, y) = (x + t, y + r t)$ [each component taken modulo 1] for $r > 0$ non rational and $0 \leq x, y < 1$. It can be shown that such a trajectory returns arbitrarily close and infinitely often to itself, but it never closes up exactly.
[24] E. N. Lorenz, Journal of the Atmospheric Sciences 20, 130 (1963), arXiv:NIHMS150003.
[25] R. M. May, Nature 261, 459 (1976).
[26] D. Ruelle and F. Takens, Les rencontres physiciens-mathématiciens de Strasbourg-RCP25 12, 1 (1971).
[27] The characteristic function of a set $A$ returns 1 if $x$ belongs to $A$ and 0, otherwise.
[28] If $\mu$ assigns positive measure to sets when, and only when, the sets have positive volume (and then $\mu$ is said to be compatible with Lebesgue), a unique density $\delta$ of $\mu$ is guaranteed to exist by the Radon-Nikodym theorem [52]. In the case of dynamical systems, axiom-A maps are known to posses a unique invariant measure compatible with Lebesgue [29].
[29] R. Bowen, Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms, Lecture Notes in Mathematics, Vol. 470 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1975); R. Bowen and D. Ruelle, Inventiones Mathematicae 29, 181 (1975); D. Ruelle, American Journal of Mathematics 98, 619 (1976).
[30] S. Smale, Bulletin of the American Mathematical Society 73, 747 (1967).
[31] G. Gallavotti, Journal of Statistical Physics 84, 899 (1996); Gallavotti and E. G. D. Cohen, ibid. 80, 931 (1995).
[32] S. M. Ulam, Problems in Modern Mathematics (Interscience, New York, 1964) p. 150.
[33] F. Takens, Lecture Notes in Mathematics 598, 366 (1981); T. Sauer, J. A. Yorke, and M. Casdagli, Journal of Statistical Physics 65, 579 (1991); E. R. Deyle and G. Sugihara, PLoS ONE 6, e18295 (2011).
[34] With mild modifications in the derivation of the main result of chapter 4 in [15], it can be shown that the estimate of the invariant measure is independent of the embedding.
[35] A. M. Fraser and H. L. Swinney, Physical Review A 33, 1134 (1986), arXiv:1511.02086; W. Liebert and H. G. Schuster, Physics Letters A 142, 107 (1989); H. Kim, R. Eykholt, and J. Salas, Physica D: Nonlinear Phenomena 127, 48 (1999).
[36] This number is likely system dependent.
[37] M. Dellnitz, G. Froyland, and O. Junge, in Ergodic theory, analysis, and efficient simulation of dynamical systems (Springer, 2001) pp. 145–174.
[38] Built-in routines for Delaunay triangulation in arbitrary dimensions can be found in standard numerical softwares, for instance the Qhull library.
[39] B. Bürer, A. Enge, and K. Fukuda, in Polytopes: Combinatorics and Computation, edited by G. et al. Kalai (Birkhäuser Basel, Basel, 2000) Chap. 6, pp. 131–154.
[40] We expect this number to be dependent on the system and the embedding dimension.
[41] H. Edelsbrunner and D. R. Grayson, Discrete and Computational Geometry 24, 707 (2000).

[42] A. Krakovská, J. Jakubík, M. Chvosteková, D. Coufal, N. Jajcay, and M. Paluš, Physical Review E 97, 042207 (2018).

[43] R. Marschinski and H. Kantz, The European Physical Journal B-Condensed Matter and Complex Systems 30, 275 (2002).

[44] B. Gourévitch and J. J. Eggermont, Journal of Neurophysiology 97, 2533 (2007).

[45] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, An introduction to transfer entropy (Springer, 2016).

[46] O. E. Rössler, Physics Letters A 57, 397 (1976).

[47] R. Q. Quiroga, J. Arnhold, and P. Grassberger, Physical Review E 61, 5142 (2000).

[48] CausalityTools.jl is a registered Julia package and the source code is found at https://github.com/kahaaga/CausalityTools.jl.

[49] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, SIAM review 59, 65 (2017).

[50] D. Kugiumtzis, Physical Review E 87, 062918 (2013).

[51] B. Grünbaum, “Convex polytopes, volume 221 of graduate texts in mathematics,” (2003).

[52] D. L. Cohn, Measure Theory, 2nd ed. (Springer, New York, 2013) p. 466.