entropy

MDPI

Article

Neural Estimator of Information for Time-Series Data with Dependency

Sina Molavipour [1,*], Hamid Ghourchian [1], Germán Bassi [2] and Mikael Skoglund [1]

[1] School of Electrical Engineering and Computer Science (EECS), KTH Royal Institute of Technology, 100 44 Stockholm, Sweden; hamidgh@kth.se (H.G.); skoglund@kth.se (M.S.)
[2] Ericsson Research, 164 83 Stockholm, Sweden; german.bassi@ericsson.com
[*] Correspondence: sinmo@kth.se

Abstract: Novel approaches to estimate information measures using neural networks are well-celebrated in recent years both in the information theory and machine learning communities. These neural-based estimators are shown to converge to the true values when estimating mutual information and conditional mutual information using independent samples. However, if the samples in the dataset are not independent, the consistency of these estimators requires further investigation. This is of particular interest for a more complex measure such as the directed information, which is pivotal in characterizing causality and is meaningful over time-dependent variables. The extension of the convergence proof for such cases is not trivial and demands further assumptions on the data. In this paper, we show that our neural estimator for conditional mutual information is consistent when the dataset is generated with samples of a stationary and ergodic source. In other words, we show that our information estimator using neural networks converges asymptotically to the true value with probability one. Besides universal functional approximation of neural networks, a core lemma to show the convergence is Birkhoff's ergodic theorem. Additionally, we use the technique to estimate directed information and demonstrate the effectiveness of our approach in simulations.

Keywords: neural networks; conditional mutual information; directed information; Markov source; variational bound

check for updates

Citation: Molavipour, S.; Ghourchian, H.; Bassi, G.; Skoglund, M. Neural Estimator of Information for Time-Series Data with Dependency. *Entropy* **2021**, *23*, 641. https://doi.org/10.3390/e23060641

Academic Editor: Friedhelm Schwenker

Received: 26 April 2021
Accepted: 18 May 2021
Published: 21 May 2021

1. Introduction

In recent decades, a tremendous effort has been done to explore capabilities of feed-forward networks and their application in various areas. Novel machine learning (ML) techniques go beyond conventional classification and regression tasks and enable revisiting well-known problems in fundamental areas such as information theory. The functional approximation power of neural networks is a compelling tool to be used for estimating information-theoretic quantities such as entropy, KL-divergence, mutual information (MI), and conditional mutual information (CMI). As an example, MI is estimated with neural networks in [1] where numerical results show notable improvements compared to the conventional methods for high-dimensional, correlated data.

Information-theoretic quantities are characterized by probability densities and most classical approaches aim at estimating the densities. These techniques may vary depending on whether the random variables are discrete or continuous. In this paper, we focus on continuous random variables. Examples of conventional non-parametric methods to estimate these quantities are histogram and partitioning techniques, where the densities are approximated and plugged-in into the definitions of the quantities, or methods based on the distance of the k-th nearest neighbor [2]. Despite vast applications of nearest neighbor methods for estimation of information-theoretic quantities, such as the proposed technique in [3], recent studies advocate using neural networks while simulations demonstrate that the accuracy of the estimations improves in several scenarios [1,4]. In particular, the results indicate that by increasing the dimension of the data, the bias of the estimation deteriorates less with neural estimators. In addition to superior performance, a neural estimator of information can be considered to be a stand-alone block and coupled in a larger network. The estimator can then be trained simultaneously with the rest of the network and measure the flow of information among variables of the network. Therefore, it facilitates the implementation of ML setups with constraints on information measures (e.g., information bottleneck [5] and representation learning [6]). These compelling features motivate exploring the benefits of neural networks to estimate other information measures and more complex data structures.

The cornerstone of neural estimators for MI is to approximate bounds on the relative entropy instead of computing it directly. These bounds are referred to as variational bounds and recently have gained attention due to their applications in ML problems. Examples are the lower bounds proposed originally in [7] by Donsker and Varadhan, and in [8] by Nguyen, Wainwright, and Jordan that are referred to as DV bound and NWJ bound, respectively. Several variants of these bounds have been reviewed in [9]. Variational bounds are tight, and the estimators proposed in [1,4,10,11] leverage this property and use neural networks to approximate the bounds and correspondingly the desired information measure. These estimators were shown to be consistent (i.e., the estimation converges asymptotically to the true value) and suitably estimate MI and CMI when the samples are independently and identically distributed (i.i.d.). However, in several applications such as time series analysis, natural language processing, or estimating information rates in communication channels with feedback, there exists a dependency among samples in the data. In this paper, we investigate analytically the convergence of our neural estimator and verify the performance of the method in estimating several information quantities.

Consider several random processes such that their realizations are dependent in time. In addition to common information-theoretic measures such as MI and CMI, more complex quantities can be studied that are paramount in representing these processes. For instance, the (temporal) causal relationship between two random processes has been expressed with quantities such as directed information (DI) [12,13] and transfer entropy (TE) [14]. Both DI and TE have a variety of applications in different areas. In communication systems, DI characterizes the capacity of a channel with feedback [15], while it has several other applications in venues including portfolio theory [16], source coding [17], and control theory [18] where DI is exploited as a measure of privacy in a cloud-based control setup. Additionally, DI was introduced as a measure of causal dependency in [19] which led to a series of works in that direction with applications in neuroscience [20,21] and social networks [22,23]. TE is also a well-celebrated measure in neuroscience [24,25], and the physics community [26,27] to quantify causality for time series. In this paper, we investigate capability of the neural estimator proposed in [11] to be used when the samples in the data are not generated independently.

Conventional approaches to estimate KL-divergence and MI such as nearest neighbor methods can be used for non-i.i.d. data; for example to estimate DI [28] and TE [29,30]. However, it is possible to leverage the benefits of neural estimators highlighted in [1] even though the data are generated from a source with dependency among its realizations. In a recent work [31], the authors estimate TE using the neural estimator for CMI introduced in [4]. Additionally, recurrent neural networks (RNN) are proposed in [32] to capture the time dependency to estimate DI. However, showing convergence of these estimators requires further theoretical investigation. Although the neural estimators are shown to be consistent in [1,4,11] for i.i.d. data, the extension of the proofs to dependent data needs to be addressed. In [32], the authors address the consistency of the estimation of DI by referring to universal approximation of RNN [33] and Breiman's ergodic theorem [34]. Because RNNs are more complicated to be implemented and tuned, in this paper, we assume simple feed-forward neural networks, which were also proposed in [1,4,11] and in this paper. A conventional step to go beyond i.i.d. processes is to investigate stationary and ergodic Markov processes which have numerous applications in modeling real-world systems. Many convergence results for i.i.d. data such as the law of large numbers can be extended to ergodic processes; however, this generalization is not always trivial. The estimator proposed in [11] exhibits major improvements in estimating the CMI. Nevertheless, it is based on a k-nearest neighbors (k-NN) sampling technique which makes the extension of the convergence proofs to non-i.i.d. data more involved. The main contribution of this paper is to provide convergence results and consistency proofs for this neural estimator when the data are stationary and ergodic Markov.

The paper is organized as follows. Notations and basic definitions are introduced in Section 2. Then, in Section 3, the neural estimator and procedures are explained. Additionally, the convergence of the estimator is studied when the data are generated from a Markov source. Next, we provide simulation results in Section 4 for synthetic scenarios and verify the effectiveness of our technique in estimating CMI and DI. Finally, we conclude the paper in Section 5 and suggest potential future directions.

2. Preliminaries

We begin by describing the notation used throughout the paper, and the main definitions are explained afterwards. Then we review variational bounds which are the basis of our neural estimator.

2.1. Notation

Random variables and their realizations are denoted by capital and lower case letters, respectively. Given two integers i and j, a sequence of random variables $X_i, X_{i+1}, \dots, X_j$ is shown as X_i^j, or simply X^j when $i = 1$. For a stochastic processes $\mathbf{Z}$, a randomly generated sample is denoted by random variable Z. We indicate sets with calligraphic notation (e.g., $\mathcal{X}$). The space of d-dimensional real vectors is shown as $\mathbb{R}^d$. The probability density function (PDF) of a random variable X at $X = x$ is denoted by $p_X(x)$ or equivalently $p(x)$, and the distribution of X, by P_X or simply P. The PDF of multiple random variables $X_1, \dots, X_i$ is $p_{X_1 \dots X_i}(x_1, \dots, x_i)$ and for simplicity it is represented by $p(x_1, \dots, x_i)$ in the paper. For the distribution P, $\mathbb{E}_P[\cdot]$ denotes the expectation with respect to its density $p(\cdot)$. All the logarithms are in base e.

The convergence of the sequence X_n almost surely (or with probability one) to X is denoted by $X_n \overset{\text{a.s.}}{\to} X$ and is defined as:

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X \right) = 1.$$

2.2. Information Measures

The information-theoretic quantities of interest for this work can be written in terms of a KL-divergence, and the available neural estimators originally aim to estimate this quantity. For a random variable X with support $\mathcal{X} \subseteq \mathbb{R}^d$, the KL-divergence between two PDFs $p(x)$ and $q(x)$ is defined as:

$$D(p(x) \parallel q(x)) := \mathbb{E}_P\left[\log \frac{p(X)}{q(X)} \right]. \tag{1}$$

Then, CMI can be defined using KL-divergence as below:

$$I(X;Y|Z) := D(p(x,y,z) \parallel p(x|z)p(y,z)). \tag{2}$$

where Y and Z are random variables with support on $\mathcal{Y}$ and $\mathcal{Z}$, which are subsets of $\mathbb{R}^d$. In this paper, we are focused on extending the estimators for CMI with non-i.i.d. data, where samples in time-series data might not be independently and identically distributed (e.g., generated from a Markov process); nonetheless, our method and consistency proofs are fairly general and can be applied for estimating KL-divergence as well. Consider a sequence of random samples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ generated from the joint process $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, where the samples are not necessarily i.i.d.. A simple step toward this extension is to verify that the previous neural estimators, e.g., [11], can be used to estimate $I(X;Y|Z)$, where $(X,Y,Z) \sim p(x,y,z)$ and the processes $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ are Markov, as in the following assumption.

Assumption 1. $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ *are jointly stationary and ergodic 1-st order Markov with marginal density* $p(x,y,z)$*. The extension of the results to d-th order Markov is straightforward.*

To explore further in generalizing the neural estimators, it is possible to investigate their capability for information measures that rely on dependent random variables. Consider the pairs $\{(X_i,Y_i)\}_{i=1}^n$ to be samples of the processes $(\mathbf{X},\mathbf{Y})$. If the generated samples are dependent in time, it is possible to measure the causal relationship between the processes with quantities such as DI and TE, defined as below:

$$I(X^n \to Y^n) := \sum_{i=1}^{n} I(X^i;Y_i|Y^{i-1}) \tag{3}$$

$$T_{\mathbf{X}\to\mathbf{Y}}(i) := I(X_{i-J}^{i-1};Y_i|Y_{i-L}^{i-1}), \tag{4}$$

where J and L are parameters of the TE that determine the length of memory to consider for $\mathbf{X}$ and $\mathbf{Y}$, respectively. Both quantities are functions of the CMI and Figure 1 visualizes the corresponding variables in each CMI term for DI and TE. In particular, each CMI term in (3) quantifies the amount of shared information between X^i and Y_i conditioned on Y^{i-1}, i.e., it excludes the effect of the causal history of $\mathbf{Y}$. In a general form, to express the causal effect of the process $\mathbf{X}$ on $\mathbf{Y}$ conditioning causally on $\mathbf{Z}$, DI is normalized with respect to n which is defined below and denoted as directed information rate (DIR):

$$\begin{aligned} I(\mathbf{X}\to\mathbf{Y}\parallel\mathbf{Z}) &:= \lim_{n\to\infty}\frac{1}{n}I(X^n\to Y^n\parallel Z^n) \\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I(X^i;Y_i|Y^{i-1},Z^i). \end{aligned} \tag{5}$$

By assuming the processes to be Markov, (5) can be simplified (see [23,35,36]). To be explicit, if both $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ and $(\mathbf{Y},\mathbf{Z})$ are stationary and ergodic 1-st order Markov, from (5) the DIR can be simplified as:

$$I(\mathbf{X}\to\mathbf{Y}\parallel\mathbf{Z}) = I(X^2;Y_2|Y_1,Z^2), \tag{6}$$

where the CMI is with respect to the stationary density $p(x^2,y^2,z^2)$ of the Markov model. To generalize this approach, let us define the *maximum Markov order* ($o_{\max}$) of a set of processes to be the minimum number o such that the Markov order of the joint random variables of any subset of the processes is less than or equal to o. So if $o_{\max} = l$ for $(\mathbf{X},\mathbf{Y},\mathbf{Z})$, then from (5) we can simplify the DIR term as:

$$I(\mathbf{X}\to\mathbf{Y}\parallel\mathbf{Z}) = I(X^{l+1};Y_{l+1}|Y^l,Z^{l+1}). \tag{7}$$

The following example shows how DIR can be computed for a linear data model, and emphasizes on the difference when DIR is conditioned causally on another process.

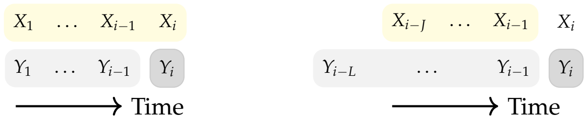

Figure 1. The memory considered for conditional mutual information terms in directed information (**left**) and transfer entropy (**right**) at time instance i. To compute directed information (**left**), the effect of X^i (i.e., X_i and all its past samples) on Y_i is considered, while the history of Y_i is excluded. However, for transfer entropy (**right**), the effect of X_{i-J}^{i-1} (i.e., the previous J samples before X_i) on Y_i is accounted for, while we exclude the history of Y_i. Note that the length of memories (J and L) for transfer entropy may differ.

Example 1. *Consider the following linear model where* $\{W_i\}_{i=1}^{\infty}$, $\{W_i'\}_{i=1}^{\infty}$, *and* $\{W_i''\}_{i=1}^{\infty}$ *are uncorrelated white Gaussian noises with variances* σ_x^2,σ_y^2, *and* σ_z^2 *respectively:*

$$\begin{cases} X_i = W_i \\ Y_i = a\,Y_{i-1} + Z_{i-1} + W_i' \\ Z_i = X_i + W_i'' \end{cases}$$

for some $|a| < 1$, *and* (X_0, Y_0, Z_0) *are distributed according to the stationary distribution of the processes* $\boldsymbol{X}$, $\boldsymbol{Y}$, *and* $\boldsymbol{Z}$. *This model holds in Assumption 1 and* $o_{\max} = 1$, *so* $I(\boldsymbol{X} \to \boldsymbol{Y})$ *can be computed as:*

$$I(\boldsymbol{X} \to \boldsymbol{Y}) = I(X_1^2; Y_2 | Y_1) = \frac{1}{2} \log\left(1 + \frac{\sigma_x^2}{\sigma_y^2 + \sigma_z^2}\right),$$

while from (7):

$$I(\boldsymbol{X} \to \boldsymbol{Y} \,\|\, \boldsymbol{Z}) = I(X_1^2; Y_2 | Y_1, Z_1^2) = 0. \tag{8}$$

As emphasized earlier, (7) holds when $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ and $(\mathbf{Y}, \mathbf{Z})$ are Markov with order l. Then the CMI estimators can be used potentially to estimate the DIR. However, the consistency of the estimation still needs to be investigated since the samples are not independent. Before introducing our technique, we review the basics for estimating information measures with neural networks.

2.3. Estimating the Variational Bound

The estimators proposed in [1,4,11] are all based on tight lower bounds on the KL-divergence, such as the *DV bound*, introduced in [7]:

$$D(p(x) \,\|\, q(x)) \geq \sup_{f \in \mathcal{F}} \mathbb{E}_P\big[f(X)\big] - \log \mathbb{E}_Q\big[\exp(f(X))\big], \tag{9}$$

where p and q are two PDFs defined over $\mathcal{X}$ with corresponding distributions P and Q, respectively, and $\mathcal{F}$ is any class of functions such that $f : \mathcal{X} \to \mathbb{R}$, and the two expectations exist and are finite. Consider a neural network with parameters $\theta \in \Theta$, then $\mathcal{F}$ can be to the class of all functions constructed with this neural network by choosing different values for the parameters θ. In more details, let $f(x)$ to be the end-to-end function of a neural network with parameters $\theta \in \Theta$ and the optimization in the right hand side (RHS) of (9) is equivalent to optimizing over Θ (as performed in [1]). Nevertheless, we can leverage from the fact that the DV bound is tight when the function is chosen as:

$$f^*(x) = \log \frac{p(x)}{q(x)} \qquad \forall x \in \mathcal{X}. \tag{10}$$

Thus, the neural network can approximate $f^*(x)$ directly and the lower bound can be computed accordingly (as performed in [4,11]).

Definition 1. *For the PDFs $p(x,y,z)$ and $p(x|z)p(y,z)$, define the corresponding distributions on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ to be $\tilde{P}$ and $\tilde{Q}$, respectively.*

Since the CMI can be stated as a KL-divergence (2), the DV bound can be defined for CMI as bellow:

$$I(X;Y|Z) \geq \sup_{f \in \mathcal{F}} \mathbb{E}_{\tilde{P}}\big[f(X,Y,Z)\big] - \log \mathbb{E}_{\tilde{Q}}\big[\exp(f(X,Y,Z))\big], \tag{11}$$

and the bound is tight by choosing

$$f^*(x,y,z) = \log \frac{p(x,y,z)}{p(x|z)p(y,z)} \qquad \forall x,y,z \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}. \tag{12}$$

The main barrier to compute this bound for $f^*(x,y,z)$ is that the densities are unknown. This challenge is addressed in [4,11] by proposing neural classifiers that can approximate $f^*(x,y,z)$ without knowing the densities. Below we review the steps of the estimation technique provided in [11]:

(1) Construct the joint batch, containing samples generated according to $p(x,y,z)$.
(2) Construct the product batch, containing samples generated according to $p(x|z)p(y,z)$.
(3) Train the neural network with a particular loss function, which we explain later, to approximate $f^*(x,y,z)$, i.e., the density ratio of $\frac{p(x,y,z)}{p(x|z)p(y,z)}$.
(4) Compute (11) using the batches and the approximated function.

To show the consistency of the estimation with this approach, it is crucial to verify if the empirical average with respect to each sample batch converges asymptotically to the corresponding expectations. Additionally, the neural network should be designed and trained to be capable of approximating the density ratio. For i.i.d. data samples, the authors in [4,11] provided the proofs in the form of concentration bounds. In this paper, we extend these proofs for non-i.i.d. data by providing convergence results for the special case of stationary and ergodic Markov processes. In the remainder of the paper, we denote the data by $\{(X_i,Y_i,Z_i)\}_{i=1}^n$ which are consecutive samples of the stationary Markov processes $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ with marginal PDF $p(x,y,z)$.

3. Main Results

In this section, we describe our proposed neural estimator in detail. To create the batches, the estimator is equipped with a k-NN sampling block such that the empirical average over the samples converges to the expected mean. Next, we describe the roadmap to show the convergence of the estimation to the true value (i.e., consistency analysis).

3.1. Batch Construction

To create the joint batch it is sufficient to take (X_i,Y_i,Z_i) randomly from the available data. Below we define the joint batch formally using an auxiliary random variable that indicates whether an instance is selected or not (see also Algorithm 1 for the implementation).

Algorithm 1: Construction of the joint batch

```
Function CreateJointBatch($Data = \{(x_i, y_i, z_i)\}_{i=1}^n$, $\alpha$):
    Initialize $\mathcal{B}^{\alpha}_{\text{joint}}$
    for $i = 1, \ldots, n$ do
        $w_i \leftarrow$ Draw a random sample from a Bernoulli distribution with parameter $\alpha$
        if $w_i = 1$ then
            Add $(x_i, y_i, z_i)$ to $\mathcal{B}^{\alpha}_{\text{joint}}$
        end
    end
    return $\mathcal{B}^{\alpha}_{\text{joint}}$
```

Definition 2 (Joint batch)**.** *Let $W_i \sim \text{Ber}(\alpha)$ for $i = 1, \ldots, n$ be independent random variables, and $\mathcal{I}_{\alpha,n}(W^n) := \{i \mid i \in \{1, \ldots, n\}, W_i = 1\}$. Then $\mathcal{B}^{\alpha}_{\text{joint}}$ is defined as*

$$\mathcal{B}^{\alpha}_{\text{joint}} := \{(X_i, Y_i, Z_i) \mid i \in \mathcal{I}_{\alpha,n}\}, \tag{13}$$

where we use $\mathcal{I}_{\alpha,n}$ to simplify the notation.

Please note that by the law of large numbers, the length of the joint batch is asymptotically αn. Next, to construct the product batch we use the method based on the k-NN technique, which is introduced in [11]. Below we define our method denoted by *isolated k-NN* technique, and explain how the product batch is constructed (see also Algorithm 2).

Algorithm 2: Construction of the product batch

```
Function CreateProdBatch($Data = \{(x_i, y_i, z_i)\}_{i=1}^n$, $\alpha'$, $s$, $k$):
    Initialize $\mathcal{B}^{\alpha',s}_{\text{prod}}$, $\mathcal{I}_{\alpha',s}$, $\mathcal{I}^c_{\alpha',s}$
    for $i = 1, \ldots, s$ do
        $w_i \leftarrow$ Draw a random sample from a Bernoulli distribution with parameter $\alpha'$
        if $w_i = 1$ then
            Add $i$ to $\mathcal{I}_{\alpha',s}$
        end
    end
    $\mathcal{I}^c_{\alpha',s} \leftarrow \{1, \ldots, n\} \setminus \mathcal{I}_{\alpha',s}$
    $\mathcal{Z}^c \leftarrow [z_l \mid l \in \mathcal{I}^c_{\alpha',s}]$
    for $i \in \mathcal{I}_{\alpha',s}$ do
        $\mathcal{A}$ = List of $k$ nearest neighbors of $z_i$ in the set $\mathcal{Z}^c$
        for $j \in \mathcal{A}$ do
            Add $(x_j, y_i, z_i)$ to $\mathcal{B}^{\alpha',s}_{\text{prod}}$
        end
    end
    return $\mathcal{B}^{\alpha',s}_{\text{prod}}$
```

Definition 3 (Product batch)**.** *For $s < n$, let $W_i \sim \text{Bernoulli}(\alpha')$ for $i = 1, \ldots, s$ be independent random variables, and*

$$\mathcal{I}_{\alpha',s}(W^s) := \{i \mid i \in \{1, \ldots, s\}, W_i = 1\} \quad \& \quad \mathcal{I}^c_{\alpha',s}(W^s) := \{1, \ldots, n\} \setminus \mathcal{I}_{\alpha',s}(W^s).$$

Then for any $\zeta \in \mathcal{Z}$ and given the data $\{(x_i, y_i, z_i)\}_{i=1}^{n}$, define $\mathcal{A}^{\alpha',k,n,s}(\zeta, z^n, w^s)$ as the set of indices of the k nearest neighbors of ζ (by Euclidean distance) among $\{z_i\}$ for $i \in \mathcal{I}^c_{\alpha',s}(w^s)$. Formally, let $\pi\colon \{1, \dots, n-s\} \to \mathcal{I}^c_{\alpha',s}(W^s)$ be a bijection such that $\|\zeta - z_{\pi(1)}\|_2 \le \dots \le \|\zeta - z_{\pi(n-s)}\|_2$. Then, $\mathcal{A}^{\alpha',k,n,s}(\zeta, z^n, w^s) := \{\pi(1), \dots, \pi(k)\}$. So the product batch can be defined as:

$$\mathcal{B}^{\alpha',s}_{\text{prod}} := \Big\{(X_{j(i)}, Y_i, Z_i) \mid i \in \mathcal{I}_{\alpha',s}(W^s),\, j(i) \in \mathcal{A}^{\alpha',k,n,s}(Z_i, Z^n, W^s)\Big\}. \tag{14}$$

Hereafter we use $\mathcal{I}_{\alpha',s}$, $\mathcal{I}^c_{\alpha',s}$, and $\mathcal{A}^{\alpha'}(\zeta)$ instead as the remaining parameters can be understood from the context. We refer to this sampling technique as isolated k-NN *in the sequel. An example is also provided in Figure 2 for the case of $k = 2$.*

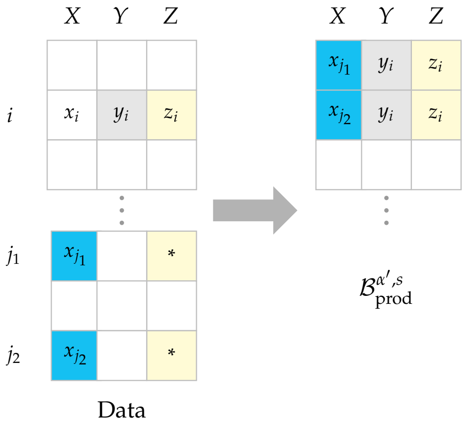

Figure 2. Construction of the product batch from the data set which is expressed as the left table. Let $w_i = 1$, and the z component of the rows denoted with '*' (indexed with j_1 and j_2) are in the k nearest neighborhood of z_i for $k = 2$. So we pack the triples (x_{j_1}, y_i, z_i) and (x_{j_2}, y_i, z_i) in the product batch as in the right table.

Remark 1. *Here we emphasize that the isolated indices are selected from the first s indices of samples while the neighbors can be searched among all n indices of data except the ones in $\mathcal{I}_{\alpha',s}(w^s)$. Additionally, note that the length of the product batch is $\alpha' sk$ asymptotically as $n \to \infty$ because sk also tends to ∞ as we see later in the assumptions of Proposition 2.*

3.2. Training the Classifier

As explained earlier, the optimal function for a tight lower bound on the CMI is obtained by the density ratio and to compute that we use the functional approximation power of neural networks. Consider a feedforward neural network with the last layer equipped with the sigmoid function. The network is parameterized with $\theta \in \Theta \subseteq \mathbb{R}^h$ where h is the number of parameters, and the neural network function is denoted by $\omega_\theta\colon \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to [0,1]$. For an input (X, Y, Z) of the network, let $C \in \{0,1\}$ denote the class of the input which determines that the tuple is generated according to $p(x,y,z)$ or $p(x|z)p(y,z)$. To be explicit, the input is either picked from the joint batch (class $C = 1$) or the product batch (class $C = 0$), and the goal is to learn the network parameters such that it can distinguish the class of new (unseen) queries. Let the loss function be the binary cross-entropy function. So for ω to be any function with inputs (x,y,z) and ranging between $[0,1]$, the expected loss is defined as:

$$L(\omega) := -\mathbb{E}\Big[C \log \omega(X,Y,Z) + (1-C)\log(1 - \omega(X,Y,Z))\Big]. \tag{15}$$

It is well-established that by minimizing $L(\omega)$, the solution ω^* would represent the probability of classifying the input in the class $C = 1$ given the input data, i.e., $\mathbb{P}(C = 1|x,y,z)$. In fact, as shown in [11] (Lemma 1) if the prior distribution on the classes is unbiased, by taking the derivative in (15) we have:

$$\Gamma(x,y,z) = \frac{p(x,y,z)}{p(x|z)p(y,z)} = \frac{\omega^*(x,y,z)}{1-\omega^*(x,y,z)}. \tag{16}$$

So from (12) the optimal function can be expressed with $\Gamma(x,y,z)$ as:

$$f^*(x,y,z) = \log\Gamma(x,y,z) \qquad \forall x,y,z \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}. \tag{17}$$

Therefore, by training the neural network, we can approximate the optimal function $f^*(x,y,z)$ and estimate the lower bound for CMI.

Consider the neural network ω_θ, then the empirical loss function is defined as:

$$\begin{aligned} L_{\text{emp}}(\omega_\theta) := &-\frac{1}{2\left|\mathcal{B}^{\alpha}_{\text{joint}}\right|}\sum_{(X,Y,Z)\in\mathcal{B}^{\alpha}_{\text{joint}}}\log\omega_\theta(X,Y,Z) \\ &-\frac{1}{2\left|\mathcal{B}^{\alpha',s}_{\text{prod}}\right|}\sum_{(X,Y,Z)\in\mathcal{B}^{\alpha',s}_{\text{prod}}}\log(1-\omega_\theta(X,Y,Z)), \end{aligned} \tag{18}$$

and the optimal parameters are obtained by solving the following problem:

$$\hat{\theta} := \arg\min_{\theta} L_{\text{emp}}(\omega_\theta). \tag{19}$$

Consequently, we can approximate the density ratio $\Gamma(x,y,z)$ from (16):

$$\hat{\Gamma}(x,y,z) = \frac{\omega_{\hat{\theta}}(x,y,z)}{1-\omega_{\hat{\theta}}(x,y,z)}. \tag{20}$$

To avoid having boundary values (i.e., $\omega_{\hat{\theta}}(x,y,z)$ close to zero or 1), the output of the neural network is clipped between $[\tau, 1-\tau]$ for some small $\tau > 0$.

Remark 2. *Please note that $\hat{\Gamma}(x,y,z)$ approximates the density ratio, if the batch sizes $\left|\mathcal{B}^{\alpha}_{\text{joint}}\right|$ and $\left|\mathcal{B}^{\alpha',s}_{\text{prod}}\right|$ are balanced. Otherwise, (20) requires a correction coefficient (see [11]). To fulfill this, given the number of samples n, one can choose the parameters such that $\alpha n = \alpha' sk$. Then, by the law of large numbers, the batches will asymptotically be balanced.*

3.3. Estimation of the DV Bound

The final step in the estimation of CMI is to compute the lower bound (11) empirically using $\hat{\Gamma}(x,y,z)$. So by substituting the expectations with empirical averages with respect to samples in the joint and the product batch, the CMI estimator is defined as:

$$\hat{I}^n_{DV}(X;Y|Z) := \frac{1}{\left|\mathcal{B}^{\alpha}_{\text{joint}}\right|}\sum_{(x,y,z)\in\mathcal{B}^{\alpha}_{\text{joint}}}\log\hat{\Gamma}(x,y,z) + \log\frac{1}{\left|\mathcal{B}^{\alpha',s}_{\text{prod}}\right|}\sum_{(x,y,z)\in\mathcal{B}^{\alpha',s}_{\text{prod}}}\hat{\Gamma}(x,y,z). \tag{21}$$

In practice, to mitigate the induced inaccuracy due to sampling from the original data, the training and estimation is repeated for several sampling trials. The steps for implementing the estimator are described in Algorithm 3. In the next part, we provide the convergence results for our estimator to validate substitution of the expectations in (11) with empirical averages with respect to the joint and the product batch. Then we show the convergence of the overall estimation to the true CMI value.

```
Algorithm 3: Estimation of CMI
    Input: $Data = \{(x_i, y_i, z_i)\}_{i=1}^{n}, \alpha, \alpha', s, k, T$
    Output: Estimation of $I(X;Y|Z)$
    for $t = 1, \ldots, T$ do
        $\mathcal{B}^{\alpha}_{\text{joint}} \leftarrow$ CreateJointBatch($Data, \alpha$) ;                    // Algorithm 1
        $\mathcal{B}^{\alpha',s}_{\text{prod}} \leftarrow$ CreateProdBatch($Data, \alpha', s, k$) ;                 // Algorithm 2
        $\mathcal{B}^{\alpha}_{\text{joint, train}}, \mathcal{B}^{\alpha}_{\text{joint, eval}} \leftarrow$ Split the batch $\mathcal{B}^{\alpha}_{\text{joint}}$ into train and evaluation sets
        $\mathcal{B}^{\alpha',s}_{\text{prod, train}}, \mathcal{B}^{\alpha',s}_{\text{prod, eval}} \leftarrow$ Split the batch $\mathcal{B}^{\alpha',s}_{\text{prod}}$ into train and evaluation sets
        Initialize the network $\omega_\theta$
        $\hat{\theta} \leftarrow$Train the network $\omega_\theta$ using $\mathcal{B}^{\alpha}_{\text{joint, train}}, \mathcal{B}^{\alpha',s}_{\text{prod, train}}$ with corresponding labels 1, 0 respectively
        $\hat{I}^{n,(t)}_{DV}(X;Y|Z) \leftarrow$ Compute (21) according to $\omega_{\hat{\theta}}$ and the data $\mathcal{B}^{\alpha}_{\text{joint, eval}}, \mathcal{B}^{\alpha',s}_{\text{prod, eval}}$
    end
    return $\hat{I}^{n,T}_{DV}(X;Y|Z) := \frac{1}{T}\sum_t \hat{I}^{n,(t)}_{DV}(X;Y|Z)$
```

3.4. Consistency Analysis

The consistency of our neural estimator (i.e., showing that the estimator converges to its true value) is based on the universal functional approximation power of neural networks and concentration results for the samples collected in the joint batch and in the product batch using the *isolated k-NN*. Informally, Hornik's functional approximation theorem [37] guarantees that feedforward neural networks are capable of fitting any continuous function. So depending on the true density of the data, there exists a choice of parameters $\tilde{\theta}$ that enables approximating the desired function with any arbitrary accuracy. Next, we show that the empirical loss function $L_{\text{emp}}(\omega_\theta)$ is concentrated around its mean $L(\omega_\theta)$ for any θ. Combining these tools, we are able to minimize the empirical loss function as in (19) and we expect $\hat{\theta}$ to be close to $\tilde{\theta}$ asymptotically; thus, eventually $\hat{\Gamma}(x,y,z)$ properly approximates $\Gamma(x,y,z)$. Additionally, the empirical computation of the DV bound is concentrated around the expected value which concludes the consistency of the end-to-end estimation of the CMI.

In this paper, we put the main focus on extending the concentration results provided in [11] (Proposition 1) with Markov assumption on data. Although conventionally many asymptotic results for i.i.d. data are assumed to hold for Markov data as well, the required extensions here are more involved due to the additional complexity of the k-NN method. In the following, we first show the convergence of the empirical average for the joint batch,

$$\left|\mathcal{B}^{\alpha}_{\text{joint}}\right|^{-1} \sum_{(X,Y,Z)\in\mathcal{B}^{\alpha}_{\text{joint}}} g(X,Y,Z) \to \mathbb{E}_{\tilde{P}}[g(X,Y,Z)],$$

where $g(\cdot)$ is any measurable function such that the expectation exists and is finite. As the product batch collects samples corresponding to the k nearest neighbors, convergence results for nearest neighbor regression are invoked to show that the empirical average for the product batch converges to the expectation with respect to the product distribution $\tilde{Q}$,

$$\left|\mathcal{B}^{\alpha',s}_{\text{prod}}\right|^{-1} \sum_{(X,Y,Z)\in\mathcal{B}^{\alpha',s}_{\text{prod}}} g(X,Y,Z) \to \mathbb{E}_{\tilde{Q}}[g(X,Y,Z)].$$

Then, we conclude the consistency of the overall estimation.

3.4.1. Convergence for the Joint Batch

One well-known extension to the law of large numbers for non-i.i.d. processes is Birkhoff's ergodic theorem, and is the basis of our proof to show the following proposition on the convergence of the sample average over the joint batch.

Proposition 1. *Consider the sequence of random variables* $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ *generated under Assumption 1. Consider the distribution* $\tilde{P}$ *in Definition 1, for any measurable function* $g(\cdot)$ *such that* $\mathbb{E}_{\tilde{P}}\big[g(X,Y,Z)\big]$ *exists and is finite,*

$$\frac{1}{\left|\mathcal{B}^{\alpha}_{\text{joint}}\right|} \sum_{(X,Y,Z)\in\mathcal{B}^{\alpha}_{\text{joint}}} g(X,Y,Z) \overset{a.s.}{\to} \mathbb{E}_{\tilde{P}}\big[g(X,Y,Z)\big]. \tag{22}$$

Proof. See Appendix A. □

3.4.2. Convergence for the Product Batch

From Definition 3, the empirical summation over all samples in the product batch is equivalent to averaging $\left|\mathcal{I}_{\alpha',s}\right|$ times k-NN regressions. Considering a sequence of pairs $\{(U_i, V_i)\}_{i=1}^n$ generated from stationary ergodic processes $(\mathbf{U}, \mathbf{V})$, the k-NN regression denotes the problem of estimating $m(u) := \mathbb{E}[V|U=u]$ with $m_n(u) := \frac{1}{k(n)}\sum_{j=1}^{k(n)} V_{r_j}$ where r_j refers to the j-th nearest neighbor of u among $U_1, \ldots, U_n$. This problem has been well studied when the pairs (U_i, V_i) are generated i.i.d.. For example in [38], the authors show the convergence of $m_n(u)$ as:

$$\mathbb{P}\bigg(\int \left|m_n(u) - m(u)\right| p(u)du \geq \epsilon\bigg) \leq \exp(-n\,a\,\epsilon^2), \tag{23}$$

for some positive constant a, when $k(n) \to \infty$ and $\frac{k(n)}{n} \to 0$. However, if the pairs are not independent, convergence results require a more advanced condition denoted geometric ϕ-mixing condition or geometric ergodicity condition [39,40]. As argued in [39], the geometric ergodicity is not a restrictive statement and holds for a wide range of processes (see also [41]). For instance, linear autoregressive processes are geometrically ergodic [41] (Ch. 15.5.2). Below we review the ϕ-mixing condition.

Definition 4 (ϕ-mixing condition)**.** *A process* $\mathbf{U}$ *is* ϕ*-mixing if for a sequence* $\{\phi_n\}_{n\in\mathbb{N}}$ *of positive numbers satisfying* $\phi_n \to 0$ *as* $n \to \infty$*, for any integer* $i > 0$ *we have:*

$$\left|P(A\cap B) - P(A)P(B)\right| \leq \phi_i P(A), \tag{24}$$

for all $n > 0$ *and all sets* A *and* B *which are members of* $\sigma(U_1, \ldots, U_n)$ *and* $\sigma(U_{n+i}, U_{n+i+1}, \ldots)$*, respectively. If* $\{\phi_n\}$ *is a geometric sequence,* $\mathbf{U}$ *is called geometrically* ϕ*-mixing.*

To show the convergence of the empirical average over the product batch, we make the following assumptions.

Assumption 2. *The sequence* $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ *is geometrically* ϕ*-mixing.*

Assumption 3. *We assume that* $\mathcal{Y}$ *and* $\mathcal{Z}$ *are compact.*

Proposition 2. *Let the sequence of random variables* $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ *be generated under Assumptions 1–3, and we choose* $k(n)$ *and* $s(n)$ *such that:*

$$\begin{aligned} &s(n)k(n) = n \\ &k(n) \to \infty \\ &s(n) \to \infty \\ &k(n)/(\log n)^2 \to \infty. \end{aligned} \tag{25}$$

Consider $\tilde{Q}$ defined in Definition 1. Then, for any function $g(\cdot)$ such that $\mathbb{E}_{\tilde{Q}}[g(X,Y,Z)]$ exists and is finite, and additionally,

$$|g(x,y_1,z) - g(x,y_2,z)| < L_g|y_1 - y_2| \qquad \forall x \in \mathcal{X},\, z \in \mathcal{Z},\, y_1, y_2 \in \mathcal{Y}, \tag{26}$$

where $L_g > 0$ is the Lipschitz constant, we have that:

$$\frac{1}{\left|\mathcal{B}^{\alpha',s}_{\text{prod}}\right|} \sum_{(X,Y,Z)\in\mathcal{B}^{\alpha',s}_{\text{prod}}} g(X,Y,Z) \overset{a.s.}{\to} \mathbb{E}_{\tilde{Q}}[g(X,Y,Z)]. \tag{27}$$

Proof. See Appendix B. □

Remark 3. *Examples of choices for $k(n)$ and $s(n)$ satisfying (25) are for instance $k(n) = n^{\frac{1}{2}}$ and $k(n) = (\log n)^{2+\epsilon}$ for some $\epsilon > 0$. Please note that in [11], the consistencies are shown when $k(n) = \Theta(n^{\frac{1}{2}})$. However, the convergence result in [11] (Theorem 1) is an explicit bound, so the condition on $k(n)$ can be relaxed (choosing a smaller $k(n)$) when we are only interested in the asymptotic behavior.*

3.4.3. Convergence of the Overall Estimation

To complete our analysis on the consistency of the neural estimator, it is required to show that the loss function is properly approximated and it converges to the optimal loss as n increase. The following assumptions on the neural network and the densities enable us to show this convergence.

Assumption 4. *For a network ω_θ parameterized with $\theta \in \Theta$, the assumption holds if Θ is closed, $\Theta \subseteq \{\theta | \|\theta\|_2 \le K\}$ for some constant $K > 0$ and ω_θ is B-Lipschitz, for some constant $B > 0$, regarding θ, for all (x,y,z), i.e.,*

$$|\omega_{\theta_1}(x,y,z) - \omega_{\theta_2}(x,y,z)| \le B\|\theta_1 - \theta_2\|_2, \qquad \forall \theta_1, \theta_2 \in \Theta, (x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}.$$

Assumption 5. *There exist $0 < p_{\min} < p_{\max} < \infty$ such that for all $x, y, z \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, the values of $p(x,y,z)$ and $p(x|z)p(y,z)$ are both in the interval $[p_{\min}, p_{\max}]$, and it holds that*

$$\frac{p_{\min}}{p_{\max} + p_{\min}} \ge \tau, \tag{28}$$

to guarantee that $\tau \le \omega^* \le 1 - \tau$.

The following theorem concludes the consistency of the end-to-end estimator.

Theorem 1. *Let Assumptions 1, 2, 3, 4, and 5 hold and $k(n)$ and $s(n)$ satisfy (25). Then the CMI estimator $\hat{I}^n_{DV}(X;Y|Z)$ (defined in (21)), converges strongly to $I(X;Y|Z)$, i.e.,*

$$\hat{I}^n_{DV}(X;Y|Z) \overset{a.s.}{\to} I(X;Y|Z). \tag{29}$$

Proof. See Appendix D. □

In the next section, we apply our estimator in several synthetic scenarios to verify its capability in estimating CMI and DI.

4. Simulation Results

In this section, we experiment with our proposed estimator of CMI and DI in the following auto-regressive model which is widely used in different applications, including wireless communications [42], defining causal notions in econometrics [43], and modeling traffic flow [44], among others:

$$\begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} = A \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} + B \begin{bmatrix} X_{i-1} \\ Y_{i-1} \\ Z_{i-1} \end{bmatrix} + \begin{bmatrix} N_i^x \\ N_i^y \\ N_i^z \end{bmatrix}, \tag{30}$$

where A and B are 3×3 matrices and the rest of variables are d-dimensional row vectors. A models the instantaneous effect of X_i, Y_i, and Z_i on each other and its diagonal elements are zero, while B models the effect of previous time instance. N_i^x, N_i^y, and N_i^z (denoted as noise in some contexts) are independent and generated i.i.d. according to zero-mean Gaussian distributions with covariance matrices $\sigma_x^2 I_d$, $\sigma_y^2 I_d$, and $\sigma_z^2 I_d$, respectively (i.e., the dimensions are d and components are uncorrelated). Please note that this model fulfills Assumptions 1 and 2 by setting appropriate initial random variables. Although the Gaussian random variables do not range in a compact set and thus, Assumption 3 does not hold, we could use truncated Gaussian distributions. Such adjustment does not significantly change the statistics of the generated dataset since the probability of finding a value far away from the mean is negligible.

In the following section, we test the capability of our estimator in estimating both conditional mutual information (CMI) and directed information (DI). In both cases, n samples are generated from the model and the estimations are performed according to Algorithms 1 and 2. Then according to Algorithm 3, the joint and product batches are split randomly in half to construct train and evaluation sets. Then the parameters of the classifier are trained with the train set and the final estimation is computed with the evaluation set (Codes are available at https://github.com/smolavipour/Neural-Estimator-of-Information-non-i.i.d, accessed on 20 May 2021).

To verify the performance of our technique, we also compared it with the approach taken in [4,31] which is as follows. Conditional mutual information can be computed by subtracting two mutual information terms, i.e.,

$$I(X;Y|Z) = I(X;Y,Z) - I(X;Z). \tag{31}$$

So instead of estimating the CMI term directly, one can use a neural estimator such as the classifier based estimator in [4] or the MINE estimator [1], and estimate each MI term in (31) to estimate the CMI. In what follows, we refer to this technique as MI-diff since it computes the difference between two MI terms.

4.1. Estimating Conditional Mutual Information

In this scenario, we estimate $I(X_1;Y_1|Z_1)$ when A and B are chosen to be:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then from (30), the CMI can be computed as below:

$$\begin{aligned} I(X_1;Y_1|Z_1) &= h(X_1|Z_1) - h(X_1|Y_1,Z_1) \\ &= h(Y_0 + Y_1 + N_1^x|Z_1) - h(Y_0 + N_1^x|Y_1,Z_1) \\ &= h(Y_0 + Y_1 + N_1^x) - h(Y_0 + N_1^x) \\ &= \frac{d}{2}\log\left(1 + \frac{\sigma_y^2 + \sigma_z^2}{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}\right). \end{aligned} \tag{32}$$

Each estimated value is an average of $T = 20$ estimations, where in each round the batches are re-selected while having a fixed dataset. This procedure is repeated for

10 Monte Carlo trials and the data are re-generated for each trial. The hyper-parameters and settings of the experiment are provided in Table 1. In Figure 3, the CMI is estimated (as $\hat{I}^{n,T}_{DV}(X_1;Y_1|Z_1)$ in Algorithm 3) with $n = 2 \times 10^4$ samples with dimension $d = 1$ when $\sigma_y = 2$, $\sigma_z = 2$ and by varying σ_x. It can be observed that the estimator can properly estimate the CMI while the variance of the estimation is also small. The latter can be inferred from the shaded region, which indicates the range of estimated CMI for a particular σ_x over all Monte Carlo trials. Next, the experiment is repeated for $d = 10$ and the results are depicted in Figure 4, where we compare our estimation of CMI with the *MI-diff* approach, which is explained in (31) and each MI term is estimated with the classifier-based estimator proposed in [4]. It can be observed that the means of both estimators are similar; nonetheless, estimating the CMI directly is more accurate and has less variation compared to the *MI-diff* approach. Additionally, our method is faster since it computes the information term only once, while in the *MI-diff* approach, two different classifiers are trained to estimate each MI term.

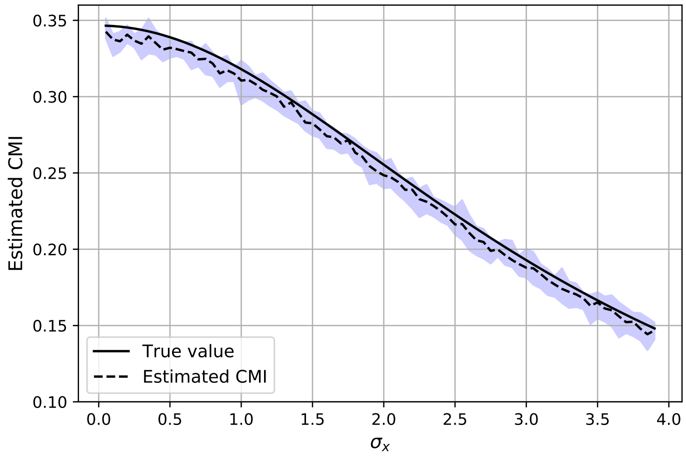

Figure 3. Estimated CMI for AR-1 model in (30) using $n = 2 \times 10^4$ samples with $d = 1$. The shaded region shows the range of the estimated values over the Monte Carlo trials.

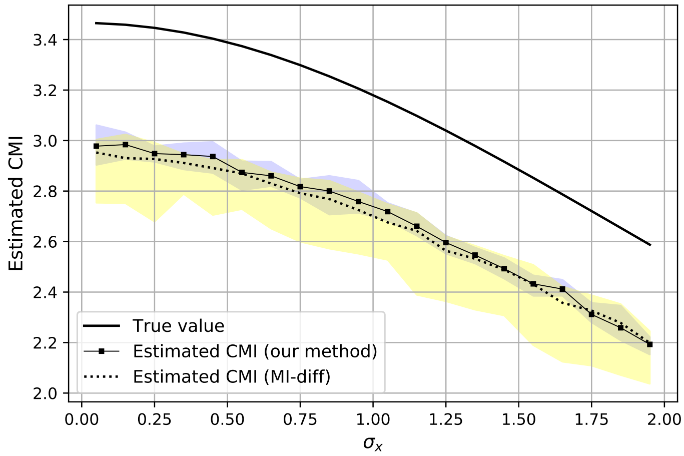

Figure 4. Estimated CMI for AR-1 model in (30) using $n = 2 \times 10^4$ samples with $d = 10$. The shaded region shows the range of the estimated values over the Monte Carlo trials. Blue shades correspond to estimation with our method, yellow shades correspond to estimation with MI-diff approach and the green shade is the overlap of the areas.

Table 1. Hyper-parameters.

Hidden units	64
Hidden layers	2 (64×64)
Activation	ReLU
τ	10^{-3}
Optimizer	Adam
Learning rate	10^{-3}
Epochs	200

4.2. Estimating Directed Information

DI can explain the underlying causal relationship among processes. This notion has wide applications in various areas. For example, consider a social network where the activities of users are monitored (e.g., the messages times as studied in [23]). The DI between these time-series data expresses how the activity of one user can affect the activity of the others. In addition, to such data analytic applications, DI characterizes the capacity of communication channels with feedback and by estimating the capacity, rates and powers of transmission can be adjusted in radio communications (see for example [32]). Now in this experiment, consider a network of three processes **X**, **Y**, and **Z**, such that the time-series data are modeled with (30) with $d = 1$ where

$$A = 0, \qquad B = \begin{bmatrix} 0 & 0 & 0 \\ b_1 & 0 & 0 \\ 0 & b_2 & 0 \end{bmatrix}. \tag{33}$$

In this model, where the relations are depicted in Figure 5, the process **X** is affecting **Y** with a delay and similarly the signal of **Y** appears on **Z** in the next time instance while an independent noise is accumulated on both steps. The DIR from **X** → **Y** in this network can be computed as follows:

$$\begin{aligned} I(\mathbf{X} \to \mathbf{Y}) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X^i, Y^{i-1}) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_{i-1}) \\ &= \frac{1}{2} \log\left(1 + \frac{b_1^2 \sigma_x^2}{\sigma_y^2}\right). \end{aligned} \tag{34}$$

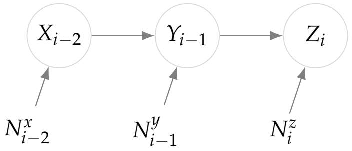

Figure 5. Causal relationship of the processes.

Similarly, for the link $\mathbf{Y} \to \mathbf{Z}$, we have:

$$\begin{aligned} I(\mathbf{Y} \to \mathbf{Z}) &= \frac{1}{n}\sum_{i=1}^{n} I(Y^i; Z_i|Z^{i-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} H(Z_i|Z^{i-1}) - H(Z_i|Y^i, Z^{i-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} H(Z_i) - H(Z_i|Y_{i-1}) \\ &= \frac{1}{2}\log\left(1 + \frac{b_1^2 b_2^2 \sigma_x^2 + b_2^2 \sigma_y^2}{\sigma_z^2}\right). \end{aligned} \tag{35}$$

Next we can compute the true DIR for the link $\mathbf{X} \to \mathbf{Z}$ as:

$$\begin{aligned} I(\mathbf{X} \to \mathbf{Z}) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(X^i; Z_i|Z^{i-1}) \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(Z_i|Z^{i-1}) - H(Z_i|X^i, Z^{i-1}) \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(Z_i) - H(Z_i|X_{i-2}) \\ &= \frac{1}{2}\log\left(1 + \frac{b_1^2 b_2^2 \sigma_x^2}{b_2^2 \sigma_y^2 + \sigma_z^2}\right). \end{aligned} \tag{36}$$

Please note that the DIR corresponding to other links (i.e., the above links in the reverse direction) is zero by similar computations. Suppose we represent the causal relationships with a directed graph, where a link between two nodes exists if the corresponding DIR is non-zero. Then according to (34)–(36), the causal relationships are described with the graph of Figure 6a.

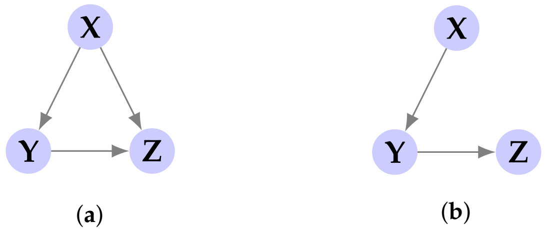

Figure 6. Graphical representation of the causal influences between the processes using pairwise directed information (**a**), and causally conditioned directed information (**b**).

To estimate the DIR, note that the processes are Markov and the *maximum Markov order* ($o_{\max}$) for the set of all processes is $o_{\max} = 2$ according to (30) and (33). Hence by (7), we can estimate DIR with the CMI estimator. For instance the DIR for processes $(\mathbf{X}, \mathbf{Y})$ can be obtained by:

$$\hat{I}^n_{DV}(\mathbf{X} \to \mathbf{Y}) := \hat{I}^n_{DV}(X^3; Y_3|Y^2),$$

where the right hand side is computed similar to (21). We performed the experiment with $n = 2 \times 10^5$ samples of dimension $d = 1$ generated according to the model (30) and (33) with $b_1 = 1$, $b_2 = 2$, $\sigma_x = 3$, $\sigma_y = 2$, and $\sigma_z = 1$, while the settings of the neural network were chosen as in Table 1. The estimated values are stated in Table 2. It can be seen that the bias of the estimator is fairly small while the variance of the estimations is negligible. This is inline with the observations in [11] when estimating CMI for i.i.d. case.

Table 2. True and estimated DIR.

	True DIR	Estimation with Our Method (Mean ± Std)
$I(\mathbf{X} \to \mathbf{Y})$	0.59	0.57 ± 0.00
$I(\mathbf{X} \to \mathbf{Z})$	0.57	0.55 ± 0.00
$I(\mathbf{Y} \to \mathbf{Z})$	1.99	1.92 ± 0.01
$I(\mathbf{Y} \to \mathbf{X})$	0	0.00 ± 0.00
$I(\mathbf{Z} \to \mathbf{X})$	0	0.00 ± 0.00
$I(\mathbf{Z} \to \mathbf{Y})$	0	0.00 ± 0.00

Although $I(\mathbf{X} \to \mathbf{Z}) > 0$, intuitively $\mathbf{X}$ is only affecting $\mathbf{Z}$ causally through $\mathbf{Y}$, which suggests that $I(\mathbf{X} \to \mathbf{Z} \parallel \mathbf{Y}) = 0$. This event is referred to as *proxy effect* when studying directed information graph (see [45]). In fact the graphical representation of causal relationships can be simplified using the notion of causally conditioned DIR as depicted in Figure 6b. To see this formally, note that from (30) it yields that:

$$\begin{aligned} I(\mathbf{X} \to \mathbf{Z} \parallel \mathbf{Y}) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(X^i; Z_i | Y^i, Z^{i-1}) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(Z_i | Y^i, Z^{i-1}) - H(Z_i | X^i, Y^i, Z^{i-1}) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(Z_i | Y_{i-1}) - H(Z_i | Y_{i-1}) \\ &= 0. \end{aligned} \tag{37}$$

Considering $o_{\max} = 2$, the causally conditioned DIR terms can be estimated with our CMI estimator according to (7); for instance,

$$\hat{I}^n_{DV}(\mathbf{X} \to \mathbf{Y} \parallel \mathbf{Z}) := \hat{I}^n_{DV}(X^3; Y_3 | Y^2, Z^3).$$

The estimation results are provided in Table 3 for all the links, where for each link we averaged over $T = 20$ estimations (as in Algorithm 3); then the procedure is repeated for 10 Monte Carlo trials in which we generate a new dataset according to the model.

Table 3. True and estimated DIR.

	True DIR	Estimation with Our Method (Mean ± Std)
$I(\mathbf{X} \to \mathbf{Y} \parallel \mathbf{Z})$	0.59	0.57 ± 0.00
$I(\mathbf{X} \to \mathbf{Z} \parallel \mathbf{Y})$	0	0.00 ± 0.00
$I(\mathbf{Y} \to \mathbf{Z} \parallel \mathbf{X})$	1.42	1.52 ± 0.01
$I(\mathbf{Y} \to \mathbf{X} \parallel \mathbf{Z})$	0	0.01 ± 0.00
$I(\mathbf{Z} \to \mathbf{X} \parallel \mathbf{Y})$	0	0.01 ± 0.00
$I(\mathbf{Z} \to \mathbf{Y} \parallel \mathbf{X})$	0	0.01 ± 0.00

In this experiment, we did not explore the effect of higher dimensions for data, although one should note that for the causally conditioned DIR estimation, with $d = 1$ the neural network is fed with data of size 9. Nevertheless, the performance of higher dimensions for this estimator with i.i.d. data has been studied in [11] and the challenges of dealing with high dimensions when data has dependency can be considered to be a future direction of this work. Additionally, although the information about $o_{\max}$ may not always be available in practice, it can be approximated by data-driven approaches similar to the method described in [45].

5. Conclusions and Future Directions

In this paper, we explored the potentials of a neural estimator for information measures when there exist time dependencies among the samples. We extended the analysis on the convergence of the estimation and provided experimental results to show the performance of the estimator in practice. Furthermore, we compared our estimation method with a similar approach taken in [4,31] (which we denoted as MI-diff), and demonstrations on synthetic scenarios show that the variances of our estimations are smaller. However, the main contribution is the derivation of proofs of convergence when the data are generated from a Markov source. Our estimator is based on a k-NN method to re-sample the dataset such that the empirical average over the samples converges to the expectation with certain density. The convergence result derived for the re-sampling technique is stand-alone and can be adopted in other sampling application.

Our proposed estimator can be used potentially in the areas of information theory, communication systems, and machine learning. For instance, the capacity of channels with feedback can be characterized with directed information and estimated with our estimator and can be investigated as a future direction. Furthermore, in machine learning applications where the data has some form of dependency (either spatial of temporal), regularizing the training with information flow requires the estimator of information to capture causality which is considered in our technique. Finally, information measures can be used in modeling and controlling a complex system and the results in this work can provide meaningful measures such as conditional dependence and causal influence.

Author Contributions: Conceptualization, S.M.; methodology, S.M., H.G., and G.B.; software, S.M.; validation, S.M., H.G., G.B. and M.S.; formal analysis, S.M., H.G., and G.B.; investigation, S.M. and G.B.; resources, M.S.; data curation, S.M.; writing—original draft preparation, S.M. and H.G.; writing—review and editing, S.M., H.G., G.B. and M.S.; visualization, S.M.; supervision, G.B. and M.S.; project administration, M.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Research Council under contract 2019-03606.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PDF	Probability density function
IID	Independent and identically distributed
MI	Mutual information
CMI	Conditional mutual information
DI	Directed information
DIR	Directed information rate
TE	Transfer entropy
DV	Donsker-Varadhan
NWJ	Nguyen-Wainwright-Jordan
k-NN	k nearest neighbors
ML	Machine learning
RNN	Recurrent neural network

Appendix A. Proof of Proposition 1

To show the convergence stated in the Proposition, let us first introduce the following lemma which is a variant of Birkhoff's ergodic theorem for the case where the samples are not necessarily subsequent.

Lemma A1. *Let U^n be n observations of a stationary and ergodic Markov process where $U_i \in \mathcal{U}$ and $\mathcal{U} \subseteq \mathcal{R}^d$. Then if $\mathbb{E}[g(U)]$ exists and is finite,*

$$\frac{1}{|\mathcal{I}_{\alpha,n}|}\sum_{j\in\mathcal{I}_{\alpha,n}} g(U_j) \overset{a.s.}{\to} \mathbb{E}[g(U)], \tag{A1}$$

where $\mathcal{I}_{\alpha,n}$ is defined in Definition 2 and the empirical average is considered to be zero when $|\mathcal{I}_{\alpha,n}| = 0$.

Proof. Consider $W_1, \dots, W_n$ generated i.i.d. and $W_i \sim \text{Bernoulli}(\alpha)$. From the definition of $\mathcal{I}_{\alpha,n}$, we can write the summation equivalently as

$$\sum_{j\in\mathcal{I}_{\alpha,n}} g(U_j) = \sum_{i=1}^{n} W_i\, g(U_i). \tag{A2}$$

Since the W_i's are independent of $g(U_i)$, the pairs $(W_i, g(U_i))$ are also stationary and ergodic Markov, so from Birkhoff's ergodic theorem,

$$\frac{1}{n}\sum_{i=1}^{n} W_i\, g(U_i) - \mathbb{E}\big[Wg(U)\big] \overset{\text{a.s.}}{\to} 0, \tag{A3}$$

and since $\mathbb{E}\big[Wg(U)\big] = \mathbb{E}[W]\mathbb{E}\big[g(U)\big] = \alpha\mathbb{E}\big[g(U)\big]$,

$$\frac{1}{n}\sum_{i=1}^{n} W_i\, g(U_i) \overset{\text{a.s.}}{\to} \alpha\mathbb{E}\big[g(U)\big]. \tag{A4}$$

On the other hand, from the strong law of large numbers:

$$\frac{|\mathcal{I}_{\alpha,n}|}{n} = \frac{1}{n}\sum_{i=1}^{n} W_i \overset{\text{a.s.}}{\to} \alpha. \tag{A5}$$

From (A4) and (A5), and since the summation in (A5) is bounded,

$$\frac{1}{|\mathcal{I}_{\alpha,n}|}\sum_{j\in\mathcal{I}_{\alpha,n}} g(U_j) \overset{\text{a.s.}}{\to} \mathbb{E}\big[g(U)\big]$$

and the proof is complete. □

Using Lemma A1, the proof of Proposition 1 becomes trivial by letting $U_i = (X_i, Y_i, Z_i)$ since the triple is a sample of a jointly stationary ergodic Markov process. Noting that $|\mathcal{I}_{\alpha,n}| = \left|\mathcal{B}^{\alpha}_{\text{joint}}\right|$ concludes the proof of the Proposition.

Appendix B. Proof of Proposition 2

To show the convergence of the empirical average over samples in the product batch, we begin by reviewing convergence results for k-NN regression.

Lemma A2 ([39] (Theorem 2-a))**.** *Consider the sequence $\{(U_i, V_i)\}_{i=1}^n$ is stationary and geometrically ϕ-mixing (see Definition 4). If $\frac{k(n)}{n} \to 0$ and $\frac{k(n)}{(\log n)^2} \to \infty$, then*

$$\sup_u \left| m_n(u) - m(u) \right| \overset{a.s.}{\to} 0. \tag{A6}$$

Now to extend Lemma A2 to the case where the samples are randomly selected for the regression, we show the following lemmas.

Lemma A3. *Let $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ be generated under Assumptions 1–3. If $\frac{k(n)}{n} \to 0$ and $\frac{k(n)}{(\log n)^2} \to \infty$, and for any $y \in \mathcal{Y}$, $\mathbb{E}_{P_{X|Z}}[g(X, y, Z) \mid Z = z]$ exists and is finite, then we have that, for all y:*

$$\sup_z \left| \tilde{g}(y, z) - \mathbb{E}_{P_{X|Z}}[g(X, y, Z) \mid Z = z] \right| \overset{a.s.}{\to} 0, \tag{A7}$$

where

$$\tilde{g}(y, z) := \frac{1}{k(n)} \sum_{j=1}^{k(n)} g(X_{r_j}, y, z),$$

and r_j refers to the index of the j-th nearest neighbor of z among $\{Z_i\}_{i=1}^n$.

Proof. The proof follows directly from Lemma A2 as y is fixed in (A7). □

Lemma A4. *Let $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ be generated under Assumptions 1–3. Then, if $k(n)$ and $s(n)$ fulfill the assumptions in (25), and for any $y \in \mathcal{Y}$, $\mathbb{E}_{P_{X|Z}}[g(X, y, Z) \mid Z = z]$ exists and is finite, for all y:*

$$\sup_z \left| \bar{g}(y, z, W^{s(n)}) - \mathbb{E}_{P_{X|Z}}[g(X, y, Z) \mid Z = z] \right| \overset{a.s.}{\to} 0, \tag{A8}$$

where

$$\bar{g}(y, z, W^{s(n)}) := \frac{1}{k(n)} \sum_{l \in \mathcal{A}^{\alpha', k(n), n, s(n)}(z, Z^n, W^{s(n)})} g(X_l, y, z), \tag{A9}$$

and $\mathcal{A}^{\alpha', k(n), n, s(n)}(z, Z^n, W^{s(n)})$ and $W^{s(n)}$ are defined in Definition 3.

Proof. See Appendix C. □

Lemma A5. *For the sequence $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ defined in Lemma A4:*

$$\left| \bar{g}(Y_{s(n)}, Z_{s(n)}, W^{s(n)}) - \mathbb{E}_{P_{X|Z}}\left[g(X, Y, Z) \mid Y = Y_{s(n)}, Z = Z_{s(n)} \right] \right| \overset{a.s.}{\to} 0, \tag{A10}$$

where $s(n) < n$ and the convergence occurs according to the random variables $Y_{s(n)}, Z_{s(n)}, W^{s(n)}$, and the sequence.

Proof. To simplify the notation, we use $\bar{g}(y, z)$ instead of $\bar{g}(y, z, W^{s(n)})$ in this proof. Since $\mathcal{Y}$ is compact, for any $\epsilon > 0$, there exist M finite balls with radius ϵ / L_g and centers $\tilde{y}_j$ for $j = 1, \ldots, M$, that cover $\mathcal{Y}$. Then, from the triangle inequality, we have:

$$\begin{aligned} &\mathbb{P}\left(\lim_{n \to \infty} \sup_{y,z} \left| \bar{g}(y, z) - \mathbb{E}_{P_{X|Z}}[g(X, y, Z) \mid Z = z] \right| \leq 2\epsilon \right) \\ &\quad \geq \mathbb{P}\left(\lim_{n \to \infty} \sup_{y,z} \left| \Delta_{(1)}(y, z) \right| + \left| \Delta_{(2)}(y, z) \right| + \left| \Delta_{(3)}(y, z) \right| \leq 2\epsilon \right), \end{aligned} \tag{A11}$$

where

$$\Delta_{(1)}(y,z) := \bar{g}(y,z) - \bar{g}(\tilde{y}_j,z) \tag{A12}$$

$$\Delta_{(2)}(y,z) := \bar{g}(\tilde{y}_j,z) - \mathbb{E}_{P_{X|Z}}[g(X,\tilde{y}_j,Z) \mid Z=z] \tag{A13}$$

$$\Delta_{(3)}(y,z) := \mathbb{E}_{P_{X|Z}}[g(X,\tilde{y}_j,Z) \mid Z=z] - \mathbb{E}_{P_{X|Z}}[g(X,y,Z) \mid Z=z], \tag{A14}$$

and $\tilde{y}_j$ is the center of the ball containing y. Note that

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{y,z}\left(\left|\Delta_{(1)}(y,z)\right| + \left|\Delta_{(2)}(y,z)\right| + \left|\Delta_{(3)}(y,z)\right|\right) \\
&\quad\le \lim_{n\to\infty}\sup_{y,z}\left|\Delta_{(1)}(y,z)\right| + \lim_{n\to\infty}\sup_{y,z}\left|\Delta_{(2)}(y,z)\right| + \lim_{n\to\infty}\sup_{y,z}\left|\Delta_{(3)}(y,z)\right| && \text{(A15)} \\
&\quad\le 2\epsilon + \lim_{n\to\infty}\sup_{y,z}\left|\Delta_{(2)}(y,z)\right|, && \text{(A16)}
\end{aligned}$$

where (A16) follows from (26) and the radius of the balls being ϵ/L_g. Thus (A11) yields:

$$\begin{aligned}
&\mathbb{P}\left(\lim_{n\to\infty}\sup_{y,z}\left|\bar{g}(y,z) - \mathbb{E}_{P_{X|Z}}[g(X,y,Z) \mid Z=z]\right| \le 2\epsilon\right) \\
&\quad\ge \mathbb{P}\left(\lim_{n\to\infty}\sup_{y,z}\left|\Delta_{(2)}(y,z)\right| \le 0\right) \\
&\quad\ge \mathbb{P}\left(\lim_{n\to\infty}\max_{\tilde{y}_j}\sup_{z}\left|\Delta_{(2)}(\tilde{y}_j,z)\right| \le 0\right) && \text{(A17)} \\
&\quad= \mathbb{P}\left(\max_{\tilde{y}_j}\lim_{n\to\infty}\sup_{z}\left|\Delta_{(2)}(\tilde{y}_j,z)\right| \le 0\right) && \text{(A18)} \\
&\quad\ge 1 - \sum_{j=1}^{M}\mathbb{P}\left(\lim_{n\to\infty}\sup_{z}\left|\Delta_{(2)}(\tilde{y}_j,z)\right| > 0\right) && \text{(A19)} \\
&\quad= 1, && \text{(A20)}
\end{aligned}$$

where (A17) holds by the definition (A13), (A18) follows since $\tilde{y}_j$ is independent of n, and the last step is due to Lemma A4. Finally since (A20) holds for any $\epsilon > 0$, according to [46] (Prop 1.13) it is concluded that:

$$\mathbb{P}\left(\lim_{n\to\infty}\sup_{y,z}\left|\bar{g}(y,z) - \mathbb{E}_{P_{X|Z}}[g(X,y,Z) \mid Z=z]\right| = 0\right) = 1. \tag{A21}$$

Consider now the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$, $\bar{g}(y,z)$ can be expressed equivalently as $\bar{g}(y,z;\psi) : \Omega \to \mathbb{R}$. Consider the functions $Y_{s(n)}(\psi) : \Omega \to \mathcal{Y}$ and $Z_{s(n)}(\psi) : \Omega \to \mathcal{Z}$, then from (A21):

$$\begin{aligned}
&\mathbb{P}\left(\psi \in \Omega : \lim_{n\to\infty}\left|\bar{g}(Y_{s(n)}(\psi), Z_{s(n)}(\psi);\psi) - \mathbb{E}_{P_{X|Z}}[g(X,Y,Z) \mid Y = Y_{s(n)}(\psi), Z = Z_{s(n)}(\psi)]\right| = 0\right) \\
&\quad\ge \mathbb{P}\left(\psi \in \Omega : \lim_{n\to\infty}\sup_{y,z}\left|\bar{g}(y,z;\psi) - \mathbb{E}_{P_{X|Z}}[g(X,y,Z) \mid Z=z]\right| = 0\right) && \text{(A22)} \\
&\quad= 1\,,
\end{aligned}$$

which implies that:

$$\left|\bar{g}(Y_{s(n)}, Z_{s(n)}, W^{s(n)}) - \mathbb{E}_{P_{X|Z}}\Big[g(X,Y,Z) \mid Y = Y_{s(n)}, Z = Z_{s(n)}\Big]\right| \overset{\text{a.s.}}{\to} 0, \tag{A23}$$

and the proof of Lemma A5 is concluded. □

Now that the required tools were introduced, we can continue the proof of Proposition 2. From Definition 3 and (A9), the LHS of (27) can be expressed as below:

$$\frac{1}{k(n)\left|\mathcal{I}_{\alpha',s(n)}\right|}\sum_{(X,Y,Z)\in\mathcal{B}^{\alpha',s}_{\text{prod}}} g(X,Y,Z) = \frac{1}{\left|\mathcal{I}_{\alpha',s(n)}\right|}\sum_{i=1}^{s(n)} W_i\,\bar{g}(Y_i,Z_i,W^{s(n)}). \tag{A24}$$

Let us define:

$$\Delta_i := \bar{g}(Y_i,Z_i,W^{s(n)}) - \mathbb{E}_{P_{X|Z}}\left[g(X,Y,Z)\mid Y=Y_i, Z=Z_i\right],$$

and from Lemma A5, we obtain that:

$$\left|\Delta_{s(n)}\right| \overset{\text{a.s.}}{\to} 0. \tag{A25}$$

As a result we can show the following strong convergence:

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to\infty}\frac{1}{s(n)}\sum_{i=1}^{s(n)} W_i\Delta_i = 0\right) &\geq \mathbb{P}\Big(\lim_{n\to\infty} W_{s(n)}\Delta_{s(n)} = 0\Big) &\text{(A26)}\\ &\geq \mathbb{P}\Big(\lim_{n\to\infty}\left|\Delta_{s(n)}\right| = 0\Big) &\text{(A27)}\\ &= 1, &\text{(A28)}\end{aligned}$$

where (A26) holds since $s(n)\to\infty$ by (25) and using Cesáro mean ([47] (Theorem 4.2.3)), (A27) holds since $W_{s(n)}\in\{0,1\}$, and the equality in the last step follows from (A25). In other words,

$$\frac{1}{s(n)}\sum_{i=1}^{s(n)} W_i\,\bar{g}(Y_i,Z_i,W^{s(n)}) - \frac{1}{s(n)}\sum_{i=1}^{s(n)} W_i\,\mathbb{E}_{P_{X|Z}}[g(X,Y,Z)\mid Y=Y_i, Z=Z_i] \overset{\text{a.s.}}{\to} 0. \tag{A29}$$

Next since the sequence $\{(W_i,Y_i,Z_i)\}_{i=1}^{s(n)}$ is stationary and ergodic, using Birkhoff's ergodic theorem we have:

$$\begin{aligned}&\frac{1}{s(n)}\sum_{i=1}^{s(n)} W_i\,\mathbb{E}_{P_{X|Z}}\left[g(X,Y,Z)\mid Y=Y_i, Z=Z_i\right]\\ &\qquad\overset{\text{a.s.}}{\to}\mathbb{E}_{P_W P_{YZ}}\Big[W\,\mathbb{E}_{P_{X|Z}}[g(X,Y,Z)\mid Y,Z]\Big].\end{aligned} \tag{A30}$$

As W is generated independently

$$\mathbb{E}_{P_W P_{YZ}}\Big[W\,\mathbb{E}_{P_{X|Z}}[g(X,Y,Z)\mid Y,Z]\Big] = \mathbb{E}[W]\,\mathbb{E}_{\tilde{Q}}\left[g(X,Y,Z)\right]. \tag{A31}$$

To complete the proof, note that

$$\frac{\left|\mathcal{I}_{\alpha',s(n)}\right|}{s(n)} \overset{\text{a.s.}}{\to} \mathbb{E}[W]. \tag{A32}$$

Therefore, from (A24) and (A29)–(A32), and $\left|\mathcal{B}^{\alpha',s}_{\text{prod}}\right| = k(n)\left|\mathcal{I}_{\alpha',s(n)}\right|$ we conclude that:

$$\frac{1}{\left|\mathcal{B}^{\alpha',s}_{\text{prod}}\right|}\sum_{(X,Y,Z)\in\mathcal{B}^{\alpha',s}_{\text{prod}}} g(X,Y,Z) \overset{\text{a.s.}}{\to} \mathbb{E}_{\tilde{Q}}\left[g(X,Y,Z)\right], \tag{A33}$$

and the proof is complete. □

Appendix C. Proof Lemma A4

According to Definition 3, the index set $\mathcal{I}_{\alpha',s(n)}$ is determined by the sequence $W^{s(n)}$. Therefore, $\mathcal{A}^{\alpha',k(n),n,s(n)}(z, Z^n, W^{s(n)})$ denotes the set of indices of the $k(n)$ nearest neighbors of z among $\{Z_i \mid i \in \mathcal{I}^c_{\alpha',s(n)}\}$, unlike in Lemma A3 where the neighbors can be chosen among the whole sequence $\{Z_i\}_{i=1}^n$. Hence, the first step is to verify the *ϕ-mixing* condition for the *isolated k-NN* method where some indices are excluded. Intuitively, if $\{X_i, Y_i, Z_i\}_{i=1}^n$ is *ϕ-mixing*, then the sequence $\{(X_i, Y_i, Z_i)\}_{i\in\mathcal{I}^c_{\alpha',s(n)}}$ is also *ϕ-mixing* since the random jumps make the asymptotic independence (see Definition 4) happen with a faster rate. Nonetheless, we can show that the sequence $\{(X_i, Y_i, Z_i)\}_{i\in\mathcal{I}^c_{\alpha',s(n)}}$ satisfy the mixing condition for Lemma A3 which is expressed in the following.

The basis of the proof for Lemma A2 and thus Lemma A3, is Collomb's inequality [48] (Theorem 2.2.1) which provides a concentration bound similar to Hoeffding's inequality for *ϕ-mixing* variables. For instance if $\mathbf{U}$ is a ϕ-mixing process where $\mathbb{E}[U_i] = 0, |U_i| \le a_1$, $\mathbb{E}[U_i^2] \le a_2$, and $\mathbb{E}[|U_i|] \le a_3$, the inequality states that:

$$\mathbb{P}\left(\left|\sum_{i=1}^n U_i\right| > \epsilon\right) \le \exp\left(3\sqrt{e}n\frac{\phi_t}{t} - a_4\epsilon + 6a_4^2 n\left(a_2 + 4a_1a_3\sum_{i=1}^t \phi_i\right)\right), \tag{A34}$$

for some integer $t < n$ and real a_4 such that $a_1a_4t \le 1/4$. In order to show a similar inequality for $\{U_i\}_{i\in\mathcal{I}^c_{\alpha',s(n)}}$, we have that:

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i\in\mathcal{I}^c_{\alpha',s(n)}} U_i\right| > \epsilon\right) &= \mathbb{P}\left(\left|\sum_{i=1}^n U_i - \sum_{i=1}^{s(n)} W_iU_i\right| > \epsilon\right)\\ &\le \mathbb{P}\left(\left|\sum_{i=1}^n U_i\right| > \epsilon/2\right) + \mathbb{P}\left(\left|\sum_{i=1}^{s(n)} W_i\,U_i\right| > \epsilon/2\right),\end{aligned} \tag{A35}$$

where both terms in (A35) are bounded with exponential terms and can be dominated by either of them. Thus, as $n \to \infty$ and $s(n) \to \infty$ (by assumption (25)) both terms tend to zero and Collomb's inequality applies to the summation over the sub-sequence of samples remained after the isolation. In other words, the required mixing condition holds for the new sequence $\{(X_i, Y_i, Z_i)\}_{i\in\mathcal{I}^c_{\alpha',s(n)}}$ and the result in Lemma A2 can be extended to this Lemma.

Next it remains to verify the conditions of Lemma A2 on $k(n)$. From (25) we have,

$$\frac{k(n)}{\left|\mathcal{I}^c_{\alpha',s(n)}\right|} \le \frac{k(n)}{n - s(n)} = \frac{1}{s(n)(1 - \frac{1}{k(n)})} \overset{\text{a.s.}}{\to} 0,$$

which yields that:

$$\frac{k(n)}{\left(\log\left|\mathcal{I}^c_{\alpha',s(n)}\right|\right)^2} \ge \frac{k(n)}{(\log n)^2} \overset{\text{a.s.}}{\to} \infty. \tag{A36}$$

Therefore, the conditions of Lemma A2 hold and from Lemma A3 it follows that for all $y \in \mathcal{Y}$:

$$\sup_{z} \left| \bar{g}(y, z, W^{s(n)}) - \mathbb{E}_{P_{X|Z}}[g(X, y, Z) \mid Z = z] \right| \overset{\text{a.s.}}{\to} 0, \tag{A37}$$

which concludes the proof of the Lemma. □

Appendix D. Proof Theorem 1

Based on the universal functional approximation theory of neural networks [37], ref. [4] (Lemma 4) implies that for any $\epsilon_0 > 0$, there exists $\tilde{\theta} \in \Theta$ such that:

$$\left| L(\omega_{\tilde{\theta}}) - L^* \right| < \frac{\epsilon_0}{2}, \tag{A38}$$

where $L^* := L(\omega^*)$ and $L(\omega)$ and ω^* were defined in (15). Moreover, from Propositions 1 and 2, for any $\theta \in \Theta$, the empirical loss $L_{\text{emp}}(\omega_\theta)$ defined in (18) converges asymptotically to the expected loss $L(\omega_\theta)$. This is obtained by letting $g(x, y, z) = \log(\omega_\theta(x, y, z))$ and $g(x, y, z) = \log(1 - \omega_\theta(x, y, z))$ in Propositions 1 and 2, respectively, and noting Remark 2. Thus we have:

$$L_{\text{emp}}(\omega_\theta) \overset{\text{a.s.}}{\to} L(\omega_\theta). \tag{A39}$$

Since $\Theta \subset \mathbb{R}^h$ and $\|\theta\|_2 \leq K, \forall \theta \in \Theta$, Θ can be covered with finite $N(\Theta, r)$ number of balls of radius r, where $N(\Theta, r)$ is bounded [49]:

$$N(\Theta, r) \leq \left(\frac{2K\sqrt{h}}{r} \right)^h. \tag{A40}$$

Let $\{\theta_1, \ldots, \theta_{N(\Theta,r)}\}$ denote the centers of the covering balls. Let j_n be the index of the ball that $\hat{\theta}$ belongs to, then from the triangle inequality we have:

$$\begin{aligned} \left| L_{\text{emp}}(\omega_{\hat{\theta}}) - L(\omega_{\hat{\theta}}) \right| &\leq \left| L_{\text{emp}}(\omega_{\hat{\theta}}) - L_{\text{emp}}(\omega_{\theta_{j_n}}) \right| + \left| L_{\text{emp}}(\omega_{\theta_{j_n}}) - L(\omega_{\theta_{j_n}}) \right| \\ &\quad + \left| L(\omega_{\theta_{j_n}}) - L(\omega_{\hat{\theta}}) \right| \\ &\leq \left| L_{\text{emp}}(\omega_{\theta_{j_n}}) - L(\omega_{\theta_{j_n}}) \right| + \frac{2Br}{\tau} \end{aligned} \tag{A41}$$

where the second inequality holds due to the Lipschitz continuity of ω_θ stated in Assumption 4. From the union bound and for any $\epsilon' > 0$, we have:

$$\begin{aligned} &\mathbb{P}\left(\lim_{n \to \infty} \left| L_{\text{emp}}(\omega_{\hat{\theta}}) - L(\omega_{\hat{\theta}}) \right| > \frac{\epsilon'}{2} \right) \\ &\leq N(\Theta, r)\, \mathbb{P}\left(\lim_{n \to \infty} \left| L_{\text{emp}}(\omega_{\theta_{j_n}}) - L(\omega_{\theta_{j_n}}) \right| > \frac{\epsilon'}{2} - \frac{2Br}{\tau} \right) && \text{(A42)} \\ &= 0, && \text{(A43)} \end{aligned}$$

where (A42) holds due to (A41), applying a union bound over all centers θ_j, and choosing $r < \frac{\epsilon'\tau}{4B}$, and the last step follows by exploiting the strong convergence in (A39). As a result, with probability one:

$$\begin{aligned}
\lim_{n\to\infty} L(\omega_{\hat{\theta}}) &\leq \lim_{n\to\infty} L_{\text{emp}}(\omega_{\hat{\theta}}) + \frac{\epsilon'}{2} && \text{(A44)}\\
&\leq \lim_{n\to\infty} L_{\text{emp}}(\omega_{\tilde{\theta}}) + \frac{\epsilon'}{2} && \text{(A45)}\\
&= L(\omega_{\tilde{\theta}}) + \frac{\epsilon'}{2} && \text{(A46)}\\
&\leq L^* + \epsilon', && \text{(A47)}
\end{aligned}$$

where (A44) is obtained from (A43), and (A45) holds since $\hat{\theta}$ minimizes $L_{\text{emp}}(\omega_\theta)$, and (A46) follows from (A39). Finally, the last step is derived using (A38) and choosing $\epsilon_0 = \epsilon'$.

To conclude the proof, note that if Assumption 5 holds, from [4] (Lemma 6) and taking similar steps as in [11] (Lemma 8), it is implied that for any given $\epsilon' > 0$, with probability one as $n \to \infty$:

$$\begin{aligned}
\mathbb{E}_{\tilde{P}}\Big[\Big|\omega^*(X,Y,Z) - \omega_{\hat{\theta}}(X,Y,Z)\Big| \mid \hat{\theta}\Big] &\leq \eta,\\
\mathbb{E}_{\tilde{Q}}\Big[\Big|\omega^*(X,Y,Z) - \omega_{\hat{\theta}}(X,Y,Z)\Big| \mid \hat{\theta}\Big] &\leq \eta,
\end{aligned} \quad \text{(A48)}$$

where $\eta := (1-\tau)p_{\max}\sqrt{2\lambda\epsilon'/p_{\min}}$, with λ being the Lebesgue measure corresponding to $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Note that the expectations in (A48) are random variables due to $\hat{\theta}$. Let us define $I^n_{DV}(X;Y|Z)$ as:

$$I^n_{DV}(X;Y|Z) := \mathbb{E}_{\tilde{P}}\Big[\log\hat{\Gamma}(X,Y,Z) \mid \hat{\theta}\Big] - \log \mathbb{E}_{\tilde{Q}}\Big[\hat{\Gamma}(X,Y,Z) \mid \hat{\theta}\Big]. \quad \text{(A49)}$$

Thus by the triangle inequality we have:

$$\begin{aligned}
&\Big|\hat{I}^n_{DV}(X;Y|Z) - I(X;Y|Z)\Big|\\
&\qquad\leq \Big|\hat{I}^n_{DV}(X;Y|Z) - I^n_{DV}(X;Y|Z)\Big| + \big|I^n_{DV}(X;Y|Z) - I(X;Y|Z)\big|.
\end{aligned} \quad \text{(A50)}$$

where $\hat{I}^n_{DV}(X;Y|Z)$ was defined in (21).

To bound the first term, note that by the triangle inequality

$$\Big|\hat{I}^n_{DV}(X;Y|Z) - I^n_{DV}(X;Y|Z)\Big| \leq \Delta_{DV} + \Delta'_{DV}, \quad \text{(A51)}$$

where

$$\Delta_{DV} := \bigg|\Big|\mathcal{B}^\alpha_{\text{joint}}\Big|^{-1} \textstyle\sum_{X,Y,Z\in\mathcal{B}^\alpha_{\text{joint}}} \log\hat{\Gamma}(X,Y,Z) - \mathbb{E}_{\tilde{P}}\Big[\log\hat{\Gamma}(X,Y,Z) \mid \hat{\theta}\Big]\bigg|$$

and

$$\Delta'_{DV} := \bigg|\log\Big|\mathcal{B}^{\alpha',s}_{\text{prod}}\Big|^{-1} \textstyle\sum_{X,Y,Z\in\mathcal{B}^{\alpha',s}_{\text{prod}}} \hat{\Gamma}(X,Y,Z) - \log\mathbb{E}_{\tilde{Q}}\Big[\hat{\Gamma}(X,Y,Z) \mid \hat{\theta}\Big]\bigg|.$$

Since $\hat{\Gamma}(\cdot)$ is bounded as:

$$\frac{\tau}{1-\tau} \leq \hat{\Gamma}(X,Y,Z) \leq \frac{1-\tau}{\tau},$$

by the Lipschitz continuity of $\log(\cdot)$ it follows that:

$$\Big|\hat{I}^n_{DV}(X;Y|Z) - I^n_{DV}(X;Y|Z)\Big| \leq \Delta_{DV} + \Delta''_{DV}, \quad \text{(A52)}$$

where

$$\Delta''_{DV} := \frac{1-\tau}{\tau}\left|\left|\mathcal{B}^{\alpha',s}_{\text{prod}}\right|^{-1}\sum\nolimits_{X,Y,Z\in\mathcal{B}^{\alpha',s}_{\text{prod}}} \hat{\Gamma}(X,Y,Z) - \mathbb{E}_{\tilde{Q}}\left[\hat{\Gamma}(X,Y,Z) \mid \hat{\theta}\right]\right|.$$

Both Δ_{DV} and Δ''_{DV} converge strongly to zero from Propositions 1 and 2, respectively, i.e., for any given $\epsilon > 0$, we have that:

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to\infty}\Delta_{DV} > \epsilon/4\right) &= 0,\\ \mathbb{P}\left(\lim_{n\to\infty}\Delta''_{DV} > \epsilon/4\right) &= 0.\end{aligned} \tag{A53}$$

To bound the second term in (A50), using the triangle inequality it yields that:

$$\begin{aligned}\left|I^n_{DV}(X;Y|Z) - I(X;Y|Z)\right| \le &\left|\mathbb{E}_{\tilde{P}}\left[\log\hat{\Gamma}(X,Y,Z) - \log\Gamma(X,Y,Z) \mid \hat{\theta}\right]\right|\\ &+\left|\log\mathbb{E}_{\tilde{Q}}\left[\hat{\Gamma}(X,Y,Z) \mid \hat{\theta}\right] - \log\mathbb{E}_{\tilde{Q}}\left[\Gamma(X,Y,Z)\right]\right|.\end{aligned} \tag{A54}$$

Thus from (A48) and the Lipschitz continuity of Γ, $\hat{\Gamma}$, and $\log(\cdot)$, it follows that:

$$\begin{aligned}&\mathbb{P}\left(\lim_{n\to\infty}\left|\mathbb{E}_{\tilde{P}}\left[\log\hat{\Gamma}(X,Y,Z) - \log\Gamma(X,Y,Z) \mid \hat{\theta}\right]\right| > \frac{\eta}{\tau(1-\tau)}\right) = 0,]\\ &\mathbb{P}\left(\lim_{n\to\infty}\left|\log\mathbb{E}_{\tilde{Q}}\left[\hat{\Gamma}(X,Y,Z) \mid \hat{\theta}\right] - \log\mathbb{E}_{\tilde{Q}}\left[\Gamma(X,Y,Z)\right]\right| > \frac{\eta}{\tau^2}\right) = 0.\end{aligned} \tag{A55}$$

Then combining (A50) and (A52)–(A55), it is concluded that with probability one as $n \to \infty$

$$\left|\hat{I}^n_{DV}(X;Y|Z) - I(X;Y|Z)\right| \le \Delta_{DV} + \Delta''_{DV} + \frac{\eta}{\tau(1-\tau)} + \frac{\eta}{\tau^2} \tag{A56}$$

$$\le \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{2} = \epsilon, \tag{A57}$$

where the last step holds by choosing $\eta = \tau^2(1-\tau)\frac{\epsilon}{2}$, and ϵ' and ϵ_0 accordingly. In other words,

$$\hat{I}^n_{DV}(X;Y|Z) \overset{\text{a.s.}}{\to} I(X;Y|Z),$$

and the proof of Theorem 1 is completed. □

References

1. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. MINE: Mutual Information Neural Estimation. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 531–540.
2. Wang, Q.; Kulkarni, S.R.; Verdú, S. Universal estimation of information measures for analog sources. *Found. Trends Commun. Inf. Theory* **2009**, *5*, 265–353. [CrossRef]
3. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef] [PubMed]
4. Mukherjee, S.; Asnani, H.; Kannan, S. CCMI: Classifier based Conditional Mutual Information Estimation. In Proceedings of the Uncertainty in Artificial Intelligence, Tel Aviv, Israel, 22–25 July 2019.
5. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 22–24 September 1999; pp. 368–377.
6. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
7. Donsker, M.D.; Varadhan, S.R.S. Asymptotic evaluation of certain markov process expectations for large time, I. *Comm. Pure Appl. Math.* **1975**, *28*, 1–47. [CrossRef]
8. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861. [CrossRef]

9. Poole, B.; Ozair, S.; van den Oord, A.; Alemi, A.A.; Tucker, G. On variational lower bounds of mutual information. In Proceedings of the NeurIPS Workshop on Bayesian Deep Learning, Montréal, QC, Canada, 7–8 December 2018.
10. Molavipour, S.; Bassi, G.; Skoglund, M. Conditional Mutual Information Neural Estimator. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 5025–5029.
11. Molavipour, S.; Bassi, G.; Skoglund, M. Neural Estimators for Conditional Mutual Information Using Nearest Neighbors Sampling. *IEEE Trans. Signal Process.* **2021**, *69*, 766–780. [CrossRef]
12. Marko, H. The bidirectional communication theory-a generalization of information theory. *IEEE Trans. Commum.* **1973**, *21*, 1345–1351. [CrossRef]
13. Massey, J. Causality, Feedback and Directed Information. In Proceedings of the International Symposium on Information Theory and Its Applications (ISITA), Honolulu, HI, USA, 27–30 November 1990; pp. 303–305.
14. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461. [CrossRef] [PubMed]
15. Kramer, G. Directed Information for Channels with Feedback. Ph.D. Thesis, Department of Information Technology and Electrical Engineering, ETH Zurich, Zürich, Switzerland, 1998.
16. Permuter, H.H.; Kim, Y.H.; Weissman, T. Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *IEEE Trans. Inf. Theory* **2011**, *57*, 3248–3259. [CrossRef]
17. Venkataramanan, R.; Pradhan, S.S. Source coding with feed-forward: Rate-distortion theorems and error exponents for a general source. *IEEE Trans. Inf. Theory* **2007**, *53*, 2154–2179. [CrossRef]
18. Tanaka, T.; Skoglund, M.; Sandberg, H.; Johansson, K.H. Directed information and privacy loss in cloud-based control. In Proceedings of the American Control Conference (ACC), Seattle, WD, USA, 24–26 May 2017; pp. 1666–1672.
19. Rissanen, J.; Wax, M. Measures of mutual and causal dependence between two time series (Corresp.). *IEEE Trans Inf. Theory* **1987**, *33*, 598–601. [CrossRef]
20. Quinn, C.J.; Coleman, T.P.; Kiyavash, N.; Hatsopoulos, N.G. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. Comput. Neurosci.* **2011**, *30*, 17–44. [CrossRef] [PubMed]
21. Cai, Z.; Neveu, C.L.; Baxter, D.A.; Byrne, J.H.; Aazhang, B. Inferring neuronal network functional connectivity with directed information. *J. Neurophysiol.* **2017**, *118*, 1055–1069. [CrossRef] [PubMed]
22. Ver Steeg, G.; Galstyan, A. Information transfer in social media. In Proceedings of the 21st international conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 509–518.
23. Quinn, C.J.; Kiyavash, N.; Coleman, T.P. Directed information graphs. *IEEE Trans. Inf. Theory* **2015**, *61*, 6887–6909. [CrossRef]
24. Vicente, R.; Wibral, M.; Lindner, M.; Pipa, G. Transfer entropy—A model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* **2011**, *30*, 45–67. [CrossRef] [PubMed]
25. Chávez, M.; Martinerie, J.; Le Van Quyen, M. Statistical assessment of nonlinear causality: Application to epileptic EEG signals. *J. Neurosci. Meth.* **2003**, *124*, 113–128. [CrossRef]
26. Spinney, R.E.; Lizier, J.T.; Prokopenko, M. Transfer entropy in physical systems and the arrow of time. *Phys. Rev. E* **2016**, *94*, 022135. [CrossRef]
27. Runge, J. Quantifying information transfer and mediation along causal pathways in complex systems. *Phys. Rev. E* **2015**, *92*, 062829. [CrossRef]
28. Murin, Y. *k*-NN Estimation of Directed Information. *arXiv* **2017**, arXiv:1711.08516.
29. Faes, L.; Kugiumtzis, D.; Nollo, G.; Jurysta, F.; Marinazzo, D. Estimating the decomposition of predictive information in multivariate systems. *Phys. Rev. E* **2015**, *91*, 032904. [CrossRef]
30. Baboukani, P.S.; Graversen, C.; Alickovic, E.; Østergaard, J. Estimating Conditional Transfer Entropy in Time Series Using Mutual Information and Nonlinear Prediction. *Entropy* **2020**, *22*, 1124. [CrossRef]
31. Zhang, J.; Simeone, O.; Cvetkovic, Z.; Abela, E.; Richardson, M. ITENE: Intrinsic Transfer Entropy Neural Estimator. *arXiv* **2019**, arXiv:1912.07277.
32. Aharoni, Z.; Tsur, D.; Goldfeld, Z.; Permuter, H.H. Capacity of Continuous Channels with Memory via Directed Information Neural Estimator. *arXiv* **2020**, arXiv:2003.04179.
33. Schäfer, A.M.; Zimmermann, H.G. Recurrent neural networks are universal approximators. *Int. J. Neural Syst.* **2007**, *17*, 253–263. [CrossRef] [PubMed]
34. Breiman, L. The individual ergodic theorem of information theory. *Ann. Math. Stat.* **1957**, *28*, 809–811. [CrossRef]
35. Kontoyiannis, I.; Skoularidou, M. Estimating the directed information and testing for causality. *IEEE Trans. Inf. Theory* **2016**, *62*, 6053–6067. [CrossRef]
36. Molavipour, S.; Bassi, G.; Skoglund, M. Testing for directed information graphs. In Proceedings of the Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 3–6 October 2017; pp. 212–219.
37. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
38. Devroye, L.; Gyorfi, L.; Krzyzak, A.; Lugosi, G. On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Stat.* **1994**, *22*, 1371–1385. [CrossRef]
39. Collomb, G. Nonparametric time series analysis and prediction: Uniform almost sure convergence of the window and k-NN autoregression estimates. *Statistics* **1985**, *16*, 297–307. [CrossRef]
40. Yakowitz, S. Nearest-neighbour methods for time series analysis. *J. Time Ser. Anal.* **1987**, *8*, 235–247. [CrossRef]

41. Meyn, S.P.; Tweedie, R.L. *Markov Chains and Stochastic Stability*; Springer Science & Business Media: Dordrecht, The Netherlands, 2012.
42. Raleigh, G.G.; Cioffi, J.M. Spatio-temporal coding for wireless communication. *IEEE Trans. Inf. Theory* **1998**, *46*, 357–366. [CrossRef]
43. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
44. Kamarianakis, Y.; Prastacos, P. Space–time modeling of traffic flow. *Comput. Geosci.* **2005**, *31*, 119–133. [CrossRef]
45. Molavipour, S.; Bassi, G.; Čičić, M.; Skoglund, M.; Johansson, K.H. Causality Graph of Vehicular Traffic Flow. *arXiv* **2020**, arXiv:2011.11323.
46. Ross, S.M.; Peköz, E.A. A Second Course in Probability. 2007. Available online: www.bookdepository.com/publishers/Pekozbooks (accessed on 20 May 2021).
47. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
48. Györfi, L.; Härdle, W.; Sarda, P.; Vieu, P. *Nonparametric Curve Estimation from Time Series*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 60.
49. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.