Justification of Logarithmic Loss via the Benefit of Side Information

Jiantao Jiao, *Student Member, IEEE*, Thomas A. Courtade, *Member, IEEE*, Kartik Venkat, *Student Member, IEEE*, and Tsachy Weissman, *Fellow, IEEE*

***Abstract*— We consider a natural measure of relevance: the reduction in optimal prediction risk in the presence of side information. For any given loss function, this relevance measure captures the benefit of side information for performing inference on a random variable under this loss function. When such a measure satisfies a natural data processing property, and the random variable of interest has alphabet size greater than two, we show that it is uniquely characterized by the mutual information, and the corresponding loss function coincides with logarithmic loss. In doing so, our work provides a new characterization of mutual information, and justifies its use as a measure of relevance. When the alphabet is binary, we characterize the only admissible forms the measure of relevance can assume while obeying the specified data processing property. Our results naturally extend to measuring the causal influence between stochastic processes, where we unify different causality measures in the literature as instantiations of directed information.**

***Index Terms*— Axiomatic characterizations, causality measures, data processing, directed information, logarithmic loss.**

I. Introduction

IN STATISTICAL decision theory, it is often a controversial issue to choose the appropriate loss function. One popular loss function is called *logarithmic loss*, defined as follows. Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ be a finite set with $|\mathcal{X}| = n$, let Γ_n denote the set of probability measures on $\mathcal{X}$, and let $\bar{\mathbb{R}}$ denote the extended real line.

Definition 1 (Logarithmic Loss): Logarithmic loss $\ell_{\log}: \mathcal{X} \times \Gamma_n \to \bar{\mathbb{R}}$ *is defined by*

$$\ell_{\log}(x, P) = \log \frac{1}{P(x)}, \tag{1}$$

where $P(x)$ *denotes the probability of* x *under measure* P.

Logarithmic loss has enjoyed numerous applications in various fields. For instance, its usage in statistics dates back to Good [1], and it has found a prominent role in learning and prediction (cf. Cesa–Bianchi and Lugosi [2, Ch. 9]). Logarithmic loss also assumes an important role in information theory, where many of the fundamental quantities (e.g., entropy, relative entropy, etc.) can be interpreted as the optimal prediction risk or regret under logarithmic loss. Recently, Courtade and Weissman [3] showed that the long-standing open problem of multiterminal source coding could be completely solved under logarithmic loss, providing yet another concrete example of its special nature. The use of the logarithm in defining entropy can be justified by its various axiomatic characterizations, the first of which dates back to Shannon [4].

The main contribution of this paper is in providing fundamental justification for inference using logarithmic loss. In particular, we show that a single and natural Data Processing requirement mandates the use of logarithmic loss. We begin by posing the following:

Question 1 (Benefit of Side Information): Suppose X, Y *are dependent random variables. How relevant is* Y *for inference on* X*?*

II. Problem Formulation and Main Results

Toward answering Question 1, let $\ell: \mathcal{X} \times \hat{\mathcal{X}} \to \bar{\mathbb{R}}$ be an arbitrary loss function with reconstruction alphabet $\hat{\mathcal{X}}$, where $\hat{\mathcal{X}}$ is arbitrary. Given $(X, Y) \sim P_{XY}$, it is natural to quantify the benefit of additional side information Y by computing the difference between the expected losses in predicting $X \in \mathcal{X}$ with and without side information Y, respectively. This motivates the following definition:

$$C(\ell, P_{XY}) \triangleq \inf_{\hat{x}_1 \in \hat{\mathcal{X}}} \mathbb{E}_P[\ell(X, \hat{x}_1)] - \inf_{\hat{X}_2(Y)} \mathbb{E}_P[\ell(X, \hat{X}_2)], \tag{2}$$

where $\hat{x}_1 \in \hat{\mathcal{X}}$ is deterministic, and $\hat{X}_2 = \hat{X}_2(Y) \in \hat{\mathcal{X}}$ is any measurable function of Y. In the following discussions, we require that indeterminate forms like $\infty - \infty$ do not appear in the definition of $C(\ell, P_{XY})$. By taking Y to be independent of X, this requirement implies that for all $P \in \Gamma_n$,

$$\left| \inf_{\hat{x}_1 \in \hat{\mathcal{X}}} \mathbb{E}_P[\ell(X, \hat{x}_1)] \right| < \infty. \tag{3}$$

The formulation (2) has appeared previously in the statistics literature. DeGroot [5] in 1962 defined the *information* contained in an experiment, which turns out to be equivalent to (2). Later, Dawid [6] defined the *coherent dependence function*, which is equivalent to (2), and used it to quantify the dependence between two random variables X, Y. Our framework

Manuscript received March 20, 2015; revised May 21, 2015; accepted July 7, 2015. Date of publication July 30, 2015; date of current version September 11, 2015. This work was supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. Jiantao Jiao and Kartik Venkat were partially supported by two Stanford Graduate Fellowships. The material in this paper was presented in part at the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA.

J. Jiao, K. Venkat, and T. Weissman are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: jiantao@stanford.edu; kvenkat@stanford.edu; tsachy@stanford.edu).

T. Courtade is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: courtade@eecs.berkeley.edu).

Communicated by H. H. Permuter, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2015.2462848

of quantifying the predictive benefit of side information is closely connected to the notion of proper scoring rules and the literature on probability forecasting in statistics. The survey by Gneiting and Raftery [7] provides a good overview.

Having introduced the yardstick in (2), we now reformulate the question of interest: *Which loss function(s) ℓ can be used to define $C(\ell, P_{XY})$ in a meaningful way?* Of course, "meaningful" is open to interpretation, but it is desirable that $C(\ell, P_{XY})$ be well-defined, at minimum. This motivates the following axiom:

Data Processing Axiom: For all distributions P_{XY}, the quantity $C(\ell, P_{XY})$ satisfies

$$C(\ell, P_{TY}) \le C(\ell, P_{XY})$$

whenever $T(X) \in \mathcal{X}$ is a statistically sufficient transformation of X for Y.

We remind the reader that the statement 'T is a statistically sufficient transform of X for Y' means that the following two Markov chains hold:

$$T - X - Y, \quad X - T - Y \tag{4}$$

That is, $T(X)$ preserves all of the information X contains about Y.

In words, the Data Processing Axiom stipulates that processing the data $X \to T$ cannot boost the predictive benefit of the side information.[1]

To convince the reader that the Data Processing Axiom is a natural requirement, suppose instead that the Data Processing Axiom did not hold. Since X and T are mutually sufficient statistics for Y, this would imply that there is *no* unique value which quantifies the benefit of side information Y for the random variable of interest. Thus, the Data Processing Axiom is needed for the benefit of side information to be well-defined.

Although the Data Processing Axiom may seem to be a benign requirement, it has far-reaching implications for the form $C(\ell, P_{XY})$ can take. This is captured by our first main result:

Theorem 1: Let $n \ge 3$. Under the Data Processing Axiom, the function $C(\ell, P_{XY})$ is uniquely determined by the mutual information,

$$C(\ell, P_{XY}) = I(X; Y), \tag{5}$$

up to a multiplicative factor.

The following corollary immediately follows from Theorem 1.

Corollary 1: Let $n \ge 3$. Under the Data Processing Axiom, the benefit of additional side information Y for inference on X with common side information W, i.e.

$$\inf_{\hat{X}_1(W)} \mathbb{E}_P[\ell(X, \hat{X}_1)] - \inf_{\hat{X}_2(Y,W)} \mathbb{E}_P[\ell(X, \hat{X}_2)], \tag{6}$$

is uniquely determined by the conditional mutual information,

$$I(X; Y|W), \tag{7}$$

up to a multiplicative factor.

[1] In fact, the Data Processing Axiom is weaker than this general data processing statement since it only addresses statistically sufficient transformations of X.

Thus, up to a multiplicative factor, we see that logarithmic loss generates the *only* measure of predictive benefit (defined according to (2)) which satisfies the Data Processing Axiom. In other words, Theorem 1 provides a definitive answer to Question 1 under the framework we have described, and also highlights the special role played by logarithmic loss.

Theorem 1 shows that mutual information uniquely quantifies the reduction of prediction risk due to side information. Note that the characterization of mutual information afforded by Theorem 1 does not explicitly require any of the mathematical properties of mutual information, such as the chain rule, or invariance to one-to-one transformations of both X and Y. Thus, beyond the operational implications of our result, Theorem 1 has strong implications for axiomatic characterization of information measures from a mathematical standpoint. On this point, we note that Csiszár, in his survey [8] names the axiomatic result of Aczél *et al.* [9] as "*intuitively most appealing*" in characterizing the entropy in terms of symmetry, expansibility, additivity, and subadditivity, whereas most other known characterizations require recursivity or the sum property. For details we refer to Csiszár [8].

Theorem 1 provides a partial explanation for why mutual information is widely used as an inferential tool across various applications in science and engineering, and is deeply imbued in fundamental concepts in various disciplines. In statistics, one of the popular criteria for objective Bayesian modeling [10] is to design a prior on the parameter to maximize the mutual information between the parameter and the observations. In machine learning, the so-called *infomax* [11] criterion states that the function that maps a set of input values to a set of output values should be chosen or learned so as to maximize the mutual information between the input and output, subject to a set of specified constraints. This principle has been widely adopted in practice, for example, in decision tree based algorithms in machine learning such as C4.5 [12], one tries to select the feature at each step of tree splitting to maximize the mutual information (called *information gain* principle [13]) between the output and the feature conditioned on previous chosen features. In some applications, mutual information arises naturally as the only answer, for example, the well known Chow–Liu algorithm [14] for learning tree graphical models relies on estimation of the mutual information, which is a natural consequence of maximum likelihood estimation. We also mention genetics [15], image processing [16], computer vision [17], secrecy [18], ecology [19], and physics [20] as fields in which mutual information is widely used. Erkip and Cover [21] argued that mutual information is a natural quantity in the context of portfolio theory, where it emerges as the increase in growth rate due to the presence of side information.

Mutual information and related information theoretic measures are instrumental in various applications. This motivates investigating optimal estimators for these quantities based on data. There exist extensive literature on this subject, and we refer to [22] for a detailed review, as well as the theory and Matlab/Python implementations of entropy and mutual information estimators that achieve the minimax rates in all the regimes of sample size and support size pairs. For the recent growing literature on information measure estimation in the high-dimensional regime, we refer to [22]–[29].

Interestingly, the assumption that $n \geq 3$ in Theorem 1 is essential. When the alphabet of X is binary, i.e. $n = 2$, the Data Processing Axiom no longer mandates the use of logarithmic loss. We have an explicit characterization for the form the measure of relevance (2) can take. The class of solutions for the binary alphabet setting is characterized by the following theorem.

Theorem 2: Let $n = 2$. Under the Data Processing Axiom, $C(\ell, P_{XY})$ must be of the form

$$C(\ell, P_{XY}) = \sum_y P_Y(y) G(P_{X|Y=y}) - G(P_X), \tag{8}$$

where $G((p, 1-p)) : \Gamma_2 \to \mathbb{R}$ is a symmetric (invariant to permutations), convex function. Moreover, for any symmetric convex function $G((p, 1-p)) : \Gamma_2 \to \mathbb{R}$, there exists a loss function ℓ whose corresponding $C(\ell, P_{XY})$ satisfies the Data Processing Axiom and is given by (8)*.*

It is worth mentioning that there is an interesting set of observations surrounding the characterization of information measures which is sensitive to the alphabet size being binary or larger. This phenomenon is explored further in [30].

The rest of this paper is organized as follows. In Section III, we explore the connections between our results and the existing literature on causal analysis, including Granger and Sims causality, Geweke's measure, transfer entropy, and directed information. The proofs of Theorems 1 and 2 are provided in Section IV. Proofs of some auxiliary lemmas are deferred to the appendix.

III. Causality Measures: An Axiomatic Viewpoint

Inferring causal relationships from observed data plays an indispensable part in scientific discovery. Granger, in his seminal work [31], proposed a predictive test for inferring causal relationships. To state his test, let X_t, Y_t, U_t be stochastic processes, where X_t, Y_t are the processes of interest, and U_t contains all information in the universe accumulated up to time t. Granger's causality test asserts that Y_t causes X_t, denoted by $Y_t \Rightarrow X_t$, if we are better able to predict X_t using the past information of U_t, than by using all past information in U_t apart from Y_t. In Granger's definition, the quality of prediction is measured by the squared error risk achieved by the optimal unbiased least-squares predictor.

In his 1980 paper, Granger [32] introduced a set of operational definitions which made it possible to derive practical testing procedures. For example, he assumes that we must be able to specify U_t in order to perform causality tests, which is slightly different from his original definition which required knowledge of all information in the universe (which is usually unavailable).

Later, Sims [33] introduced a related concept of causality, which was proved to be equivalent to Granger's definition in Sims [33], Hosoya [34], and Chamberlain [35] in a variety of settings.

Motivated by Granger's framework for testing causality using linear prediction, Geweke [36], [37] proposed a causality measure to quantify the extent to which Y is causing X. Quoting Geweke (emphasis ours):

> *"The empirical literature abounds with <u>tests</u> of independence and unidirectional causality for various pairs of time series, but there have been virtually no investigations of the <u>degree</u> of dependence or the extent of various kinds of feedback. The latter approach is more realistic in the typical case in which the hypothesis of independence of unidirectional causality is not literally entertained, but it requires that one be able to measure linear dependence and feedback."*

In other words, Geweke makes the important distinction between a *causality test* which makes a binary decision on whether one process causes another, and a *causality measure* which quantifies the degree to which one process causes another. Geweke proposed the following measure as a natural starting point:

$$F_{Y \Rightarrow X} \triangleq \ln \frac{\sigma^2(X_t | X^{t-1})}{\sigma^2(X_t | X^{t-1}, Y^{t-1})}, \tag{9}$$

where $\sigma^2(X_t | X^{t-1}, Y^{t-1})$ is the variance of the prediction residue when predicting X_t via the optimal linear predictor constructed from observation X^{t-1}, Y^{t-1}. Note that if $F_{Y \Rightarrow X} > 0$, we could conclude $Y_t \Rightarrow X_t$ according to Granger's test.

It has long been observed that the restriction to optimal linear predictors in testing causality is not necessary. In fact, Chamberlain [35] proved a general equivalence between Granger and Sims' causality tests by replacing linear predictors with conditional independence tests. However, the natural generalization of (9) wasn't clear until Gourieroux *et al.* [38] proposed the so-called *Kullback causality measures* in 1987. It is now well-known that Kullback causality measures are equivalent to (9) under linear Gaussian models (e.g. Barnett *et al.* [39]).

Using information theoretic terms, Kullback causality measures are nothing but the directed information introduced by Massey [40], and motivated by Marko [41]. Using modern notation, the directed information from X^n to Y^n is defined as

$$I(X^n \to Y^n) \triangleq \sum_{i=1}^n I(X^i; Y_i | Y^{i-1}) \tag{10}$$

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

where $H(Y^n \| X^n)$ is the *causally conditional entropy*, defined by

$$H(Y^n \| X^n) \triangleq \sum_{i=1}^n H(Y_i | Y^{i-1}, X^i). \tag{12}$$

Massey and Massey [42] established the pleasing conservation law of directed information:

$$I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n) \tag{13}$$

$$= I(X^{n-1} \to Y^n) + I(Y^{n-1} \to X^n) + \sum_{i=1}^n I(X_i; Y_i | X^{i-1}, Y^{i-1}), \tag{14}$$

which implies that the extent to which process X_t influences process Y_t and vice-versa always sum to the total mutual information between the two processes. Since $I(Y^{n-1} \to X^n)$ can be expressed as

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

X_i being conditionally independent of Y^{i-1} given X^{i-1} is equivalent to $I(Y^{n-1} \to X^n) = 0$. This corresponds precisely to the definition of general Granger non-causality. Permuter *et al.* [43] showed various applications of directed information in portfolio theory, data compression, and hypothesis testing in the presence of causality constraints. Amblard and Michel [44] reviewed the intimate connections between Granger causality and directed information theory.

We remark that, for practical applications, the directed information between stochastic processes can be computed using the universal estimators proposed in [45], which exhibit near-optimal statistical properties.

Finally, we note that the notion of *transfer entropy* in the physics literature, which was proposed by Schreiber [46] in 2000, turns out to be equivalent to directed information.

To connect our present discussion on causality measures to Theorem 1, we recall that the directed information rate [47] between a pair of jointly stationary finite-alphabet processes X_t, Y_t can be written as:

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{n} I(Y^{n-1} \to X^n) \\ &= \inf_{T_1(X_{-\infty}^{-1})} \mathbb{E}[\ell_{\log}(X_0, T_1)] - \inf_{T_2(X_{-\infty}^{-1}, Y_{-\infty}^{-1})} \mathbb{E}[\ell_{\log}(X_0, T_2)].\end{aligned}$$

In light of this, we can conclude from Theorem 1 and Corollary 1 that the directed information rate is the *unique* measure of causality which assumes the form (2) and satisfies the Data Processing Axiom. Thus, our axiomatic viewpoint explains why the same causality measure has appeared so often in varied fields including economics, statistics, information theory, and physics. Except in the binary case, we roughly have the following: *All 'reasonable' causality measures defined by a difference of predictive risks must coincide*.[2]

IV. Proof of Main Results

In this section, we provide complete proofs of Theorems 1 and 2 and highlight the key ideas.

To begin, we show that the measure of relevance defined in (2) is equivalently characterized by a bounded convex function defined on the $\mathcal{X}$-simplex. The following lemma achieves this goal.

Lemma 1: There exists a bounded convex function $V : \Gamma_n \to \mathbb{R}$, depending on ℓ, such that

$$C(\ell, P_{XY}) = \left(\sum_y P_Y(y) V(P_{X|Y=y})\right) - V(P_X). \quad (15)$$

[2]Here, the authors' interpretation of "reasonable" is reflected by the Data Processing Axiom. In the context of this section, the Data Processing Axiom stipulates that any reasonable causality measure should be invariant under statistically sufficient transformations of the data – a desirable property and natural criterion.

The proof of Lemma 1 follows from defining $V(P)$ by

$$V(P) = -\inf_{\hat{x}\in\hat{\mathcal{X}}} \mathbb{E}_P[\ell(X, \hat{x})], \quad (16)$$

and its details are deferred to the appendix. In the statistics literature, the quantity $-V(P)$ is usually called the *generalized entropy* or the *Bayes envelope*. We refer to Dawid [48] for details.

In the literature of concentration inequalities, the following functional

$$H_\Phi(Z) = \mathbb{E}\Phi(Z) - \Phi(\mathbb{E}Z), \quad (17)$$

where Φ is a convex function, is called Φ-entropy, which in fact quantifies the gap in Jensen's inequality. As shown in Lemma 1, functional $C(\ell, P_{XY})$ is closely related to the notion of Φ-entropy. We refer to Boucheron *et al.* [49, Ch. 14] for a nice survey on the usage of Φ-entropies in proving concentration inequalities.

The next lemma asserts that we only need to consider symmetric (invariant to permutations) functions $V(P)$.

Lemma 2: Under the Data Processing Axiom, there exists a symmetric finite convex function $G : \Gamma_n \to \mathbb{R}$, such that

$$C(\ell, P_{XY}) = \left(\sum_y P_Y(y) G(P_{X|Y=y})\right) - G(P_X), \quad (18)$$

and $G(\cdot)$ is equal to $V(\cdot)$ in Lemma 1 up to a linear translation:

$$G(P) = V(P) + \langle c, P\rangle, \quad (19)$$

where $c \in \mathbb{R}^n$ is a constant vector.

The proof of Lemma 2 follows by applying a permutation to the space $\mathcal{X}$ and applying the Data Processing Axiom. Details are deferred to the appendix.

Now we are in the position to begin the proof of Theorem 1 in earnest.

A. The Case $n \geq 3$

It suffices to consider the case when the side information Y is binary valued, i.e., $Y \in \{1, 2\}$. We will show that the Data Processing Axiom mandates the usage of the logarithmic loss even when we constrain ourselves to this situation.

Define $\alpha \triangleq \mathbb{P}\{Y = 1\}$. Take $P_{\lambda_1}^{(t)}, P_{\lambda_2}^{(t)}$ to be two probability distribution on $\mathcal{X}$ parametrized in the following way:

$$P_{\lambda_1}^{(t)} = (\lambda_1 t, \lambda_1(1-t), r - \lambda_1, p_4, \ldots, p_n) \quad (20)$$

$$P_{\lambda_2}^{(t)} = (\lambda_2 t, \lambda_2(1-t), r - \lambda_2, p_4, \ldots, p_n), \quad (21)$$

where $r \triangleq 1 - \sum_{i\geq 4} p_i$, $t \in [0, 1]$, $0 \leq \lambda_1 < \lambda_2 \leq r$.

Taking $P_{X|1} \triangleq P_{\lambda_1}^{(t)}$, $P_{X|2} \triangleq P_{\lambda_2}^{(t)}$, it follows from Lemma 1 that

$$\begin{aligned}C(\ell, P_{XY}) = \alpha V(P_{\lambda_1}^{(t)}) + (1-\alpha) V(P_{\lambda_2}^{(t)}) \\ - V(\alpha P_{\lambda_1}^{(t)} + (1-\alpha) P_{\lambda_2}^{(t)}).\end{aligned} \quad (22)$$

Note that the following transformation $T(X)$ is a statistically sufficient transformation of X for Y:

$$T(X) = \begin{cases} x_1 & X \in \{x_1, x_2\}, \\ X & \text{otherwise}. \end{cases} \quad (23)$$

The Data Processing Axiom implies that for all $\alpha \in [0,1]$, $t \in [0,1]$ and legitimate $\lambda_2 > \lambda_1 \geq 0$,

$$\alpha V(P_{\lambda_1}^{(t)}) + (1-\alpha)V(P_{\lambda_2}^{(t)}) - V(\alpha P_{\lambda_1}^{(t)} + (1-\alpha)P_{\lambda_2}^{(t)}) = \alpha V(P_{\lambda_1}^{(1)}) + (1-\alpha)V(P_{\lambda_2}^{(1)}) - V(\alpha P_{\lambda_1}^{(1)} + (1-\alpha)P_{\lambda_2}^{(1)}). \quad (24)$$

We now define the function

$$R(\lambda, t) \triangleq V(P_\lambda^{(t)}), \quad (25)$$

where we note that the bi-variate function $R(\lambda, t)$ implicitly depends on the parameters $p_4, p_5, \ldots, p_n$, which we shall fix for the rest of this proof. Thus, $R(\lambda, t) = R(\lambda, t; p_4, p_5, \ldots, p_n)$.

Note that by definition,

$$R(\alpha\lambda_1 + (1-\alpha)\lambda_2, t) = V(\alpha P_{\lambda_1}^{(t)} + (1-\alpha)P_{\lambda_2}^{(t)}), \quad (26)$$

hence we know that

$$\alpha R(\lambda_1, t) + (1-\alpha)R(\lambda_2, t) - R(\alpha\lambda_1 + (1-\alpha)\lambda_2, t) = \alpha R(\lambda_1, 1) + (1-\alpha)R(\lambda_2, 1) - R(\alpha\lambda_1 + (1-\alpha)\lambda_2, 1). \quad (27)$$

Taking $\lambda_1 = 0, \lambda_2 = r = 1 - \sum_{i\geq 4} p_i$. We define $\tilde{R}(\lambda, t) \triangleq R(\lambda, t) - \lambda U(t)$, where

$$U(t) = \frac{R(r, t)}{r}. \quad (28)$$

It follows that

$$\tilde{R}(0, t) = V(P_0^{(t)}), \qquad \tilde{R}(r, t) = 0, \quad \forall t \in [0,1], \quad (29)$$

and we note that $V(P_0^{(t)})$ in fact does not depend on t.

With the help of (29), we plug $R(\lambda, t) = \tilde{R}(\lambda, t) + \lambda U(t)$ into (27), and obtain

$$\tilde{R}((1-\alpha)r, t) = \tilde{R}((1-\alpha)r, 1), \quad \forall \alpha \in [0,1], t \in [0,1]. \quad (30)$$

In other words, there exists a function $E : [0,1] \to \mathbb{R}$, such that

$$\tilde{R}(\lambda, t) = E(\lambda). \quad (31)$$

Since $R(\lambda, t) = \tilde{R}(\lambda, t) + \lambda U(t)$, we know that there exist real-valued functions E, U (indexed by $p_4, \ldots, p_n$) such that

$$R(\lambda, t) = \lambda U(t) + E(\lambda). \quad (32)$$

Expressing λ, t in terms of p_1, p_2, we have

$$\lambda = p_1 + p_2, \quad t = \frac{p_1}{p_1 + p_2}. \quad (33)$$

By definition of $R(\lambda, t)$, we can re-write (32) as

$$V(p_1, p_2, p_3, p_4, \ldots, p_n) = (p_1 + p_2) U\left(\frac{p_1}{p_1 + p_2}; p_4, \ldots, p_n\right) + E(p_1 + p_2; p_4, \ldots, p_n). \quad (34)$$

By Lemma 2, we know that there exists a symmetric (permutation invariant) finite convex function $G : \Gamma_n \to \mathbb{R}$, such that

$$G(P) = V(P) + \langle c, P \rangle. \quad (35)$$

In other words, we have proved that G is of the form

$$G(P) = (p_1 + p_2) U\left(\frac{p_1}{p_1 + p_2}; p_4, \ldots, p_n\right) + E(p_1 + p_2; p_4, \ldots, p_n) + \langle c, P \rangle. \quad (36)$$

For notational simplicity, we define

$$Y(p_1, p_2) \triangleq G(P), \quad (37)$$

where we again note that $Y(p_1, p_2; p_4, \ldots, p_n)$ is a bi-variate function parametrized by $p_4, p_5, \ldots, p_n$. This gives

$$Y(p_1, p_2) = (p_1 + p_2) U\left(\frac{p_1}{p_1 + p_2}\right) + E(p_1 + p_2) + c_1 p_1 + c_2 p_2 + c_3 (r - p_1 - p_2). \quad (38)$$

Since $G(P)$ is a symmetric function, we know that if we exchange p_1 and p_3 in $G(P)$, the value of $G(P)$ will not change. In other words, for $r = p_1 + p_2 + p_3$, we have

$$(r - p_3) U\left(\frac{p_1}{r - p_3}\right) + E(r - p_3) + c_1 p_1 + c_2 p_2 + c_3 p_3 = (r - p_1) U\left(\frac{p_3}{r - p_1}\right) + E(r - p_1) + c_1 p_3 + c_2 p_2 + c_3 p_1, \quad (39)$$

which is equivalent to

$$(r - p_3) U\left(\frac{p_1}{r - p_3}\right) + E(r - p_3) + (c_3 - c_1) p_3 = (r - p_1) U\left(\frac{p_3}{r - p_1}\right) + E(r - p_1) + (c_3 - c_1) p_1. \quad (40)$$

Defining $\tilde{E}(x) \triangleq E(r - x) + (c_3 - c_1)x$, we have

$$(r - p_3) U\left(\frac{p_1}{r - p_3}\right) + \tilde{E}(p_3) = (r - p_1) U\left(\frac{p_3}{r - p_1}\right) + \tilde{E}(p_1). \quad (41)$$

Interestingly, we can solve for general solutions of the above functional equation, which has connections to the so-called *fundamental equation of information theory*:

Lemma [50]–[52]: The most general measurable solution of

$$f(x) + (1 - x) g\left(\frac{y}{1 - x}\right) = h(y) + (1 - y) k\left(\frac{x}{1 - y}\right), \quad (42)$$

for $x, y \in [0, 1)$ with $x + y \in [0, 1]$, where $f, h : [0, 1) \to \mathbb{R}$ and $g, k : [0, 1] \to \mathbb{R}$, has the form

$$f(x) = aH_2(x) + b_1 x + d, \quad (43)$$

$$g(y) = aH_2(y) + b_2 y + b_1 - b_4, \quad (44)$$

$$h(x) = aH_2(x) + b_3 x + b_1 + b_2 - b_3 - b_4 + d, \quad (45)$$

$$k(y) = aH_2(y) + b_4 y + b_3 - b_2, \quad (46)$$

for $x \in [0, 1), y \in [0, 1]$, where $H_2(x) = -x \ln x - (1 - x) \ln(1 - x)$ is the binary Shannon entropy and a, b_1, b_2, b_3, b_4, and d are arbitrary constants.

Remark 1: If $f = g = h = k$ in (43)-(46), the corresponding functional equation is called the 'fundamental equation of information theory.'

In order to apply the above lemma to our setting, we define

$$q_i = p_i/r, \quad i = 1, 2, 3 \tag{47}$$

and $h(x) = \tilde{E}(rx)/r$. Then we know

$$(1-q_3)U\left(\frac{q_1}{1-q_3}\right) + h(q_3) = (1-p_1)U\left(\frac{q_3}{1-q_1}\right) + h(q_1). \tag{48}$$

Applying the general solution of (42), setting $f = h$, $g = k = U$, we have

$$b_1 = b_3, b_2 = b_4. \tag{49}$$

Thus,

$$h(x) = aH_2(x) + b_1 x + d, \tag{50}$$

$$U(y) = aH_2(y) + b_2 y + b_1 - b_2. \tag{51}$$

By the definition of $h(x)$ and $\tilde{E}(x)$, we have that

$$E(x) = raH_2(x/r) + (b_1 + c_1 - c_3)(r - x) + d. \tag{52}$$

Plugging the general solutions to $U(x)$, $E(x)$ into (38), and redefining the constants, we have

$$\begin{aligned} Y(p_1, p_2) &= A\Big(p_1 \ln p_1 + p_2 \ln p_2 + (r - p_1 - p_2)\ln(r - p_1 - p_2)\Big) \\ &\quad + Bp_1 + Cp_2 + D. \end{aligned} \tag{53}$$

Note that the constants A, B, C, D are functions of $p_4, \ldots, p_n$. Therefore, we have the following general representation of the symmetric function $G(P)$:

$$\begin{aligned} G(P) = {} & A(p_4, \ldots, p_n)\,(p_1 \ln p_1 + p_2 \ln p_2 + p_3 \ln p_3) \\ & + B(p_4, \ldots, p_n)p_1 + C(p_4, \ldots, p_n)p_2 \\ & + D(p_4, \ldots, p_n), \end{aligned} \tag{54}$$

where we have made the dependence on $p_4 \ldots p_n$ explicit. Now we utilize the property that $Y(p_1, p_2)$ is invariant to permutations. Exchanging p_1, p_2, we obtain that $B \equiv C$. Exchanging p_1, p_3, we obtain that $B \equiv C \equiv 0$. Doing an arbitrary permutation on $p_4, \ldots, p_n$, since p_1, p_2, p_3 enjoy two degrees of freedom, we know that $A(p_4, \ldots, p_n), D(p_4, \ldots, p_n)$ are symmetric functions.

Exchanging p_1, p_4 and comparing the coefficients for $p_2 \ln p_2$, we know that

$$A(p_4, p_5, \ldots, p_n) = A(p_1, p_5, \ldots, p_n), \tag{55}$$

since A is symmetric, and thus we can conclude that A is a constant. Now exchanging p_1, p_4 gives us

$$\begin{aligned} Ap_1 \ln p_1 - Ap_4 \ln p_4 = {} & D(p_1, p_5, \ldots, p_n) \\ & - D(p_4, p_5, \ldots, p_n). \end{aligned} \tag{56}$$

Taking partial derivatives with respect to p_1 (we vary p_2 simultaneously to ensure P still lies on the simplex) on both sides of (56), we obtain

$$A\,(\ln p_1 + 1) = \frac{\partial}{\partial p_1} D(p_1, p_5, \ldots, p_n). \tag{57}$$

Integrating on both sides with respect to p_1, we know there exists a function f such that

$$D(p_1, p_5, \ldots, p_n) = Ap_1 \ln p_1 + f(p_5, \ldots, p_n). \tag{58}$$

Since D is symmetric, we further know that

$$D(p_4, \ldots, p_n) = \sum_{i \geq 4} Ap_i \ln p_i. \tag{59}$$

To sum up, we have

$$G(P) = A \sum_{i=1}^{n} p_i \ln p_i. \tag{60}$$

To guarantee that $G(P)$ is convex, we need $A > 0$.

Plugging (60) into Lemma 2, the proof is complete.

B. The Case $n = 2$

Under the Data Processing Axiom, Lemma 2 implies the corresponding representation. On the other hand, for an arbitrary convex function G, the Savage representation of proper scoring rules [7] gives the construction of the corresponding loss function ℓ. Indeed, the Savage representation asserts, for a convex function G, we can define a loss function $\ell_G(x, Q) : \mathcal{X} \times \Gamma_n \to \bar{\mathbb{R}}$ by

$$\ell_G(x, Q) \triangleq \langle G'(Q), Q \rangle - G(Q) - G'_x(Q), \tag{61}$$

where $G'(Q)$ denotes a sub-gradient of $G(Q)$ at Q, and $G'_x(Q)$ is the component of $G'(Q)$ corresponding to $Q(x)$ (see, e.g., [7] for details). The loss function $\ell_G(x, Q)$ also satisfies

$$P \in \inf_{Q \in \Gamma_n} \mathbb{E}_P[\ell_G(X, Q)]. \tag{62}$$

Substituting loss function $\ell_G(x, Q)$ into (2) defines a valid $C(\ell, P_{XY})$. The proof is completed via noting that the only non-trivial statistically sufficient transform on a binary alphabet is permutation transform, and the function G is assumed to be invariant to permutations.

Appendix
Proof of Lemmas

A. Proof of Lemma 1

It follows from (3) that if we define

$$V(P) = -\inf_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{E}_P[\ell(X, \hat{x})], \tag{63}$$

then $V(P)$ cannot take values in $\{\infty, -\infty\}$.

Since $\mathbb{E}_P[\ell(X, \hat{x})]$ is linear in P, and $V(P)$ is the pointwise supremum over a family of linear functions of P, we know $V(P)$ is convex and lower semi-continuous on Γ_n.

Since Γ_n is a compact set, we know that the lower semi-continuous function $V(P)$ attains its minimum on Γ_n.

At the same time, since Γ_n is a polytope, we know $\forall P = (p_1, p_2, \ldots, p_n) \in \Gamma_n$, we have $P = \sum_{i=1}^{n} p_i \delta_i$, where $\delta_i = (0, 0, \ldots, 1, 0, \ldots, 0)$ is a distribution that puts mass one at symbol i.

Since $V(P)$ is convex, we have

$$V(P) = V(\sum_{i=1}^{n} p_i\delta_i) \leq \sum_{i=1}^{n} p_i V(\delta_i) \leq \max\{V(\delta_i), 1 \leq i \leq n\}. \quad (64)$$

That is to say, the function $V(P)$ attains its maximum at one of the boundary points δ_i. Thus, we know that $V(P)$ is bounded.

Now we proceed to show that

$$\inf_{\hat{X}(Y)} \mathbb{E}_P[\ell(X, \hat{X}(Y))] = -\sum_y P_Y(y)V(P_{X|Y=y}). \quad (65)$$

First, for any estimator $\hat{X}(Y)$, by the law of iterated expectation, we have

$$\mathbb{E}_P[\ell(X, \hat{X}(Y))] = \mathbb{E}_P[\mathbb{E}_P[\ell(X, \hat{X}(Y))|Y]] \quad (66)$$

$$\geq \mathbb{E}_P[-V(P_{X|Y=y})] \quad (67)$$

$$= -\sum_y P_Y(y)V(P_{X|Y=y}). \quad (68)$$

Hence,

$$\inf_{\hat{X}(Y)} \mathbb{E}_P[\ell(X, \hat{X}(Y))] \geq -\sum_y P_Y(y)V(P_{X|Y=y}). \quad (69)$$

Second, by the definition of infimum, for any $\epsilon > 0$, there exists an estimator $\hat{x}_\epsilon(y) \subset \hat{\mathcal{X}}$ such that

$$-V(P_{X|Y=y}) > \sum_{x\in\mathcal{X}} P_{X|Y=y}(x)\ell(x, \hat{x}_\epsilon(y)) - \epsilon. \quad (70)$$

Now define an estimator $\hat{X}(Y) = \hat{x}_\epsilon(Y)$. We have

$$\mathbb{E}_P[\ell(X, \hat{X}(Y))] = \mathbb{E}_P[\mathbb{E}_P[\ell(X, \hat{X}(Y))|Y]] \quad (71)$$

$$= \mathbb{E}_P[\mathbb{E}_P[\ell(X, \hat{x}_\epsilon(Y))|Y]] \quad (72)$$

$$< \mathbb{E}_P[-V(P_{X|Y=y}) + \epsilon] \quad (73)$$

$$= -\sum_y P_Y(y)V(P_{X|Y=y}) + \epsilon. \quad (74)$$

By the arbitrariness of ϵ we have

$$\inf_{\hat{X}(Y)} \mathbb{E}_P[\ell(X, \hat{X}(Y))] \leq -\sum_y P_Y(y)V(P_{X|Y=y}). \quad (75)$$

Combining it with (69), we know that (65) holds. The claim follows from plugging (63) and (65) into the definition of $C(\ell, P_{XY})$.

B. Proof of Lemma 2

By Lemma 1, we know there exists a convex function $V : \Gamma_n \to \mathbb{R}$, such that

$$C(\ell, P_{XY}) = \left(\sum_y P_Y(y)V(P_{X|Y=y})\right) - V(P_X). \quad (76)$$

Let $\delta_i \triangleq (0, 0, \ldots, 1, \ldots, 0)$ be a distribution in Γ_n that puts mass one on the i-th symbol of $\mathcal{X}$. Define $a_i \triangleq V(\delta_i)$. We know that $a_i \in \mathbb{R}, \forall i = 1, 2, \ldots, n$.

Define the convex function $G : \Gamma_n \to \mathbb{R}$ as

$$G(P) = V(P) - \sum_{i=1}^{n} a_i p_i. \quad (77)$$

Now it is easy to verify that $G(\delta_i) = 0, \forall i = 1, 2, \ldots, n$. After some algebra we can show that

$$C(\ell, P_{XY}) = \left(\sum_y P_Y(y)G(P_{X|Y=y})\right) - G(P_X). \quad (78)$$

Taking $Y \in \mathcal{X}$, and $P_Y = (p_1, p_2, \ldots, p_n)$ to be an arbitrary probability distribution. Setting $P_{X|Y=y} = \delta_y$, then we have

$$C(\ell, P_{XY}) = -G(P_X) = -G((p_1, p_2, \ldots, p_n)). \quad (79)$$

Define $T = \pi(X)$ to be a permutation of X, which is sufficient for Y. The Data Processing Axiom implies that

$$C(\ell, P_{XY}) = C(\ell, P_{TY}), \quad (80)$$

By construction, we have

$$C(\ell, P_{XY}) = -G((p_1, p_2, \ldots, p_n)), \quad (81)$$

$$C(\ell, P_{TY}) = -G((p_{\pi^{-1}(1)}, p_{\pi^{-1}(2)}, \ldots, p_{\pi^{-1}(n)})), \quad (82)$$

which implies that the function G is invariant to permutations. We take $c = -(a_1, a_2, \ldots, a_n)$ to finish the proof.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor, two anonymous reviewers, and Yanjun Han, whose comments have significantly improved the presentation of the paper.

REFERENCES

[1] I. J. Good, "Rational decisions," *J. Roy. Statist. Soc. B (Methodological)*, vol. 14, no. 1, pp. 107–114, 1952.
[2] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[3] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.
[4] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423 and 623–656, 1948.
[5] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *Ann. Math. Statist.*, vol. 33, no. 2, pp. 404–419, Jun. 1962. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177704567
[6] A. P. Dawid. (1998). "Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design," Dept. Statist. Sci., Univ. College London, London, U.K., Tech. Rep. 139. [Online]. Available: http://www.ucl.ac.uk/Stats/research/abs94.html
[7] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Statist. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.
[8] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–273, 2008.
[9] J. Aczél, B. Forte, and C. T. Ng, "Why the Shannon and Hartley entropies are 'natural,'" *Adv. Appl. Probab.*, vol. 6, no. 1, pp. 131–146, 1974.
[10] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, vol. 31. New York, NY, USA: Springer-Verlag, 1998.
[11] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.
[12] J. R. Quinlan, *C4.5: Programs for Machine Learning*, vol. 1. San Mateo, CA, USA: Morgan Kaufmann, 1993.
[13] S. Nowozin, "Improved information gain estimates for decision tree induction," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 297–304.

[14] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.

[15] C. Olsen, P. E. Meyer, and G. Bontempi, "On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information," *EURASIP J. Bioinformat. Syst. Biol.*, vol. 2009, Art. ID 308959, 2009.

[16] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[17] P. Viola and W. M. Wells, III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.

[18] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon, "Mutual information analysis: A comprehensive study," *J. Cryptol.*, vol. 24, no. 2, pp. 269–291, 2011.

[19] M. O. Hill, "Diversity and evenness: A unifying notation and its consequences," *Ecology*, vol. 54, no. 2, pp. 427–432, 1973.

[20] F. Franchini, A. R. Its, and V. E. Korepin, "Rényi entropy of the XY spin chain," *J. Phys. A, Math. Theoretical*, vol. 41, no. 2, p. 025302, 2008.

[21] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[22] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2835–2885, May 2015.

[23] G. Valiant and P. Valiant, "Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proc. 43rd Annu. ACM Symp. Theory Comput.*, 2011, pp. 685–694.

[24] G. Valiant and P. Valiant, "The power of linear estimators," in *Proc. IEEE 52nd Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2011, pp. 403–412.

[25] P. Valiant and G. Valiant, "Estimating the unseen: Improved estimators for entropy and other properties," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2157–2165.

[26] Y. Wu and P. Yang. (2014). "Minimax rates of entropy estimation on large alphabets via best polynomial approximation." [Online]. Available: http://arxiv.org/abs/1407.0381

[27] J. Jiao, K. Venkat, Y. Han, and T. Weissman. (Jun. 2014). "Maximum likelihood estimation of functionals of discrete distributions." [Online]. Available: http://arxiv.org/abs/1406.6959

[28] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi. (2014). "The complexity of estimating Rényi entropy." [Online]. Available: http://arxiv.org/abs/1408.1000v1

[29] J. Jiao, K. Venkat, Y. Han, and T. Weissman. (2014). "Beyond maximum likelihood: From theory to practice." [Online]. Available: http://arxiv.org/abs/1409.7458

[30] J. Jiao, T. A. Courtade, A. No, K. Venkat, and T. Weissman, "Information measures: The curious case of the binary alphabet," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7616–7626, Dec. 2014.

[31] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[32] C. W. J. Granger, "Testing for causality: A personal viewpoint," *J. Econ. Dyn. Control*, vol. 2, no. 1, pp. 329–352, 1980.

[33] C. A. Sims, "Money, income, and causality," *Amer. Econ. Rev.*, vol. 62, no. 4, pp. 540–552, 1972.

[34] Y. Hosoya, "On the Granger condition for non-causality," *Econometrica*, vol. 45, no. 7, pp. 1735–1736, 1977.

[35] G. Chamberlain, "The general equivalence of Granger and Sims causality," *Econometrica*, vol. 50, no. 3, pp. 569–581, 1982.

[36] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *J. Amer. Statist. Assoc.*, vol. 77, no. 378, pp. 304–313, 1982.

[37] J. F. Geweke, "Measures of conditional linear dependence and feedback between time series," *J. Amer. Statist. Assoc.*, vol. 79, no. 388, pp. 907–915, 1984.

[38] C. Gourieroux, A. Monfort, and E. Renault, "Kullback causality measures," *Ann. Econ. Statist.*, nos. 6–7, pp. 369–410, 1987.

[39] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Phys. Rev. Lett.*, vol. 103, no. 23, p. 238701, 2009.

[40] J. L. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl.*, Honolulu, HI, USA, Nov. 1990, pp. 303–305.

[41] H. Marko, "The bidirectional communication theory—A generalization of information theory," *IEEE Trans. Commum.*, vol. 21, no. 12, pp. 1345–1351, Dec. 1973.

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

[43] 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. 6, pp. 3248–3259, Jun. 2011.

[44] P.-O. Amblard and O. J. J. Michel, "The relation between Granger causality and directed information theory: A review," *Entropy*, vol. 15, no. 1, pp. 113–143, 2012.

[45] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6220–6242, Oct. 2013.

[46] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, p. 461, 2000.

[47] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, Swiss Federal Inst. Technol. Zürich, Zürich, Switzerland, 1998.

[48] A. P. Dawid, "The geometry of proper scoring rules," *Ann. Inst. Statist. Math.*, vol. 59, no. 1, pp. 77–93, 2007.

[49] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.

[50] P. Kannappan and C. T. Ng, "Measurable solutions of functional equations related to information theory," *Proc. Amer. Math. Soc.*, vol. 38, no. 2, pp. 303–310, 1973.

[51] G. Maksa, "Solution on the open triangle of the generalized fundamental equation of information with four unknown functions," *Utilitas Math.*, vol. 21, pp. 267–282, 1982.

[52] J. Aczél and C. T. Ng, "Determination of all semisymmetric recursive information measures of multiplicative type on n positive discrete probability distributions," *Linear Algebra Appl.*, vols. 52–53, pp. 1–30, Jul. 1983.

Jiantao Jiao (S'13) received his B.Eng. degree with the highest honor in Electronic Engineering from Tsinghua University, Beijing, China in 2012, and a M.S. degree in Electrical Engineering from Stanford University in 2014. He is currently working towards the Ph.D. degree in the Department of Electrical Engineering at Stanford University. He is a recipient of the Stanford Graduate Fellowship (SGF). His research interests include information theory and statistical signal processing, with applications in communication, control, computation, networking, data compression, and learning.

Thomas A. Courtade (S'06–M'13) is an Assistant Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. Prior to joining UC Berkeley in 2014, he was a postdoctoral fellow supported by the NSF Center for Science of Information. He received his Ph.D. and M.S. degrees from UCLA in 2012 and 2008, respectively, and he graduated summa cum laude with a B.Sc. in Electrical Engineering from Michigan Technological University in 2007.

His honors include a Distinguished Ph.D. Dissertation award and an Excellence in Teaching award from the UCLA Department of Electrical Engineering, and a Jack Keil Wolf Student Paper Award for the 2012 International Symposium on Information Theory.

Kartik Venkat (S'12) is a Ph.D. candidate in the Department of Electrical Engineering at Stanford University. His research interests include statistical inference, information theory, machine learning, and their applications in genomics, wireless networks, neuroscience, and quantitative finance. Kartik received a Bachelors degree in Electrical Engineering from the Indian Institute of Technology, Kanpur in 2010, and a Master's degree in Electrical Engineering from Stanford University in 2012. His honors include a Stanford Graduate Fellowship for Engineering and Sciences, the Numerical Technologies Founders Prize, and a Jack Keil Wolf ISIT Student Paper Award at the 2012 International Symposium on Information Theory.

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 currently 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.

Tsachys research is focused on information theory, compression, communication, statistical signal processing, the interplay between them, and their applications. He is recipient of several best paper awards, and prizes for excellence in research and teaching. He served on the editorial board of the IEEE TRANSACTIONS ON INFORMATION THEORY from Sept. 2010 to Aug. 2013, and currently serves on the editorial board of *Foundations and Trends in Communications and Information Theory*. He is Founding Director of the Stanford Compression Forum.