Interpretations of Directed Information in Portfolio Theory, Data Compression, and Hypothesis Testing

Haim H. Permuter, *Member, IEEE*, Young-Han Kim, *Member, IEEE*, and Tsachy Weissman, *Senior Member, IEEE*

***Abstract*—We investigate the role of directed information in portfolio theory, data compression, and statistics with causality constraints. In particular, we show that directed information is an upper bound on the increment in growth rates of optimal portfolios in a stock market due to causal side information. This upper bound is tight for gambling in a horse race, which is an extreme case of stock markets. Directed information also characterizes the value of causal side information in instantaneous compression and quantifies the benefit of causal inference in joint compression of two stochastic processes. In hypothesis testing, directed information evaluates the best error exponent for testing whether a random process Y causally influences another process X or not. These results lead to a natural interpretation of directed information $I(Y^n \to X^n)$ as the amount of information that a random sequence $Y^n = (Y_1, Y_2, \ldots, Y_n)$ causally provides about another random sequence $X^n = (X_1, X_2, \ldots, X_n)$. A new measure, *directed lautum information*, is also introduced and interpreted in portfolio theory, data compression, and hypothesis testing.**

***Index Terms*—Causal conditioning, causal side information, directed information, hypothesis testing, instantaneous compression, Kelly gambling, lautum information, portfolio theory.**

I. Introduction

MUTUAL information $I(X;Y)$ between two random variables X and Y arises as the canonical answer to a variety of questions in science and engineering. Most notably, Shannon [1] showed that the capacity C, the maximal data rate for reliable communication, of a discrete memoryless channel $p(y\,|\,x)$ with input X and output Y is given by

$$C = \max_{p(x)} I(X;Y). \tag{1}$$

Shannon's channel coding theorem leads naturally to the operational interpretation of mutual information $I(X;Y) = H(X) - H(X\,|\,Y)$ as the amount of uncertainty about X that can be reduced by observing Y, or equivalently, the amount of information that Y can provide about X. Indeed, mutual information $I(X;Y)$ plays a central role in Shannon's random coding argument, because the probability that independently drawn sequences $X^n = (X_1, X_2, \ldots, X_n)$ and $Y^n = (Y_1, Y_2, \ldots, Y_n)$ "look" as if they were drawn jointly decays exponentially in n with $I(X;Y)$ as the first order in the exponent. Shannon also proved a dual result [2] that the rate distortion function $R(D)$, the minimum compression rate to describe a source X by its reconstruction $\hat{X}$ within average distortion D, is given by $R(D) = \min_{p(\hat{x}\,|\,x)} I(X;\hat{X})$. In another duality result (the Lagrange duality this time) to (1), Gallager [3] proved the minimax redundancy theorem, connecting the redundancy of the universal lossless source code to the maximum mutual information (capacity) of the channel with conditional distribution consisting of the set of possible source distributions (cf. [4]).

It has been shown that mutual information plays important roles in problems beyond those related to describing sources or transferring information through channels. Among the most celebrated of such examples is the use of mutual information in gambling. In 1956, Kelly [5] showed that if a horse race outcome can be represented as an independent and identically distributed (i.i.d.) random variable X, and the gambler has some side information Y relevant to the outcome of the race, then the mutual information $I(X;Y)$ captures the difference between growth rates of the optimal gambler's wealth with and without side information Y. Thus, Kelly's result provides an interpretation of mutual information $I(X;Y)$ as the financial value of side information Y for gambling in the horse race X.

In order to tackle problems arising in information systems with causally dependent components, Massey [6], inspired by Marko's work [7] on bidirectional communication, coined the notion of "directed information" from X^n to Y^n, defined as

$$I(X^n \to Y^n) := \sum_{i=1}^{n} I(X^i; Y_i\,|\,Y^{i-1}) \tag{2}$$

and showed that the normalized maximum directed information upper bounds the capacity of channels with feedback. Subsequently, it was shown that directed information, as defined by Massey, indeed characterizes the capacity of channels with feedback [8]–[16] and the rate distortion function with feedforward [17]. Note that directed information (2) can also be rewritten as

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

each term of which corresponds to the achievable rate at time i given side information (X^{i-1}, Y^{i-1}) (cf. [11] for the details). Recently, directed information has also been used in models of computational biology [18]–[21] and in the context linear prediction representation for the rate distortion function of a stationary Gaussian source [22].

Manuscript received December 24, 2009; revised November 24, 2010; accepted November 27, 2010. Date of current version May 25, 2011. This work was supported in part by NSF Grant CCF-0729195 and in part by BSF Grant 2008402. H. H. Permuter was supported in part by the Marie Curie Reintegration fellowship.

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).

Y.-H. Kim is with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093-0407 USA (e-mail: yhk@ucsd.edu).

T. Weissman is with the Information Systems Lab, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.ed).

Communicated by I. Kontoyiannis, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2011.2136270

The main contribution of this paper is showing that directed information has natural interpretations in portfolio theory, compression, and statistics when causality constraints exist. In stock market investment (Section III), directed information between the stock price X and side information Y is an upper bound on the increase in growth rates due to *causal* side information. This upper bound is tight when specialized to gambling in horse races. In data compression (Section IV), we show that directed information characterizes the value of causal side information in instantaneous compression and it also quantifies the role of causal inference in joint compression of two stochastic processes. In hypothesis testing (Section V), we show that directed information is the exponent of the minimum type II error probability when one is to decide whether Y_i has causal influence on X_i or not. Finally, we introduce the notion of directed lautum[1] information (Section VI), which extends the notion of lautum information by Palomar and Verdú [23] to accommodate causality. We briefly discuss its role in horse race gambling, data compression, and hypothesis testing.

II. Preliminaries: Directed Information and Causal Conditioning

Throughout this paper, we use the *causal conditioning* notation $(\cdot\|\cdot)$ developed by Kramer [8]. We denote by $p(x^n \,\|\, y^{n-d})$ the probability mass function of $X^n = (X_1, \ldots, X_n)$ *causally conditioned* on Y^{n-d} for some integer $d \geq 0$, which is defined as

$$p(x^n \,\|\, y^{n-d}) := \prod_{i=1}^{n} p(x_i \,|\, x^{i-1}, y^{i-d}). \tag{4}$$

By convention, if $i < d$, then y^{i-d} is set to null, i.e., if $i < d$ then $p(x_i \,|\, x^{i-1}, y^{i-d})$ is just $p(x_i \,|\, x^{i-1})$. In particular, we use extensively the cases $d = 0, 1$

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

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

Using the chain rule, it is easily verified that

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

The *causally conditional* entropy $H(X^n \,\|\, Y^n)$ and $H(X^n \,\|\, Y^{n-1})$ are defined respectively as

$$\begin{aligned} H(X^n \,\|\, Y^n) &:= E\left[\log \frac{1}{p(X^n \,\|\, Y^n)}\right] \\ &= \sum_{i=1}^{n} H(X_i \,|\, X^{i-1}, Y^i) \\ H(X^n \,\|\, Y^{n-1}) &:= E\left[\log \frac{1}{p(X^n \,\|\, Y^{n-1})}\right] \\ &= \sum_{i=1}^{n} H(X_i \,|\, X^{i-1}, Y^{i-1}). \end{aligned} \tag{8}$$

[1]*Lautum* ("elegant" in Latin) is the reverse spelling of "mutual" as aptly coined in [23].

Under this notation, directed information defined in (2) can be rewritten as

$$I(X^n \to Y^n) = H(Y^n) - H(Y^n \,\|\, X^n) \tag{9}$$

which hints, in a rough analogy to mutual information, a possible interpretation of directed information $I(Y^n \to X^n) = H(X^n) - H(X^n \,\|\, Y^n)$ as the amount of information that causally available side information Y^n can provide about X^n.

Note that the channel capacity results involve the term $I(X^n \to Y^n)$, which measures the amount of information transfer over the forward link from X^n to Y^n. In gambling, however, the increase in growth rate is due to the side information (the backward link), and, therefore, the expression $I(Y^n \to X^n)$ appears. Throughout the paper we also use the notation $I(Y^{n-1} \to X^n)$ which denotes the directed information between the n-tuple $(\emptyset, Y^{n-1})$, i.e., the null symbol followed by Y^{n-1}, and X^n, that is,

$$I(Y^{n-1} \to X^n) = \sum_{i=2}^{n} I(Y^{i-1}; X_i \,|\, X^{i-1}). \tag{10}$$

Using the causal conditioning notation, given in (8), the directed information $I(Y^{n-1} \to X^n)$ can be written as

$$I(Y^{n-1} \to X^n) = H(X^n) - H(X^n \,\|\, Y^{n-1}). \tag{11}$$

Directed information and mutual information obey the conservation law

$$I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n) \tag{12}$$

which was established by Massey and Massey [24]. The conservation law is a direct consequence of the chain rule (7), and we show later in Section IV-B that it has a natural interpretation as a conservation of a mismatch cost in data compression.

The causally conditional entropy rate of a random process X given another random process Y and the directed information rate from X to Y are defined respectively as

$$\mathcal{H}(X \,\|\, Y) := \lim_{n\to\infty} \frac{H(X^n \,\|\, Y^n)}{n} \tag{13}$$

$$\mathcal{I}(X \to Y) := \lim_{n\to\infty} \frac{I(X^n \to Y^n)}{n} \tag{14}$$

when these limits exist. In particular, when (X, Y) is stationary ergodic, both quantities are well-defined, namely, the limits in (13) and (14) exist [8, Properties 3.5 and 3.6].

III. Portfolio Theory

Here we show that directed information is an upper bound on the increment in growth rates of optimal portfolios in a stock market due to causal side information. We start by considering a special case of gambling in a horse race market and show that the upper bound is tight. Then we consider the general case of stock market investment.

A. Horse Race Gambling With Causal Side Information

Assume that there are m racing horses and let X_i denote the winning horse at time i, i.e., $X_i \in \mathcal{X} := \{1, 2, \ldots, m\}$. At time

i, the gambler has some side information relevant to the performance of the horses, which we denote as Y_i. For instance, a possible side information may be the health condition of the horses and the jockeys on the day of the race. We assume that the gambler invests all his/her capital in the horse race as a function of the previous horse race outcomes X^{i-1} and side information Y^i up to time i. Let $b(x_i \,|\, x^{i-1}, y^i)$ be the portion of wealth that the gambler bets on horse x_i given $X^{i-1} = x^{i-1}$ and $Y^i = y^i$. Obviously, the gambling scheme (without short or margin) should satisfy $b(x_i \,|\, x^{i-1}, y^i) \geq 0$ and $\sum_{x_i \in \mathcal{X}} b(x_i \,|\, x^{i-1}, y^i) = 1$ for any history (x^{i-1}, y^i). Let $\alpha(x_i \,|\, x^{i-1})$ denote the odds of horse x_i given the previous outcomes x^{i-1}, which is the amount of capital that the gambler receives for each unit capital that the gambler invested in the horse. We denote by $S(x^n \,\|\, y^n)$ the gambler's wealth after n races with outcomes x^n and causal side information y^n. Finally, $\frac{1}{n} W(X^n \,\|\, Y^n)$ denotes the *growth rate* of wealth, where the growth $W(X^n \,\|\, Y^n)$ is defined as the expectation over the logarithm (base 2) of the gambler's wealth, i.e.,

$$W(X^n \,\|\, Y^n) := E[\log S(X^n \,\|\, Y^n)]. \tag{15}$$

Without loss of generality, we assume that the gambler's initial wealth S_0 is 1. We assume that at any time n the gambler invests all his/her capital, and, therefore, we have

$$\begin{aligned} &S(X^n \,\|\, Y^n) \\ &\quad = b(X_n \,|\, X^{n-1}, Y^n)\alpha(X_n \,|\, X^{n-1}) S(X^{n-1} \,\|\, Y^{n-1}). \end{aligned} \tag{16}$$

By simple recursion, this implies that

$$S(X^n \,\|\, Y^n) = \prod_{i=1}^{n} b(X_i \,|\, X^{i-1}, Y^i)\alpha(X_i \,|\, X^{i-1}). \tag{17}$$

Let $W^*(X^n \,\|\, Y^n)$ denote the maximum growth, i.e.,

$$W^*(X^n \,\|\, Y^n) := \max_{\{b(x_i \,|\, x^{i-1}, y^{i-1})\}_{i=1}^n} W(X^n \,\|\, Y^n). \tag{18}$$

The following theorem characterizes the investment strategy that maximizes the growth.

Theorem 1 (Optimal Causal Gambling): For any finite horizon n, the maximum growth is

$$W^*(X^n \,\|\, Y^n) = E[\log \alpha(X^n)] - H(X^n \,\|\, Y^n) \tag{19}$$

which is achieved when the gambler invests money proportional to the causally conditional distribution of the horse race outcome, i.e.,

$$\begin{aligned} b(x_i \,|\, x^{i-1}, y^i) &= p(x_i \,|\, x^{i-1}, y^i) \\ &\forall i \in \{1, \ldots, n\}, x^i \in \mathcal{X}^i, y^i \in \mathcal{Y}^i. \end{aligned} \tag{20}$$

Note that since $\{p(x_i \,|\, x^{i-1}, y^i)\}_{i=1}^n$ uniquely determines $p(x^n \,\|\, y^n)$, and since $\{b(x_i \,|\, x^{i-1}, y^i)\}_{i=1}^n$ uniquely determines $b(x^n \,\|\, y^n)$, then (20) is equivalent to

$$b(x^n \,\|\, y^n) \equiv p(x^n \,\|\, y^n). \tag{21}$$

Furthermore, note that the best strategy is greedy, namely, at any time i the best strategy is to maximize $W(X^i \,\|\, Y^i)$ regardless of the horizon n. In other words, at any time i the gambler should maximize the expected growth rate, i.e., $E[\log S(X^i \,\|\, Y^i)]$.

Proof: Consider

$$\begin{aligned} &W^*(X^n \,\|\, Y^n) \\ &\quad = \max_{b(x^n \,\|\, y^n)} E[\log b(X^n \,\|\, Y^n)\alpha(X^n)] \\ &\quad = \max_{b(x^n \,\|\, y^n)} E[\log b(X^n \,\|\, Y^n)] + E[\log \alpha(X^n)] \\ &\quad = -H(X^n \,\|\, Y^n) + E[\log \alpha(X^n)]. \end{aligned} \tag{22}$$

Here the last equality is achieved by choosing $b(x^n \,\|\, y^n) = p(x^n \,\|\, y^n)$, and it is justified by the following upper bound:

$$\begin{aligned} &E[\log b(X^n \,\|\, Y^n)] \\ &\quad = \sum_{x^n, y^n} p(x^n, y^n) \left[\log p(x^n \,\|\, y^n) + \log \frac{b(x^n \,\|\, y^n)}{p(x^n \,\|\, y^n)} \right] \\ &\quad = -H(X^n \,\|\, Y^n) + \sum_{x^n, y^n} p(x^n, y^n) \log \frac{b(x^n \,\|\, y^n)}{p(x^n \,\|\, y^n)} \\ &\quad \overset{(a)}{\leq} -H(X^n \,\|\, Y^n) + \log \sum_{x^n, y^n} p(x^n, y^n) \frac{b(x^n \,\|\, y^n)}{p(x^n \,\|\, y^n)} \\ &\quad \overset{(b)}{=} -H(X^n \,\|\, Y^n) + \log \sum_{x^n, y^n} p(y^n \,\|\, x^{n-1}) b(x^n \,\|\, y^n) \\ &\quad \overset{(c)}{=} -H(X^n \,\|\, Y^n). \end{aligned} \tag{23}$$

where (a) follows from Jensen's inequality, (b) from the chain rule, and (c) from the fact that $\sum_{x^n, y^n} p(y^n \,\|\, x^{n-1}) b(x^n \,\|\, y^n) = 1$. ∎

In case that the odds are fair, i.e., $\alpha(X_i \,|\, X^{i-1}) = 1/m$,

$$W^*(X^n \,\|\, Y^n) = n \log m - H(X^n \,\|\, Y^n) \tag{24}$$

and thus, the sum of the growth rate and the entropy of the horse race process conditioned causally on the side information is constant, and one can see a duality between $H(X^n \,\|\, Y^n)$ and $W^*(X^n \,\|\, Y^n)$.

Let us define $\Delta W(X^n \,\|\, Y^n)$ as the increase in the growth due to the causal side information, i.e.,

$$\Delta W(X^n \,\|\, Y^n) = W^*(X^n \,\|\, Y^n) - W^*(X^n) \tag{25}$$

where $W^*(X^n)$ denotes the maximum growth when side information is not available.

Corollary 1 (Increase in the Growth Rate): The increase in growth rate due to a causal side information sequence Y^n for a horse race sequence X^n is

$$\frac{1}{n} \Delta W(X^n \,\|\, Y^n) = \frac{1}{n} I(Y^n \to X^n). \tag{26}$$

As a special case, if the horse race outcome and side information are pairwise i.i.d., then the (normalized) directed information $\frac{1}{n} I(Y^n \to X^n)$ becomes the single-letter mutual information $I(X; Y)$, which coincides with Kelly's result [5].

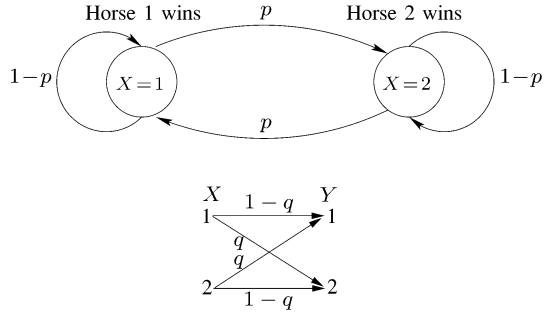

Fig. 1. Setting of Example 1. The winning horse X_i is represented as a Markov process with two states. In state 1, horse number 1 wins, and in state 2, horse number 2 wins. The side information Y_i is a noisy observation of the winning horse X_i.

Proof: From the definition of directed information (9) and Theorem 1, we obtain

$$W^*(X^n \,\|\, Y^n) - W^*(X^n) = -H(X^n \,\|\, Y^n) + H(X^n) = I(Y^n \to X^n).$$

■

Example 1 (Gambling in a Markov Horse Race Process With Causal Side Information): Consider the case in which two horses are racing, and the winning horse X_i behaves as a Markov process, as shown in Fig. 1. A horse that won will win again with probability p and lose with probability $1-p (0 \le p \le 1)$. At time zero, we assume that the two horses have equal probability of wining. The side information revealed to the gambler at time i is Y_i, which is a noisy observation of the horse race outcome X_i. It has probability $1-q$ of being equal to X_i, and probability q of being different from X_i. In other words, $Y_i = X_i + V_i \bmod 2$, where V_i is a Bernoulli(q) process independent of X_i.

For this example, the increase in growth rate due to side information is

$$\Delta W := \frac{1}{n}\Delta W(X^n \,\|\, Y^n) = h(p * q) - h(q) \qquad (27)$$

where $h(x) := -x \log x - (1-x)\log(1-x)$ is the binary entropy function, and $p * q = (1-p)q + (1-q)p$ denotes the parameter of a Bernoulli distribution that results from convolving two Bernoulli distributions with parameters p and q.

The increment of ΔW in growth rate can be readily derived using the identity in (3) as follows:

$$\begin{aligned}\frac{1}{n} I(Y^n \to X^n) &\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i; X_i^n \,|\, X^{i-1}, Y^{i-1}\right) \\ &\overset{(b)}{=} H(Y_1 \,|\, X_0) - H(Y_1 \,|\, X_1) \qquad (28)\end{aligned}$$

where equality (a) is the identity from (3), which can be easily verified by the chain rule for mutual information [11, (9)], and (b) is due to the stationarity of the process.

If the side information is known with some lookahead $k \ge 0$, meaning that at time i the gambler knows Y^{i+k}, then the optimal growth after n steps, $W^*(X^n \,\|\, Y^{n+k})$, is given by

$$W^*(X^n \,\|\, Y^{n+k}) = -H(X^n \,\|\, Y^{n+k}) + E[\log \alpha(X^n)] \qquad (29)$$

where $H(X^n \,\|\, Y^{n+k}) = \sum_{i=1}^{n} H(X_i \,|\, X^{i-1}, Y^{i+k})$. The identity (29) follows from similar steps as (22), just replacing $(Y^1, Y^2, \ldots, Y^n)$ by $(Y^{1+k}, Y^{2+k}, \ldots, Y^{n+k})$, respectively. Hence, the increase in the growth rate after n gambling rounds due to side information with lookahead k is

$$\begin{aligned}&W^*(X^n \,\|\, Y^{n+k}) - W^*(X^n) \\ &\qquad = -H(X^n \,\|\, Y^{n+k}) + H(X^n) = I(Y^{n+k} \to X^n).\end{aligned}$$

As n tends to infinity the increase in the growth rate is given by

$$\begin{aligned}\Delta W &= \lim_{n\to\infty} \frac{1}{n} I(Y^{n+k} \to X^n) \\ &\overset{(a)}{=} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(Y_{i+k}; X_i^n \,|\, X^{i-1}, Y^{i+k-1}) \\ &\overset{(b)}{=} \lim_{n\to\infty} \frac{n}{n} H(Y_{k+1} \,|\, Y^k, X_0) - \frac{n-k}{n} H(Y_1 \,|\, X_1) \\ &\quad + \frac{1}{n}\sum_{i=n-k+1}^{n} H(Y_{i+k} \,|\, X^n, Y^{i+k-1}) \\ &= H(Y_{k+1} \,|\, Y^k, X_0) - H(Y_1 \,|\, X_1) \qquad (30)\end{aligned}$$

where steps (a) and (b) follow from the same arguments as in (28). As more side information $(Y_1, Y_2, \ldots)$ becomes available to the gambler ahead of time, the increase in the optimal growth rate converges to the mutual information [5] instead of directed information. This is due to

$$\begin{aligned}&\lim_{k\to\infty} \lim_{n\to\infty} (W^*(X^n \,\|\, Y^{n+k}) - W^*(X^n)) \\ &\quad \overset{(a)}{=} \lim_{k\to\infty} H(Y_{k+1} \,|\, Y^k, X_0) - H(Y_1 \,|\, X_1) \\ &\quad \overset{(b)}{=} \lim_{k\to\infty} \frac{H(Y^k)}{k} - H(Y_1 \,|\, X_1) \\ &\quad = \lim_{k\to\infty} \frac{1}{k} I(Y^k; X^k) \qquad (31)\end{aligned}$$

where step (a) follows from (30) and step (b) follows from the fact that the sequence $H(Y_{k+1} \,|\, Y^{k-1}, X_0)$ converges to the entropy rate of the process, i.e., $\lim_{k\to\infty} H(Y_{k+1} \,|\, Y^k, X_0) = \lim_{k\to\infty} \frac{H(Y^k)}{k}$.

B. Investment in a Stock Market With Causal Side Information

We use notation similar to that in [25, ch. 16]. A stock market at time i is represented by a vector $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{im})$, where m is the number of stocks, and the *price relative* X_{ik} is the ratio of the price of stock at the end of day $i-1$ to the price of the stock at the end of day i. Note that gambling in a horse race is an extreme case of stock market investment—for horse races, the price relatives are all zero except for one stock.

We assume that at time i there is side information Y^i that is known to the investor. A *portfolio* is an allocation of wealth

across the stocks. A nonanticipating or causal portfolio strategy with causal side information at time i is denoted as $\mathbf{b}(\mathbf{x}^{i-1}, y^i)$, and it satisfies $\sum_{k=1}^{m} b_k(\mathbf{x}^{i-1}, y^i) = 1$ and $b_k(\mathbf{x}^{i-1}, y^i) \geq 0$ for all possible $(\mathbf{x}^{i-1}, y^i)$. We define $S(\mathbf{x}^n \,\|\, y^n)$ to be the wealth at the end of day n for a stock sequence $\mathbf{x}^n$ and causal side information y^n. We have

$$S(\mathbf{x}^n \,\|\, y^n) = (\mathbf{b}^t(\mathbf{x}^{n-1}, y^n) \cdot \mathbf{x}_n) S(\mathbf{x}^{n-1} \,\|\, y^{n-1}) \tag{32}$$

where $\mathbf{b}^t \cdot \mathbf{x}$ denotes inner product between the two (column) vectors $\mathbf{b}$ and $\mathbf{x}$. The goal is to maximize the growth

$$W(\mathbf{X}^n \,\|\, Y^n) := E[\log S(\mathbf{X}^n \,\|\, Y^n)]. \tag{33}$$

The justification for maximizing the growth rate is due to [26, Theorem 5]; such a portfolio strategy will exceed the wealth of any other strategy to the first order in the exponent for almost every sequence of outcomes from the stock market, namely, if $S^*(\mathbf{X}^n \,\|\, Y^n)$ is the wealth corresponding to the growth rate optimal return, then

$$\limsup_n \frac{1}{n} \log \left(\frac{S(\mathbf{X}^n \,\|\, Y^n)}{S^*(\mathbf{X}^n \,\|\, Y^n)} \right) \leq 0 \quad \text{a.s.} \tag{34}$$

Let us define

$$W(\mathbf{X}_n \,|\, \mathbf{X}^{n-1}, Y^n) := E[\log(\mathbf{b}^t(\mathbf{X}^{n-1}, Y^n)\mathbf{X}_n)]. \tag{35}$$

From this definition follows the chain rule:

$$W(\mathbf{X}^n \,\|\, Y^n) = \sum_{i=1}^{n} W(\mathbf{X}_i \,|\, \mathbf{X}^{i-1}, Y^i) \tag{36}$$

from which we obtain

$$\begin{aligned}
&\max_{\{\mathbf{b}(\mathbf{x}^{i-1}, y^i)\}_{i=1}^n} W(\mathbf{X}^n \,\|\, Y^n) \\
&\quad \overset{(a)}{=} \sum_{i=1}^{n} \max_{\mathbf{b}(\mathbf{x}^{i-1}, y^i)} W(\mathbf{X}_i \,|\, X^{i-1}, Y^i) \\
&\quad = \sum_{i=1}^{n} \max_{\mathbf{b}(\mathbf{x}^{i-1}, y^i)} \int_{\mathbf{x}^{i-1}, y^i} f(\mathbf{x}^{i-1}, y^i) W(\mathbf{X}_i \,|\, \mathbf{x}^{i-1}, y^i) \\
&\quad = \sum_{i=1}^{n} \int_{\mathbf{x}^{i-1}, y^i} f(\mathbf{x}^{i-1}, y^i) \max_{\mathbf{b}(\mathbf{x}^{i-1}, y^i)} W(\mathbf{X}_i \,|\, \mathbf{x}^{i-1}, y^i)
\end{aligned} \tag{37}$$

where $f(\mathbf{x}^{i-1}, y^i)$ denotes the probability density function of $(\mathbf{x}^{i-1}, y^i)$. Here step (a) follows since $\max_{\alpha,\beta} f_1(\alpha) + f_2(\beta) = \max_\alpha f_1(\alpha) + \max_\beta f_2(\beta)$, and $W(\mathbf{X}_j \,|\, X^{j-1}, Y^j)$ depends only on the jth component of the sequence $\{\mathbf{b}(\mathbf{x}^{i-1}, y^i)\}_{i=1}^n$. The maximization in (37), namely, $\max_{\mathbf{b}(\mathbf{x}^{i-1}, y^i)} W(\mathbf{X}_i \,|\, \mathbf{x}^{i-1}, y^i)$, is equivalent to the maximization of the growth rate for the memoryless case where the cumulative distribution function of the stock vector $\mathbf{X}$ is $\Pr(\mathbf{X} \leq \mathbf{x}) = \Pr(\mathbf{X}_i \leq \mathbf{x} \,|\, \mathbf{x}^{i-1}, y^i)$ and the portfolio $\mathbf{b} = \mathbf{b}(\mathbf{x}^{i-1}, y^i)$ is a function of (x^{i-1}, y^i), i.e.,

$$\begin{aligned}
&\text{maximize } E[\log(\mathbf{b}^t\mathbf{X}) \,|\, X^{i-1} = x^{i-1}, Y^i = y^i] \\
&\text{subject to } \sum_{k=1}^{m} b_k = 1 \\
&\qquad b_k \geq 0, \quad \forall k \in \{1, 2, \ldots, m\}.
\end{aligned} \tag{38}$$

The objective $E[\log(\mathbf{b}^t\mathbf{X}) \,|\, X^{i-1} = x^{i-1}, Y^i = y^i]$ is simply $W(\mathbf{X}_i \,|\, \mathbf{x}^{i-1}, y^i)$, and the constraints are due to the fact that we invest all the money without short or margin.

In order to upper bound the difference in growth rate due to causal side information, we recall the following result that bounds the loss in growth rate incurred by optimizing the portfolio with respect to a wrong distribution $g(\mathbf{x})$ rather than the true distribution $f(\mathbf{x})$.

Theorem 2 ([27, Theorem 1]): Let $f(\mathbf{x})$ be the probability density function of a stock vector $\mathbf{X}$, i.e., $\mathbf{X} \sim f(\mathbf{x})$. Let $\mathbf{b}_f$ be the growth rate portfolio corresponding to $f(\mathbf{x})$, and let $\mathbf{b}_g$ be the growth rate portfolio corresponding to another density $g(\mathbf{x})$. Then the increase ΔW in optimal growth rate due to using $\mathbf{b}_f$ instead of $\mathbf{b}_g$ is upper bounded by

$$\Delta W = E[\log(\mathbf{b}_f^t\mathbf{X})] - E[\log(\mathbf{b}_g^t\mathbf{X})] \leq D(f \,\|\, g) \tag{39}$$

where $D(f \,\|\, g) := \int f(x) \log \frac{f(x)}{g(x)} dx$ denotes the Kullback-Leibler divergence between the probability density functions f and g.

Using Theorem 2, we can upper bound the increase in growth rate due to causal side information by directed information as shown in the following theorem. This is due to the fact that directed information can be written as divergence between $p(x^n, y^n)$ and $p(x^n)p(y^n \,\|\, x^{n-1})$.

Theorem 3 (Upper Bound on Increase in Growth Rate): The increase in the optimal growth rate for a stock market sequence $\mathbf{X}^n$ due to a causal side information sequence Y^n is upper bounded as

$$W^*(\mathbf{X}^n \,\|\, Y^n) - W^*(\mathbf{X}^n) \leq I(Y^n \to \mathbf{X}^n) \tag{40}$$

where $W^*(\mathbf{X}^n \,\|\, Y^n) \triangleq \max_{\{\mathbf{b}(\mathbf{X}^{i-1}, Y^i)\}_{i=1}^n} W(\mathbf{X}^n \,\|\, Y^n)$ and $W^*(\mathbf{X}^n) := \max_{\{\mathbf{b}(\mathbf{X}^{i-1})\}_{i=1}^n} W(\mathbf{X}^n)$.

Proof: Consider

$$\begin{aligned}
&W^*(\mathbf{X}^n \,\|\, Y^n) - W^*(\mathbf{X}^n) \\
&\quad = \sum_{i=1}^{n} \int_{\mathbf{x}^{i-1}, y^i} f(\mathbf{x}^{i-1}, y^i) \\
&\qquad \times \left[\max_{\mathbf{b}(\mathbf{x}^{i-1}, y^i)} W(\mathbf{X}_i \,|\, \mathbf{x}^{i-1}, y^i) - \max_{\mathbf{b_i}(\mathbf{x}^{i-1})} W(\mathbf{X}_i \,|\, \mathbf{x}^{i-1}) \right] \\
&\quad \overset{(a)}{\leq} \sum_{i=1}^{n} \int_{\mathbf{x}^{i-1}, y^i} f(\mathbf{x}^{i-1}, y^i) \\
&\qquad \times \left[\int_{\mathbf{x}_i} f(\mathbf{x}_i \,|\, \mathbf{x}^{i-1}, y^i) \log \frac{f(\mathbf{x}_i \,|\, \mathbf{x}^{i-1}, y^i)}{f(\mathbf{x}_i \,|\, \mathbf{x}^{i-1})} \right] \\
&\quad = \sum_{i=1}^{n} E\left[\log \frac{f(\mathbf{X}_i \,|\, \mathbf{X}^{i-1}, Y^i)}{f(\mathbf{X}_i \,|\, \mathbf{X}^{i-1})} \right] \\
&\quad = \sum_{i=1}^{n} h(\mathbf{X}_i \,|\, \mathbf{X}^{i-1}) - h(\mathbf{X}_i \,|\, \mathbf{X}^{i-1}, Y^i) \\
&\quad = I(Y^n \to \mathbf{X}^n)
\end{aligned} \tag{41}$$

where the inequality (a) follows from Theorem 2. ∎

Note that the upper bound in Theorem 3 is tight for gambling in horse races (Corollary 1). Hence, we can conclude from Corollary 1 and Theorem 3 that directed information is the upper

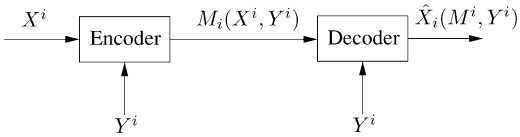

Fig. 2. Instantaneous data compression with causal side information.

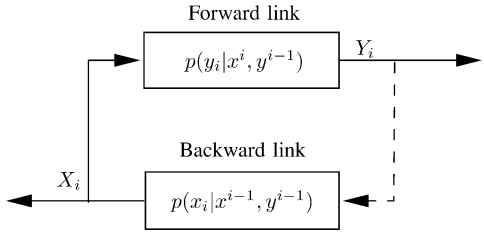

Fig. 3. Compression of two correlated sources $\{X_i, Y_i\}_{i\geq 1}$. Since any joint distribution can be decomposed as $p(x^n, y^n) = p(x^n \,\|\, y^{n-1})p(y^n \,\|\, x^n)$, each link embraces the existence of a forward or feedback channel (chemical reaction). We investigate the influence of the link knowledge on joint compression of $\{X_i, Y_i\}_{i\geq 1}$.

bound on the increase of growth rate due to causal side information, and this upper bound is achieved in a horse race market.

IV. Data Compression

In this section we investigate the role of directed information in data compression and find two interpretations:

1) Directed information characterizes the value of causal side information in instantaneous compression.
2) It also quantifies the role of causal inference in joint compression of two stochastic processes.

A. Instantaneous Lossless Compression With Causal Side Information

Let $X_1, X_2 \ldots$ be a source and $Y_1, Y_2, \ldots$ be side information about the source. The source is to be encoded losslessly by an instantaneous code with causally available side information, as depicted in Fig. 2. More formally, an *instantaneous lossless source encoder with causal side information* consists of a sequence of mappings $\{M_i\}_{i\geq 1}$ such that each $M_i : \mathcal{X}^i \times \mathcal{Y}^i \mapsto \{0,1\}^*$ has the property that for every x^{i-1} and y^i, $M_i(x^{i-1}\cdot, y^i)$ is an instantaneous (prefix) code.

An instantaneous lossless source encoder with causal side information operates sequentially, emitting the concatenated bit stream $M_1(X_1, Y_1)M_2(X^2, Y^2)\ldots$. The defining property that $M_i(x^{i-1}\cdot, y^i)$ is an instantaneous code for every x^{i-1} and y^i is a necessary and sufficient condition for the existence of a decoder that can losslessly recover x^i based on y^i and the bit stream $M_1(x_1, y_1)M_2(x^2, y^2)\ldots$ as soon as it receives $M_1(x_1, y_1)M_2(x^2, y^2)\ldots M_i(x^i, y^i)$ for all sequence pairs $(x_1, y_1), (x_2, y_2), \ldots$, and all $i \geq 1$. Let $\ell(x^n \,\|\, y^n)$ denote the length of the concatenated string $M_1(x_1, y_1)M_2(x^2, y^2)\ldots M_n(x^n, y^n)$. Then the following result is established by using Kraft's inequality and adapting Huffman coding to the case where causal side information is available.

Theorem 4 (Lossless Source Coding With Causal Side Information): Any instantaneous lossless source encoder with causal side information satisfies

$$\frac{1}{n}E\ell(X^n \,\|\, Y^n) \geq \frac{1}{n}\sum_{i=1}^{n} H(X_i \,|\, X^{i-1}, Y^i) \quad \forall\, n \geq 1. \quad (42)$$

Conversely, there exists an instantaneous lossless source encoder with causal side information satisfying

$$\frac{1}{n}E\ell(X^n \,\|\, Y^n) \leq \frac{1}{n}\sum_{i=1}^{n} r_i + H(X_i \,|\, X^{i-1}, Y^i) \;\forall\; n \geq 1 \quad (43)$$

where $r_i = \sum_{x^{i-1}, y^i} p(x^{i-1}, y^i) \min(1, \max_{x_i} p(x_i \,|\, x^{i-1}, y^{i-1}) + 0.086)$.

Proof: The lower bound follows from Kraft's inequality [25, Theorem 5.3.1] and the upper bound follows from Huffman coding on the conditional probability $p(x_i \,|\, x^{i-1}, y^i)$. The redundancy term r_i follows from Gallager's redundancy bound [28], $\min(1, P_i + 0.086)$, where P_i is the probability of the most likely source letter at time i, averaged over side information sequence (X^{i-1}, Y^i). ∎

Since the Huffman code achieves the entropy rate for dyadic probability, it follows that if the conditional probability $p(x_i \,|\, x^{i-1}, y^{i-1})$ is dyadic, i.e., if each conditional probability equals to 2^{-k} for some integer k, then (42) can be achieved with equality.

Combined with the identity $\sum_{j=i}^{n} H(X_i \,|\, X^{i-1}, Y^i) = H(X^n) - I(Y^n \to X^n)$, Theorem 4 implies that the compression rate saved in optimal sequential lossless compression due to the causal side information is upper bounded by $\frac{1}{n}I(Y^n \to X^n) + 1$, and lower bounded by $\frac{1}{n}I(Y^n \to X^n) - 1$. If all the probabilities are dyadic, then the compression rate saving is exactly equal to the directed information rate $\frac{1}{n}I(Y^n \to X^n)$. This saving should be compared to $\frac{1}{n}I(X^n; Y^n)$, which is the saving in the absence of causality constraint.

B. Cost of Mismatch in Data Compression

It is well known [25, Thm. 5.4.3] that if we design an optimal lossless compression code according to $p(x)p(y)$ where the actual distribution is $p(x, y)$, then an additional rate that is needed (redundancy) is equal to $I(X; Y)$. In this section we quantify the value of knowing causal influence between two processes when designing the optimal joint compression.

Suppose we compress a pair of correlated sources $\{(X_i, Y_i)\}$ jointly with an optimal lossless variable length code (such as the Huffman code), and we denote by $E(\ell(X^n, Y^n))$ the average length of the optimal code. Assume further that Y_i is generated randomly by a forward link $p(y_i \,|\, y^{i-1}, x^i)$ as in a communication channel or a chemical reaction, and X_i is generated by a backward link $p(x_i \,|\, y^{i-1}, x^{i-1})$ such as in the case of an encoder or a controller with feedback. By the chain rule (7), for causally conditional probabilities (7), any joint distribution can be modeled according to Fig. 3.

Recall that the optimal variable-length lossless code, in which both links are taken into account, has the average length [25, Ch 5.4]

$$H(X^n,Y^n) \le E(\ell(X^n,Y^n)) < H(X^n,Y^n)+1.$$

What happens when the compression scheme design fails to take the forward link into account?

Lemma 1: If a lossless code is erroneously designed to be optimal for the case in which the forward link does not exist, namely, the code is designed for the joint distribution $p(y^n)p(x^n \,\|\, y^{n-1})$, then the redundancy (up to $\frac{1}{n}$ bit) is $\frac{1}{n}I(X^n \to Y^n)$.

Proof: Let $E(\ell_b(X^n,Y^n))$ be the average code designed optimally for the joint distribution $p(y^n)p(x^n \,\|\, y^{n-1})$. Then $E(\ell_b(X^n,Y^n))$ is lower bounded [25, Ch 5.4] by

$$E(\ell_b(X^n,Y^n)) \ge \sum_{x^n,y^n} p(x^n,y^n)\log\frac{1}{p(y^n)p(x^n \,\|\, y^n-1)} \quad (44)$$

and upper bounded [25, Ch 5.4] by

$$\begin{aligned} E(\ell_b(X^n,Y^n)) &\le \sum_{x^n,y^n} p(x^n,y^n)\left\lceil \log\frac{1}{p(y^n)p(x^n \,\|\, y^n-1)}\right\rceil \\ &\le \sum_{x^n,y^n} p(x^n,y^n)\log\frac{1}{p(y^n)p(x^n \,\|\, y^n-1)}+1. \end{aligned} \quad (45)$$

Now note that

$$\begin{aligned} &\sum_{x^n,y^n} p(x^n,y^n)\log\frac{1}{p(y^n)p(x^n \,\|\, y^n-1)} \\ &\quad= \sum_{x^n,y^n} p(x^n,y^n)\log\frac{p(x^n,y^n)}{p(y^n)p(x^n \,\|\, y^n-1)} \\ &\qquad+ H(X^n,Y^n) \\ &\quad= \sum_{x^n,y^n} p(x^n,y^n)\log\frac{p(y^n \,\|\, x^n)}{p(y^n)} + H(X^n,Y^n) \\ &\quad= I(X^n \to Y^n) + H(X^n,Y^n). \end{aligned} \quad (46)$$

Hence, the redundancy (the gap from the minimum average code length) is

$$\begin{aligned} \frac{1}{n}I(X^n \to Y^n) &\le \frac{E(\ell_b(X^n,Y^n)) - E(\ell(X^n,Y^n))}{n} \\ &\le \frac{1}{n}I(X^n \to Y^n) + \frac{1}{n}. \end{aligned} \quad (47)$$

■

Similarly, if the backward link is ignored, we have the following redundancy.

Lemma 2: If a lossless code is erroneously designed to be optimal for the case in which the backward link does not exist, namely, the code is designed for the joint distribution $p(y^n \,\|\, x^n)p(x^n)$, then the redundancy (up to $\frac{1}{n}$ bit) is $\frac{1}{n}I(Y^{n-1} \to X^n)$.

The proof of the lemma follows similar steps as the proof of Lemma 1 with (46) replaced by

$$\begin{aligned} &\sum_{x^n,y^n} p(x^n,y^n)\log\frac{1}{p(y^n \,\|\, x^n)p(x^n)} \\ &\quad= \sum_{x^n,y^n} p(x^n,y^n)\log\frac{p(x^n,y^n)}{p(y^n \,\|\, x^n)p(x^n)} + H(X^n,Y^n) \\ &\quad= \sum_{x^n,y^n} p(x^n,y^n)\log\frac{p(x^n \,\|\, y^{n-1})}{p(x^n)} + H(X^n,Y^n) \\ &\quad= I(Y^{n-1} \to X^n) + H(X^n,Y^n). \end{aligned} \quad (48)$$

Note that the redundancy due to ignoring both links is the sum of the redundancies from ignoring each link. At the same time, this redundancy is for the code designed for the joint distribution $p(y^n)p(x^n)$ and, hence, is equal to $\frac{1}{n}I(X^n;Y^n)$ (up to $\frac{1}{n}$ bit). This recovers the conservation law (12) operationally.

V. Directed Information and Statistics: Hypothesis Testing

Consider a system with an input sequence $(X_1,X_2,\ldots,X_n)$ and output sequence $(Y_1,Y_2,\ldots,Y_n)$, where the input is generated by a stimulation mechanism or a controller, which observes the previous outputs, and the output may be generated either causally from the input according to $\{p(y_i \,|\, y^{i-1},x^i)\}_{i=1}^n$ (the null hypothesis H_0) or independently from the input according to $\{p(y_i \,|\, y^{i-1})\}_{i=1}^n$ (the alternative hypothesis H_1). For instance, this setting occurs in communication or biological systems, where we wish to test whether the observed system output Y^n is in response to one's own stimulation input X^n or to some other input that uses the same stimulation mechanism and, therefore, induces the same marginal distribution $p(y^n)$. The stimulation mechanism $p(x^n \,\|\, y^{n-1})$, the output generator $p(y^n \,\|\, x^n)$, and the sequences X^n and Y^n are assumed to be known.

An *acceptance region* A is the set of all sequences (x^n,y^n) for which we accept the null hypothesis H_0. The complement of A, denoted by A^c, is the rejection region, namely, the set of all sequences (x^n,y^n) for which we reject the null hypothesis H_0 and accept the alternative hypothesis H_1. Let

$$\alpha := \Pr(A^c \,|\, H_0) \quad \text{and} \quad \beta := \Pr(A \,|\, H_1) \quad (49)$$

denote the probabilities of *type I error* and *type II error*, respectively.

The following theorem interprets the directed information rate $\mathcal{I}(X \to Y)$ as the best error exponent of β that can be achieved while α is less than some constant $\epsilon > 0$.

Theorem 5 (Chernoff–Stein Lemma for the Causal Dependence Test: Type II Error): Let $(X,Y) = \{X_i,Y_i\}_{i=1}^\infty$ be a stationary and ergodic random process. Let $A_n \subseteq \mathcal{X}^n \times \mathcal{Y}^n$ be an acceptance region, and let α_n and β_n be the corresponding probabilities of type I and type II errors (49). For $0 < \epsilon < \frac{1}{2}$, let

$$\beta_n^{(\epsilon)} = \min_{A_n \subseteq \mathcal{X}^n \times \mathcal{Y}^n, \alpha_n < \epsilon} \beta_n. \quad (50)$$

Then

$$\lim_{n\to\infty} -\frac{1}{n}\log\beta_n^{(\epsilon)} = \mathcal{I}(X\to Y) \tag{51}$$

where the directed information rate is the one induced by the joint distribution from H_0, i.e., $p(x^n \,\|\, y^{n-1})p(y^n \,\|\, x^n)$.

Theorem 5 is reminiscent of the achievability proof in the channel coding theorem. In the random coding achievability proof [25, ch 7.7] we check whether the output Y^n is resulting from a message (or equivalently, from an input sequence X^n) and we would like the error exponent, which is, according to Theorem 5, $I(X^n \to Y^n)$, to be as large as possible so we can distinguish between as many messages as possible.

The proof of Theorem 5 combines arguments from the Chernoff–Stein lemma [25, Theorem 11.8.3] with the Shannon–McMillan–Breiman theorem for directed information [17, Lemma 3.1], which implies that for a jointly stationary ergodic random process

$$\frac{1}{n}\log\frac{p(Y^n \,\|\, X^n)}{p(Y^n)} \to \mathcal{I}(X\to Y) \quad \text{in probability.}$$

Proof: Achievability: Fix $\delta > 0$ and let A_n be

$$A_n = \left\{(x^n, y^n) : \left|\frac{1}{n}\log\frac{p(y^n \,\|\, x^n)}{p(y^n)} - \mathcal{I}(X\to Y)\right| < \delta\right\} \tag{52}$$

By the AEP for directed information [17, Lemma 3.1] we have that $\Pr(A_n \,|\, H_0) \to 1$ in probability; hence, there exists $N(\epsilon)$ such that for all $n > N(\epsilon)$, $\alpha_n = \Pr(A_n^c \,|\, H_0) < \epsilon$. Furthermore

$$\begin{aligned}
\beta_n &= \Pr(A_n \,|\, H_1) \\
&= \sum_{x^n, y^n \in A_n} p(x^n \,\|\, y^{n-1})p(y^n) \\
&\overset{(a)}{\le} \sum_{x^n, y^n \in A_n} p(x^n \,\|\, y^{n-1})p(y^n \,\|\, x^n)2^{-n(\mathcal{I}(X\to Y)-\delta)} \\
&= 2^{-n(\mathcal{I}(X\to Y)-\delta)} \sum_{x^n, y^n \in A_n} p(x^n \,\|\, y^{n-1})p(y^n \,\|\, x^n) \\
&\overset{(b)}{\le} 2^{-n(\mathcal{I}(X\to Y)-\delta)}(1-\alpha_n)
\end{aligned} \tag{53}$$

where inequality (a) follows from the definition of A_n and (b) from the definition of α_n. We conclude that

$$\limsup_{n\to\infty} \frac{1}{n}\log\beta_n \le -\mathcal{I}(X\to Y) + \delta \tag{54}$$

establishing the achievability since $\delta > 0$ is arbitrary.

Converse: Let $B_n \subseteq \mathcal{X}^n \times \mathcal{Y}^n$ such that $\Pr(B_n^c \,|\, H_0) < \epsilon < \frac{1}{2}$. Consider

$$\begin{aligned}
&\Pr(B_n \,|\, H_1) \\
&\quad \ge \Pr(A_n \cap B_n \,|\, H_1) \\
&\quad = \sum_{(x^n, y^n) \in A_n \cap B_n} p(x^n \,\|\, y^{n-1})p(y^n) \\
&\quad \ge \sum_{(x^n, y^n) \in A_n \cap B_n} p(x^n \,\|\, y^{n-1})p(y^n \,\|\, x^{n-1})2^{-n(\mathcal{I}(X\to Y)+\delta)} \\
&\quad = 2^{-n(\mathcal{I}(X\to Y)+\delta)} \Pr(A_n \cap B_n \,|\, H_0) \\
&\quad = 2^{-n(\mathcal{I}(X\to Y)+\delta)} \left(1 - \Pr\left(A_n^c \cup B_n^c \,|\, H_0\right)\right) \\
&\quad \ge 2^{-n(\mathcal{I}(X\to Y)+\delta)} \left(1 - \Pr\left(A_n^c \,|\, H_0\right) - \Pr\left(B_n^c \,|\, H_0\right)\right).
\end{aligned} \tag{55}$$

Since $\Pr(A_n^c \,|\, H_0) \to 0$ and $\Pr(B_n^c \,|\, H_0) < \epsilon < \frac{1}{2}$, we obtain

$$\liminf_{n\to\infty} \frac{1}{n}\log\beta_n \ge -(\mathcal{I}(X\to Y) + \delta). \tag{56}$$

Finally, since $\delta > 0$ is arbitrary, the proof of the converse is completed. ∎

VI. Directed Lautum Information

Recently, Palomar and Verdú [23] have defined the lautum information $L(X^n; Y^n)$ as

$$L(X^n; Y^n) := \sum_{x^n, y^n} p(x^n)p(y^n)\log\frac{p(y^n)}{p(y^n \,|\, x^n)} \tag{57}$$

and showed that it has operational interpretations in statistics, compression, gambling, and portfolio theory, when the true distribution is $p(x^n)p(y^n)$ but, mistakenly, a joint distribution $p(x^n, y^n)$ is assumed. In this section we show that if causal relations between the sequences are mistakenly assumed, then two new measures we refer to as *directed lautum information* of the first and second types emerge as a penalty for the mistaken assumptions. We first present the definitions and basic properties of these new measures. The operational interpretations of the directed lautum information are given in compression, statistics and portfolio theory. The proofs of the theorems and lemmas in this section are deferred to the Appendix.

Directed lautum information is defined similarly as lautum information but regular conditioning is replaced by causal conditioning as follows.

Definition 1 (Directed Lautum Information): We define *directed lautum information* of the first type by

$$L_1(X^n \to Y^n) := \sum_{x^n, y^n} p(x^n)p(y^n)\log\frac{p(y^n)}{p(y^n \,\|\, x^n)} \tag{58}$$

and of the second type by

$$L_2(X^n \to Y^n) := \sum_{x^n, y^n} p(x^n \,\|\, y^{n-1})p(y^n)\log\frac{p(y^n)}{p(y^n \,\|\, x^n)}. \tag{59}$$

When $p(x^n \,\|\, y^{n-1}) = p(x^n)$ (no feedback), the two definitions coincide. We will shortly see that directed lautum information of the first type has operational meanings in scenarios where the true distribution is $p(x^n)p(y^n)$ and, mistakenly, a joint distribution of the form $p(x^n)p(y^n \,\|\, x^n)$ is assumed. The second type occurs when the true distribution is $p(x^n \,\|\, y^{n-1})p(y^n)$, but a joint distribution of the form $p(x^n \,\|\, y^{n-1})p(y^n \,\|\, x^n)$ is assumed.

We have the following conservation law for the first-type directed lautum information:

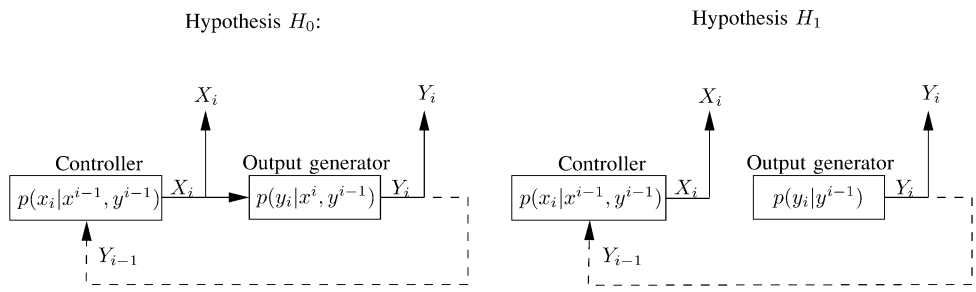

Fig. 4. Hypothesis testing. H_0: The input sequence $(X_1, X_2 \ldots, X_n)$ causally influences the output sequence $(Y_1, Y_2, \ldots, Y_n)$ through the causal conditioning distribution $p(y^n \,\|\, x^n)$. H_1: The output sequence $(Y_1, Y_2, \ldots, Y_n)$ was not generated by the input sequence $(X_1, X_2, \ldots, X_n)$, but by another input from the same stimulation mechanism $p(x^n \,\|\, y^{n-1})$.

Lemma 3 (Conservation Law for Directed Lautum Information of the First Type): For any discrete jointly distributed random vectors X^n and Y^n

$$L(X^n;Y^n) = L_1(X^n \to Y^n) + L_1(Y^{n-1} \to X^n). \quad (60)$$

A direct consequence of the lemma is the following condition for the equality between two types of directed lautum information and regular lautum information.

Corollary 2: If

$$L(X^n;Y^n) = L_1(X^n \to Y^n) \quad (61)$$

then

$$p(x^n) = p(x^n \,\|\, y^{n-1}) \quad (62)$$

for all $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ with $p(x^n, y^n) > 0$. Conversely, if (62) holds, then

$$L(X^n;Y^n) = L_1(X^n \to Y^n) = L_2(X^n \to Y^n). \quad (63)$$

The *lautum information rate* and *directed lautum information rates* are respectively defined as

$$\mathcal{L}(X;Y) := \lim_{n\to\infty} \frac{1}{n} L(Y^n;X^n), \quad (64)$$

$$\mathcal{L}_j(X \to Y) := \lim_{n\to\infty} \frac{1}{n} L_j(Y^n \to X^n) \quad \text{for } j = 1, 2 \quad (65)$$

whenever the limits exist. The next lemma provides a technical condition for the existence of the limits.

Lemma 4: If the process $\{(X_i, Y_i)\}$ is stationary and Markov, i.e., $p(x_i, y_i \,|\, x^{i-1}, y^{i-1}) = p(x_i, y_i \,|\, x_{i-k}^{i-1}, y_{i-k}^{i-1})$ for some k, then $\mathcal{L}(X;Y)$ and $\mathcal{L}_2(X \to Y)$ are well defined. Similarly, if the process $\{(X_i, Y_i)\}$ is such that for all n, $p(x^n, y^n) = p(x^n)p(y^n \,\|\, x^n)$, and the process is stationary and Markov, then $\mathcal{L}_1(X \to Y)$ is well defined.

Adding causality constraints to the problems that were considered in [23], we obtain the following results for data compression, hypothesis testing, and horse race gambling. These results provide operational interpretations to the directed lautum information.

A. Compression With Joint Distribution Mismatch

In Section IV we investigated the cost of ignoring forward and backward links when compressing (X^n, Y^n) relative to the optimal lossless variable length code. Here we investigate the penalty of assuming forward and backward links when in fact neither exists. Let X^n and Y^n be independent sequences.

Lemma 5: If a lossless code is erroneously designed to be optimal for the case where the forward link exists, namely, the code is designed for the joint distribution $p(x^n)p(y^n \,\|\, x^n)$, then the per-symbol penalty is within $\frac{1}{n}$ bit of $\frac{1}{n}L_1(X^n \to Y^n)$.

Similarly, if we incorrectly assume that the backward link $p(x^n \,\|\, y^{n-1})$ exists, we have the following lemma.

Lemma 6: If a lossless code is erroneously designed to be optimal for the case where the backward link exists, namely, the code is designed for the joint distribution $p(x^n \,\|\, y^{n-1})p(y^n)$, then the per-symbol penalty is within $\frac{1}{n}$ bit of $\frac{1}{n}L_1(Y^{n-1} \to X^n)$.

If both links are mistakenly assumed, the penalty [23] is lautum information $L(X^n;Y^n)$. Note that the penalty due to erroneously assuming both links is the sum of the penalty from erroneously assuming each link. This recovers the conservation law (60) for lautum information operationally.

B. Hypothesis Testing

We revisit the hypothesis testing problem in Section V, which is described in Fig. 4. As a dual to Theorem 6, we characterize the minimum type I error exponent given the type II error probability:

Theorem 6 (Chernoff–Stein Lemma for the Causal Dependence Test: Type I Error): Let $(X, Y) = \{X_i, Y_i\}_{i=1}^{\infty}$ be stationary, ergodic, and Markov of some order such that $p(x^n, y^n) = 0$ implies $p(x^n \,\|\, y^{n-1})p(y^n) = 0$. Let $A_n \subseteq \mathcal{X}^n \times \mathcal{Y}^n$ be an acceptance region, and let α_n and β_n be the corresponding probabilities of type I and type II errors (49). For $0 < \epsilon < \frac{1}{2}$, let

$$\alpha_n^{(\epsilon)} = \min_{A_n \subseteq \mathcal{X}^n \times \mathcal{Y}^n, \beta_n < \epsilon} \alpha_n. \quad (66)$$

Then

$$\lim_{n\to\infty} -\frac{1}{n} \log \alpha_n^{(\epsilon)} = \mathcal{L}_2(X \to Y) \quad (67)$$

where the directed lautum information rate is the one induced by the joint distribution from H_0, i.e., $p(x^n \,\|\, y^{n-1})p(y^n \,\|\, x^n)$.

The proof of Theorem 6 follows very similar steps as in the proof of Theorem 5 upon taking

$$A_n^c = \left\{ x^n, y^n : \left| \frac{1}{n} \log \frac{p(y^n)}{p(y^n \,\|\, x^n)} - \mathcal{L}_2(X \to Y) \right| < \delta \right\} \quad (68)$$

analogously to (52), and using the Markov assumption to guarantee the AEP; we omit the details.

C. Horse Race Gambling With Mismatched Causal Side Information

Consider the horse race setting in Section III-A where the gambler has causal side information. The joint distribution of horse race outcomes X^n and the side information Y^n is given by $p(x^n)p(y^n \,\|\, x^{n-1})$, namely, $X_i \to X^{i-1} \to Y^i$ form a Markov chain, and, therefore, the side information does not increase the growth rate. The gambler mistakenly assumes a joint distribution $p(x^n \,\|\, y^n)p(y^n \,\|\, x^{n-1})$, and, therefore, uses a gambling scheme $b^*(x^n \,\|\, y^n) = p(x^n \,\|\, y^n)$.

Theorem 7: If the gambling scheme $b^*(x^n \,\|\, y^n) = p(x^n \,\|\, y^n)$ is applied to the horse race described above, then the penalty in the growth with respect to the optimal gambling scheme $b^*(x^n)$ that uses no side information is $L_2(Y^n \to X^n)$. For the special case where the side information is independent of the horse race outcomes, the penalty is $L_1(Y^n \to X^n)$.

This result can be readily extended to the general stock market, for which the penalty is *upper bounded* by $L_2(Y^n \to X^n)$.

VII. Concluding Remarks

We have established the role of directed information in portfolio theory, data compression, and hypothesis testing. Put together with its key role in communications [8], [10]–[14], [17], [29] and in estimation [30], directed information is a key quantity in scenarios where causality and the arrow of time are crucial to the way a system operates. Among other things, these findings suggest that the estimation of directed information can be an effective tool for inferring causal relationships and related properties in a wide array of problems. This direction is under current investigation; see, for example, [31].

Appendix

Proof of Lemma 3: Consider

$$\begin{aligned} L(X^n;Y^n) &\overset{(a)}{=} \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{p(y^n)p(x^n)}{p(y^n,x^n)} \\ &\overset{(b)}{=} \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{p(y^n)p(x^n)}{p(y^n \,\|\, x^n)p(x^n \,\|\, y^{n-1})} \\ &= \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{p(y^n)}{p(y^n \,\|\, x^n)} \\ &\quad + \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{p(x^n)}{p(x^n \,\|\, y^{n-1})} \\ &= L_1(X^n \to Y^n) + L_1(Y^{n-1} \to X^n) \end{aligned} \quad (69)$$

where (a) follows from the definition of lautum information and (b) follows from the chain rule $p(y^n, x^n) = p(y^n \,\|\, x^n)p(x^n \,\|\, y^{n-1})$. ∎

Proof of Corollary 2: The proof of the first part follows from the conservation law (69) and the nonnegativity of Kullback-Leibler divergence [25, Theorem 2.6.3] (i.e., $L_1(Y^{n-1} \to X^n) = 0$ implies that $p(x^n) = p(x^n \,\|\, y^{n-1})$). The second part follows from the definitions of regular and directed lautum information. ∎

Proof of Lemma 4: It is easy to see the sufficiency of the conditions for $\mathcal{L}_1(X \to Y)$ from the following identity:

$$\begin{aligned} L_1(X^n \to Y^n) &= \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{p(x^n)p(y^n)}{p(x^n)p(y^n \,\|\, x^n)} \\ &= -H(X^n) - H(Y^n) \\ &\quad - \sum_{x^n,y^n} p(x^n)p(y^n) \log p(x^n)p(y^n \,\|\, x^n). \end{aligned}$$

Since the process is stationary the limits $\lim_{n\to\infty} \frac{1}{n} H(X^n)$ and $\lim_{n\to\infty} \frac{1}{n} H(Y^n)$ exist. Furthermore, since the pmf is of the form $p(x^n, y^n) = p(x^n)p(y^n \,||\, x^n)$ and since the process is stationary and Markov (i.e., $p(x_i, y_i \,|\, x^{i-1}, y^{i-1}) = p(x_i, y_i \,|\, x_{i-k}^{i-1}, y_{i-k}^{i-1})$ for some finite k), the limit $\lim_{n\to\infty} \frac{1}{n} \sum_{x^n,y^n} p(x^n)p(y^n) \log p(x^n)p(y^n \,\|\, x^n)$ exists. The sufficiency of the condition can be proved for $\mathcal{L}_2(X \to Y)$ and the lautum information rate using a similar argument. ∎

Proof of Lemma 5: Let $E(\ell_f(X^n, Y^n))$ be the average code designed optimally for the joint distribution $p(x^n)p(y^n \,\|\, x^n)$. Then $E(\ell_b(X^n, Y^n))$ is lower bounded [25, Ch 5.4] by

$$E(\ell_f(X^n, Y^n)) \geq \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{1}{p(y^n \,\|\, x^n)p(x^n)} \quad (70)$$

and upper bounded [25, Ch 5.4] by

$$\begin{aligned} E(\ell_f(X^n, Y^n)) &\leq \sum_{x^n,y^n} p(x^n)p(y^n) \left\lceil \log \frac{1}{p(y^n \,\|\, x^n)p(x^n)} \right\rceil \\ &\leq \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{1}{p(y^n \,\|\, x^n)p(x^n)} + 1. \end{aligned} \quad (71)$$

Now note that

$$\begin{aligned} &\sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{1}{p(y^n \,||\, x^n)p(x^n)} \\ &\quad = \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{p(x^n)p(y^n)}{p(y^n \,||\, x^n)p(x^n)} \\ &\qquad + H(X^n) + H(Y^n) \\ &\quad = \sum_{x^n,y^n} p(x^n)p(y^n) \log \frac{p(y^n)}{p(y^n \,||\, x^n)} + H(X^n) + H(Y^n) \\ &\quad = L_1(X^n \to Y^n) + H(X^n) + H(Y^n). \end{aligned} \quad (72)$$

Hence, the penalty (the gap from the minimum average length code if there was no mismatch) is

$$\frac{1}{n}L_1(X^n \to Y^n) \le \frac{E(\ell_b(X^n, Y^n)) - E(l(X^n, Y^n))}{n} \le \frac{1}{n}L_1(X^n \to Y^n) + \frac{1}{n}. \quad (73)$$

Proof of Lemma 6: The proof of the lemma follows similar steps as the proof of Lemma 5 with (72) was replaced by

$$\begin{aligned} E(l(X^n, Y^n)) &= \sum_{x^n, y^n} p(x^n)p(y^n) \log \frac{1}{p(x^n \,\|\, y^{n-1})p(y^n)} \\ &= \sum_{x^n, y^n} p(x^n)p(y^n) \log \frac{p(x^n)p(y^n)}{p(x^n \,\|\, y^{n-1})p(y^n)} \\ &\quad + H(X^n) + H(Y^n) \\ &= \sum_{x^n, y^n} p(x^n)p(y^n) \log \frac{p(x^n)}{p(x^n \,\|\, y^{n-1})} + H(X^n) + H(Y^n) \\ &= L_1(Y^{n-1} \to X^n) + H(X^n) + H(Y^n). \end{aligned} \quad (74)$$

■

Proof of Theorem 7: The optimal growth rate where the joint distribution is $p(x^n)p(y^n \,\|\, x^{n-1})$ is $W^*(X^n) = E[\log \alpha(X^n)] - H(X^n)$. Let $E_{p(x^n)p(y^n \,\|\, x^{n-1})}$ denotes the expectation with respect to the joint distribution $p(x^n)p(y^n \,\|\, x^{n-1})$. The growth rate for the gambling strategy $b(x^n \,\|\, y^n) = p(x^n \,\|\, y^n)$ is

$$\begin{aligned} W^*(X^n \,\|\, Y^n) &= E_{p(x^n)p(y^n \,\|\, x^{n-1})}[\log b(X^n \,\|\, Y^n)\alpha(X^n)] \\ &= E_{p(x^n)p(y^n \,\|\, x^{n-1})}[\log p(X^n \,\|\, Y^n)] \\ &\quad + E_{p(x^n)p(y^n \,\|\, x^{n-1})}[\log \alpha(X^n)]; \end{aligned} \quad (75)$$

hence $W^*(X^n) - W^*(X^n \,\|\, Y^n) = L_2(Y^n \to X^n)$. In the special case, where the side information is independent of the horse outcome, namely, $p(y^n \,\|\, x^{n-1}) = p(y^n)$, then $L_2(Y^n \to X^n) = L_1(Y^n \to X^n)$. ■

ACKNOWLEDGMENT

The authors would like to thank I. Kontoyiannis for helpful discussions.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379, 623–423, 656, 1948.

[2] C. E. Shannon, "Coding theorems for a discrete source with fidelity criterion," in *Inf. Decision Processes*, R. E. Machol, Ed. : McGraw-Hill, 1960, pp. 93–126.

[3] R. G. Gallager, Source Coding With Side Information and Universal Coding, Sep. 1976, unpublished manuscript.

[4] B. Y. Ryabko, "Encoding a source with unknown but ordered probabilities," *Probl. Inf. Transmission*, pp. 134–139, 1979.

[5] J. Kelly, "A new interpretation of information rate," *Bell Syst. Tech. J.*, vol. 35, pp. 917–926, 1956.

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

[7] H. Marko, "The bidirectional communication theory: A generalization of information theory," *IEEE Trans. Commun.*, vol. 21, pp. 1335–1351, 1973.

[8] G. Kramer, "Directed Information for Channels With Feedback," Ph.D. dissertation, Swiss Fed. Inst. Technol. (ETH), Zürich, Switzerland, 1998.

[9] S. C. Tatikonda, "Control Under Communication Constraints," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, 2000.

[10] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.

[11] Y.-H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory.*, vol. 25, no. 4, pp. 1488–1499, Apr. 2008.

[12] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[13] 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.

[14] H. H. Permuter, T. Weissman, and J. Chen, "Capacity region of the finite-state multiple access channel with and without feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2455–2477, Jun. 2009.

[15] B. Shrader and H. H. Permuter, "Feedback capacity of the compound channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3629–3644, Aug. 2009.

[16] R. Dabora and A. Goldsmith, "The capacity region of the degraded finite-state broadcast channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1828–1851, Apr. 2010.

[17] R. Venkataramanan and S. S. Pradhan, "Source coding with feedforward: Rate-distortion theorems and error exponents for a general source," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2154–2179, Jun. 2007.

[18] A. Rao, A. Hero, D. States, and J. Engel, "Inference of biologically relevant gene influence networks using the directed information criterion," presented at the ICASSP, 2006.

[19] M. Lungarella and O. Sporns, "Mapping information flow in sensorimotor networks," *PLoS Comput. Biol.*, vol. 2, no. 10, p. e144, Oct. 2006.

[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.*, pp. 1–28, 2010.

[21] O. Rivoire and S. Leibler, *The Value of Information for Populations in Varying Environments*, Oct. 2010.

[22] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate-distortion function by prediction," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, Jul. 2008.

[23] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Trans. Inf. Theory*, vol. 54, pp. 964–975, 2008.

[24] J. Massey and P. C. Massey, "Conservation of mutual and directed information," in *Proc. Int. Symp. Inf. Theory*, Adelaide, Australia, Sep. 2005, pp. 157–158.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.

[26] P. H. Algoet and T. M. Cover, "A sandwich proof of the Shannon–McMillan–Breiman theorem," *Ann. Probab.*, vol. 16, pp. 899–909, 1988.

[27] A. R. Barron and T. M. Cover, "A bound on the financial value of information," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 1097–1100, Sep. 1988.

[28] R. Gallager, "Variations on a theme by Huffman," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 6, pp. 668–674, Nov. 1978.

[29] B. Shrader and H. H. Permuter, "On the compound finite state channel with feedback," presented at the IEEE Int. Symp. Inf. Theory, Nice, France, 2007.

[30] H. H. Permuter, Y. H. Kim, and T. Weissman, "Directed information, causal estimation and communication in continuous time," presented at the Control Over Communication Channels (ConCom), Seoul, South Korea, Jun. 2009.

[31] L. Zhao, T. Weissman, Y.-H. Kim, and H. H. Permuter, "Universal estimation of directed information," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, Jun. 2010, pp. 1433–1437.

Haim H. Permuter (M'08) received the 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. degrees in electrical engineering from Stanford University, Stanford, CA, 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 lecturer at Ben-Gurion university.

Dr. Permuter is a recipient of the Fulbright Fellowship, the Stanford Graduate Fellowship (SGF), the Allon Fellowship, and the Bergmann award.

Young-Han Kim (S'99–M'06) received the B.S. degree (Hons.) in electrical engineering from Seoul National University, Seoul, Korea, in 1996, the M.S. degrees in electrical engineering and statistics, and the Ph.D. degree in electrical engineering, both 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 Assistant 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 and the 2009 U.S.-Israel Binational Science Foundation Bergmann Memorial Award.

Tsachy Weissman (S'99–M'02–SM'07) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Technion–Israel Institute of Technology, Haifa, Israel.

During 2002–2003, he was with the Information Theory Research Group at Hewlett Packard Labs. He has been on the Faculty of the Electrical Engineering Department, Stanford University, Stanford, CA, since 2003, spending the two academic years 2007–2009 with the Technion Department of Electrical Engineering. His research is focused on information theory, statistical signal processing, the interplay between them, and their applications. He is the inventor of several patents and involved in a number of high-tech companies as a researcher or member of the technical board.

Dr. Weissman received the IT/COM Societies Best Paper Award, a Horev Fellowship for Leaders in Science and Technology, and a Henry Taub Prize for Excellence in Research, in addition to other awards. He currently serves as Associate Editor for Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY.