Universal Estimation of Directed Information

Lei Zhao, Haim Permuter, Young-Han Kim, and Tsachy Weissman

***Abstract*— In this paper, we develop a universal algorithm to estimate Massey's directed information for stationary ergodic processes. The sequential probability assignment induced by a universal source code plays the critical role in the estimation. In particular, we use context tree weighting to implement the algorithm. Some numerical results are provided to illustrate the performance of the proposed algorithm.**

I. Introduction

First introduced by Massey in [1], directed information arises as a natural counterpart of mutual information for channel capacity when feedback is present. In [2] and [3], Kramer extended the use of directed information to discrete memoryless networks with feedback, including the two-way channel and the multiple access channel. 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 [4] proved that the feedback capacity is equal to the limit of the supremum of the normalized directed information from the input to the output. Tatikonda and Mitter [5] used directed information to prove a general feedback channel coding theorem for channels with memory. In [6], Permuter et al. considered the capacity of discrete-time channels with feedback where the feedback is a time-invariant deterministic function of the output. Under mild conditions, they showed that the capacity is the maximum of the normalized directed information between the input and output sequence in the limit. Recently, Permuter et al. [7] showed that directed information plays an important role in portfolio theory, data compression, and hypothesis testing, where causality constraints exist.

Besides information theory, directed information is shown to be a valuable tool in biology, when inference about causality is needed. In [8], directed information was used to identify pairwise influence. The authors in [9] used directed information to test the inference of influence in gene networks. Thus it is of both theoretical and practical interests to develop a way to estimate directed information efficiently.

As we were completing this paper, [10] was brought to our attention, in which the authors used directed information to infer causal relationships in ensemble neural spike train recordings. At the heart of both our estimation framework and theirs is the estimation of causally conditional entropy. The main difference is that they took a parametric approach [10, Assumption 3, Page 10], while our approach is based on non-parametric universal data compressors, and therefore leading to stronger universality properties.

Lei Zhao and Tsachy Weissman are with the Department of Electrical Engineering, Stanford University. Email: {leiz,tsachy}@stanford.edu

Haim Permuter is with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel. Email: haimp@bgu.ac.il

Young-Han Kim is with the Department of Electrical and Computer Engineering, University of California, San Diego. Email: yhk@ucsd.edu

Notation: We use capital letter X to denote a random variable and use small letter x to denote the corresponding realization or constant. Calligraphic letter $\mathcal{X}$ denotes the alphabet of X and $|\mathcal{X}|$ denotes the cardinality of the alphabet.

II. Preliminaries

We first give the mathematical definitions of directed information and causally conditional entropy, and then discuss the relation between universal sequential probability assignment and universal source coding.

A. Directed information

Directed information from X^n to Y^n is defined as

$$I\left(X^n \to Y^n\right) = H(Y^n) - H(Y^n||X^n), \tag{1}$$

where $H(Y^n||X^n)$ is the *causally conditional entropy* [2], defined as

$$H(Y^n||X^n) = \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^i). \tag{2}$$

Compared with the definition of mutual information

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

the conditional entropy is replaced by the causal conditioning. And unlike mutual information, directed information is not symmetric, i.e., $I\left(Y^n \to X^n\right) \neq I\left(X^n \to Y^n\right)$ in general. Other interesting properties of directed information such as the conservation law can be found in [2], [11].

For random processes $\mathbf{X}$,$\mathbf{Y}$, which are jointly stationary, we can define directed information rate [2] as follows:

$$\mathrm{H}(\mathbf{Y}||\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} H(Y^n||X^n), \tag{3}$$

$$\mathrm{I}\left(\mathbf{X} \to \mathbf{Y}\right) = \lim_{n\to\infty} \frac{1}{n} I\left(X^n \to Y^n\right), \tag{4}$$

The existence of the limit can be checked as follows:

$$\begin{aligned}
&\mathrm{I}\left(\mathbf{X} \to \mathbf{Y}\right) \\
&= \lim_{n\to\infty} \frac{1}{n} I(X^n \to Y^n) \\
&= \lim_{n\to\infty} \frac{1}{n} \left(H(Y^n) - H(Y^n||X^n)\right) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^i) \\
&= H(Y_0|Y_{-\infty}^{-1}) - H(Y_0|X_{-\infty}^{0}, Y_{-\infty}^{-1}),
\end{aligned}$$

where the last equality is obtained via the property of Cesáro mean [12] and standard martingale arguments [13]. Note that the entropy rate $\mathrm{H}(\mathbf{Y})$ of the process $\mathbf{Y}$ is equal to $H(Y_0|Y_{-\infty}^{-1})$, and $\mathrm{H}(\mathbf{Y}||\mathbf{X}) = H(Y_0|X_{-\infty}^{0}, Y_{-\infty}^{-1})$. Thus

$$\mathrm{I}\left(\mathbf{X} \to \mathbf{Y}\right) = \mathrm{H}(\mathbf{Y}) - \mathrm{H}(\mathbf{Y}||\mathbf{X}). \tag{5}$$

B. Universal sequential probability assignment and universal source coding

A sequential probability assignment Q consists of a set of conditional probabilities $\{Q_{X_i|x^{i-1}}(\cdot), \forall x^{i-1} \in \mathcal{X}^{i-1}\}_{i=1}^{\infty}$. Note that Q induces a probability distribution on X^n in the obvious way.

Definition 1 *A sequential probability assignment Q is universal if*

$$\limsup_{n\to\infty} \frac{1}{n} D(P_{X^n}||Q_{X^n}) = 0 \tag{6}$$

for any stationary probability measure P.

A source code for an n-block source sequence, $\mathcal{C}_n$, is defined as a mapping from a source sequence x^n to a binary sequence of finite length, i.e.,

$$\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*. \tag{7}$$

More explicitly,

$$\mathcal{C}_n(x^n) = b_1, b_2, \cdots, b_{l_n}, \tag{8}$$

where $l_n = l_n(x^n)$ is the code length. Furthermore, $\mathcal{C}_n$ is said to be non-singular if $\mathcal{C}_n(x^n) \neq \mathcal{C}_n(y^n), \forall x^n \neq y^n$. It is said to be *uniquely decodable* if all its extensions are non-singular [12]. The codeword lengths $l_n(\cdot)$ of any uniquely decodable code must satisfy the Kraft inequality

$$\sum_{x^n \in \mathcal{X}^n} 2^{-l_n(x^n)} \leq 1. \tag{9}$$

See [12] for a proof.

For any uniquely decodable code, we have

$$\frac{1}{n}\mathsf{E} l_n(X^n) = \frac{1}{n}H(X^n) + \frac{1}{n}D(P_{X^n}||Q_{X^n}) - \frac{1}{n}\log k_n, \tag{10}$$

where $k_n = \sum_{x^n} 2^{-l(x^n)}$, and $Q(x^n) = \frac{2^{-l(x^n)}}{k_n}$. $Q(x^n)$ induces a probability measure on $\mathcal{X}^n$. With slight abuse of notation, call this measure Q.

Definition 2 *The sequential probability assignment induced by a source code $\mathcal{C}_n$ is the set of conditional probabilities $\{Q_{X_i|x^{i-1}}\}_{i=1}^n$, where*

$$Q_{X_i|x^{i-1}}(a) = \frac{Q(ax^{i-1})}{Q(x^{i-1})}. \tag{11}$$

where, ax^{i-1} is a concatenation of symbol a and sequence x^{i-1}.

Definition 3 *A source coding scheme is a sequence of source codes. It is said to be universal if each code is uniquely decodable and*

$$\limsup_{n\to\infty} \frac{1}{n}\mathsf{E}\left[l_n(X^n)\right] = \mathrm{H}(\mathbf{X}). \tag{12}$$

for every stationary source $\mathbf{X}$.

The per-symbol expected number of bits based on the universal source coding scheme is a good estimate of $\mathrm{H}(\mathbf{X})$.

By the Kraft inequality, $-\frac{1}{n}\log k_n \geq 0$. Given a universal source coding scheme and a stationary source $\mathbf{X}$, by (10),

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} D(P_{X^n}||Q_{X^n}) \\
&\leq \limsup_{n\to\infty} \left(\frac{1}{n}\mathsf{E} l_n(X^n) - \frac{1}{n}H(X^n)\right) \\
&= 0.
\end{aligned} \tag{13}$$

Thus we can construct a universal sequential probability assignment from a universal coding scheme.

III. Estimation of $\mathrm{I}(\mathbf{X} \to \mathbf{Y})$

As we have seen, $\mathrm{I}(\mathbf{X} \to \mathbf{Y}) = \mathrm{H}(\mathbf{Y}) - \mathrm{H}(\mathbf{Y}||\mathbf{X})$. In this section, we will show an estimate of $\mathrm{H}(\mathbf{Y}||\mathbf{X})$ based on a universal sequential probability assignments. Similar method applies to the estimate of $\mathrm{H}(\mathbf{Y})$.

Let $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ be the set of all distribution on $\mathcal{X} \times \mathcal{Y}$. Define f as the function that maps a joint distribution $P_{X,Y}$ of a random vector (X, Y) to the corresponding conditional entropy $H(Y|X)$, i.e.,

$$f(P_{X,Y}) = -\sum_{x,y} P_{X,Y}(x,y) \log P_{Y|X}(y|x), \tag{14}$$

where $P_{Y|X}(\cdot|\cdot)$ is the conditional distribution induced by $P_{X,Y}$.

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

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

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

Proof: 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 $f_{\max} \triangleq \log|\mathcal{X}| + \log|\mathcal{Y}|$. We have

$$\begin{aligned}
|f(P) - f(Q)| &\leq \epsilon \mathbf{1}_{\{\|P-Q\|_1 \leq \delta_\epsilon\}} + f_{\max}\mathbf{1}_{\{\|P-Q\|_1 > \delta_\epsilon\}} \\
&\leq \epsilon + f_{\max}\frac{\|P-Q\|_1}{\delta_\epsilon} \\
&\leq \epsilon + \frac{f_{\max}}{\delta_\epsilon}\|P-Q\|_1 \\
&= \epsilon + K_\epsilon \|P-Q\|_1,
\end{aligned} \tag{15}$$

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

Lemma 2 *[15] Let $\mathbf{X}$ be a stationary ergodic process. If $\lim_{k\to\infty} g_k(\mathbf{X}) \to g(\mathbf{X})$, w.p. 1, and $\{g_k(\cdot)\}$ are bounded, then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} g_k\left(T^k(\mathbf{X})\right) = \mathsf{E} g(\mathbf{X}), \text{ w.p. 1 and in } L_1, \tag{16}$$

where $T(\cdot)$ is the shift operator.

Now define $g_k(\mathbf{X}, \mathbf{Y}) = f(P_{X_0,Y_0|X_{-k}^{-1},Y_{-k}^{-1}})$ for a jointly stationary and ergodic process $(\mathbf{X}, \mathbf{Y})$. Note that, by martingale convergence [13], $g_k(\mathbf{X}, \mathbf{Y}) \to g(\mathbf{X}, \mathbf{Y})$, w.p. 1, where $g(\mathbf{X}, \mathbf{Y}) = f(P_{X_0,Y_0|X_{-\infty}^{-1},Y_{-\infty}^{-1}})$. Noting further that

$\mathsf{E}g(\mathbf{X},\mathbf{Y}) = \mathrm{H}(\mathbf{Y}||\mathbf{X})$, we can apply Lemma 2 and get the following corollary:

Corollary 1

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(P_{X_k,Y_k|X^{k-1},Y^{k-1}}) = \mathrm{H}(\mathbf{Y}||\mathbf{X}) \text{ w.p. 1, and in } L_1. \tag{17}$$

Let Q be a universal sequential probability assignment on $(\mathcal{X}\times\mathcal{Y})^\infty$. Denote the normalized casual conditional entropy estimate induced by Q by:

$$\widehat{H}_Q(X^{n-1},Y^{n-1}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f\left(Q_{X_i,Y_i|X^{i-1},Y^{i-1}}\right) \tag{18}$$

Note that $\widehat{H}_Q(X^{n-1},Y^{n-1})$ is a *random variable*, which is a function of the pair (X^{n-1},Y^{n-1}), not to be confused with deterministic objects such as $H(X^{n-1},Y^{n-1})$, that depend on the *distribution* of (X^{n-1},Y^{n-1}). Our goal is to prove that $\widehat{H}_Q$ is a good estimator of $\mathrm{H}(\mathbf{Y}||\mathbf{X})$ if Q is a universal sequential probability assignment.

Theorem 1 *Let Q be a universal sequential probability assignment on on $(\mathcal{X}\times\mathcal{Y})^\infty$. Then*

$$\lim_{n\to\infty}\widehat{H}_Q(X^{n-1},Y^{n-1}) = \mathrm{H}(\mathbf{Y}||\mathbf{X}) \text{ in } L_1. \tag{19}$$

Proof: Fix an arbitrary $\epsilon > 0$.

$$\begin{aligned}
&\mathsf{E}\left|\widehat{H}_Q(X^{n-1},Y^{n-1}) - \frac{1}{n}\sum_{k=1}^{n} f\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}\right)\right| \\
&= \mathsf{E}\left|\frac{1}{n}\sum_{k=1}^{n}\left(f\left(Q_{X_k,Y_k|X^{k-1},Y^{k-1}}\right) - f\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}\right)\right)\right| \\
&\le \frac{1}{n}\mathsf{E}\sum_{k=1}^{n}\left|f\left(Q_{X_k,Y_k|X^{k-1},Y^{k-1}}\right) - f\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}\right)\right| \\
&\overset{(a)}{\le} \frac{1}{n}\sum_{k=1}^{n}\mathsf{E}\left(\epsilon + K_\epsilon\left\|Q_{X_k,Y_k|X^{k-1},Y^{k-1}} - P_{X_k,Y_k|X^{k-1},Y^{k-1}}\right\|_1\right) \\
&\overset{(b)}{\le} \frac{K_\epsilon}{n}\sum_{k=1}^{n}\mathsf{E}\left[\sqrt{\frac{1}{2}D\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}||Q_{X_k,Y_k|X^{k-1},Y^{k-1}}\right)}\right] + \epsilon \\
&\overset{(c)}{\le} \frac{K_\epsilon}{n}\sum_{k=1}^{n}\sqrt{\frac{1}{2}\mathsf{E}\left[D\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}||Q_{X_k,Y_k|X^{k-1},Y^{k-1}}\right)\right]} + \epsilon \\
&= \epsilon + \frac{K_\epsilon}{n}\sum_{k=1}^{n} \\
&\qquad \sqrt{\frac{1}{2}D\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}||Q_{X_k,Y_k|X^{k-1},Y^{k-1}}\middle|X^{k-1},Y^{k-1}\right)} \\
&\overset{(d)}{\le} \epsilon + K_\epsilon\sqrt{\frac{1}{2n}}\times \\
&\qquad \sqrt{\sum_{k=1}^{n} D\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}||Q_{X_k,Y_k|X^{k-1},Y^{k-1}}\middle|X^{k-1},Y^{k-1}\right)} \\
&\overset{(e)}{=} \epsilon + K_\epsilon\sqrt{\frac{1}{2n}D\left(P_{X^n,Y^n}||Q_{X^n,Y^n}\right)}
\end{aligned} \tag{20}$$

where

- (a) comes from Lemma 1,
- (b) is due to Pinsker's inequality. Note that both $P_{X_k,Y_k|X^{k-1},Y^{k-1}}$ and $Q_{X_k,Y_k|X^{k-1},Y^{k-1}}$ are functions of (X^{k-1},Y^{k-1}). Thus $D\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}||Q_{X_k,Y_k|X^{k-1},Y^{k-1}}\right)$ is a random variable,
- (c) and (d) come from the concavity of $\sqrt{\cdot}$,
- (e) is because of the chain rule of the Kullback-Leibler divergence.

$$\begin{aligned}
&\limsup_{n\to\infty}\mathsf{E}\left|\widehat{H}_Q(X^{n-1},Y^{n-1}) - \mathrm{H}(\mathbf{Y}||\mathbf{X})\right| \\
&\le \limsup_{n\to\infty}\mathsf{E}\left|\widehat{H}_Q(X^{n-1},Y^{n-1}) - \frac{1}{n}\sum_{k=1}^{n} f\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}\right)\right| \\
&\quad + \limsup_{n\to\infty}\mathsf{E}\left|\frac{1}{n}\sum_{k=1}^{n} f\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}\right) - \mathrm{H}(\mathbf{Y}||\mathbf{X})\right| \\
&\overset{(f)}{=} \limsup_{n\to\infty}\mathsf{E}\left|\widehat{H}_Q(X^{n-1},Y^{n-1}) - \frac{1}{n}\sum_{k=1}^{n} f\left(P_{X_k,Y_k|X^{k-1},Y^{k-1}}\right)\right| \\
&\overset{(g)}{\le} \epsilon + \limsup_{n\to\infty} K_\epsilon\sqrt{\frac{1}{2n}D\left(P_{X^n,Y^n}||Q_{X^n,Y^n}\right)} \\
&\overset{(h)}{=} \epsilon
\end{aligned} \tag{21}$$

where (f) is because of Corollary 1; (g) comes from (20); (h) is due to Definition 1. Now we can use the arbitrariness of ϵ to complete the proof. ∎

Set $\mathbf{X} = \emptyset$. If $Q_{\mathbf{Y}}$ is a universal probability assignment for $\mathcal{Y}^\infty$, we have

$$\lim_{n\to\infty}\widehat{H}_{Q_{\mathbf{Y}}}(Y^{n-1}) = \mathrm{H}(\mathbf{Y}) \text{ in } L_1. \tag{22}$$

To sum up, provided with universal probability assignments $Q_{\mathbf{Y}}$ and $Q_{\mathbf{X},\mathbf{Y}}$ on $\mathcal{Y}^\infty$ and $(\mathcal{X},\mathcal{Y})^\infty$, respectively, we can construct an estimate of $\mathrm{I}(\mathbf{X}\to\mathbf{Y})$ based on a realization of $(\mathbf{X},\mathbf{Y})$, which converges to the true value in L_1. As we have shown in the previous section, a universal source coding scheme induces a universal probability assignment. Thus, using two universal compressors on $\mathbf{Y}$ and $(\mathbf{X},\mathbf{Y})$, respectively (Actually, it can be the same source coding method. Just apply it to different alphabets.), we can have a good estimate of the directed information rate from $\mathbf{X}$ to $\mathbf{Y}$.

IV. Algorithm and Numerical examples

Given a source realization x^n, y^n and universal source code $\mathcal{C}_n^Y$ on $\mathbf{Y}$ and $\mathcal{C}_n^{X,Y}$ on $(\mathbf{X},\mathbf{Y})$, we use the sequential probability assignments $Q_{\mathbf{Y}}$ and $Q_{\mathbf{X},\mathbf{Y}}$ induced by $\mathcal{C}_n^Y$ and $\mathcal{C}_n^{X,Y}$ to calculate $\widehat{H}_{Q_{\mathbf{Y}}}(y^{n-1})$ and $\widehat{H}_{Q_{\mathbf{X},\mathbf{Y}}}(x^{n-1},y^{n-1})$. The first one is our estimate of $\mathrm{H}(\mathbf{Y})$ and the second one is our estimate of $\mathrm{H}(\mathbf{Y}||\mathbf{X})$. The difference of the two gives the estimate of $\mathrm{I}(\mathbf{X}\to\mathbf{Y})$. We have proved that when n is large, our estimates will be close to the true value (in the L_1 sense.).

Although our theorem holds true for any universal source coding scheme, in practice, a scheme with low complexity and fast rates of convergence is preferred. We emphasize here that the computation of the sequential probability assignment

(11) induced by an arbitrary source code is computationally non-trivial in general.

In our implementation, we choose the context tree weighting as our universal source coding scheme. One advantage is that the complexity of the algorithm is linear in the block length n and the algorithm provides the sequential probability assignments $\{Q_{X_i|x^{i-1}}\}_{i=1}^n$ directly [16]. Note that while the original context tree weighting was tuned for binary sequence, it has been extended for larger alphabet [17]. As a first step experiment, we assume a depth D of the context tree, which is larger than the memory of the source. This short coming can be overcome by the method introduced in [18], although we did not implement it here. Our algorithm of estimating $\mathrm{H}(\mathbf{Y}||\mathbf{X})$ is explained in Algorithm 1. $\mathrm{H}(\mathbf{Y})$ can be obtained similarly as we have discussed.

Algorithm 1 Universal estimation algorithm based on context tree weighting

Fix block length n and context tree depth D.
$\hat{H} \leftarrow 0$
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 $i \leftarrow D+1, n$ **do**
  Gather the context z_{i-D}^{i-1} for the ith symbol z_i.
  Update the context tree. The estimated probability $\hat{P}_{Z_i|z^{i-1}}(\cdot)$ is obtained along the way.
  Update $\hat{H}$ as $\hat{H} \leftarrow \hat{H} + f\left(\hat{P}_{X_i,Y_i|y^{i-1},x^{i-1}}\right)$.
end for
$\hat{H} \leftarrow \frac{1}{n}\hat{H}$
$\hat{H}$ is the estimation of $\mathrm{H}(\mathbf{Y}||\mathbf{X})$.

A. Stationary process passing through a DMC

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

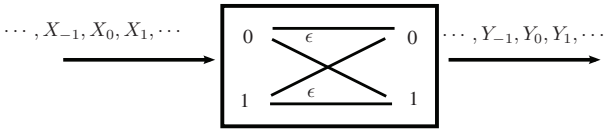

Fig. 1. Example 1 setup: $\mathbf{X}$ is a binary first order Markov process.

We run our universal estimation algorithm to estimate the directed information rate $\mathrm{I}(\mathbf{Y} \to \mathbf{X})$ with $p = 0.1$ and $\epsilon = 0.4$. The results are shown in Fig. 2. The depth of the context tree is set to be 3. As data length grows, the estimated vale is getting closer to the true value.

B. Presence of shifts

We use the same setup in Section IV-A. Instead of observing X^n directly, we assume that a shifted version of X^n is actually

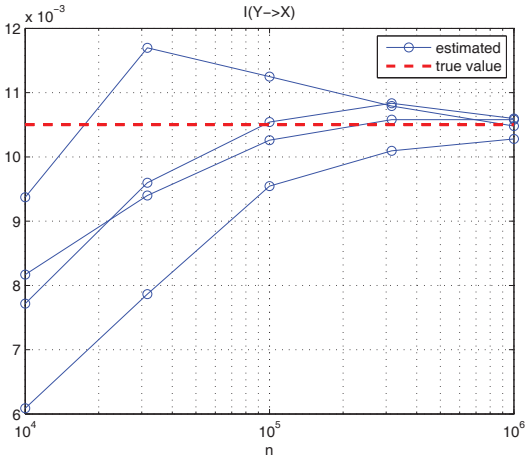

Fig. 2. $\mathbf{X}$ is a binary first order Markov process with transition probability 0.1. $\mathbf{Y}$ is the result of $\mathbf{X}$ passing through a binary symmetric channel with parameter 0.4. The simulation was performed four times, each with data length $10^4, 10^{4.5}, 10^5, 10^{5.5}, 10^6$ and context tree depth 3.

observed, i.e. $Z_i = X_{i+d}$, where d is the unit of shifts. We use our algorithm to estimate $H(\mathbf{Y}||\mathbf{Z})$ for different d values. Note that $H(Y_i|Y^{i-1}, Z^i) = H(Y_i|X_i)$ for $d \geq 0$. For $d < 0$, $H(Y_i|Y^{i-1}, Z^i)$ increases with d for large i.

We run the algorithm for $p = 0.2$ and $\epsilon = 0.4$ with block length 200000 and context tree depth 3. Note that $H(Y_i|X_i) = H_2(0.2)$ when $i \to \infty$. In Fig. 3, the simulation results is plotted for different values of d.

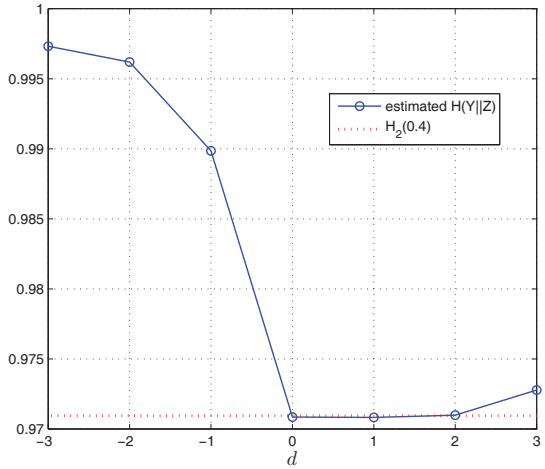

Fig. 3. $\mathbf{X}$ is a binary first order Markov process with transition probability 0.2. $\mathbf{Y}$ is the result of $\mathbf{X}$ passing through a binary symmetric channel with parameter 0.4. $\mathbf{Z}$ is a shifted version of $\mathbf{X}$ with d units of shift. The algorithm estimates $\mathrm{H}(\mathbf{Y}||\mathbf{Z})$ with $n = 200000$ and context tree depth 6.

C. Detection of the presence of feedback.

The following example is used by Massey to demonstrate the difference between mutual information (MI) and directed

information (DI) when feedback is present in a discrete memoryless channel. The input of a binary symmetric channel is the delayed output of the channel, Fig. 4. The mutual information

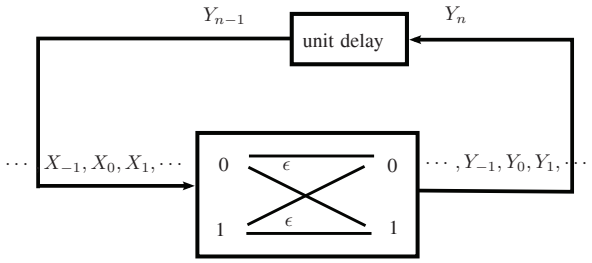

Fig. 4. Example 2 setup: The output of a binary symmetric channel is used as the input to the channel with unit delay.

rate and directed information can be easily computed in the following way:

$$\begin{aligned}\mathrm{I}(\mathbf{X};\mathbf{Y}) &= \lim_{n\to\infty}\frac{1}{n}I(X^n;Y^n)\\ &= \lim_{n\to\infty}\frac{1}{n}H(Y^n) - \lim_{n\to\infty}\frac{1}{n}H(Y^n|X^n)\\ &= \lim_{n\to\infty}\frac{1}{n}H(Y^n) - \lim_{n\to\infty}\frac{1}{n}H(Y_n|X^n)\\ &= H_2(\epsilon), \end{aligned} \tag{23}$$

where the second last equality is because $X_i = Y_{i-1}$; the last equality comes from the fact $H(Y_n|X^n)$ is bounded by 1. $H_2(t) \triangleq -t\log_2 t - (1-t)\log_2(1-t)$. Similarly,

$$\begin{aligned}&\mathrm{I}(\mathbf{X}\to\mathbf{Y})\\ &= \lim_{n\to\infty}\frac{1}{n}I(X^n\to Y^n)\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}H(Y_i|Y^{i-1}) - \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}H(Y_i|Y^{i-1},X^i)\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}H(Y_i|Y_{i-1}) - \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}H(Y_i|Y_{i-1})\\ &= 0 \end{aligned} \tag{24}$$

In this setup, the difference between the estimated MI rate and DI rate can be used as an indicator of the presence of feedback. For different ϵ value, we estimate the MI rate and DI rate for data size $n = 10^5$, shown in Fig. 5.

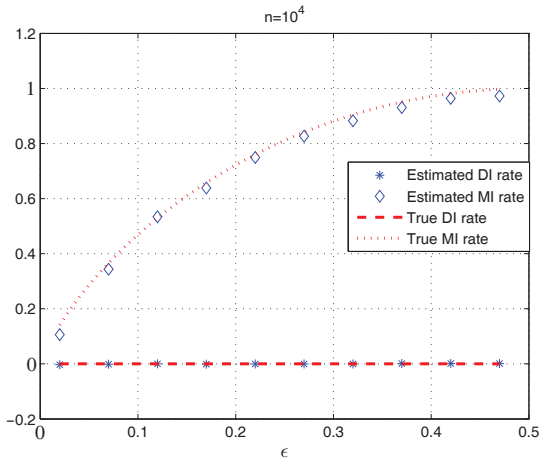

Fig. 5. Estimation of MI rate and DI rate when feedback is present.

REFERENCES

[1] J. Massey. "Causality, feedback and directed information",*Proc. Int. Symp. Inf. Theory Applic. (ISITA-90)*, pp. 303–305, Nov. 1990.

[2] G. Kramer. *Directed information for channels with feedback*. Ph.D. dissertation, Swiss Federal Institute of Technology (ETH) Zurich, 1998.

[3] G. Kramer. Capacity results for the discrete memoryless network. *IEEE Trans. Inf. Theory*, IT-49:4–21, 2003.

[4] Y.-H. Kim. A coding theorem for a class of stationary channels with feedback. *IEEE Trans. Inf. Theory.*, 25:1488–1499, April, 2008.

[5] S. Tatikonda and S. Mitter. The capacity of channels with feedback. *IEEE Trans. Inf. Theory*, 55:323–349, 2009.

[6] H. H. Permuter, T. Weissman, and A. J. Goldsmith. Finite state channels with time-invariant deterministic feedback. *IEEE Trans. Inf. Theory*, 55(2):644–662, 2009.

[7] H. Permuter, Y. H. Kim, and T. Weissman, "Interpretations of Directed Information in Portfolio Theory, Data Compression, and Hypothesis Testing", *IEEE Trans. on Inf. Th.*, submitted, `http://arxiv.org/abs/0912.4872`

[8] P. Mathai, N. Martins, B. Shapiro "On the Detection of Gene Network Interconnections using Directed Mutual Information", ITA 2007.

[9] A. Rao, A. O. Hero III, D. J. States, and J. D. Engel, "Using Directed Information to Build Biologically Relevant Influence Networks," *Journal on Bioinformatics and Computational Biology*, vol. 6, no.3, pp. 493-519, June 2008.

[10] C. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings", *Journal of Computational Neuroscience: Special Issue on Methods of Information Theory in Computational Neuroscience*, submitted Dec 15, 2009,

[11] J. Massey and P.C. Massey. Conservation of mutual and directed information. *Proc. Int. Symp. Information Theory (ISIT-05)*, pages 157–158, 2005.

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New-York, 2nd edition, 2006.

[13] Leo Breiman, *Probability*, SIAM: Society for Industrial and Applied Mathematics, May, 1992

[14] M. Weinberger, N. Merhav, and M. Feder "Optimal Sequential Probabilty Assignment for Individual Sequences",*IEEE Trans. on Inf. Th.*, Vol. 40, No. 2, March 1994, pp 384 –pp396.

[15] R. Durrett, *Probability: Theory and Example*, Duxbury Press, 3 edition, 2004.

[16] F. Willems, Y. Shtarkov, and T. Tjalken, "The Context-Tree Weighting Method: Basic Properties",*IEEE Trans. on Inf. Th.*. Vol. 41, No. 3, May 1995, pp 653 – 664.

[17] Tj. J. Tjalkens, Y. M. Shtarkov, and F. M. J. Willems, "Sequential Weighting Algorithms for Multi-Alphabet Sources," *6th Joint Swedish-Russian International Workshop on Information Theory*, Molle, Sweden, 1993, pp. 230 – 234.

[18] F. Willems, "The Context-Tree Weighting Method: Extensions",*IEEE Trans. on Inf. Th.*. Vol. 44, No. 2, Mar. 1998, pp 792 – 798.