

Universal Estimation of Directed Information

Jiantao Jiao, *Student Member, IEEE*, Haim H. Permuter, *Member, IEEE*, Lei Zhao, Young-Han Kim, *Senior Member, IEEE*, and Tsachy Weissman, *Fellow, IEEE*

Abstract—Four estimators of the directed information rate between a pair of jointly stationary ergodic finite-alphabet processes are proposed, based on universal probability assignments. The first one is a Shannon–McMillan–Breiman-type estimator, similar to those used by Verdú in 2005 and Cai *et al.* in 2006 for estimation of other information measures. We show the almost sure and L_1 convergence properties of the estimator for any underlying universal probability assignment. The other three estimators map universal probability assignments to different functionals, each exhibiting relative merits such as smoothness, nonnegativity, and boundedness. We establish the consistency of these estimators in almost sure and L_1 senses, and derive near-optimal rates of convergence in the minimax sense under mild conditions. These estimators carry over directly to estimating other information measures of stationary ergodic finite-alphabet processes, such as entropy rate and mutual information rate, with near-optimal performance and provide alternatives to classical approaches in the existing literature. Guided by these theoretical results, the proposed estimators are implemented using the context-tree weighting algorithm as the universal probability assignment. Experiments on synthetic and real data are presented, demonstrating the potential of the proposed schemes in practice and the utility of directed information estimation in detecting and measuring causal influence and delay.

Index Terms—Causal influence, context-tree weighting (CTW), directed information, rate of convergence, universal probability assignment.

I. INTRODUCTION

FIRST introduced by Marko [1] and Massey [2], directed information arises as a natural counterpart of mutual information for channel capacity when causal feedback from the

receiver to the sender is present. In [3] and [4], Kramer extended the use of directed information to discrete memoryless networks with feedback, including the two-way channel and the multiple access channel. Tatikonda and Mitter [5] used directed information spectrum to establish a general feedback channel coding theorem for channels with memory. For a class of stationary channels with feedback, where the output is a function of the current and past m inputs and channel noise, Kim [6] proved that the feedback capacity is equal to the limit of the maximum normalized directed information from the input to the output. Permuter *et al.* [7] considered the capacity of discrete-time finite-state channels with feedback where the feedback is a time-invariant function of the output. Under mild conditions, they showed that the capacity is again the limit of the maximum normalized directed information. Recently, Permuter *et al.* [8] showed that directed information plays an important role in portfolio theory, data compression, and hypothesis testing under causality constraints.

Beyond information theory, directed information is a valuable tool in biology, for it provides an alternative to the notion of Granger causality [9], which has been perhaps the most widely used means of identifying causal influence between two random processes. For example, Mathai *et al.* [10] used directed information to identify pairwise influence in gene networks. Similarly, Rao *et al.* [11] used directed information to test the direction of influence in gene networks.

Since directed information has significance in various fields, it is of both theoretical and practical importance to develop efficient methods of estimating it. The problem of estimating information measures, such as entropy, relative entropy, and mutual information, has been extensively studied in the literature. Verdú [12] gave an overview of universal estimation of information measures. Wyner and Ziv [13] applied the idea of Lempel–Ziv parsing to estimate entropy rate, which converges in probability for all stationary ergodic processes. Ziv and Merhav [14] used Lempel–Ziv parsing to estimate relative entropy (Kullback–Leibler divergence) and established consistency under the assumption that the observations are generated by independent Markov sources. Cai *et al.* [15] proposed two universal relative entropy estimators for finite-alphabet sources, one based on the Burrows–Wheeler transform (BWT) [16] and the other based on the context-tree weighting (CTW) algorithm [17]. The BWT-based estimator was applied in universal entropy estimation by Cai *et al.* [18], while the CTW-based one was applied in universal erasure entropy estimation by Yu and Verdú [19].

For the problem of estimating directed information, Quinn *et al.* [20] developed an estimator to infer causality in an ensemble of neural spike train recordings. Assuming a parametric generalized linear model and stationary ergodic Markov processes, they established strong consistency results. Compared to [20], Zhao *et al.* [21] focused on universal methods for arbitrary sta-

Manuscript received January 11, 2012; revised October 16, 2012; accepted October 23, 2012. Date of publication July 09, 2013; date of current version September 11, 2013. This work was supported in part by the Center for Science of Information (CSOI), in part by the NSF Science and Technology Center under Grant CCF-0939370, in part by the US–Israel Binational Science Foundation (BSF) under Grant 2008402, in part by the NSF Grant CCF-0939370, and in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-10-1-0124. H. H. Permuter was supported in part by the Marie Curie Reintegration fellowship and in part by the GIF Young Scientists’ Program. This paper was presented in part at the 2010 IEEE International Symposium on Information Theory and the 2012 IEEE International Symposium on Information Theory.

J. Jiao is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: jiantao@stanford.edu).

H. H. Permuter is with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: haimp@bgu.ac.il).

L. Zhao was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is now with Jump Operations, Chicago, IL 60654 USA (e-mail: zhaolei122@gmail.com).

Y.-H. Kim is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: yhk@ucsd.edu).

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu).

Communicated by I. Kontoyiannis, Associate Editor At Large.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIT.2013.2267934

tionary ergodic processes with finite alphabet and showed their L_1 consistencies.

As an improvement and extension of [21], the main contribution of this paper is a general framework for estimating information measures of stationary ergodic finite-alphabet processes, using “single-letter” information-theoretic functionals. Although our methods can be applied in estimating a number of information measures, for concreteness and relevance to emerging applications we focus on estimating the directed information rate between a pair of jointly stationary ergodic finite-alphabet processes.

The first proposed estimator is adapted from the universal relative entropy estimator in [15] using the CTW algorithm, and we provide a refined analysis yielding strong consistency results. We further propose three additional estimators in a unified framework, present both weak and strong consistency results, and establish near-optimal rates of convergence under mild conditions. We then employ our estimators on both simulated and real data, showing their effectiveness in measuring channel delays and causal influences between different processes. In particular, we use these estimators on the daily stock market data from 1990 to 2011 to observe a significant level of causal influence from the Dow Jones Industrial Average to the Hang Seng Index, but relatively low causal influence in the reverse direction.

The rest of the paper is organized as follows. Section II reviews preliminaries on directed information, universal probability assignments, and the CTW algorithm. Section III presents our proposed estimators and their basic properties. Section IV is dedicated to performance guarantees of the proposed estimators, including their consistencies and minimax-optimal rates of convergence. Section V shows experimental results in which we apply the proposed estimators to simulated and real data. Section VI concludes the paper. The proofs of the main results are given in the Appendixes.

II. PRELIMINARIES

We begin with mathematical definitions of directed information and causally conditional entropy. We also define universal and pointwise universal probability assignments. We then introduce the CTW algorithm used in our implementations of the universal estimators that are introduced in the next section.

Throughout the paper, we use uppercase letters $X, Y, \dots$ to denote random variables and lowercase letters $x, y, \dots$ to denote values they assume. By convention, $X = \emptyset$ means that X is a degenerate random variable (unspecified constant) regardless of its support. We denote the n -tuple $(X_1, X_2, \dots, X_n)$ as X^n and $(x_1, x_2, \dots, x_n)$ as x^n . Calligraphic letters $\mathcal{X}, \mathcal{Y}, \dots$ denote alphabets of $X, Y, \dots$, and $|\mathcal{X}|$ denotes the cardinality of $\mathcal{X}$. Boldface letters $\mathbf{X}, \mathbf{Y}, \dots$ denote stochastic processes, and throughout this paper, they are finite-alphabet.

Given a probability law P , $P(x^i) = P\{X^i = x^i\}$ denotes the probability mass function (pmf) of X^i and $P(x_i|x^{i-1})$ denotes the conditional pmf of X_i given $\{X^{i-1} = x^{i-1}\}$, i.e., with slight abuse of notation, x_i here is a dummy variable and $P(x_i|x^{i-1})$ is an element of $\mathcal{M}(\mathcal{X})$, the probability simplex on $\mathcal{X}$, representing the said conditional pmf. Accordingly,

$P(x_i|X^{i-1})$ denotes the conditional pmf $P(x_i|x^{i-1})$ evaluated for the random sequence X^{i-1} , which is an $\mathcal{M}(\mathcal{X})$ -valued random vector, while $P(X_i|X^{i-1})$ is the random variable denoting the X_i th component of $P(x_i|X^{i-1})$. Throughout this paper, $\log(\cdot)$ is base 2 and $\ln(\cdot)$ is base e .

A. Directed Information

Given a pair of random sequences X^n and Y^n , the *directed information* from X^n to Y^n is defined as

$$I(X^n \rightarrow Y^n) \triangleq \sum_{i=1}^n I(X^i; Y_i | Y^{i-1}) \quad (1)$$

$$= H(Y^n) - H(Y^n \| X^n), \quad (2)$$

where $H(Y^n \| X^n)$ is the *causally conditional entropy* [3], defined as

$$H(Y^n \| X^n) \triangleq \sum_{i=1}^n H(Y_i | Y^{i-1}, X^i). \quad (3)$$

Compared to mutual information

$$I(X^n; Y^n) = H(Y^n) - H(Y^n | X^n), \quad (4)$$

directed information in (2) has the causally conditional entropy in place of the conditional entropy. Thus, unlike mutual information, directed information is not symmetric, i.e., $I(Y^n \rightarrow X^n) \neq I(X^n \rightarrow Y^n)$, in general.

The following notation of *causally conditional pmfs* will be used throughout

$$p(x^n \| y^n) = \prod_{i=1}^n p(x_i | x^{i-1}, y^i), \quad (5)$$

$$p(x^n \| y^{n-1}) = \prod_{i=1}^n p(x_i | x^{i-1}, y^{i-1}). \quad (6)$$

It can be easily verified that

$$p(x^n, y^n) = p(y^n \| x^n) p(x^n \| y^{n-1}) \quad (7)$$

and that we have the *conservation laws*

$$I(X^n; Y^n) = I(X^n \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n), \quad (8)$$

$$I(X^n; Y^n) = I(X^{n-1} \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n) + \sum_{i=1}^n I(X_i; Y_i | X^{i-1}, Y^{i-1}), \quad (9)$$

where

$$I(Y^{n-1} \rightarrow X^n) = I((\emptyset, Y^{n-1}) \rightarrow X^n) \quad (10)$$

$$= H(X^n) - \sum_{i=1}^n H(X_i | X^{i-1}, Y^{i-1}) \quad (11)$$

denotes the *reverse directed information*. Other interesting properties of directed information can be found in [3], [22], and [23].

The *directed information rate* [3] between a pair of jointly stationary finite-alphabet processes $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} I(X^n \rightarrow Y^n). \quad (12)$$

The existence of the limit can be checked [3] as

$$\bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) = \lim_{n \rightarrow \infty} \frac{1}{n} I(X^n \rightarrow Y^n) \quad (13)$$

$$= \lim_{n \rightarrow \infty} \frac{1}{n} (H(Y^n) - H(Y^n \| X^n)) \quad (14)$$

$$= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(Y_i | Y^{i-1}) \\ - \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(Y_i | Y^{i-1}, X^i) \quad (15)$$

$$= H(Y_0 | Y_{-\infty}^{-1}) - H(Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1}), \quad (16)$$

where the last equality is obtained via the property of Cesàro mean and standard martingale arguments; see [24, Chs. 4 and 16]. Note that the entropy rate $\bar{H}(\mathbf{Y})$ of the process $\mathbf{Y}$ is equal to $H(Y_0 | Y_{-\infty}^{-1})$. In a similar vein, the *causally conditional entropy rate* is defined as

$$\bar{H}(\mathbf{Y} \| \mathbf{X}) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} H(Y^n \| X^n) \quad (17)$$

$$= H(Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1}). \quad (18)$$

Thus,

$$\bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) = \bar{H}(\mathbf{Y}) - \bar{H}(\mathbf{Y} \| \mathbf{X}). \quad (19)$$

This identity shows that if we estimate $\bar{H}(\mathbf{Y})$ and $\bar{H}(\mathbf{Y} \| \mathbf{X})$ separately and if both estimates converge, we have a convergent estimate of the directed information rate.

B. Universal Probability Assignment

A probability assignment Q consists of a set of conditional pmfs $Q(x_i | x^{i-1})$ for every $x^{i-1} \in \mathcal{X}^{i-1}$ and $i = 1, 2, \dots$. Note that Q induces a probability measure on a random process $\mathbf{X}$ and the pmf $Q(x^n) = Q(x_1)Q(x_2|x_1) \cdots Q(x_n|x^{n-1})$ on X^n for each n .

Definition 1 (Universal Probability Assignment): Let P be a class of probability measures. A probability assignment Q is said to be *universal for the class P* if the normalized relative entropy satisfies

$$\lim_{n \rightarrow \infty} \frac{1}{n} D(P(x^n) \| Q(x^n)) = 0 \quad (20)$$

for every probability measure P in P . A probability assignment Q is said to be *universal* (without a qualifier) if it is universal for the class of stationary probability measures.

Definition 2 (Pointwise Universal Probability Assignment): A probability assignment Q is said to be *pointwise universal for P* if

$$\limsup_{n \rightarrow \infty} \left(\frac{1}{n} \log \frac{1}{Q(X^n)} - \frac{1}{n} \log \frac{1}{P(X^n)} \right) \leq 0 \quad P\text{-a.s.} \quad (21)$$

for every $P \in P$. A probability assignment Q is said to be *pointwise universal* if it is pointwise universal for the class of stationary ergodic probability measures.

It is well known that there exist universal and pointwise universal probability assignments. Ornstein[25] constructed a pointwise universal probability assignment, which was generalized to Polish spaces by Algoet[26]. Morvai *et al.* [27] used universal source codes to induce a probability assignment and established its universality. Since the quantity

$(1/n) \log(1/Q(X^n))$ is generally unbounded, a pointwise universal probability assignment is not necessarily universal. However, if we have a pointwise universal probability assignment, it is easy to construct a probability assignment that is both pointwise universal and universal. Let $Q_1(x^n)$ be a pointwise universal probability assignment and $Q_2(x^n)$ be the i.i.d. uniform distribution, then it is easy to verify that

$$\tilde{Q}(x^n) = a_n Q_2(x^n) + (1 - a_n) Q_1(x^n) \quad (22)$$

is both universal and pointwise universal provided that a_n decays subexponentially, for example, $a_n = 1/n$. For more discussions on universal probability assignments, see, for example, [28] and the references therein.

C. Context-Tree Weighting

The sequential probability assignment we use in the implementations of our directed information estimators is the celebrated CTW algorithm by Willems *et al.* [17]. One of the main advantages of the CTW algorithm is that its computational complexity is linear in the block length n , and the algorithm provides the probability assignments Q directly; see [17] and [29]. Note that while the original CTW algorithm was tuned for binary processes, it has been extended for larger alphabets in [30], an extension that we use in this paper. In our experiments with simulated data, we assume that the depth of the context tree is larger than the memory of the source. This assumption can be alleviated by the algorithm introduced by Willems [31], which we will not implement in this paper.

An example of a context tree of input sequence x_{-2}^8 with a binary alphabet is shown in Fig. 1. In general, each node in the tree corresponds to a context, which is a string of symbols preceding the symbol that follows. For concreteness, assume the alphabet is $\{0, 1, \dots, M-1\}$. With a slight abuse of notation, we use s to represent both a node in the context tree and a specific context. At every node s , we use a length- M array $(a_{0,s}, a_{1,s}, \dots, a_{M-1,s})$ to count the numbers of different values emitted with context s in sequence x^n . In Fig. 1, the counts $(a_{0,s}, a_{1,s})$ are marked near each node s , and they are simply numbers of zeros and ones emitted from node s .

Take any sequence z^n whose alphabet is $\{0, 1, \dots, M-1\}$. If z^n contains b_0 zeros, b_1 ones, b_2 twos, and so on, the Krichevsky–Trofimov probability estimate of z^n [32], i.e., $P_e(z^n) = P_e(b_0, b_1, \dots, b_{M-1})$ can be computed sequentially. We let $P_e(0, 0, \dots, 0) = 1$, and for $b_0 \geq 0, b_1 \geq 0, \dots, b_{M-1} \geq 0, 0 \leq i \leq M-1$, we have

$$P_e(b_0, b_1, \dots, b_{i-1}, b_i + 1, b_{i+1}, \dots, b_{M-1}) \\ \triangleq \frac{b_i + 1/2}{b_0 + b_1 + \dots + b_{M-1} + M/2} \\ \times P_e(b_0, b_1, \dots, b_{i-1}, b_i, b_{i+1}, \dots, b_{M-1}). \quad (23)$$

We denote the Krichevsky–Trofimov probability estimate of the M -array counts at node s of sequence x^n as $P_e^s(x^n)$. The weighted probability P_w^s at node s of sequence x^n in the CTW algorithm is calculated as

$$P_w^s(x^n) = \begin{cases} \frac{1}{2} P_e^s(x^n) + \frac{1}{2} \prod_{i=0}^{M-1} P_w^{is}(x^n) & 0 \leq l(s) < D \\ P_e^s(x^n) & l(s) = D \end{cases} \quad (24)$$

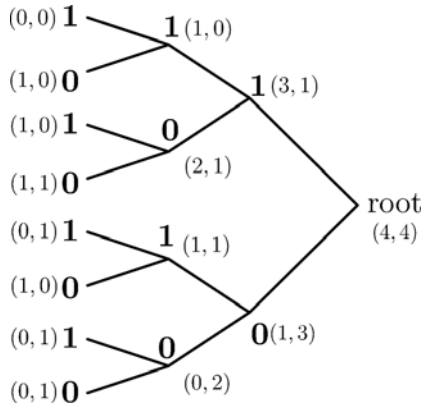


Fig. 1. Illustration of the CTW algorithm when $D = 3$ and $(x_{-2}, x_{-1}, x_0, x_1, \dots, x_8) = 00011010010$. The count starts at x_1 . For example, there are 3 zeros and 1 one with context 1, represented by count $(3, 1)$ at the node of context 1 in the upper right.

where the node is is the i th child of node s and $l(s)$ is the depth of node s . When we build the context tree from sequence x^n , we add symbols one by one. In adding symbol x_t , $1 \leq t \leq n$, we have to update the counts $(a_{0,s}, a_{1,s}, \dots, a_{M-1,s})$, the estimated probability P_e^s , and the weighted probability P_w^s for each context s of x_t . The order of updates is from the context of the longest depth (a leaf node) to the root.

Let λ denote the root node of the context tree, then $P_w^\lambda(x^n)$ is the universal probability assignment in the CTW algorithm, which will be denoted as $Q(x^n)$ in Section III. We compute the sequential probability assignments as

$$Q(x_{n+1}|x^n) = \frac{Q(x^{n+1})}{Q(x^n)} = \frac{P_w^\lambda(x^{n+1})}{P_w^\lambda(x^n)}. \quad (25)$$

In [29, Ch. 5], Willems and Tjalkens introduced a factor $\beta^s(x^n)$ at every node s to simplify the calculation of the sequential probability assignment, which could also help understand how the weighted probabilities are updated when the input sequence x^n grows to x^{n+1} . For each node s , we define factor $\beta^s(x^n)$ as

$$\beta^s(x^n) \triangleq \frac{P_e^s(x^n)}{\prod_{i=0}^{M-1} P_w^{is}(x^n)}. \quad (26)$$

Assuming js is a context of x_{n+1} , where $0 \leq j \leq M-1$. Obviously, any other node ks , $k \neq j$ cannot be a context of x_{n+1} . We

express $P_w^s(X_{n+1} = q|x^n)$, $q = 0, 1, \dots, M-1$ in (33) given at the bottom of this page, which shows that the sequential probability assignment $Q(x_{n+1}|x^n)$ is a weighted summation of the Krichevsky–Trofimov sequential probability assignments, i.e., $P_e^s(X_{n+1} = q|x^n)$ at all nodes of the context tree.

By (23), for any node s ,

$$P_e^s(X_{n+1} = q|x^n) \geq \frac{1/2}{n + |\mathcal{X}|/2} \geq \frac{1}{2n + |\mathcal{X}|}. \quad (27)$$

Thus, $Q(x_{n+1}|x^n) = \Omega(1/n)$, or more precisely,

$$Q(x_{n+1}|x^n) \geq \frac{1}{2n + |\mathcal{X}|}. \quad (28)$$

The probability assignment Q in the CTW algorithm is both universal and pointwise universal for the class of stationary irreducible aperiodic finite-alphabet Markov processes. For the proof of universality, see [17]. The pointwise universality is proved in Lemma 2 in Appendix A.

III. FOUR ESTIMATORS

In this section, we introduce four estimators of the directed information rate $\bar{I}(\mathbf{X} \rightarrow \mathbf{Y})$ of a pair $(\mathbf{X}, \mathbf{Y})$ of jointly stationary ergodic processes with finite alphabets. Let $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ be the set of all probability distributions on $\mathcal{X} \times \mathcal{Y}$. Define f to be the function that maps a joint pmf $P(x, y)$ of a random pair (X, Y) to the corresponding conditional entropy $H(Y|X)$, i.e.,

$$f(P) \triangleq - \sum_{x,y} P(x, y) \log P(y|x), \quad (29)$$

where $P(y|x)$ is the conditional pmf induced by $P(x, y)$. Take Q as a universal probability assignment, either on processes with $(\mathcal{X} \times \mathcal{Y})$ -valued components, or with $\mathcal{Y}$ -valued components, as will be clear from the context. See (30) and (33), at the bottom of the page.

Recall the definition of the directed information from X^n to Y^n :

$$I(X^n \rightarrow Y^n) = \sum_{i=1}^n I(X^i; Y_i | Y^{i-1}) = H(Y^n) - H(Y^n \| X^n), \quad (34)$$

$$P_w^s(X_{n+1} = q|x^n) = \frac{P_w^s(X_{n+1} = q, x^n)}{P_w^s(x^n)} \quad (30)$$

$$= \frac{\frac{1}{2} P_e^s(X_{n+1} = q, x^n) + \frac{1}{2} \prod_{i=0}^{M-1} P_w^{is}(X_{n+1} = q, x^n)}{P_w^s(x^n)} \quad (31)$$

$$= \frac{\frac{1}{2} P_e^s(x^n) P_e^s(X_{n+1} = q|x^n) + \frac{1}{2} P_w^{js}(X_{n+1} = q|x^n) \prod_{i=0}^{M-1} P_w^{is}(x^n)}{P_w^s(x^n)} \quad (32)$$

$$= \frac{\beta^s(x^n)}{1 + \beta^s(x^n)} P_e^s(X_{n+1} = q|x^n) + \frac{1}{1 + \beta^s(x^n)} P_w^{js}(X_{n+1} = q|x^n) \quad (33)$$

we define the four estimators as follows:

$$\hat{I}_1(X^n \rightarrow Y^n) \triangleq \hat{H}_1(Y^n) - \hat{H}_1(Y^n \| X^n), \quad (35)$$

$$\hat{I}_2(X^n \rightarrow Y^n) \triangleq \hat{H}_2(Y^n) - \hat{H}_2(Y^n \| X^n), \quad (36)$$

$$\hat{I}_3(X^n \rightarrow Y^n) \triangleq \frac{1}{n} \sum_{i=1}^n D(Q(y_i | X^i, Y^{i-1}) \| Q(y_i | Y^{i-1})), \quad (37)$$

$$\hat{I}_4(X^n \rightarrow Y^n) \triangleq \frac{1}{n} \sum_{i=1}^n D(Q(x_{i+1}, y_{i+1} | X^i, Y^i) \| Q(y_{i+1} | Y^i) Q(x_{i+1} | X^i, Y^i)), \quad (38)$$

where

$$\hat{H}_1(Y^n \| X^n) \triangleq -\frac{1}{n} \log Q(Y^n \| X^n), \quad (39)$$

$$\hat{H}_2(Y^n \| X^n) \triangleq \frac{1}{n} \sum_{i=1}^n f(Q(x_{i+1}, y_{i+1} | X^i, Y^i)), \quad (40)$$

$$\hat{H}_2(Y^n) \triangleq \frac{1}{n} \sum_{i=1}^n \sum_{y_{i+1}} Q(y_{i+1} | Y^i) \log \frac{1}{Q(y_{i+1} | Y^i)}, \quad (41)$$

$$\hat{H}_1(Y^n) \triangleq \hat{H}_1(Y^n \| \emptyset^n). \quad (42)$$

Recall that $Q(y_i | X^i, Y^{i-1})$ denotes the conditional pmf $Q(y_i | x^i, y^{i-1})$ evaluated for the random sequence (X^i, Y^{i-1}) , and $Q(Y^n \| X^n)$ denotes the causally conditional pmf $Q(y^n \| x^n)$ evaluated for (X^n, Y^n) . Thus, an entropy estimate such as $\hat{H}_1(Y^n \| X^n)$ is a *random variable* (since it is a function of (X^n, Y^n)), as opposed to the entropy terms such as $H(Y^n \| X^n)$, which are deterministic and depend on the *distribution* of (X^n, Y^n) .

Note that in (37) and (38) the universal probability assignments conditioned on different data are calculated separately. For example, $Q(y_i | Y^{i-1})$ is not computed from $Q(x_i, y_i | X^{i-1}, Y^{i-1})$, but from running the universal probability assignment algorithm again on dataset Y^{i-1} . In the case of $Q(Y_i | X^i, Y^{i-1})$, which is inherent in the computation of $Q(Y^n \| X^n)$, the estimate is computed from pmf $Q(x_i, y_i | X^{i-1}, Y^{i-1})$ via $Q(Y_i | X^i, Y^{i-1}) = Q(X_i, Y_i | X^{i-1}, Y^{i-1}) / \sum_{y_i} Q(X_i, y_i | X^{i-1}, Y^{i-1})$.

We can express $\hat{I}_4$ in another form which might be enlightening

$$\hat{I}_4 = G_n - \hat{H}_2(Y^n \| X^n), \quad (43)$$

where G_n is

$$G_n = \frac{1}{n} \sum_{i=1}^n \sum_{(x_{i+1}, y_{i+1})} Q(x_{i+1}, y_{i+1} | X^i, Y^i) \log \frac{1}{Q(y_{i+1} | Y^i)}. \quad (44)$$

It is also worthwhile to note that $\hat{I}_4$ involves an average of x_{i+1} in the relative entropy term for each i , which makes it analytically different from $\hat{I}_3$.

Here is the big picture of the general ideas behind these estimators. The first estimator $\hat{I}_1$ is calculated through the difference of two terms, each of which takes the form of

(39). Since the Shannon–McMillan–Breiman theorem guarantees the asymptotic equipartition property (AEP) of entropy rate [24] as well as directed information rate [33], it is natural to believe that $\hat{I}_1$ would converge to the directed information rate. This is indeed the case, which is proved in Appendix B. The Shannon–McMillan–Breiman-type estimators have been widely applied in the literature of information-theoretic measure estimation, for example, relative entropy estimation by Cai *et al.* [15], and erasure entropy estimation by Yu and Verdú [19].

Equation (39) can be rewritten in the Cesàro mean form, i.e.,

$$-\frac{1}{n} \log Q(Y^n \| X^n) = \frac{1}{n} \sum_{i=1}^n \log \frac{1}{Q(Y_i | Y^{i-1}, X^i)}, \quad (45)$$

and estimators $\hat{I}_2$ through $\hat{I}_4$ are derived by changing every term in the Cesàro mean to other functionals of probability assignments Q . For concreteness, estimator $\hat{I}_2$ uses conditional entropy as the functional, and estimators $\hat{I}_3$ and $\hat{I}_4$ use relative entropy.

One disadvantage of $\hat{I}_1$ is that it has a nonzero probability of being very large, since it only averages over logarithms of estimated conditional probabilities, while the directed information rate that it estimates is always bounded by $\log |\mathcal{Y}|$.

The estimator $\hat{I}_2$ is the universal directed information estimator introduced in [21]. Thanks to the use of information-theoretic functionals to “smooth” the estimate, the absolute value of $\hat{I}_2(X^n \rightarrow Y^n)$ is upper bounded by $\log |\mathcal{Y}|$ on any realization, a clear advantage over $\hat{I}_1$.

The common disadvantage of $\hat{I}_1$ and $\hat{I}_2$ is that they are computed by subtraction of two nonnegative quantities. When there is insufficient data, or the stationarity assumption is violated, $\hat{I}_1$ and $\hat{I}_2$ may generate negative outputs, which is clearly undesirable. In order to overcome this, $\hat{I}_3$ and $\hat{I}_4$ are introduced, which take the form of a (random) relative entropy and are always nonnegative. Section V-D gives an example where $\hat{I}_1$ and $\hat{I}_2$ give negative estimates, which might be caused by the fact that the underlying process (stock market) is not stationary, at least in a short term.

IV. PERFORMANCE GUARANTEES

In this section, we establish the consistency of the proposed estimators, mainly in the almost sure and L_1 senses. Under some mild conditions, we derive near-optimal rates of convergence in the minimax sense. The proofs of the stated results are given in the Appendixes.

Theorem 1: Let Q be a universal probability assignment and $(\mathbf{X}, \mathbf{Y})$ be a pair of jointly stationary ergodic finite-alphabet processes. Then,

$$\lim_{n \rightarrow \infty} \hat{I}_1(X^n \rightarrow Y^n) = \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \quad \text{in } L_1. \quad (46)$$

Furthermore, if Q is also a pointwise universal probability assignment, then the limit in (46) holds almost surely as well.

The proof of Theorem 1 is in Appendix B-A. If $(\mathbf{X}, \mathbf{Y})$ is a stationary irreducible aperiodic finite-alphabet Markov process, we can say more about the performance of $\hat{I}_1$ using the probability assignment in the CTW algorithm.

Proposition 1: Let Q be the CTW probability assignment and let $(\mathbf{X}, \mathbf{Y})$ be a jointly stationary irreducible aperiodic finite-alphabet Markov process whose order is bounded by the prescribed tree depth in the CTW algorithm, and let $\mathbf{Y}$ be a stationary irreducible aperiodic finite-alphabet Markov process with the same order as $(\mathbf{X}, \mathbf{Y})$. Then, there exists a constant C_1 such that

$$\mathbb{E} \left| \hat{I}_1(X^n \rightarrow Y^n) - \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \right| \leq C_1 n^{-1/2} \log n, \quad (47)$$

and $\forall \epsilon > 0$, P -a.s.

$$\left| \hat{I}_1(X^n \rightarrow Y^n) - \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \right| = o(n^{-1/2} (\log n)^{5/2+\epsilon}). \quad (48)$$

The proof of Proposition 1 is given in Appendix B-B.

We can establish similar consistency results for the second estimator $\hat{I}_2$ in (36).

Theorem 2: Let Q be a universal probability assignment, and finite-alphabet process $(\mathbf{X}, \mathbf{Y})$ be jointly stationary ergodic. Then,

$$\lim_{n \rightarrow \infty} \hat{I}_2(X^n \rightarrow Y^n) = \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \text{ in } L_1. \quad (49)$$

The proof of Theorem 2 is given in Appendix B-C. As was the case for $\hat{I}_1$, if the process $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary irreducible aperiodic finite-alphabet Markov process, we can say more about the performance of $\hat{I}_2$ using the CTW algorithm as follows:

Proposition 2: Let Q be the probability assignment in the CTW algorithm. If $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary irreducible aperiodic finite-alphabet Markov process whose order does not exceed the prescribed tree depth in the CTW algorithm, and $\mathbf{Y}$ is also a stationary irreducible aperiodic finite-alphabet Markov process with the same order as $(\mathbf{X}, \mathbf{Y})$, then

$$\lim_{n \rightarrow \infty} \hat{I}_2(X^n \rightarrow Y^n) = \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \quad P\text{-a.s. and in } L_1, \quad (50)$$

and there exists a constant C_2 such that

$$\mathbb{E} \left| \hat{I}_2(X^n \rightarrow Y^n) - \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \right| \leq C_2 n^{-1/2} (\log n)^{3/2}. \quad (51)$$

The proof of Proposition 2 is given in Appendix B-D.

We also investigate the minimax lower bound of estimating directed information rate, and show the rates of convergence for the first two estimators are optimal within a logarithmic factor. Note that entropy rate is a special case of directed information rate if we take process $\mathbf{Y} = \mathbf{X}$, so the minimax lower bound also applies in the universal entropy estimation situation. Actually in the proof of proposition 3, we indeed reduce the general problem to entropy estimation problem to show the minimax lower bound.

Proposition 3: Let $P(\mathbf{X}, \mathbf{Y})$ be any class of processes that include the class of i.i.d. processes. Then, there exists a positive constant C_3 such that

$$\inf_{\hat{I}} \sup_{P(\mathbf{X}, \mathbf{Y})} \mathbb{E} |\hat{I} - \bar{I}(\mathbf{X} \rightarrow \mathbf{Y})| \geq C_3 n^{-1/2}, \quad (52)$$

where the infimum is over all estimators $\hat{I}$ of the directed information rate based on (X^n, Y^n) .

The proof of Proposition 3 is given in Appendix B-E. Evidently, convergence rates better than $O(n^{-1/2})$ is not attainable even with respect to the class of i.i.d. sources and thus, *a fortiori*, in our setting of a much larger uncertainty set.

For the third and fourth estimators, we establish the following consistency results using the CTW algorithm.

Theorem 3: Let Q be the probability assignment in the CTW algorithm. If $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary irreducible aperiodic finite-alphabet Markov process whose order does not exceed the prescribed tree depth in the CTW algorithm, and $\mathbf{Y}$ is also a stationary irreducible aperiodic finite-alphabet Markov process with the same order as $(\mathbf{X}, \mathbf{Y})$, then

$$\lim_{n \rightarrow \infty} \hat{I}_3(X^n \rightarrow Y^n) = \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \quad P\text{-a.s. and in } L_1. \quad (53)$$

Theorem 4: Let Q be the probability assignment in the CTW algorithm. If $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary irreducible aperiodic finite-alphabet Markov process whose order does not exceed the prescribed tree depth in the CTW algorithm, and $\mathbf{Y}$ is also a stationary irreducible aperiodic finite-alphabet Markov process with the same order as $(\mathbf{X}, \mathbf{Y})$, then

$$\lim_{n \rightarrow \infty} \hat{I}_4(X^n \rightarrow Y^n) = \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) \quad P\text{-a.s. and in } L_1. \quad (54)$$

The proofs of Theorems 3 and 4 are given in Appendixes B-F and B-G.

Remark 1: The properties of the CTW probability assignment we use in the proofs of Theorems 3 and 4 are not only universality and pointwise universality, but also lower boundedness (recall Section II-C).

Remark 2: Note that the assumption that $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary irreducible aperiodic finite-alphabet Markov process does not imply $\mathbf{Y}$ also has these properties. Suppose that $\mathbf{X}$ is a Markov process of order m , $\mathbf{Y}$ is a hidden Markov process whose internal process is $\mathbf{X}$, then it is obvious that joint process $(\mathbf{X}, \mathbf{Y})$ is Markov with order m , but $\mathbf{Y}$ is not a Markov process. In applications, it is sensible to assume that a process $\mathbf{Z}$ can be approximated by Markov processes better and better as the Markov order increases, i.e., there exists constants $C' > 0, 0 \leq \rho < 1$, such that

$$0 \leq H(Z_0 | Z_{-k}^{-1}) - \bar{H}(\mathbf{Z}) \leq \frac{C'}{\ln(2)} \rho^k. \quad (55)$$

It deserves mentioning that the exponentially fast convergence in (55) can be satisfied under mild conditions. For example, as shown in Birch [34], let $\mathbf{G}$ be a Markov process with strictly positive transition probabilities, and $Z_n = \psi(G_n)$, then (55) holds. For more on this “exponential forgetting” property, please refer to Gland and Mevel [35] and Hochwald and Jelenkovic [36].

The properties established for the proposed estimators are summarized in Table I.

V. ALGORITHMS AND NUMERICAL EXAMPLES

In this section, we use the CTW algorithm as the universal probability assignment to describe the corresponding directed

TABLE I
PROPERTIES OF THE PROPOSED ESTIMATORS

	Support	Rates of convergence
$\hat{I}_1$	$(-\infty, \infty)$	$O(n^{-1/2} \log n)$
$\hat{I}_2$	$[-\log \mathcal{Y} , \log \mathcal{Y}]$	$O(n^{-1/2} (\log n)^{3/2})$
$\hat{I}_3$	$[0, \infty)$	—
$\hat{I}_4$	$[0, \infty)$	—

information estimation algorithms and perform experiments on simulated as well as real data. The CTW algorithm [17] has a linear computational complexity in the block length n , and it provides the probability assignment Q directly. A brief introduction on how the CTW works can be found in Section II-C.

For simplicity and concreteness, we explicitly describe the algorithm for computing $\hat{I}_2$. The algorithms for the other estimators are identical, except for the update rule, which is given, respectively, by (35) to (38).

Algorithm 1 Universal Estimator $\hat{I}_2$ Based on the CTW Algorithm

Fix block length n and context tree depth D .

$\hat{I}_2 \leftarrow 0$

for $i \leftarrow 1, n$ **do**

$z_i = (x_i, y_i)$ Make a super symbol with alphabet size $|\mathcal{X}||\mathcal{Y}|$

end for

for $i \leftarrow D + 1, n + 1$ **do**

Gather the context z_{i-D}^{i-1} for the i th symbol z_i .

Update the context tree for every possible value of z_i .

The estimated pmf $Q(z_i|Z^{i-1})$ is obtained along the way.

Gather the context y_{i-D}^{i-1} for the i th symbol y_i .

Update the context tree for every possible value of y_i .

The estimated pmf $Q(y_i|Y^{i-1})$ is obtained along the way.

Update $\hat{I}_2$ as $\hat{I}_2 \leftarrow \hat{I}_2 + f(Q(x_i, y_i|X^{i-1}, Y^{i-1})) - f(Q(y_i|Y^{i-1}))$ where $f(\cdot)$ is defined in (29).

end for

$\hat{I}_2 \leftarrow \hat{I}_2 / (n - D)$

We now present the performance of the estimators on synthetic and real data. The synthetic data are generated using Markov processes that are passed through simple channels such as discrete memory channels (DMC), or channels with intersymbol interference. We compare the performances of the estimators to each other, as well as the ground truth, which we are able to analytically compute. We also extend the proposed methods to estimation of directed information with delay, and to estimation of mutual information. Further, we show how one can use the directed information estimator to detect delay of a



Fig. 2. Section V-A setup: $\mathbf{X}$ is a binary first order Markov process with transition probability p , and $\mathbf{Y}$ is the output of a binary symmetric channel with crossover probability ϵ corresponding to the input $\mathbf{X}$.

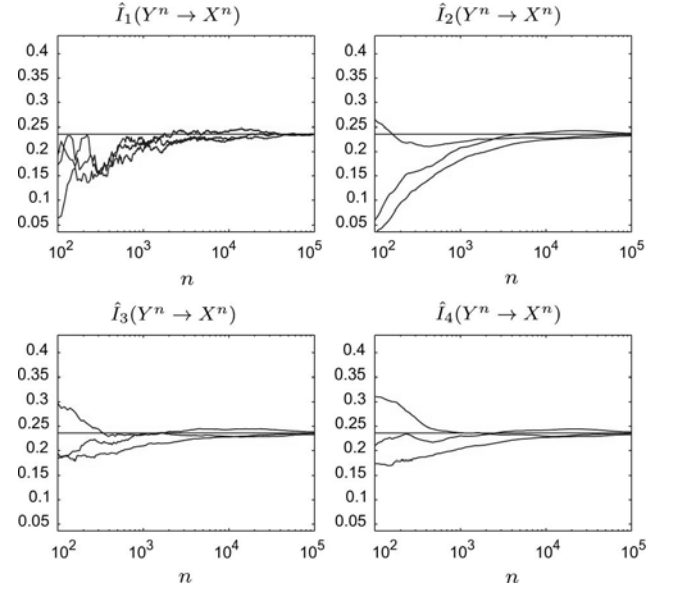


Fig. 3. Estimation of $\bar{I}(\mathbf{Y} \rightarrow \mathbf{X})$: The straight line is the analytical value.

channel, and to detect the “causal influence” of one sequence on another. Finally, we apply our estimators on real stock market data to detect the causal influence that exists between the Chinese and the US stock markets.

A. Stationary Hidden Markov Processes

Let $\mathbf{X}$ be a binary symmetric first-order Markov process with transition probability p , i.e., $\mathbb{P}(X_n \neq X_{n-1} | X_{n-1}) = p$. Let $\mathbf{Y}$ be the output of a binary symmetric channel with crossover probability ϵ , corresponding to the input process $\mathbf{X}$, as depicted in Fig. 2.

We use the four algorithms presented to estimate the directed information rate $\bar{I}(\mathbf{Y} \rightarrow \mathbf{X})$ for the case where $p = 0.3$ and $\epsilon = 0.2$. The depth of the context tree is set to be 3. The simulation was performed three times. The results are shown in Fig. 3. As the data length grows, the estimated value approaches the true value for all four algorithms.

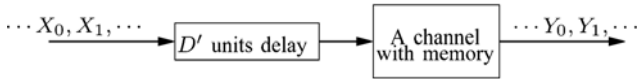
The true value can be simply computed analytically as

$$I(Y^n \rightarrow X^n) = H(X^n) - H(X^n | Y^n) \quad (56)$$

$$= \sum_{i=1}^n H(X_i | X^{i-1}) - H(X_i | X^{i-1}, Y^i) \quad (57)$$

$$= \sum_{i=1}^n H(X_i | X_{i-1}) - H(X_i | X_{i-1}, Y_i) \quad (58)$$

$$= \sum_{i=1}^n H_b(p) - (p\epsilon + \bar{p}\bar{\epsilon})H_b\left(\frac{p\epsilon}{p\epsilon + \bar{p}\bar{\epsilon}}\right) - (\bar{p}\epsilon + p\bar{\epsilon})H_b\left(\frac{\bar{p}\epsilon}{\bar{p}\epsilon + p\bar{\epsilon}}\right), \quad (59)$$

Fig. 4. Using the shifted directed information estimation to find the delay D' .

where (58) follows from the Markov property of the input process and the memorylessness of the channel and in (59), and $\bar{p}$ denotes $1 - p$.

One can note from Fig. 3 that the sample paths of $\hat{I}_2$ and $\hat{I}_4$ indeed appear to be smoother, as one might expect from that fact that they use the entropy and relative entropy functionals on the pmf estimate $Q(x_i, y_i | Y^{i-1}, X^{i-1})$. The first estimator is apparently the least smooth, since it uses the probability assignments evaluated on the sample path, and is highly sensitive to its idiosyncrasies.

B. Channel Delay Estimation via Shifted Directed Information

Assume a setting similar to that in Section V-A, a stationary process that passes through a channel, but now there exists a delay in the entrance of the input to the channel, as depicted in Fig. 4.

Our goal is to find the delay D' . We use the shifted directed information $I(Y_{d+1}^n \rightarrow X^{n-d})$ to estimate D' , where $I(Y_{d+1}^n \rightarrow X^{n-d})$ is defined as

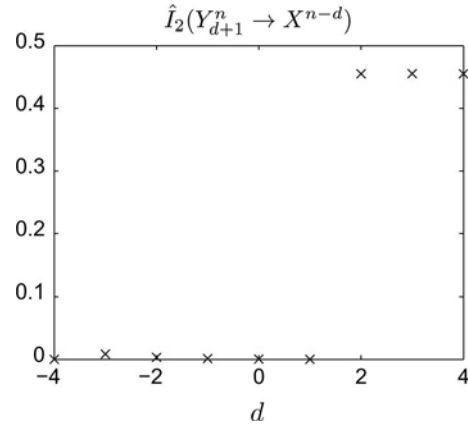
$$I(Y_{d+1}^n \rightarrow X^{n-d}) \triangleq \sum_{i=1}^{n-d} H(X_i | X^{i-1}) - H(X_i | X^{i-1}, Y_{d+1}^{d+i}). \quad (60)$$

To illustrate the idea, suppose $\mathbf{X}$ is a binary stationary process, and we define the binary process $\mathbf{Y}$ as follows:

$$Y_i = X_{i-D'} + X_{i-D'-1} + W_i, \quad (61)$$

where $W_i \sim \text{Bernouli}(\epsilon)$ and in addition (61) is modulo 2. The goal is to find the delay D' from the observations of the processes $\mathbf{Y}$ and $\mathbf{X}$. Note that the mutual information rate $\lim_{n \rightarrow \infty} \frac{1}{n} I(Y^n; X^n)$ is not influenced by D' . However, the shifted directed information rate $\lim_{n \rightarrow \infty} \frac{1}{n-d} I(Y_{d+1}^n \rightarrow X^{n-d})$ is highly influenced by D' . Assuming that there is no feedback, for $d < D'$ we have the Markov chain $Y_{d+1}^{i+d} \rightarrow X^{i-1} \rightarrow X_i$ due to (61), and therefore $I(Y_{d+1}^n \rightarrow X^{n-d}) = 0$. However, for $d \geq D'$, $I(Y_{d+1}^n \rightarrow X^{n-d}) > 0$. For instance, in the channel example (61), if $W_i = 0$ with probability 1 then for $d \geq D'$, $I(Y_{d+1}^n \rightarrow X^{n-d}) = H(X^{n-d})$. Therefore, we can use the shifted directed information $I(Y_{d+1}^n \rightarrow X^{n-d})$ to estimate D' .

Fig. 5 depicts $\hat{I}_2(Y_{d+1}^n \rightarrow X^{n-d})$ where $n = 10^6$ for the setting in Fig. 4, where the input is a binary stationary Markov process of order one and the channel is given by (61). The delay of the channel D' is equal to 2. We use $\hat{I}_2$ to estimate the shifted directed information (all algorithms perform similarly for this case) where the tree depth of the CTW algorithm is set to be 6. The result in Fig. 5 shows that for $d < D'$, $\hat{I}_2(Y_{d+1}^n \rightarrow X^{n-d})$ is very close to zero and for $d \geq D'$, $\hat{I}_2(Y_{d+1}^n \rightarrow X^{n-d})$ is significantly larger than zero.

Fig. 5. Value of $\hat{I}_2(Y_{d+1}^n \rightarrow X^{n-d})$ where $n = 10^6$ for the setting depicted in Fig. 4 with $D' = 2$. When $d < 2$, $\hat{I}_2(Y_{d+1}^n \rightarrow X^{n-d})$ is very close to zero and for $d \geq 2$, $\hat{I}_2(Y_{d+1}^n \rightarrow X^{n-d})$ is significantly larger than zero.

C. Causal Influence Measurement

There is an extensive literature on detecting and measuring causal influence. See, for example, [37] for a recent survey of some of the common tools and approaches in biomedical informatics. One particularly celebrated tool, in both the life sciences and economics, for assessing whether and to what extent one time series influences another is the Granger causality test [9]. The idea is to model $\mathbf{Y}$ first as a univariate autoregressive time series with error correction term V_i

$$Y_i = \sum_{j=1}^p a_j Y_{i-j} + V_i, \quad (62)$$

and then model it again using $\mathbf{X}$ as causal side information

$$Y_i = \sum_{j=1}^p [b_j Y_{i-j} + c_j X_{i+1-j}] + \tilde{V}_i \quad (63)$$

with $\tilde{V}_i$ as the new error correction term. The Granger causality is defined as

$$G_{\mathbf{X} \rightarrow \mathbf{Y}} \triangleq \log \frac{\text{var}(V_i)}{\text{var}(\tilde{V}_i)}, \quad (64)$$

and the bigger it is, the more inclined the practitioner is to assert that $\mathbf{X}$ is causally influencing $\mathbf{Y}$. It is a simple exercise to verify that when the process pair is jointly Gauss–Markov with evolution that obeys both (62) and (63) with $p = \infty$, the Granger causality coincides with the directed information rate (up to a multiplicative constant) [23].

In this section, we implement our universal estimators of directed information to infer causal influences in more general scenarios, where the Gauss–Markov modeling assumption inherent in Granger causality fails to adequately capture the nature of the data.

One philosophical basis for causal analysis is that when we measure causal influence between two processes, $\mathbf{X}$ and $\mathbf{Y}$, there is an underlying assumption that X_i happens earlier than Y_i for every (X_i, Y_i) . Under this assumption, we say two jointly distributed processes $\mathbf{X}$ and $\mathbf{Y}$ induce a forward channel $P(y_i | x^i, y^{i-1})$ and a backward channel $P(x_i | x^{i-1}, y^{i-1})$, as

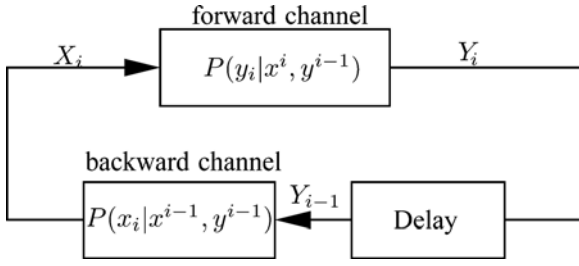


Fig. 6. Modeling any two processes using forward channel $P(y_i|x^i, y^{i-1})$ and backward channel $P(x_i|x^{i-1}, y^{i-1})$.

depicted in Fig. 6, where $\mathbf{X}$ is the input process. In this section, we present the use of directed information, reverse directed information, and mutual information to measure the causal influence between two processes.

Definition 3 (Existence of a Channel): We say that the forward channel does not exist if $P(y_i|x^i, y^{i-1}) = P(y_i|y^{i-1})$ for $i \geq 1$ and similarly the backward channel does not exist if $P(x_i|x^{i-1}, y^{i-1}) = P(x_i|x^{i-1})$ for $i \geq 1$.

We interpret the existence of the forward link as that the sequence $\mathbf{Y}$ is “influenced” or “caused” by the process $\mathbf{X}$. Similarly, the existence of the backward link is interpreted as that $\mathbf{X}$ is “influenced” or “caused” by the sequence $\mathbf{Y}$. We would like to answer the following two questions:

- 1) Does the forward channel exist?
- 2) Does the backward channel exist?

Directed information can naturally answer these questions. It is straightforward to note from the definition of directed information that the forward link exists if and only if $I(X^n \rightarrow Y^n) > 0$ and the backward link exists if and only if $I(Y^{n-1} \rightarrow X^n) > 0$. More generally, the directed information $I(X^n \rightarrow Y^n)$ quantifies how much $\mathbf{X}$ influences $\mathbf{Y}$, while the directed information in the reverse direction $I(Y^{n-1} \rightarrow X^n)$ quantifies how much $\mathbf{Y}$ influences $\mathbf{X}$. The mutual information, which is the sum of those two directed informations, [see (8)], quantifies the mutual influence of the two sequences. Therefore, using the directed information measures, it is natural to adopt terminology as follows:

- 1) *Case A:* $I(X^n \rightarrow Y^n) \gg I(Y^{n-1} \rightarrow X^n)$, we say that $\mathbf{X}$ causes $\mathbf{Y}$.
- 2) *Case B:* $I(X^n \rightarrow Y^n) \ll I(Y^{n-1} \rightarrow X^n)$, we say that $\mathbf{Y}$ causes $\mathbf{X}$.
- 3) *Case C:* $I(X^n \rightarrow Y^n) \simeq I(Y^{n-1} \rightarrow X^n) \gg 0$, we say that the processes are mutually causing each other.
- 4) *Case D:* $I(X^n; Y^n) = 0$, we say that the processes are independent of each other.

To illustrate this idea, consider processes $\mathbf{X}$ and $\mathbf{Y}$ generated by the system that is depicted in Fig. 7, where the forward channel is a BSC(α) and the backward channel is a BSC(β) where $0 \leq \alpha \leq \frac{1}{2}$ and $0 \leq \beta \leq \frac{1}{2}$. Intuitively, if α is much less than β , then the process $\mathbf{X}$ is influencing $\mathbf{Y}$, and if α is much larger than β , the process $\mathbf{Y}$ is influencing $\mathbf{X}$. If α and β have similar values then the processes mutually influence each other, and finally if they are both equal to $\frac{1}{2}$, then the processes are independent of each other. Note that the information-theoretic measures can be analytically calculated as in (65)–(67), and indeed if $I(X^n \rightarrow Y^n) > I(Y^{n-1} \rightarrow X^n)$, then $\alpha < \beta$ and vice

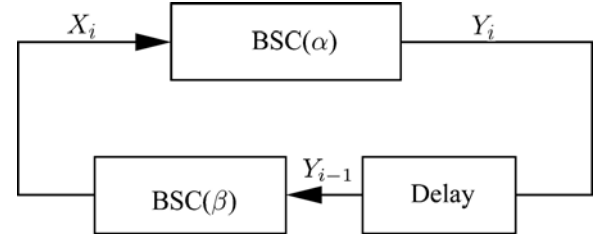


Fig. 7. Simulation of a sequence of random variables $\{X_i, Y_i\}_{i \geq 1}$ according to the relation shown in the scheme. Namely, Y_i is the output of a binary symmetric channel with parameter α and input X_i and X_i is the output of a binary symmetric channel with parameter β and input Y_{i-1} . The initial random variable X_1 is assumed to be distributed Bernoulli($\frac{1}{2}$).

versa. Hence, the intuition regarding which process influences the other is consistent with cases A through D presented above

$$\frac{1}{n}I(X^n \rightarrow Y^n) = H_b(\alpha\bar{\beta} + \bar{\alpha}\beta) - H_b(\alpha) \quad (65)$$

where the terms $\bar{\alpha}$ and $\bar{\beta}$ denote $1 - \alpha$ and $1 - \beta$, respectively. Similarly, we have

$$\frac{1}{n}I(Y^{n-1} \rightarrow X^n) = H_b(\alpha\bar{\beta} + \bar{\alpha}\beta) - H_b(\beta) \quad (66)$$

and

$$\frac{1}{n}I(Y^n; X^n) = 2H_b(\alpha\bar{\beta} + \bar{\alpha}\beta) - H_b(\beta) - H_b(\alpha). \quad (67)$$

Since the normalized reverse directed information is nothing but the normalized directed information between another pair of processes, where one is shifted, the estimators $\hat{I}_1$ to $\hat{I}_4$ can be easily adapted to this situation, and the convergence theorems (Theorem 1 through Theorem 4) apply also (with the appropriate translations) to the reverse directed information. Finally, the normalized mutual information can be estimated once we have the normalized directed information and the normalized reverse directed information simply by summing them.

Fig. 8 depicts the estimated and analytical information-theoretic measures $\frac{1}{n}I(X^n \rightarrow Y^n)$, $\frac{1}{n}I(Y^{n-1} \rightarrow X^n)$, and $\frac{1}{n}I(X^n; Y^n)$ for the case $\alpha = 0.1$ and $\beta = 0.2$. One can note that with just a few hundreds of samples, directed information and reverse directed information start strongly indicating that $\alpha < \beta$, in other words, $\mathbf{X}$ influences $\mathbf{Y}$ more than the other way around.

D. Causal Influence in Stock Markets

Here, we use the historic data of the Hang Seng Index (HSI) and the Dow Jones Index (DJIA) between 1990 and 2011 to compute the directed information rate between these two indexes. The data of those two indexes are presented in Fig. 9 on a daily time scale.

There is no time overlap between the stock market in Hong Kong and in New York, that is, when the stock market in Hong Kong is open, the stock market in New York is closed, and vice versa. Therefore, the causal influence between the markets is well defined. Since the value of the stock market is continuous, we discretize it into three values: -1 , 0 , and 1 . Value -1 means that the stock market went down in one day by more than 0.8% , value 1 means that the stock market went up in one day by more

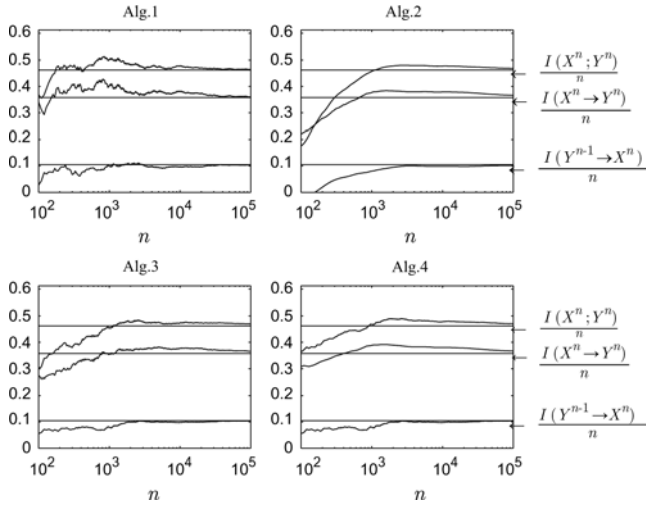


Fig. 8. Information-theoretic measures $\frac{1}{n}I(X^n \rightarrow Y^n)$, $\frac{1}{n}I(Y^{n-1} \rightarrow X^n)$, and $\frac{1}{n}I(X^n; Y^n)$ evaluated using the four algorithms. The data was generated according to the setting in Fig. 7 where $\alpha = 0.1$ and $\beta = 0.2$. The straight black line is the analytical value given by (65)–(67) and the blue lines are the estimated values.

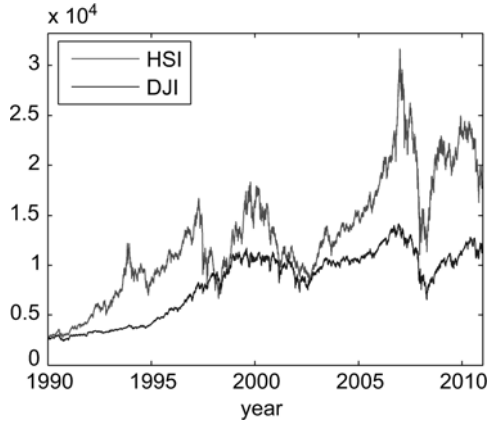


Fig. 9. Hang Seng Index (HSI) and the Dow Jones Industrial Average (DJIA) between 1990 and 2011. The goal is to determine which index is causally influencing the other.

than 0.8%, and value 0 means that the absolute change is less than 0.8%.

We denote by X_i and Y_i the (quantized ternary valued) change in the HSI and the DJIA in day i , respectively, and estimate the normalized mutual information $\frac{1}{n}I(X^n; Y^n)$, the normalized directed information $\frac{1}{n}I(X^n \rightarrow Y^n)$, and the normalized reverse directed information $\frac{1}{n}I(Y^{n-1} \rightarrow X^n)$, using all four algorithms. Fig. 10 plots our estimates of these information-theoretic measures.

Evidently, the reverse directed information is much higher than the directed information; hence there is a significant causal influence by the DJIA on the HSI, and a low influence in the reverse direction. In other words, between 1990 and 2011, it was the Chinese market that was influenced by the US market rather than the other way around.

It is also worth noting that estimators $\hat{I}_1$ and $\hat{I}_2$ do generate negative outputs as shown in Fig. 10. It may be caused by various reasons, such as data insufficiency and nonstationarity of process $(\mathbf{X}, \mathbf{Y})$. In such cases of insufficient data, we would

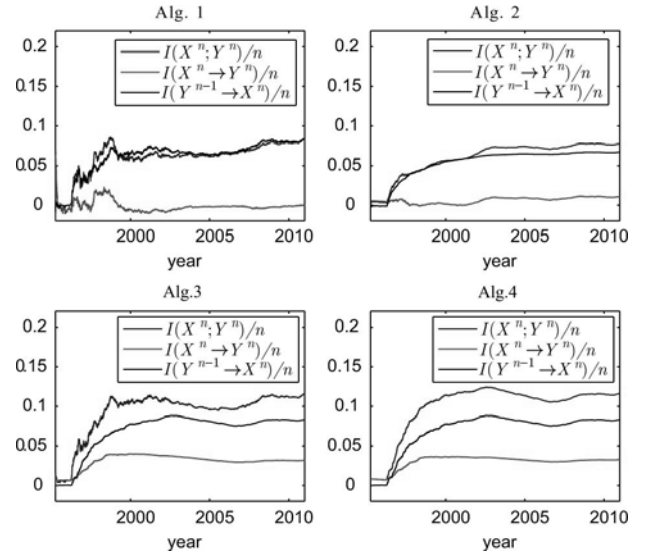


Fig. 10. Estimates of information-theoretic measures between HSI denoted by $\mathbf{X}$, and DJI denoted by $\mathbf{Y}$. It is clear that the reverse directed information is much higher than the directed information, hence it is DJI that causally influences HSI rather than the other way around.

prefer estimators $\hat{I}_3$ and $\hat{I}_4$, since they are always nonnegative, which can be sensibly interpreted in practice.

VI. CONCLUDING REMARKS

We have presented four approaches to estimating the directed information rate between a pair of jointly stationary ergodic finite-alphabet processes. Weak and strong consistency results have been established for all four estimators, in precise senses of varying strengths. For two of these estimators, we established convergence rates that are optimal to within logarithmic factors. The other two have their own merits, such as nonnegativity on every sample path. Experiments on simulated and real data substantiate the potential of the proposed approaches in practice and the efficacy of directed information estimation as a tool for detecting and quantifying causality and delay.

APPENDIX A SOME KEY LEMMAS

Here is the roadmap of the Appendixes. In Appendix A, we list some key lemmas without proofs, and in Appendix B we prove the main theorems and propositions in Section IV. Appendix C provides the proofs of the lemmas in Appendix A.

The first lemma is on the asymptotic equipartition property (AEP) for causally conditional entropy rate. It was proved in [33] that the AEP for causally conditional entropy rate holds in the almost sure sense. Here we prove that it holds in the L_1 sense as well. We also show convergence rates for jointly stationary irreducible aperiodic Markov processes.

Lemma 1: Let $(\mathbf{X}, \mathbf{Y})$ be a jointly stationary ergodic finite-alphabet process. Then,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log P(Y^n | X^n) = \overline{H}(\mathbf{Y} | \mathbf{X}) \quad P\text{-a.s. and in } L_1. \quad (68)$$

In addition, if $(\mathbf{X}, \mathbf{Y})$ is irreducible aperiodic Markov, then

$$\mathbb{E} \left| -\frac{1}{n} \log P(Y^n \| X^n) - \overline{H}(\mathbf{Y} \| \mathbf{X}) \right| = O(n^{-1/2} \log n) \quad (69)$$

and for every $\epsilon > 0$,

$$\begin{aligned} & -\frac{1}{n} \log P(Y^n \| X^n) - \overline{H}(\mathbf{Y} \| \mathbf{X}) \\ & = o(n^{-1/2} (\log n)^{5/2+\epsilon}) \quad P\text{-a.s.} \end{aligned} \quad (70)$$

The next lemma shows that the conditional probability induced by the CTW algorithm converges to the true probability of a Markov process if the dataset is sufficiently large.

Lemma 2: Let Q be the CTW probability assignment and let $\mathbf{X}$ be a stationary irreducible aperiodic finite-alphabet Markov process whose order is bounded by the prescribed tree depth of the CTW algorithm. Then,

$$\lim_{n \rightarrow \infty} Q(x_{n+1} | X^n) - P(x_{n+1} | X^n) = 0 \quad P\text{-a.s.} \quad (71)$$

Lemma 3 ([21, Lemma 1]): For any $\epsilon > 0$, there exists $K_\epsilon > 0$ such that for all P and Q in $\mathcal{M}(\mathcal{X}, \mathcal{Y})$:

$$|f(P) - f(Q)| \leq \epsilon + K_\epsilon \|P - Q\|_1, \quad (72)$$

where $\|\cdot\|_1$ is the l_1 norm (viewing P and Q as $|\mathcal{X}||\mathcal{Y}|$ -dimensional simplex vectors), and f is defined in (29).

Lemma 4: Let P, Q be two probability mass functions in $\mathcal{M}(\mathcal{X}, \mathcal{Y})$, denote $\theta = \|P - Q\|_1$. If $\theta < 1/2$, we have

$$|f(P) - f(Q)| \leq 2\theta \log \frac{|\mathcal{X}||\mathcal{Y}|}{\theta}, \quad (73)$$

where f is defined in (29).

Lemma 5: Let $\mathbf{X}$ be a stationary irreducible aperiodic finite-alphabet Markov process. For fixed $i \geq 1$, let random variable $V_i(X_{i-m}^i)$ be a deterministic function of random vector X_{i-m}^i , where m is the Markov order. Let V_i be uniformly bounded by a constant V for any i , and $\mathbb{E} V_i = 0, \forall i \geq 1$. Then, there exists a constant C_4 such that

$$\mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n V_i \right)^2 \leq C_4 V^2 n^{-1}. \quad (74)$$

Lemma 6 (Breiman's Generalized Ergodic Theorem [38]): Let $\mathbf{X}$ be a stationary ergodic process. If $\lim_{k \rightarrow \infty} g_k(\mathbf{X}) = g(\mathbf{X})$ P -a.s. and $\mathbb{E}[\sup_k |g_k|] < \infty$, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n g_k(T^k(\mathbf{X})) = \mathbb{E} g(\mathbf{X}) \quad P\text{-a.s.} \quad (75)$$

where $T(\cdot)$ is the shift operator which increases the index by 1, and T^k increases the index by k .

Here, we paraphrase a result from [30] on the redundancy bounds of the CTW probability assignment.

Lemma 7 ([30]): Let Q be the CTW probability assignment and let $\mathbf{X}$ be a stationary finite-alphabet Markov process whose order is bounded by the prescribed tree depth of the

CTW algorithm. Then, there exist constants C_5, C_6 such that the pointwise redundancy is bounded as

$$\max_{x^n} \left(\log \frac{1}{Q(x^n)} - \log \frac{1}{P(x^n)} \right) \leq C_5 \log n + C_6 \quad (76)$$

where $C_5 > 0, C_6$ depend on nothing but the parameters specifying the process $\mathbf{X}$. In particular, taking expectation over the inequality with respect to P , the redundancy is bounded as

$$D(P(x^n) \| Q(x^n)) \leq C_5 \log n + C_6. \quad (77)$$

Remark 3: The constants C_5, C_6 can be specified once the parameters of process $\mathbf{X}$ are given. For example, see [30], where

$$C_5 = \frac{(\gamma - 1)|\mathcal{S}|}{2}, \quad (78)$$

$$C_6 = \frac{(\gamma - 1)|\mathcal{S}|}{2} \log \frac{1}{|\mathcal{S}|} + |\mathcal{S}| \left(\frac{\gamma}{\gamma - 1} + \log \gamma \right) - \frac{1}{\gamma - 1}. \quad (79)$$

Here γ is the size of alphabet, in this case $\gamma = |\mathcal{X}|$. $|\mathcal{S}|$ is the number of states in the Markov process, given Markov order m , $|\mathcal{S}| \leq |\mathcal{X}|^m$.

APPENDIX B

PROOFS OF THEOREMS AND PROPOSITIONS

For brevity, in the sequel we denote $\hat{H}_1(Y^n \| X^n)$ by $\hat{H}_1$, $\hat{H}_2(Y^n \| X^n)$ by $\hat{H}_2$, $\hat{I}_i(X^n \rightarrow Y^n)$ by $\hat{I}_i, i = 1, 2, 3, 4$.

A) Proof of Theorem 1: Briefly speaking, we need to show estimator $\hat{I}_1$ converges to the corresponding directed information rate $\bar{I}(\mathbf{X} \rightarrow \mathbf{Y})$ for any jointly stationary ergodic process $(\mathbf{X}, \mathbf{Y})$. Since $\hat{I}_1$ is defined in (35) as $\hat{H}_1(Y^n) - \hat{H}_1(Y^n \| X^n)$, if we can show the corresponding convergence properties of $\hat{H}_1(Y^n \| X^n)$, then we have the desired convergence properties of $\hat{I}_1$ since $\hat{H}_1(Y^n) = \hat{H}_1(Y^n \| \emptyset)$.

Given Q is a universal probability assignment, first we show $\hat{I}_1$ converges in L_1 . Then, we show given Q is a pointwise universal probability assignment, $\hat{I}_1$ also converges almost surely.

L_1 Convergence: We decompose

$$\hat{H}_1 - \overline{H}(\mathbf{Y} \| \mathbf{X}) = C_n + D_n, \quad (80)$$

where

$$C_n = \hat{H}_1 + \frac{1}{n} \log P(Y^n \| X^n) \quad (81)$$

$$D_n = -\frac{1}{n} \log P(Y^n \| X^n) - \overline{H}(\mathbf{Y} \| \mathbf{X}). \quad (82)$$

According to Lemma 1 shown in Appendix A, we know D_n converges to zero in L_1 . Now we deal with C_n . Pinsker[39] proved the existence of a universal constant $\Gamma > 0$ such that

$$D(P \| Q) \leq \mathbb{E}_P \left\{ \left| \log \left(\frac{dP}{dQ} \right) \right| \right\} \leq D(P \| Q) + \Gamma \sqrt{D(P \| Q)}. \quad (83)$$

Barron[40] simplified Pinsker's argument and proved that the constant $\Gamma = \sqrt{2}$ is best possible when natural logarithms are

used in the definition of $D(P\|Q)$. Here, we follow Barron's arguments to bound $\mathbb{E}|C_n|$ with C_n defined in (81).

Denote the set $\{(x^n, y^n) : P(y^n\|x^n) \leq Q(y^n\|x^n)\}$ as $\mathcal{B}_n$, we have

$$\begin{aligned} \mathbb{E}|C_n| &= \sum_{(x^n, y^n) \in (\mathcal{X} \times \mathcal{Y})^n \setminus \mathcal{B}_n} P(x^n, y^n) \frac{1}{n} \log \frac{P(y^n\|x^n)}{Q(y^n\|x^n)} \\ &\quad + \sum_{(x^n, y^n) \in \mathcal{B}_n} P(x^n, y^n) \frac{1}{n} \log \frac{Q(y^n\|x^n)}{P(y^n\|x^n)} \end{aligned} \quad (84)$$

$$\begin{aligned} &= \mathbb{E} \left[\frac{1}{n} \log \frac{P(Y^n\|X^n)}{Q(Y^n\|X^n)} \right] \\ &\quad + 2 \sum_{(x^n, y^n) \in \mathcal{B}_n} P(x^n, y^n) \frac{1}{n} \log \frac{Q(y^n\|x^n)}{P(y^n\|x^n)}. \end{aligned} \quad (85)$$

Define $C_{n1} \triangleq \mathbb{E} \left[\frac{1}{n} \log \frac{P(Y^n\|X^n)}{Q(Y^n\|X^n)} \right]$, $C_{n2} \triangleq \sum_{(x^n, y^n) \in \mathcal{B}_n} P(x^n, y^n) \frac{1}{n} \log \frac{Q(y^n\|x^n)}{P(y^n\|x^n)}$, we bound

$$C_{n1} = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\log \frac{P(Y_i|X^i, Y^{i-1})}{Q(Y_i|X^i, Y^{i-1})} \right] \quad (86)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\mathbb{E} \left[\log \frac{P(Y_i|X^i, Y^{i-1})}{Q(Y_i|X^i, Y^{i-1})} \middle| X^{i-1}, Y^{i-1} \right] \right] \quad (87)$$

$$\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\mathbb{E} \left[\log \frac{P(Y_i, X_i|X^{i-1}, Y^{i-1})}{Q(Y_i, X_i|X^{i-1}, Y^{i-1})} \middle| X^{i-1}, Y^{i-1} \right] \right] \quad (88)$$

$$= \frac{1}{n} D(P(x^n, y^n) \| Q(x^n, y^n)). \quad (89)$$

Then, $\forall i$, define $\mathcal{C}_i \triangleq \mathcal{C}_i(x^i, y^{i-1}) = \{y_i : P(y_i|x^i, y^{i-1}) \leq Q(y_i|x^i, y^{i-1})\}$. We bound C_{n2} from (112) to (120), where

- 1) equation (113) follows by the log-sum inequality [24, Th. 2.7.1],
 - 2) equation (114) follows since $\forall x > -1, \log(1+x) \leq x/\ln(2)$,
 - 3) equation (115) follows since $|x| \geq x$,
 - 4) equation (116) follows by Scheffé's theorem [42, Lemma 2.1],
 - 5) equation (117) follows by Pinsker's inequality [42, Lemma 2.5],
 - 6) equation (118) follows by the concavity of $\sqrt{\cdot}$,
 - 7) equation (119) follows by data-processing inequality [24, Th. 2.8.1],
 - 8) equation (120) follows by the chain rule for relative entropy, the concavity of $\sqrt{\cdot}$, and data-processing inequality.
- Combining (89) and (120), we have

$$\begin{aligned} \mathbb{E}|C_n| &\leq \frac{1}{n} D(P(x^n, y^n) \| Q(x^n, y^n)) \\ &\quad + \sqrt{\frac{2}{\ln(2)}} \sqrt{D(P(x^{n+1}, y^{n+1}) \| Q(x^{n+1}, y^{n+1}))/n}, \end{aligned} \quad (90)$$

by definition of universal probability assignment, we show C_n converges to zero in L_1 . Since

$$\mathbb{E}|\hat{H}_1 - \overline{H}(\mathbf{Y}|\mathbf{X})| \leq \mathbb{E}|C_n| + \mathbb{E}|D_n| \rightarrow 0 \quad n \rightarrow \infty, \quad (91)$$

we know $\hat{I}_1$ converges to $\overline{I}(\mathbf{X} \rightarrow \mathbf{Y})$ in L_1 .

Almost sure convergence: Consider the probability of the following event:

$$\mathcal{A}_{n,\epsilon} = \{(x^n, y^n) : \hat{H}_1 \leq -\frac{1}{n} \log P(y^n\|x^n) - \epsilon\}, \quad (92)$$

we have

$$\mathbb{P}(\mathcal{A}_{n,\epsilon}) = \sum_{(x^n, y^n) \in \mathcal{A}_{n,\epsilon}} P(x^n, y^n) \quad (93)$$

$$= \sum_{(x^n, y^n) \in \mathcal{A}_{n,\epsilon}} P(y^n\|x^n) P(x^n\|y^{n-1}) \quad (94)$$

$$\leq \sum_{(x^n, y^n) \in \mathcal{A}_{n,\epsilon}} Q(y^n\|x^n) 2^{-n\epsilon} P(x^n\|y^{n-1}) \quad (95)$$

$$= 2^{-n\epsilon} \sum_{(x^n, y^n) \in \mathcal{A}_{n,\epsilon}} Q(y^n\|x^n) P(x^n\|y^{n-1}) \quad (96)$$

$$\leq 2^{-n\epsilon}, \quad (97)$$

where the first inequality is because of the definition of even $\mathcal{A}_{n,\epsilon}$, and the last step follows from the fact that for any two conditional distributions of the form $Q(y^n\|x^n)$ and $P(x^n\|y^{n-1})$, we have $Q(y^n\|x^n)P(x^n\|y^{n-1}) = \tilde{Q}(x^n, y^n)$ where $\tilde{Q}$ is a joint distribution. As

$$\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{A}_{n,\epsilon}) < \infty, \quad (98)$$

by the Borel–Cantelli Lemma, we have

$$\liminf_{n \rightarrow \infty} \hat{H}_1 - \left(-\frac{1}{n} \log P(Y^n\|X^n) \right) \geq 0. \quad P\text{-a.s.} \quad (99)$$

In order to get an inequality with the reverse direction, write $\hat{H}_1 - (-\frac{1}{n} \log P(Y^n\|X^n))$ explicitly as

$$\begin{aligned} &\hat{H}_1 + \frac{1}{n} \log P(Y^n\|X^n) \\ &= \frac{1}{n} \log \frac{P(Y^n\|X^n)}{Q(Y^n\|X^n)} \end{aligned} \quad (100)$$

$$= \frac{1}{n} \log \frac{P(Y^n, X^n)}{Q(Y^n, X^n)} - \frac{1}{n} \log \frac{P(X^n\|Y^{n-1})}{Q(X^n\|Y^{n-1})}, \quad (101)$$

by the definition of pointwise universality (2), we know

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{P(Y^n, X^n)}{Q(Y^n, X^n)} \leq 0, \quad P\text{-a.s.} \quad (102)$$

with a similar argument used for showing (99), we show

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log \frac{P(X^n\|Y^{n-1})}{Q(X^n\|Y^{n-1})} \leq 0, \quad P\text{-a.s.} \quad (103)$$

then we have

$$\limsup_{n \rightarrow \infty} \hat{H}_1 - \left(-\frac{1}{n} \log P(Y^n\|X^n) \right) \leq 0. \quad P\text{-a.s.} \quad (104)$$

Combining (104) with (99),

$$\lim_{n \rightarrow \infty} \hat{H}_1 - \left(-\frac{1}{n} \log P(Y^n\|X^n) \right) = 0. \quad P\text{-a.s.} \quad (105)$$

By Lemma 1 shown in Appendix A,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log P(Y^n \| X^n) - \bar{H}(\mathbf{Y} \| \mathbf{X}) = 0, \quad P\text{-a.s.} \quad (106)$$

which implies the convergence of $\hat{I}_1$ to $\bar{I}(\mathbf{X} \rightarrow \mathbf{Y})$ also holds almost surely.

B) Proof of Proposition 1: For similar reasons as shown in the proof of Theorem 1, here it suffices to show the convergence properties of $\hat{H}_1$. For convenience, we restate some arguments shown in the proof of Theorem 1. We decompose $\hat{H}_1 - \bar{H}(\mathbf{Y} \| \mathbf{X})$ as

$$\hat{H}_1 - \bar{H}(\mathbf{Y} \| \mathbf{X}) = C_n + D_n, \quad (107)$$

where

$$C_n = \hat{H}_1 + \frac{1}{n} \log P(Y^n \| X^n) \quad (108)$$

$$D_n = -\frac{1}{n} \log P(Y^n \| X^n) - \bar{H}(\mathbf{Y} \| \mathbf{X}), \quad (109)$$

and we restate (90)

$$\begin{aligned} \mathbb{E}|C_n| &\leq \frac{1}{n} D(P(x^n, y^n) \| Q(x^n, y^n)) \\ &\quad + \sqrt{\frac{2}{\ln(2)}} \sqrt{D(P(x^{n+1}, y^{n+1}) \| Q(x^{n+1}, y^{n+1}))/n}. \end{aligned} \quad (110)$$

L_1 convergence rates: We apply Lemma 7 in Appendix A. Plugging (77) of Lemma 7 in (110), we have

$$\mathbb{E}|C_n| = O((\log n)^{1/2} n^{-1/2}). \quad (111)$$

See (112)–(120) given at the bottom of the page.

Combining (111) with the L_1 convergence rates of D_n shown in Lemma 1 in Appendix A, we have

$$\mathbb{E}|\hat{H}_1 - \bar{H}(\mathbf{Y} \| \mathbf{X})| \leq \mathbb{E}|C_n| + \mathbb{E}|D_n| \quad (121)$$

$$= O(n^{-1/2} \log n), \quad (122)$$

then we know the convergence rates in Proposition 1 hold as follows:

$$\mathbb{E}|\hat{I}_1(X^n \rightarrow Y^n) - \bar{I}(\mathbf{X} \rightarrow \mathbf{Y})| = O(n^{-1/2} \log n). \quad (123)$$

Almost sure convergence rates: We look at the almost sure convergence rates of C_n (108) at first. We know the probability of event $\mathcal{A}_{n,\epsilon}$ defined in (92) is bounded as

$$\mathbb{P}(\mathcal{A}_{n,\epsilon}) \leq 2^{-n\epsilon}. \quad (124)$$

$$C_{n2} \leq \frac{1}{n} \sum_{i=1}^n \sum_{(x^i, y^{i-1})} P(x^i, y^{i-1}) \sum_{y_i \in \mathcal{C}_i} P(y_i | x^i, y^{i-1}) \log \frac{Q(y_i | x^i, y^{i-1})}{P(y_i | x^i, y^{i-1})} \quad (112)$$

$$\leq \frac{1}{n} \sum_{i=1}^n \sum_{(x^i, y^{i-1})} P(x^i, y^{i-1}) P(Y_i \in \mathcal{C}_i | x^i, y^{i-1}) \log \frac{Q(Y_i \in \mathcal{C}_i | x^i, y^{i-1})}{P(Y_i \in \mathcal{C}_i | x^i, y^{i-1})} \quad (113)$$

$$\leq \frac{1}{n} \sum_{i=1}^n \sum_{(x^i, y^{i-1})} P(x^i, y^{i-1}) \frac{1}{\ln(2)} (Q(Y_i \in \mathcal{C}_i | x^i, y^{i-1}) - P(Y_i \in \mathcal{C}_i | x^i, y^{i-1})) \quad (114)$$

$$\leq \frac{1}{n} \sum_{i=1}^n \sum_{(x^i, y^{i-1})} P(x^i, y^{i-1}) \frac{1}{\ln(2)} |Q(Y_i \in \mathcal{C}_i | x^i, y^{i-1}) - P(Y_i \in \mathcal{C}_i | x^i, y^{i-1})| \quad (115)$$

$$\leq \frac{1}{2n} \sum_{i=1}^n \sum_{(x^i, y^{i-1})} P(x^i, y^{i-1}) \frac{1}{\ln(2)} \sum_{y_i} |P(y_i | x^i, y^{i-1}) - Q(y_i | x^i, y^{i-1})| \quad (116)$$

$$\leq \frac{1}{2n} \sum_{i=1}^n \sum_{(x^i, y^{i-1})} P(x^i, y^{i-1}) \sqrt{\frac{2}{\ln(2)}} \sqrt{D(P(y_i | x^i, y^{i-1}) \| Q(y_i | x^i, y^{i-1}))} \quad (117)$$

$$\leq \frac{1}{2n} \sum_{i=1}^n \sqrt{\frac{2}{\ln(2)}} \sqrt{\mathbb{E} D(P(y_i | X^i, Y^{i-1}) \| Q(y_i | X^i, Y^{i-1}))} \quad (118)$$

$$\leq \frac{1}{2n} \sum_{i=1}^n \sqrt{\frac{2}{\ln(2)}} \sqrt{\mathbb{E} D(P(y_i, x_{i+1} | X^i, Y^{i-1}) \| Q(y_i, x_{i+1} | X^i, Y^{i-1}))} \quad (119)$$

$$\leq \sqrt{\frac{1}{2 \ln(2)}} \sqrt{D(P(x^{n+1}, y^{n+1}) \| Q(x^{n+1}, y^{n+1}))/n} \quad (120)$$

For any fixed $\delta' > \delta > 0$, taking $\epsilon = n^{-1+\delta}$ in (92), we see $\mathcal{A}_{n,\epsilon}$ is equal to the set

$$\{(x^n, y^n) : n^{1-\delta'} \left(\hat{H}_1 + \frac{1}{n} \log P(y^n \| x^n) \right) \leq -n^{\delta-\delta'}\}. \quad (125)$$

Note that

$$\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{A}_{n,\epsilon}) \leq \sum_{n=1}^{\infty} 2^{-n^{\delta}} < \infty. \quad (126)$$

By the Borel–Cantelli lemma, since $n^{\delta-\delta'}$ goes to zero as $n \rightarrow \infty$, we proved that

$$\liminf_{n \rightarrow \infty} n^{1-\delta'} \left(\hat{H}_1 + \frac{1}{n} \log P(y^n \| x^n) \right) \geq 0 \quad P\text{-a.s.} \quad (127)$$

In order to get an inequality of the reverse direction, dividing (101) by $n^{-1+\delta'}$, we have

$$n^{1-\delta'} \left(\hat{H}_1 + \frac{1}{n} \log P(Y^n \| X^n) \right) \quad (128)$$

$$= n^{1-\delta'} \left(\frac{1}{n} \log \frac{P(Y^n, X^n)}{Q(Y^n, X^n)} \right) - n^{1-\delta'} \left(\frac{1}{n} \log \frac{P(X^n \| Y^{n-1})}{Q(X^n \| Y^{n-1})} \right). \quad (129)$$

By the pointwise redundancy of the CTW algorithm restated in Lemma 7 in Appendix A, we know

$$\limsup_{n \rightarrow \infty} \frac{1}{\log n} \log \frac{P(Y^n, X^n)}{Q(Y^n, X^n)} \leq 1 \quad P\text{-a.s.} \quad (130)$$

then we have

$$\limsup_{n \rightarrow \infty} n^{1-\delta'} \left(\frac{1}{n} \log \frac{P(Y^n, X^n)}{Q(Y^n, X^n)} \right) \leq 0 \quad P\text{-a.s.} \quad (131)$$

For the second term on the right-hand side of (129), following similar argument applied to show (127), we know

$$\limsup_{n \rightarrow \infty} -n^{1-\delta'} \left(\frac{1}{n} \log \frac{P(X^n \| Y^{n-1})}{Q(X^n \| Y^{n-1})} \right) \leq 0 \quad P\text{-a.s.} \quad (132)$$

From (131) and (132), we obtain

$$\limsup_{n \rightarrow \infty} n^{1-\delta'} \left(\hat{H}_1 + \frac{1}{n} \log P(Y^n \| X^n) \right) \leq 0 \quad P\text{-a.s.} \quad (133)$$

Combining (127) and (133) together, we know $\forall \delta' > 0$,

$$\lim_{n \rightarrow \infty} \hat{H}_1 + \frac{1}{n} \log P(Y^n \| X^n) = o(n^{-1+\delta'}) \quad P\text{-a.s.} \quad (134)$$

Putting (134) and the almost sure convergence rates of D_n shown in Lemma 1 in Appendix A together, we know $\forall \epsilon > 0$,

$$\hat{I}_1(X^n \rightarrow Y^n) - \bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) = o(n^{-1/2}(\log n)^{5/2+\epsilon}). \quad P\text{-a.s.}$$

C) Proof of Theorem 2: It suffices to show the convergence properties of $\hat{H}_2$. We decompose

$$\hat{H}_2(Y^n \| X^n) - \bar{H}(\mathbf{Y} \| \mathbf{X}) = A_n + B_n, \quad (135)$$

where

$$A_n = \frac{1}{n} \sum_{k=1}^n f(P(x_{k+1}, y_{k+1} | X^k, Y^k)) - \bar{H}(\mathbf{Y} \| \mathbf{X}) \quad (136)$$

$$B_n = \hat{H}_2(Y^n \| X^n) - \frac{1}{n} \sum_{k=1}^n f(P(x_{k+1}, y_{k+1} | X^k, Y^k)). \quad (137)$$

Define $g_k(\mathbf{X}, \mathbf{Y}) \triangleq f(P(x_1, y_1 | X_{-k}^0, Y_{-k}^0))$ for a jointly stationary and ergodic process $(\mathbf{X}, \mathbf{Y})$. Note that, by martingale convergence [41], $g_k(\mathbf{X}, \mathbf{Y}) \rightarrow g(\mathbf{X}, \mathbf{Y})$, P -a.s. where $g(\mathbf{X}, \mathbf{Y}) = f(P(x_1, y_1 | X_{-\infty}^0, Y_{-\infty}^0))$. Noting further that $\mathbb{E}g(\mathbf{X}, \mathbf{Y}) = \bar{H}(\mathbf{Y} \| \mathbf{X})$ and $\forall k, g_k$ are bounded, we can apply Lemma 6 in Appendix A and obtain the following result:

$$\lim_{n \rightarrow \infty} A_n = 0 \quad P\text{-a.s. and in } L_1. \quad (138)$$

Then, we deal with B_n defined in (137) from (155) to (162), where fixing an arbitrary $\epsilon > 0$:

- 1) equation (157) follows by Lemma 3 in Appendix A,
- 2) equation (158) follow by Pinsker's inequality,
- 3) equations (159) and (161) follow by the concavity of $\sqrt{\cdot}$,
- 4) equation (162) follows by the chain rule for relative entropy.

We continue to bound

$$\lim_{n \rightarrow \infty} \mathbb{E} \left| \hat{H}_2(Y^n \| X^n) - \bar{H}(\mathbf{Y} \| \mathbf{X}) \right| \quad (139)$$

$$\leq \lim_{n \rightarrow \infty} \mathbb{E} |A_n| + \lim_{n \rightarrow \infty} \mathbb{E} |B_n| \quad (140)$$

$$= \lim_{n \rightarrow \infty} \mathbb{E} |B_n| \quad (141)$$

$$\leq \epsilon + \lim_{n \rightarrow \infty} K_\epsilon \sqrt{\frac{2 \ln(2)}{n} D(P(x^{n+1}, y^{n+1}) \| Q(x^{n+1}, y^{n+1}))} \quad (142)$$

$$= \epsilon \quad (143)$$

where (141) follows by (138), (142) follows by (162), (143) follows by Definition 1. Now we can use the arbitrariness of ϵ to complete the proof.

D) Proof of Proposition 2: It suffices to show the convergence properties of $\hat{H}_2$.

Almost Sure Convergence: For stationary ergodic process $(\mathbf{X}, \mathbf{Y})$, let

$$g_k(\mathbf{X}, \mathbf{Y}) = f(Q(x_0, y_0 | X_{-k}^{-1})) \quad (144)$$

$$g(\mathbf{X}, \mathbf{Y}) = f(P(x_0, y_0 | X_{-\infty}^{-1}, Y_{-\infty}^{-1})), \quad (145)$$

by Lemma 2 in Appendix A,

$$\lim_{k \rightarrow \infty} g_k(\mathbf{X}, \mathbf{Y}) - g(\mathbf{X}, \mathbf{Y}) = 0 \quad P\text{-a.s.} \quad (146)$$

Since $\mathbb{E}[\sup_k |g_k|] \leq \log |\mathcal{Y}|$, by Lemma 6 in Appendix A,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n g_k(T^k(\mathbf{X}, \mathbf{Y})) = \lim_{n \rightarrow \infty} \hat{H}_2 = \bar{H}(\mathbf{Y} \| \mathbf{X}), \quad (147)$$

which justifies the almost sure convergence of $\hat{H}_2$.

L₁ Convergence Rates: For convenience, we restate the definitions of A_n and B_n as follows:

$$A_n = \frac{1}{n} \sum_{k=1}^n f(P(x_{k+1}, y_{k+1}|X^k, Y^k)) - \overline{H}(\mathbf{Y}|\mathbf{X}) \quad (148)$$

$$B_n = \hat{H}_2(Y^n|X^n) - \frac{1}{n} \sum_{k=1}^n f(P(x_{k+1}, y_{k+1}|X^k, Y^k)). \quad (149)$$

Letting V_k be $f(P(x_{k+1}, y_{k+1}|X^k, Y^k)) - \overline{H}(\mathbf{Y}|\mathbf{X})$, V be $\log |\mathcal{Y}|$, and applying Lemma 5 in Appendix A, we know

$$\mathbb{E}|A_n| \leq \sqrt{\mathbb{E}A_n^2} = O(n^{-1/2}). \quad (150)$$

Then, we bound $\mathbb{E}|B_n|$ from (178) to (183).

- 1) Equation (180) is an application of Lemmas 2 and 4 in Appendix A. Indeed, Lemma 2 guarantees that when $n \rightarrow \infty$, the ℓ_1 norm of the difference of $P(x_{k+1}, y_{k+1}|X^k, Y^k)$ and $Q(x_{k+1}, y_{k+1}|X^k, Y^k)$ will be small enough so that Lemma 4 can be applied.
- 2) Equation (181) follows by Pinsker's inequality and the fact that function $t \log(1/t)$ is increasing for small t .
- 3) Equations (182) and (183) are by the concavity of $\sqrt{t} \log(1/\sqrt{t})$ and the chain rule for relative entropy.

Because of the monotonicity of $\sqrt{t} \log(1/\sqrt{t})$ when $t \approx 0$, we can plug in the redundancy bounds of the CTW algorithm in Lemma 7 in Appendix A, i.e., (77) into (183), then have

$$\mathbb{E}|B_n| = O(n^{-1/2}(\log n)^{3/2}). \quad (151)$$

Combining (151) with (150), we proved Proposition 2.

E) Proof of Proposition 3: We rephrase a general lemma showing minimax lower bounds.

Lemma 8 [42, Th. 2.2]: Let $\mathcal{F}$ be a class of models, and suppose we have observations Z distributed according to P_f , $f \in \mathcal{F}$. Let $d(\hat{f}, f)$ be the performance measure of the estimator $\hat{f}(Z)$ relative to the true model f . Assume also $d(\cdot, \cdot)$ is a semidistance, i.e., it satisfies

- 1) $d(f, g) = d(g, f) \geq 0$,
- 2) $d(f, f) = 0$,
- 3) $d(f, g) \leq d(h, f) + d(h, g)$.

Let $f_0, f_1 \in \mathcal{F}$ satisfy $d(f_0, f_1) \geq 2s > 0$, where s is fixed. Then,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} P_f(d(\hat{f}, f) \geq s) \geq \inf_{\hat{f}} \max_{j \in \{0,1\}} P_{f_j}(d(\hat{f}, f_j) \geq s) \quad (152)$$

$$\geq \frac{1}{4} \exp(-D(P_{f_1} \| P_{f_0})). \quad (153)$$

In this proof, $\mathcal{F}$ in Lemma 8 is taken to be $P(\mathbf{X}, \mathbf{Y})$. Denote the binary entropy as $H_b(p) = -p \log p - (1-p) \log(1-p)$ and the class of i.i.d. processes as $\mathcal{M}_0$. Since

$$H'_b(p) = \log \frac{1-p}{p}, \quad (154)$$

and $H'_b(p)$ is decreasing in interval $[2/8, 3/8]$, we know (155) and (162), at the bottom of the page.

Lemma 9: $\forall p, q \in [2/8, 3/8]$, we have

$$|H_b(p) - H_b(q)| \geq \log(5/3)|p - q|. \quad (163)$$

$$\mathbb{E}|B_n| = \mathbb{E} \left| \frac{1}{n} \sum_{k=1}^n (f(Q(x_{k+1}, y_{k+1}|X^k, Y^k)) - f(P(x_{k+1}, y_{k+1}|X^k, Y^k))) \right| \quad (155)$$

$$\leq \frac{1}{n} \mathbb{E} \sum_{k=1}^n |f(Q(x_{k+1}, y_{k+1}|X^k, Y^k)) - f(P(x_{k+1}, y_{k+1}|X^k, Y^k))| \quad (156)$$

$$\leq \frac{1}{n} \sum_{k=1}^n \mathbb{E} (\epsilon + K_\epsilon \|Q(x_{k+1}, y_{k+1}|X^k, Y^k) - P(x_{k+1}, y_{k+1}|X^k, Y^k)\|_1) \quad (157)$$

$$\leq \frac{K_\epsilon}{n} \sum_{k=1}^n \mathbb{E} \left[\sqrt{2 \ln(2) D(P(x_{k+1}, y_{k+1}|X^k, Y^k) \| Q(x_{k+1}, y_{k+1}|X^k, Y^k))} \right] + \epsilon \quad (158)$$

$$\leq \frac{K_\epsilon}{n} \sum_{k=1}^n \sqrt{2 \ln(2) \mathbb{E} [D(P(x_{k+1}, y_{k+1}|X^k, Y^k) \| Q(x_{k+1}, y_{k+1}|X^k, Y^k))]} + \epsilon \quad (159)$$

$$= \epsilon + \frac{K_\epsilon}{n} \sum_{k=1}^n \sqrt{2 \ln(2) \mathbb{E} D(P(x_{k+1}, y_{k+1}|X^k, Y^k) \| Q(x_{k+1}, y_{k+1}|X^k, Y^k))} \quad (160)$$

$$\leq \epsilon + K_\epsilon \sqrt{\frac{2 \ln(2)}{n}} \times \sqrt{\sum_{k=1}^n \mathbb{E} D(P(x_{k+1}, y_{k+1}|X^k, Y^k) \| Q(x_{k+1}, y_{k+1}|X^k, Y^k))} \quad (161)$$

$$= \epsilon + K_\epsilon \sqrt{\frac{2 \ln(2)}{n} D(P(x^{n+1}, y^{n+1}) \| Q(x^{n+1}, y^{n+1}))} \quad (162)$$

We also show a lemma bounding the divergence between two Bernoulli pmfs.

Lemma 10: Let P and Q be Bernoulli pmfs with parameters, respectively, $1/2-p$ and $1/2-q$. If $|p|, |q| \leq 1/4$, then $D(P\|Q) \leq 8(p-q)^2$.

Lemma 10 can be verified as follows:

$$D(P\|Q) = (1/2-p) \log \frac{1/2-p}{1/2-q} + (1/2+p) \log \frac{1/2+p}{1/2+q} \quad (164)$$

$$= (1/2-p) \log \left(1 + \frac{q-p}{1/2-q} \right) + (1/2+p) \log \left(1 + \frac{p-q}{1/2+q} \right) \quad (165)$$

$$\leq \frac{1}{\ln(2)} \left((1/2-p) \frac{q-p}{1/2-q} + (1/2+p) \frac{p-q}{1/2+q} \right) \quad (166)$$

$$= \frac{1}{\ln(2)} \frac{(p-q)^2}{1/4-q^2} \quad (167)$$

$$\leq 8(p-q)^2, \quad (168)$$

where the first inequality holds because $\log(1+x) \leq x/\ln(2) \forall x > -1$, and the second inequality holds because $|q| \leq 1/4$.

Taking the observations model as $X_i \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(q)$, $Y_i = X_i$, then we have $\bar{I}(\mathbf{X} \rightarrow \mathbf{Y}) = H(X)$. Assuming under model $f_0, q = q_0 = 1/4$, under model $f_1, q = q_1 = 1/4 + 1/\sqrt{n}$, and

$n \geq 64$. Let $\hat{I}_n$ be an arbitrary estimator of $\bar{I}(\mathbf{X} \rightarrow \mathbf{Y})$ based on (X_1^n, Y_1^n) , $d(x, y) = |x - y|$, we have

$$d(H_b(q_0), H_b(q_1)) \geq \log(5/3)|q_0 - q_1| = \log(5/3)/\sqrt{n}. \quad (169)$$

Then, we take $s = \log(5/3)/(2\sqrt{n})$ to satisfy the assumption of Lemma 8. For brevity, here we denote $\bar{I}(\mathbf{X} \rightarrow \mathbf{Y})$ as I . By Lemma 8,

$$\inf_{\hat{I}_n} \sup_{\mathcal{M}_0} P_f(d(\hat{I}_n, I) \geq s) \geq \inf_{\hat{I}_n} \max_{j \in \{0,1\}} P_{f_j}(d(\hat{I}_n, H_b(q_j)) \geq s) \quad (170)$$

$$\geq \frac{1}{4} \exp(-D(P_{f_1}\|P_{f_0})). \quad (171)$$

Then, we bound $D(P_{f_1}\|P_{f_0})$:

$$D(P_{f_1}\|P_{f_0}) = n \mathbb{E}_{f_1} \left[\log \frac{P_{f_1}(X_1, Y_1)}{P_{f_0}(X_1, Y_1)} \right] \quad (172)$$

$$\leq 8n(q_0 - q_1)^2 \quad (173)$$

$$= 8. \quad (174)$$

Thus, we have

$$\inf_{\hat{I}_n} \sup_{\mathcal{M}_0} P_f(d(\hat{I}_n, I) \geq s) \geq \frac{1}{4} e^{-8}. \quad (175)$$

Using Markov's inequality,

$$\inf_{\hat{I}_n} \sup_{P(\mathbf{X}, \mathbf{Y})} \mathbb{E}|\hat{I}_n - I| \geq \inf_{\hat{I}_n} \sup_{\mathcal{M}_0} \mathbb{E}|\hat{I}_n - I| \quad (176)$$

$$\geq \frac{1}{4} e^{-8} s = \frac{1}{8} e^{-8} \log(5/3) \frac{1}{\sqrt{n}}. \quad (177)$$

See (178)–(183) given at the bottom of the page.

$$\mathbb{E}|B_n| = \mathbb{E} \left| \frac{1}{n} \sum_{k=1}^n (f(Q(x_{k+1}, y_{k+1}|X^k, Y^k)) - f(P(x_{k+1}, y_{k+1}|X^k, Y^k))) \right| \quad (178)$$

$$\leq \frac{1}{n} \mathbb{E} \sum_{k=1}^n |f(Q(x_{k+1}, y_{k+1}|X^k, Y^k)) - f(P(x_{k+1}, y_{k+1}|X^k, Y^k))| \quad (179)$$

$$\leq \frac{1}{n} \mathbb{E} \sum_{k=1}^n 2\|P(x_{k+1}, y_{k+1}|X^k, Y^k) - Q(x_{k+1}, y_{k+1}|X^k, Y^k)\|_1 \times \log \frac{|\mathcal{X}||\mathcal{Y}|}{\|P(x_{k+1}, y_{k+1}|X^k, Y^k) - Q(x_{k+1}, y_{k+1}|X^k, Y^k)\|_1} \quad (180)$$

$$\leq \frac{1}{n} \mathbb{E} \sum_{k=1}^n 2\sqrt{2\ln(2)D(P(x_{k+1}, y_{k+1}|X^k, Y^k)\|Q(x_{k+1}, y_{k+1}|X^k, Y^k))} \times \log \frac{|\mathcal{X}||\mathcal{Y}|}{\sqrt{2\ln(2)D(P(x_{k+1}, y_{k+1}|X^k, Y^k)\|Q(x_{k+1}, y_{k+1}|X^k, Y^k))}} \quad (181)$$

$$\leq \frac{1}{n} \sum_{k=1}^n 2\sqrt{2\ln(2)\mathbb{E}D(P(x_{k+1}, y_{k+1}|X^k, Y^k)\|Q(x_{k+1}, y_{k+1}|X^k, Y^k))} \times \log \frac{|\mathcal{X}||\mathcal{Y}|}{\sqrt{2\ln(2)\mathbb{E}D(P(x_{k+1}, y_{k+1}|X^k, Y^k)\|Q(x_{k+1}, y_{k+1}|X^k, Y^k))}} \quad (182)$$

$$\leq 2\sqrt{2\ln(2)D(P(x^{n+1}, y^{n+1})\|Q(x^{n+1}, y^{n+1}))}/n \log \frac{|\mathcal{X}||\mathcal{Y}|}{\sqrt{2\ln(2)D(P(x^{n+1}, y^{n+1})\|Q(x^{n+1}, y^{n+1}))}/n} \quad (183)$$

F) *Proof of Theorem 3:* We decompose

$$\begin{aligned} \hat{I}_3 = & \frac{1}{n} \sum_{i=1}^n \sum_{y_i} Q(y_i|X^i, Y^{i-1}) \log \frac{1}{Q(y_i|Y^{i-1})} \\ & - \frac{1}{n} \sum_{i=1}^n \sum_{y_i} Q(y_i|X^i, Y^{i-1}) \log \frac{1}{Q(y_i|X^i, Y^{i-1})}. \end{aligned} \quad (184)$$

Following the proof of almost sure and L_1 convergence of $\hat{H}_2$ in that of Proposition 2, we can show that the second term on the right-hand side of (184) converges to $\overline{H}(\mathbf{Y}|\mathbf{X})$ almost surely and in L_1 under the conditions of Theorem 3. Denote the first term on the right-hand side of (184) as

$$F_n = \frac{1}{n} \sum_{i=1}^n \sum_{y_i} Q(y_i|X^i, Y^{i-1}) \log \frac{1}{Q(y_i|Y^{i-1})}. \quad (185)$$

Then, it suffices to show the almost sure and L_1 convergence of F_n to $\overline{H}(\mathbf{Y})$. Decompose $F_n - \overline{H}(\mathbf{Y})$ as

$$F_n - \overline{H}(\mathbf{Y}) = R_n + S_n,$$

where

$$\begin{aligned} R_n = & \frac{1}{n} \sum_{i=1}^n \sum_{y_i} P(y_i|X^i, Y^{i-1}) \log P(y_i|Y^{i-1}) \\ & - \frac{1}{n} \sum_{i=1}^n \sum_{y_i} Q(y_i|X^i, Y^{i-1}) \log Q(y_i|Y^{i-1}) \quad (186) \\ S_n = & -\frac{1}{n} \sum_{i=1}^n \sum_{y_i} P(y_i|X^i, Y^{i-1}) \log P(y_i|Y^{i-1}) - \overline{H}(\mathbf{Y}). \end{aligned} \quad (187)$$

Almost sure convergence: Express R_n as $\frac{1}{n} \sum_{i=1}^n Z_i$, where

$$\begin{aligned} Z_i = & - \sum_{y_i} Q(y_i|X^i, Y^{i-1}) \log Q(y_i|Y^{i-1}) \\ & + \sum_{y_i} P(y_i|X^i, Y^{i-1}) \log P(y_i|Y^{i-1}). \end{aligned} \quad (188)$$

According to Lemma 2 in Appendix A, the CTW probability assignments, $Q(y_i|X^i, Y^{i-1})$ and $Q(y_i|Y^{i-1})$ both converge almost surely to the true probability $P(y_i|X^i, Y^{i-1})$ and $P(y_i|Y^{i-1})$. Therefore,

$$\lim_{i \rightarrow \infty} Z_i = 0 \quad P\text{-a.s.} \quad (189)$$

Then, we know the Cesàro mean of $\{Z_i\}_{i=1}^n$ also converges to zero almost surely, i.e.,

$$\lim_{n \rightarrow \infty} R_n = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Z_i = 0 \quad P\text{-a.s.} \quad (190)$$

Now we show S_n converges to zero almost surely, which is implied by Birkhoff's ergodic theorem.

L_1 Convergence: We express R_n in another form in (205), and bound $\mathbb{E}|R_n|$ from (206) to (212), where

- 1) the first part of (209) is derived by (83), and the second part of (209) is implied by the fact that the CTW probability assignment is lower bounded (27);
- 2) equation (210) follows by Pinsker's inequality,
- 3) equation (211) follows by data-processing inequality,
- 4) equation (212) follows by the chain rule of relative entropy and concavity of $\sqrt{\cdot}$.

After applying Lemma 7 in Appendix A, we know R_n converges to zero in L_1 . By Birkhoff's ergodic theorem, we know the convergence of S_n is also in L_1 , which completes the proof of L_1 convergence.

G) *Proof of Theorem 4:* We decompose $\hat{I}_4$

$$\hat{I}_4 = G_n - \hat{H}_2, \quad (191)$$

where $\hat{H}_2$ is the estimator for $\overline{H}(\mathbf{Y}|\mathbf{X})$ in $\hat{I}_2$, G_n is defined as

$$G_n = \frac{1}{n} \sum_{i=1}^n \sum_{(x_{i+1}, y_{i+1})} Q(x_{i+1}, y_{i+1}|X^i, Y^i) \log \frac{1}{Q(y_{i+1}|Y^i)}. \quad (192)$$

Since G_n is in similar form as F_n , we can follow corresponding steps in the proof of Theorem 3 to establish Theorem 4 analogously.

APPENDIX C PROOFS OF TECHNICAL LEMMAS

A) *Proof of Lemma 1:*

General stationary ergodic processes: The convergence holds almost surely by the Shannon–McMillan–Breiman theorem for causally conditional entropy rate (see, for example, [33]). We now prove the AEP also holds in L_1 .

Denote

$$A_n = -\frac{1}{n} \log P(Y^n|X^n) \quad (193)$$

$$B_n = -\frac{1}{n} \log P(Y^n|X^n, X_{-\infty}^0, Y_{-\infty}^0) \quad (194)$$

$$C_n = B_n - A_n, \quad (195)$$

where $P(Y^n|X^n, X_{-\infty}^0, Y_{-\infty}^0) = \prod_{i=1}^n P(Y_i|X_{-\infty}^i, Y_{-\infty}^{i-1})$. Our goal is to show that $\mathbb{E}|A_n - \overline{H}(\mathbf{Y}|\mathbf{X})|$ converges to zero when $n \rightarrow \infty$.

Note that

$$\mathbb{E}A_n = \frac{1}{n} \sum_{i=1}^n H(Y_i|Y^{i-1}, X^i), \quad (196)$$

$$\mathbb{E}B_n = \overline{H}(\mathbf{Y}|\mathbf{X}). \quad (197)$$

By stationarity of $(\mathbf{X}, \mathbf{Y})$ and conditioning reduces entropy, we know $H(Y_i|Y^{i-1}, X^i)$ is a nonnegative, nonincreasing sequence in i , and further, it converges to $\overline{H}(\mathbf{Y}|\mathbf{X})$. Since $\mathbb{E}A_n$ is the Cesàro mean of sequence $\{H(Y_i|Y^{i-1}, X^i)\}_{i=1}^n$, it follows that $\mathbb{E}A_n$ converges to $\overline{H}(\mathbf{Y}|\mathbf{X})$ as $n \rightarrow \infty$. Thus,

$$\lim_{n \rightarrow \infty} \mathbb{E}C_n = 0. \quad (198)$$

We have

$$\mathbb{E}|A_n - \overline{H}(\mathbf{Y}|\mathbf{X})| = \mathbb{E}|A_n - \mathbb{E}B_n| \quad (199)$$

$$\leq \mathbb{E}|C_n| + \mathbb{E}|B_n - \mathbb{E}B_n|. \quad (200)$$

By Birkhoff's ergodic theorem, $\mathbb{E}|B_n - \mathbb{E}B_n|$ converges to zero when $n \rightarrow \infty$. It now suffices to show that $\lim_{n \rightarrow \infty} \mathbb{E}|C_n| = 0$. Denote the CDF of random variable C_n as $F_n(x)$, then we have

$$\mathbb{E}|C_n| = -\mathbb{E}C_n + 2 \int_0^\infty x dF_n(x), \quad (201)$$

$$= -\mathbb{E}C_n + 2 \int_0^\infty P(C_n > x) dx, \quad (202)$$

where the second step follows by integration by parts and the fact that $1 - F_n(x) = P(C_n > x)$. Let $B(X_\infty^0, Y_\infty^0) \triangleq$

$\{(x^n, y^n) : P(x^n, y^n | X_\infty^0, Y_\infty^0) > 0\}$, we have (235), then by Markov's inequality, we have

$$P\left(\frac{P(Y^n \| X^n)}{P(Y^n \| X^n, X_\infty^0, Y_\infty^0)} \geq t_n\right) \leq \frac{1}{t_n}, \quad (203)$$

for arbitrary positive t_n .

Taking $t_n = 2^{nx}$, $x \geq 0$, we have

$$P\left(\frac{1}{n} \log \frac{P(Y^n \| X^n)}{P(Y^n \| X^n, X_\infty^0, Y_\infty^0)} \geq x\right) \leq 2^{-nx}. \quad (204)$$

See (205) and (212) given at the bottom of the page.

$$\begin{aligned} R_n &= \frac{1}{n} \sum_{i=1}^n \sum_{y_i} P(y_i | X^i, Y^{i-1}) \log \frac{P(y_i | Y^{i-1})}{Q(y_i | Y^{i-1})} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \sum_{y_i} (P(y_i | X^i, Y^{i-1}) - Q(y_i | X^i, Y^{i-1})) \log Q(y_i | Y^{i-1}), \end{aligned} \quad (205)$$

$$\begin{aligned} \mathbb{E}|R_n| &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left| \sum_{y_i} P(y_i | X^i, Y^{i-1}) \log \frac{P(y_i | Y^{i-1})}{Q(y_i | Y^{i-1})} \right| \\ &\quad + \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left| \sum_{y_i} (P(y_i | X^i, Y^{i-1}) - Q(y_i | X^i, Y^{i-1})) \log Q(y_i | Y^{i-1}) \right| \end{aligned} \quad (206)$$

$$\begin{aligned} &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\sum_{y_i} P(y_i | X^i, Y^{i-1}) \left| \log \frac{P(y_i | Y^{i-1})}{Q(y_i | Y^{i-1})} \right| \right] \\ &\quad + \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left| \sum_{y_i} (P(y_i | X^i, Y^{i-1}) - Q(y_i | X^i, Y^{i-1})) \log Q(y_i | Y^{i-1}) \right| \end{aligned} \quad (207)$$

$$\begin{aligned} &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\sum_{y_i} P(y_i | Y^{i-1}) \left| \log \frac{P(y_i | Y^{i-1})}{Q(y_i | Y^{i-1})} \right| \right] \\ &\quad + \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\sum_{y_i} \log \frac{1}{Q(y_i | Y^{i-1})} |P(y_i | X^i, Y^{i-1}) - Q(y_i | X^i, Y^{i-1})| \right] \end{aligned} \quad (208)$$

$$\begin{aligned} &\leq \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}D(P(y_i | Y^{i-1}) \| Q(y_i | Y^{i-1})) + \sqrt{\frac{2}{\ln(2)}} \sqrt{\mathbb{E}D(P(y_i | Y^{i-1}) \| Q(y_i | Y^{i-1}))} \right) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\log(2i + |\mathcal{Y}|) \sum_{y_i} |P(y_i | X^i, Y^{i-1}) - Q(y_i | X^i, Y^{i-1})| \right] \end{aligned} \quad (209)$$

$$\begin{aligned} &\leq \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}D(P(y_i | Y^{i-1}) \| Q(y_i | Y^{i-1})) + \sqrt{\frac{2}{\ln(2)}} \sqrt{\mathbb{E}D(P(y_i | Y^{i-1}) \| Q(y_i | Y^{i-1}))} \right) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \log(2i + |\mathcal{Y}|) \sqrt{2 \ln(2)} \mathbb{E}D(P(y_i | X^i, Y^{i-1}) \| Q(y_i | X^i, Y^{i-1})) \end{aligned} \quad (210)$$

$$\begin{aligned} &\leq \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}D(P(y_i | Y^{i-1}) \| Q(y_i | Y^{i-1})) + \sqrt{\frac{2}{\ln(2)}} \sqrt{\mathbb{E}D(P(y_i | Y^{i-1}) \| Q(y_i | Y^{i-1}))} \right) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \log(2i + |\mathcal{Y}|) \sqrt{2 \ln(2)} \mathbb{E}D(P(x_i, y_i | X^i, Y^{i-1}) \| Q(x_i, y_i | X^i, Y^{i-1})) \end{aligned} \quad (211)$$

$$\begin{aligned} &\leq \frac{1}{n} D(P(y^n) \| Q(y^n)) + \sqrt{\frac{2}{\ln(2)}} \frac{D(P(y^n) \| Q(y^n))}{n} \\ &\quad + \log(2n + |\mathcal{Y}|) \sqrt{\frac{2 \ln(2) D(P(x^n, y^n) \| Q(x^n, y^n))}{n}} \end{aligned} \quad (212)$$

Equivalently,

$$P(C_n > x) \leq 2^{-nx}. \quad (213)$$

Plugging (213) into (202), we have

$$\mathbb{E}|C_n| = -\mathbb{E}C_n + 2 \int_0^\infty P(C_n > x) dx \quad (214)$$

$$\leq -\mathbb{E}C_n + \frac{2}{n \ln(2)}. \quad (215)$$

By (198), we know

$$\lim_{n \rightarrow \infty} \mathbb{E}|C_n| = 0. \quad (216)$$

By (200), we know the AEP for causally conditional entropy holds in L_1 .

Irreducible aperiodic Markov processes: We express $-\frac{1}{n} \log P(Y^n \| X^n) - \overline{H}(\mathbf{Y} \| \mathbf{X})$ as

$$-\frac{1}{n} \log P(Y^n \| X^n) - \overline{H}(\mathbf{Y} \| \mathbf{X}) = \frac{1}{n} \sum_{i=1}^n Z_i, \quad (217)$$

where

$$Z_i = -\log P(Y_i | X_{i-m}^i, Y_{i-m}^{i-1}) - \overline{H}(\mathbf{Y} \| \mathbf{X}), \quad (218)$$

and m is the order of the Markov process $(\mathbf{X}, \mathbf{Y})$. Let

$$g_i = -\log P(Y_i | X_{i-m}^i, Y_{i-m}^{i-1}) \quad (219)$$

and denote $\mathbb{E}g_i$ by H . Here, H does not depend on i since the Markov process is stationary.

We decompose Z_i as

$$Z_i = (g_i^L - H^L) + (g_i^{L'} - H^{L'}), \quad (220)$$

where $g_i^L = g_i \mathbf{1}_{\{|g_i| \leq L\}}$, $g_i^{L'} = g_i - g_i^L$, $H^L = \mathbb{E}g_i^L$, and $H^{L'} = \mathbb{E}g_i^{L'} = \overline{H}(\mathbf{Y} \| \mathbf{X}) - H^L$. We expand

$$\begin{aligned} \mathbb{E} \left(\sum_{i=1}^n Z_i \right)^2 &= \mathbb{E} \left(\sum_{i=1}^n g_i^L - H^L \right)^2 + \mathbb{E} \left(\sum_{i=1}^n g_i^{L'} - H^{L'} \right)^2 \\ &\quad + 2\mathbb{E} \left(\sum_{i=1}^n g_i^L - H^L \right) \left(\sum_{i=1}^n g_i^{L'} - H^{L'} \right) \end{aligned} \quad (221)$$

and bound the three terms on the right-hand side of (221) separately.

For the first term, by Lemma 5 in Appendix A with $\mathbf{X} \leftarrow (\mathbf{X}, \mathbf{Y})$, $V_i \leftarrow g_i^L - H^L$, and $V \leftarrow L$, we have

$$\mathbb{E} \left(\sum_{i=1}^n g_i^L - H^L \right)^2 = O(nL^2). \quad (222)$$

For the second term, consider

$$\mathbb{E} \left(\sum_{i=1}^n g_i^{L'} - H^{L'} \right)^2 \leq n^2 \max_i \mathbb{E}(g_i^{L'} - H^{L'})^2. \quad (223)$$

Define

$$E_{i,K} = \{(x_{i-m}^i, y_{i-m}^{i-1}) : K \leq -\log P(y_i | x_{i-m}^i, y_{i-m}^{i-1}) \leq K+1\}, \quad (224)$$

we have

$$\mathbb{E}(g_i^{L'} - H^{L'})^2 \leq \mathbb{E}(g_i^L)^2 \quad (225)$$

$$\leq \sum_{K=L}^{\infty} \int_{E_{i,K}} (\log P(Y_i | X_{i-m}^i, Y_{i-m}^{i-1}))^2 d\mu \quad (226)$$

$$\leq \sum_{K=L}^{\infty} |\mathcal{Y}|(K+1)^2 2^{-K} \quad (227)$$

$$= O(L^2 2^{-L}), \quad (228)$$

where the last inequality is an inequality developed by McMillan [43], and the last step could be intuitively understood since the terms decay rapidly, the sum is dominated by the largest term, hence the order. Now we have

$$\mathbb{E} \left(\sum_{i=1}^n g_i^{L'} - H^{L'} \right)^2 = O(n^2 L^2 2^{-L}). \quad (229)$$

See (230) and (235) given at the bottom of the page.

$$\mathbb{E} \left[\frac{P(Y^n \| X^n)}{P(Y^n \| X^n, X_{-\infty}^0, Y_{-\infty}^0)} \right] = \mathbb{E} \left[\mathbb{E} \left\{ \frac{P(Y^n \| X^n)}{P(Y^n \| X^n, X_{-\infty}^0, Y_{-\infty}^0)} \middle| X_{-\infty}^0, Y_{-\infty}^0 \right\} \right] \quad (230)$$

$$= \mathbb{E} \left[\sum_{(x^n, y^n) \in B(X_{-\infty}^0, Y_{-\infty}^0)} \frac{P(y^n \| x^n)}{P(y^n \| x^n, X_{-\infty}^0, Y_{-\infty}^0)} P(x^n, y^n | X_{-\infty}^0, Y_{-\infty}^0) \right] \quad (231)$$

$$= \mathbb{E} \left[\sum_{(x^n, y^n) \in B(X_{-\infty}^0, Y_{-\infty}^0)} P(y^n \| x^n) P(x^n \| y^{n-1}, X_{-\infty}^0, Y_{-\infty}^0) \right] \quad (232)$$

$$\leq \sum_{(x^n, y^n)} P(y^n \| x^n) P(x^n \| y^{n-1}) \quad (233)$$

$$= \sum_{(x^n, y^n)} P(x^n, y^n) \quad (234)$$

$$= 1 \quad (235)$$

For the third term, we apply the Cauchy–Schwarz inequality

$$2\mathbb{E} \left(\sum_{i=1}^n g_i^L - H^L \right) \left(\sum_{i=1}^n g_i^{L'} - H^{L'} \right) \quad (236)$$

$$\leq 2\sqrt{\mathbb{E} \left(\sum_{i=1}^n g_i^L - H^L \right)^2} \sqrt{\mathbb{E} \left(\sum_{i=1}^n g_i^{L'} - H^{L'} \right)^2} \quad (237)$$

$$= O(n^{3/2} L^2 2^{-L/2}). \quad (238)$$

Summing the three terms together and taking $L = 2 \log n$, we have

$$\mathbb{E} \left| \sum_{i=1}^n Z_i \right|^2 = O(n(\log n)^2) \quad (239)$$

and thus

$$\mathbb{E} \left| -\frac{1}{n} \log P(Y^n \| X^n) - \overline{H}(\mathbf{Y} \| \mathbf{X}) \right| = \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n Z_i \right| \quad (240)$$

$$\leq \frac{1}{n} \sqrt{\mathbb{E} \left| \sum_{i=1}^n Z_i \right|^2} \quad (241)$$

$$= O(n^{-1/2} \log n). \quad (242)$$

Now we deal with the almost sure convergence rates of AEP of causally conditional entropy rate. We restate the Gál–Koksma theorem[44] as follows.

Lemma 11 (Gál–Koksma Theorem): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(Z_n)_{n \geq 1}$ be a sequence of random variables belonging to L^p , $p \geq 1$, such that

$$\mathbb{E}|Z_{M+1} + Z_{M+2} + \cdots + Z_{M+n}|^p = O(\Psi(n)) \quad (243)$$

uniformly in M , where $\Psi(n)/n$ is a nondecreasing sequence. Then, for every $\epsilon > 0$,

$$\begin{aligned} Z_1(\omega) + Z_2(\omega) + \cdots + Z_n(\omega) \\ = o((\Psi(n)(\log n)^{p+1+\epsilon})^{\frac{1}{p}}) \quad P\text{-a.s.} \end{aligned} \quad (244)$$

The bound in (239) indicates that if we take $\Psi(n) = n(\log n)^2$ and $p = 2$ in the Gál–Koksma theorem, then for every $\epsilon > 0$,

$$-\frac{1}{n} \log P(Y^n \| X^n) - \overline{H}(\mathbf{Y} \| \mathbf{X}) = o(n^{-1/2} (\log n)^{5/2+\epsilon}) \quad (245)$$

$$P\text{-a.s.} \quad (246)$$

B) Proof of Lemma 2: Denote the alphabet size $|\mathcal{X}|$ as M . We examine the updating computation of $P_w^\lambda(X_{n+1} = q|x^n)$, $q = 0, 1, \dots, M-1$. For an internal node s in the updating path, if js is in the updating path, we have (33). For the leaf node v in the updating path,

$$P_w^v(X_{n+1} = q|x^n) = P_e^v(X_{n+1} = q|x^n). \quad (247)$$

The computation of $P_w^\lambda(X_{n+1} = q|x^n)$ starts from a leaf and is repeated recursively along the updating path, until we reach the root node λ and obtain $P_w^\lambda(X_{n+1} = q|x^n)$. Thus, $P_w^\lambda(X_{n+1} = q|x^n)$ is a weighted sum of $P_e^s(X_{n+1} = q|x^n)$, where s is any node in the updating path.

Let $\{s \rightarrow \lambda\}$ denote the set of nodes in the path from s to λ . The weight associated with $P_e^s(X_{n+1} = q|x^n)$ is

$$\beta^s(x^n) \prod_{u \in \{s \rightarrow \lambda\}} \frac{1}{\beta^u(x^n) + 1}, \quad (248)$$

where s is an internal node in the updating path. The weight associated with $P_w^v(X_{n+1} = q|x^n)$, where v is the leaf node in the updating path, is

$$\prod_{u \in \{u \rightarrow \lambda\} \setminus \{v\}} \frac{1}{\beta^u(x^n) + 1}. \quad (249)$$

The convergence properties of $P_w^\lambda(X_{n+1} = q|x^n)$ depends on the limiting behavior of $\beta^s(X^n)$ at every node s along the updating path. If s is an internal node in the tree representation of the source, we actually have $\lim_{n \rightarrow \infty} \beta^s(X^n) = 0$ almost surely. This fact was stated in [15, Lemma 4]. Here, we restate this fact and give a proof for stationary irreducible aperiodic finite-alphabet Markov processes.

Lemma 12: Let s be an internal node in the tree representation of the source. Then,

$$\lim_{n \rightarrow \infty} \beta^s(X^n) = 0 \quad P\text{-a.s.} \quad (250)$$

Proof: It suffices to show

$$\lim_{n \rightarrow \infty} \frac{\beta^s(X^n)}{\beta^s(X^n) + 1} = 0 \quad P\text{-a.s.} \quad (251)$$

We have

$$\frac{\beta^s(X^n)}{\beta^s(X^n) + 1} \quad (252)$$

$$= \frac{P_e^s(X^n)}{2P_w^s(X^n)} \quad (253)$$

$$\leq \frac{P_e^s(X^n)}{\prod_{i=0}^{M-1} P_w^{is}(X^n)} \quad (254)$$

$$\leq 2^M \frac{P_e^s(X^n)}{\prod_{i=0}^{M-1} P_e^{is}(X^n)} \quad (255)$$

$$= 2^M \exp \left\{ n_s \cdot \left(\frac{1}{n_s} \log P_e^s(X^n) - \frac{1}{n_s} \log \prod_{i=0}^{M-1} P_e^{is}(X^n) \right) \right\}, \quad (256)$$

where n_s denotes the number of symbols in X^n with context s , and the inequalities follow from applying (24) repeatedly. Here, since s is an internal node of the tree, without loss of generality, we can assume offsprings of s do not have all the same conditional distribution. If it were violated, we can simply iterate the inequalities obtain above till we reach the leaf nodes of the tree, after which we can apply the same arguments that will be shown later.

It was shown in [32] that the Krichevsky–Trofimov probability estimate of sequence X^n , i.e., $P_e(X^n)$, satisfies the following bound:

$$\left| \frac{\log P_e(X^n)}{n} - \sum_{a \in \mathcal{X}} \frac{N(a|X^n)}{n} \log \frac{N(a|X^n)}{n} \right| \leq \frac{M-1}{2} \frac{\log n}{n} + \frac{C}{n}, \quad (257)$$

where $N(a|X^n)$ denotes the number of symbol a in the sequence X^n , and C is a constant depending only on the alphabet size M .

Under the assumption of Lemma 2, Markov process $\mathbf{X}$ is ergodic, hence

$$\lim_{n \rightarrow \infty} \frac{N(a|X^n)}{n} = \pi(a), \quad (258)$$

where $\pi(\cdot)$ is the stationary distribution of $\mathbf{X}$. Equation (258) implies that

$$\lim_{n \rightarrow \infty} \frac{\log P_e(X^n)}{n} = -H(\pi). \quad (259)$$

Applying the same argument to $\frac{1}{n_s} \log P_e^s(X^n)$, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n_s} \log P_e^s(X^n) = -H(\pi_s), \quad (260)$$

where π_s is the stationary conditional distribution conditioned on context s . Analogously, for node is , we have

$$\lim_{n \rightarrow \infty} \frac{1}{n_{is}} \log P_e^{is}(X^n) = -H(\pi_{is}), \quad (261)$$

thus

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n_s} \log P_e^s(X^n) - \frac{1}{n_s} \log \prod_{i=0}^{M-1} P_e^{is}(X^n) \\ &= - \left(H(\pi_s) - \sum_{i \in \mathcal{X}} p_i H(\pi_{is}) \right), \end{aligned} \quad (262)$$

where $p_i = P(\text{context is } is | \text{context is } s)$. It is obvious that

$$\pi_s = \sum_{i \in \mathcal{X}} p_i \pi_{is}. \quad (263)$$

By the strict concavity of entropy functional and the fact that the offsprings of s do not all have the same conditional distribution, we know

$$\lim_{n \rightarrow \infty} \frac{1}{n_s} \log P_e^s(X^n) - \frac{1}{n_s} \log \prod_{i=0}^{M-1} P_e^{is}(X^n) < 0 \quad (264)$$

which implies

$$\lim_{n \rightarrow \infty} \frac{\beta^s(X^n)}{\beta^s(X^n) + 1} = 0 \quad P\text{-a.s.} \quad (265)$$

hence

$$\lim_{n \rightarrow \infty} \beta^s(X^n) = 0 \quad P\text{-a.s.} \quad (266)$$

holds. $\blacksquare$

We know $Q(q|X^n) = P_w^\lambda(X_{n+1} = q|X^n)$ can be expressed as a weighted sum of $P_e^s(X_{n+1} = q|X^n)$ for s in the updating path

$$Q(q|X^n) = \sum_s w_s P_e^s(X_{n+1} = q|X^n), \quad (267)$$

where w_s are given in (248) and (249). Lemma 12 implies that for s an internal node of the tree representation of $\mathbf{X}$, $w_s \rightarrow 0$. Hence,

$$\lim_{n \rightarrow \infty} Q(q|X^n) - \sum_{\text{leaf node}} w_s P_e^s(X_{n+1} = q|X^n) = 0 \quad P\text{-a.s.} \quad (268)$$

For s leaf node, by the property of Krichevsky–Trofimov probability estimate, we know

$$\lim_{n \rightarrow \infty} P_e^s(X_{n+1} = q|X^n) - P(q|X^n) = 0, \quad (269)$$

where $P(q|X^n)$ is the true conditional probability. Thus, we have

$$\begin{aligned} Q(x_{n+1}|X^n) - P(x_{n+1}|X^n) \\ = P_w^\lambda(x_{n+1}|X^n) - P(x_{n+1}|X^n) \end{aligned} \quad (270)$$

$$\rightarrow 0 \quad P\text{-a.s.} \quad (271)$$

C) Proof of Lemma 3: Fix $\epsilon > 0$. Since $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ is bounded and closed, $f(\cdot)$ is uniformly continuous. Thus, there exists δ_ϵ such that $|f(P) - f(Q)| \leq \epsilon$ if $\|P - Q\|_1 \leq \delta_\epsilon$. Furthermore, $f(\cdot)$ is bounded by $f_{\max} \triangleq \log |\mathcal{X}| + \log |\mathcal{Y}|$. Therefore, we have

$$|f(P) - f(Q)| \leq \epsilon \mathbf{1}_{\{\|P - Q\|_1 \leq \delta_\epsilon\}} + f_{\max} \mathbf{1}_{\{\|P - Q\|_1 > \delta_\epsilon\}} \quad (272)$$

$$\leq \epsilon + f_{\max} \frac{\|P - Q\|_1}{\delta_\epsilon} \quad (273)$$

$$= \epsilon + K_\epsilon \|P - Q\|_1, \quad (274)$$

where $K_\epsilon = f_{\max}/\delta_\epsilon$.

D) Proof of Lemma 4: Since

$$H(Y|X) = H(X, Y) - H(X), \quad (275)$$

we can bound $|f(P) - f(Q)|$ as

$$\begin{aligned} |f(P) - f(Q)| \\ = |H_P(X, Y) - H_P(X) - H_Q(X, Y) + H_Q(X)| \end{aligned} \quad (276)$$

$$\leq |H_P(X, Y) - H_Q(X, Y)| + |H_P(X) - H_Q(X)|. \quad (277)$$

Now, by [45, Lemma 2.7], we have

$$|H_P(X, Y) - H_Q(X, Y)| \leq \theta \log \frac{|\mathcal{X}||\mathcal{Y}|}{\theta}, \quad (278)$$

$$|H_P(X) - H_Q(X)| \leq \theta_X \log \frac{|\mathcal{X}|}{\theta_X}, \quad (279)$$

where $\theta = \|P_{XY} - Q_{XY}\|_1$ and $\theta_X = \|P_X - Q_X\|_1$. Since

$$\theta = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} |P(x, y) - Q(x, y)| \quad (280)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} |P(x, y) - Q(x, y)| \quad (281)$$

$$\geq \sum_{x \in \mathcal{X}} \left| \sum_{y \in \mathcal{Y}} P(x, y) - Q(x, y) \right| \quad (282)$$

$$= \sum_{x \in \mathcal{X}} |P(x) - Q(x)| \quad (283)$$

$$= \theta_X, \quad (284)$$

we have

$$|f(P) - f(Q)| \leq 2\theta \log \frac{|\mathcal{X}||\mathcal{Y}|}{\theta}. \quad (285)$$

E) Proof of Lemma 5: We first define the α -mixing coefficient of a stationary process.

Definition 4 (α -Mixing Coefficient): For a stationary process $\mathbf{X}$ adapted to the filtration $(\mathcal{F}_n)_{-\infty}^{\infty}$, the α -mixing coefficient is defined as

$$\alpha(n) \triangleq \sup |P(A \cap B) - P(A)P(B)|, \quad (286)$$

where the supremum is over all $A \in \mathcal{F}_{-\infty}^0$ and $B \in \mathcal{F}_n^{\infty}$.

According to [46], if $\mathbf{X}$ is a stationary irreducible aperiodic Markov process, $\alpha(n)$ tends to zero exponentially fast in n , i.e., there exist $C_7 > 0$ and $C_8 > 0$ such that

$$\alpha(n) \leq C_7 e^{-C_8 n}. \quad (287)$$

We bound $\mathbb{E}((1/n) \sum_{i=1}^n V_i)^2$ as follows:

$$\mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n V_i \right|^2 = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}|V_i|^2 + \frac{2}{n^2} \sum_{1 \leq i < j \leq n} \mathbb{E}V_i V_j \quad (288)$$

$$\leq \frac{V^2}{n} + \frac{2}{n^2} \sum_{1 \leq i < j \leq n} \mathbb{E}V_i V_j, \quad (289)$$

where (289) holds because $V_i \forall i$ is uniformly bounded by constant V .

By Billingsley's inequality [47, Corollary 1.1], taking into account that $\mathbb{E}V_i = 0 \forall i$, we know the following bound holds:

$$|\mathbb{E}V_i V_j| = |\text{Cov}(V_i, V_j)| \leq 4V^2 \alpha(|i - j|). \quad (290)$$

Plugging (290) into (289), we have

$$\mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n V_i \right|^2 \leq \frac{V^2}{n} + \frac{8V^2}{n^2} \sum_{1 \leq i < j \leq n} \alpha(|i - j|) \quad (291)$$

$$\leq \frac{V^2}{n} + \frac{8V^2}{n^2} C_7 \sum_{k=1}^{n-1} k e^{-C_8(n-k)} \quad (292)$$

$$\leq \frac{V^2}{n} + \frac{8C_7 V^2 e^{C_8}}{n(e^{C_8} - 1)^2}. \quad (293)$$

Thus, we show Lemma 5 holds with $C_4 = 1 + 8C_7 e^{C_8} / (e^{C_8} - 1)^2$.

ACKNOWLEDGMENT

The authors would like to thank T. Coleman for helpful discussions on the merits of nonnegative directed information estimators during H. Permuter's visit at UCSD. They would also like to thank the Associate Editor and anonymous reviewers for their very helpful suggestions that significantly improved the presentation of the results. J. Jiao would like to thank H. Kim for insightful discussions in the revision stage of the paper.

REFERENCES

- [1] H. Marko, "The bidirectional communication theory-a generalization of information theory," *IEEE Trans. Commun.*, vol. COM-21, no. 12, pp. 1345-1351, Dec. 1973.
- [2] J. L. Massey, "Causality, feedback, and directed information," in *Proc. Int. Symp. Inf. Theory Appl.*, Honolulu, HI, USA, Nov. 1990, pp. 303-305.
- [3] G. Kramer, "Directed information for channels with feedback," Dr. sc. thchn. dissertation, Swiss Federal Institute of Technology, Zurich, Switzerland, 1998.
- [4] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4-21, Jan. 2003.
- [5] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323-349, Jan. 2009.
- [6] Y.-H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1488-1499, Apr. 2008.
- [7] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 644-662, Feb. 2009.
- [8] H. H. Permuter, Y.-H. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 3248-3259, Jun. 2011.
- [9] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424-438, 1969.
- [10] P. Mathai, N. C. Martins, and B. Shapiro, "On the detection of gene network interconnections using directed mutual information," presented at the UCSD Inf. Theory Appl. Workshop, 2007.
- [11] A. Rao, A. O. Hero, D. J. States, and J. D. Engel, "Using directed information to build biologically relevant influence networks," *J. Bioinf. Comput. Biol.*, vol. 6, no. 3, pp. 493-519, 2008.
- [12] S. Verdú, "Universal estimation of information measures," presented at the IEEE Inf. Theory Workshop, 2005.
- [13] A. D. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1250-1258, Nov. 1989.
- [14] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1270-1279, Jul. 1993.
- [15] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal divergence estimation for finite-alphabet sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3456-3475, Aug. 2006.
- [16] M. Burrows and D. J. Wheeler, A block-sorting lossless data compression algorithm Digital Systems Research Center, Palo Alto, CA, USA, Tech. Rep. 124, 1994.
- [17] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653-664, May 1995.
- [18] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal entropy estimation via block sorting," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1551-1561, Jul. 2004.
- [19] J. Yu and S. Verdú, "Universal erasure entropy estimation," in *Proc. IEEE Int. Symp. Inf. Theory*, 2006, pp. 2358-2362.
- [20] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *J. Comput. Neurosci.*, vol. 30, pp. 17-44, 2011.

- [21] L. Zhao, Y.-H. Kim, H. H. Permuter, and T. Weissman, "Universal estimation of directed information," in *Proc. IEEE Int. Symp. Inf. Theory*, 2010, pp. 230–234.
- [22] J. L. Massey and P. C. Massey, "Conservation of mutual and directed information," in *Proc. IEEE Int. Symp. Inf. Theory*, 2005, pp. 157–158.
- [23] P.-O. Amblard and O. J. J. Michel, "Relating Granger causality to directed information theory for networks of stochastic processes," 2011 [Online]. Available: <http://arxiv.org/abs/0911.2873v4>
- [24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.
- [25] D. Ornstein, "Guessing the next output of a stationary process," *Israel J. Math.*, vol. 30, pp. 292–296, 1978.
- [26] P. Algoet, "Universal schemes for prediction, gambling and portfolio selection," *Annu. Probab.*, vol. 20, pp. 901–941, 1992.
- [27] G. Morvai, S. J. Yakowitz, and P. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 483–498, Mar. 1997.
- [28] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Feb. 1998.
- [29] F. Willems and T. Tjalkens, Complexity reduction of the context-tree weighting algorithm: A study for KPN research Univ. Eindhoven, Eindhoven, The Netherlands, EIDMA Rep. RS.97.01, 1997.
- [30] T. J. Tjalkens, Y. M. Shtarkov, and F. M. J. Willems, "Sequential weighting algorithms for multi-alphabet sources," in *Proc. 6th Joint Swedish-Russian Int. Workshop Inf. Theory*, 1993, pp. 230–234.
- [31] F. M. J. Willems, "The context-tree weighting method: Extensions," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 792–798, Mar. 1998.
- [32] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 2, pp. 199–207, Mar. 1981.
- [33] R. Venkataramanan and S. S. Pradhan, "Source coding with feed-forward: Rate-distortion theorems and error exponents for a general source," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2154–2179, Jun. 2007.
- [34] J. Birch, "Approximations for the entropy for functions of Markov chains," *Annu. Math. Statist.*, vol. 33, pp. 930–938, 1962.
- [35] F. L. Gland and L. Mevel, "Exponential forgetting and geometric ergodicity in hidden Markov models," *Math. Control Signals Syst.*, vol. 13, no. 1, pp. 63–93, 2000.
- [36] B. M. Hochwald and P. Jelenkovic, "State learning and mixing in entropy of hidden Markov processes and the Gilbert-Elliott channel," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 128–138, Jan. 1999.
- [37] S. Kleinberg and G. Hripesak, "A review of causal inference for biomedical informatics," *J. Biomed. Inf.*, vol. 44, no. 6, pp. 1102–1112, 2011.
- [38] L. Breiman, "The individual ergodic theorem of information theory," *Annu. Math. Statist.*, vol. 28, no. 3, pp. 809–811, 1957.
- [39] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA, USA: Holden-Day, 1964.
- [40] A. R. Barron, "Entropy and the central limit theorem," *Annu. Probab.*, vol. 14, pp. 336–342, 1986.
- [41] L. Breiman, *Probability*. Philadelphia, PA, USA: SIAM, 1992.
- [42] A. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY, USA: Springer-Verlag, 2008.
- [43] B. McMillan, "The basic theorems of information theory," *Annu. Math. Statist.*, vol. 24, no. 2, pp. 196–219, 1953.
- [44] I. S. Gál and J. F. Koksma, "Sur l'ordre de grandeur des fonctions sommables," *C. R. Acad. Sci. Paris*, vol. 227, pp. 1321–1323, 1948.
- [45] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Budapest, Hungary: Akadémiai Kiadó, 1981.
- [46] R. Bradley, "Basic properties of strong mixing conditions. A survey and some open questions," *Probab. Surv.*, vol. 2, pp. 107–144, 2005.
- [47] D. Bosq, "Nonparametric statistics for stochastic processes," *Lecture Notes Statist.*, 1996.

Jiantao Jiao (SM'13) received the B.Eng. degree with the highest honor in Electronic Engineering from Tsinghua University, Beijing, China, in 2012. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering, Stanford University. His research interests include information theory and statistical signal processing, with applications in communication, control, computation, networking, data compression, and learning.

Mr. Jiao is a recipient of the Stanford Graduate Fellowship (SGF), the highest award offered by Stanford University.

Haim H. Permuter (M'08) received his B.Sc. (summa cum laude) and M.Sc. (summa cum laude) degrees in Electrical and Computer Engineering from the Ben-Gurion University, Israel, in 1997 and 2003, respectively, and the Ph.D. degree in Electrical Engineering from Stanford University, California in 2008.

Between 1997 and 2004, he was an officer at a research and development unit of the Israeli Defense Forces. He is currently a senior lecturer at Ben-Gurion University.

Dr. Permuter is a recipient of the Fulbright Fellowship, the Stanford Graduate Fellowship (SGF), Allon Fellowship, and the 2009 U.S.-Israel Binational Science Foundation Bergmann Memorial Award.

Lei Zhao received the B.Eng. degree from Tsinghua University, China, in 2003, the M.S. degree in Electrical and Computer Engineering from Iowa State University, Ames, in 2006, and the Ph.D. degree in Electrical Engineering from Stanford University, California in 2011.

Dr. Zhao is currently working at Jump Operations, Chicago, IL, USA.

Young-Han Kim (S'99–M'06–SM'12) received the B.S. degree with honors in electrical engineering from Seoul National University, Seoul, Korea, in 1996 and the M.S. degrees in electrical engineering and in statistics, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 2001, 2006, and 2006, respectively.

In July 2006, he joined the University of California, San Diego, where he is an Associate Professor of Electrical and Computer Engineering. His research interests are in statistical signal processing and information theory, with applications in communication, control, computation, networking, data compression, and learning.

Dr. Kim is a recipient of the 2008 NSF Faculty Early Career Development (CAREER) Award the 2009 US-Israel Binational Science Foundation Bergmann Memorial Award, and the 2012 IEEE Information Theory Paper Award. He is currently on the Editorial Board of the IEEE TRANSACTIONS ON INFORMATION THEORY, serving as an Associate Editor for Shannon theory. He is also serving as a Distinguished Lecturer for the IEEE Information Theory Society.

Tsachy Weissman (S'99–M'02–SM'07–F'13) graduated summa cum laude with a B.Sc. in electrical engineering from the Technion in 1997, and earned his Ph.D. at the same place in 2001. He then worked at Hewlett-Packard Laboratories with the information theory group until 2003, when he joined Stanford University, where he is Associate Professor of Electrical Engineering and incumbent of the STMicroelectronics chair in the School of Engineering. He has spent leaves at the Technion, and at ETH Zurich.

Tsachy's research is focused on information theory, statistical signal processing, the interplay between them, and their applications.

He is recipient of several best paper awards, and prizes for excellence in research.

He currently serves on the editorial boards of the IEEE TRANSACTIONS ON INFORMATION THEORY and Foundations and Trends in Communications and Information Theory.