Justification of Logarithmic Loss via the Benefit of Side Information

Jiantao Jiao*, Thomas Courtade†, Kartik Venkat*, and Tsachy Weissman*
*Department of Electrical Engineering, Stanford University
†Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
Email: {jiantao, kvenkat, tsachy}@stanford.edu, courtade@eecs.berkeley.edu

***Abstract*—We consider a natural measure of the benefit of side information: the reduction in optimal estimation risk when side information is available to the estimator. When such a measure satisfies a natural data processing property, and the source alphabet has cardinality greater than two, we show that it is uniquely characterized by the optimal estimation risk under logarithmic loss, and the corresponding measure is equal to mutual information. Further, when the source alphabet is binary, we characterize the only admissible forms the measure of predictive benefit can assume. These results unify many causality measures in the literature as instantiations of directed information, and present a natural axiomatic characterization of mutual information without requiring the sum or recursivity property.**

I. Introduction

In statistical decision theory, it is often a controversial issue to choose the appropriate loss function in quantifying the risk for a given application. One popular loss function is the *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* $\ell_{\log}\colon \mathcal{X} \times \Gamma_n \to \bar{\mathbb{R}}$ *is defined by*

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

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 estimation risk under logarithmic loss. Recently, Courtade and Weissman[3] showed that the long-standing open problem of multiterminal source coding can be completely solved under logarithmic loss, which demonstrates the specialty of logarithmic loss in lossy source coding problems. The use of the logarithm in defining entropy arises due to its various axiomatic characterizations, the first of which dates back to Shannon [4].

This work was supported in part by the NSF Center for Science of Information under grant agreement CCF-0939370.

The main contribution of this paper is in providing fundamental justification for inference using logarithmic loss. In particular, we show that a single modest 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, Z *are jointly distributed random variables. How significant is the contribution of* Z *for inference on* X*?*

II. Problem Formulation and Main Results

Toward answering Question 1, let $\ell\colon \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$ be an arbitrary loss function with reproduction alphabet $\mathcal{Y}$, where $\mathcal{Y}$ is arbitrary. Given $(X, Z) \sim P_{XZ}$, where Z lies in an arbitrary measurable space, it's natural to quantify the benefit of additional side information Z by computing the difference between the expected losses in estimating $X \in \mathcal{X}$ with and without side information Z, respectively. This motivates the definition:

$$C(\ell, P_{XZ}) \triangleq \inf_{y_1 \in \mathcal{Y}} \mathbb{E}_P[\ell(X, y_1)] - \inf_{Y_2(Z)} \mathbb{E}_P[\ell(X, Y_2)], \tag{2}$$

where $y_1 \in \mathcal{Y}$ is deterministic, and $Y_2 = Y_2(Z) \in \mathcal{Y}$ is any measurable function of Z. The expectation is taken with respect to P_{XZ}. We require that indeterminate forms like $\infty - \infty$ do not appear in the definition of $C(\ell, P_{XZ})$. By taking Z to be independent of X, we obtain for all $P \in \Gamma_n$, $|\inf_{y_1 \in \mathcal{Y}} \mathbb{E}_P[\ell(X, y_1)]| < \infty$.

The formulation (2) has appeared previously in the statistics literature. In [5], Dawid defined the *coherent dependence function*, which is equivalent to (2), and used it to quantify the dependence between two random variables X, Z. Our framework of quantifying the predictive benefit of side information is also closely connected to the notion of proper scoring rules and the literature on probability forecasting in statistics. The survey by Gneiting and Raftery [6] provides a good overview.

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

978-1-4799-5186-4/14/$31.00 ©2014 IEEE

Data Processing Axiom. *For all distributions P_{XZ}, the quantity $C(\ell, P_{XZ})$ satisfies*

$$C(\ell, P_{TZ}) \leq C(\ell, P_{XZ})$$

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

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

$$T - X - Z, \quad X - T - Z \tag{3}$$

In other words, $T(X)$ is a lossless representation of all of the information X contains about Z.

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 Z, this would imply that there is *no* unique value which quantifies the benefit of side information Z for the random variable of interest. Thus, the Data Processing Axiom is needed for the benefit of side information to be well-defined.

Benign as the Data Processing Axiom, it has far-reaching implications for the form $C(\ell, P_{XZ})$ can take. This is captured by our first main result:

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

$$C(\ell, P_{XZ}) = I(X; Z), \tag{4}$$

up to a multiplicative factor.

The following corollary immediately follows from Theorem 1.

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

$$\inf_{Y_1(W)} \mathbb{E}_P[\ell(X, Y_1)] - \inf_{Y_2(Z,W)} \mathbb{E}_P[\ell(X, Y_2)], \tag{5}$$

is uniquely determined by the conditional mutual information,

$$I(X; Z|W), \tag{6}$$

up to a multiplicative factor.

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 that logarithmic loss plays.

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

Theorem 1 shows that mutual information is natural to measure the amount of reduction of statistical risk when we have side information. Incidentally, Erkip and Cover [7] argued that mutual information was 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.

It is worth mentioning that Theorem 1 is closely connected to existing results on axiomatic characterizations of information measures; see Csiszár [8] for a survey. To emphasize our contribution, we note that Csiszár [8] names only the axiomatic result of Acźel, Forte, and Ng [9] as a characterization of information measures as functions of the underlying probability distribution that requires neither recursivity nor the sum property. However, [9] focuses on entropy characterization, and the framework therein does not extend to the problem we consider.

Interestingly, the assumption that $n \geq 3$ in Theorem 1 is essential. The class of solutions for the binary alphabet setting is characterized by the following theorem.

Theorem 2. *Let $n = 2$. $C(\ell, P_{XZ})$ is of the form*

$$C(\ell, P_{XZ}) = \mathbb{E}_Z[G(P_{X|Z})] - G(P_X)$$

for a symmetric convex function $G((p, 1-p)) : \Gamma_2 \to \mathbb{R}$ if, and only if, the Data Processing Axiom holds.

We mention in passing that there is an interesting regime of observations surrounding the characterization of information measures, which is sensitive to the alphabet size being binary or larger. This phenomenon is explored in detail in [10].

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. Proof sketches of Theorems 1 and 2 are provided in Section IV. Details of proofs can be found in [11].

III. Causality Measures: An Axiomatic Viewpoint

Inferring causal relationships from observed data plays an indispensable part in scientific discovery. Granger, in his seminal work [12], 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 [13] 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 [14] introduced a related concept of causality, which was proved to be equivalent to Granger's definition in Sims [14], Hosoya [15], and Chamberlain [16] in a variety of settings.

Motivated by Granger's framework for testing causality using linear prediction, Geweke [17][18] proposed a causality measure to quantify the extent to which Y is causing X. Quoting Geweke (emphasis ours):

> *"The empirical literature abounds with tests of independence and unidirectional causality for various pairs of time series, but there have been virtually no investigations of the degree 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:

$$F_{Y\Rightarrow X} \triangleq \ln \frac{\sigma^2(X_t|X^{t-1})}{\sigma^2(X_t|X^{t-1},Y^{t-1})}, \tag{7}$$

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 [16] proved a general equivalence between Granger and Sims' causality tests by replacing linear predictors with conditional independence tests. However, the most natural generalization of (7) wasn't clear until Gourieroux, Monfort, and Renault [19] proposed the so-called *Kullback causality measures* in 1987. It is now well-known that Kullback causality measures are equivalent to (7) under linear Gaussian models (cf. Barnett, Barrett and Seth [20]).

Using information theoretic terms, Kullback causality measures are nothing but the directed information introduced by Massey [21], and motivated by Marko [22]. Using modern notation, the directed information from X^n to Y^n is defined as

$$I(X^n \to Y^n) \triangleq \sum_{i=1}^n I(X^i;Y_i|Y^{i-1}) \tag{8}$$

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

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

$$H(Y^n\|X^n) \triangleq \sum_{i=1}^n H(Y_i|Y^{i-1},X^i). \tag{10}$$

Massey and Massey [23] established the pleasing conservation law of directed information:

$$I(X^n;Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n) \tag{11}$$

$$= I(X^{n-1} \to Y^n) + I(Y^{n-1} \to X^n) + \sum_{i=1}^n I(X_i;Y_i|X^{i-1},Y^{i-1}), \tag{12}$$

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, Kim, and Weissman [24] showed various applications of directed information in portfolio theory, data compression, and hypothesis testing in the presence of causality constraints.

We remark that, for practical applications, the directed information between stochastic processes can be computed using the universal estimators proposed in [25], which exhibit optimal statistical and convergence properties.

Finally, we note that the notion of *transfer entropy* in the physics literature, which was proposed by Schreiber [26] 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 [27] between a pair of jointly stationary finite-alphabet processes X_t, Y_t can be written as:

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

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

Due to space constraints, we can only sketch the proof of Theorems 1 and 2 and highlight the key ideas. Complete details could be found in [11].

[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.

To begin, we need to put all estimation risk on the common footing by eliminating the arbitrary nature of the reconstruction alphabet $\mathcal{Y}$. 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*

$$C(\ell, P_{XZ}) = \left(\sum_z P_Z(z) V(P_{X|Z}) \right) - V(P_X). \tag{13}$$

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

$$V(P) = - \inf_{y \in \mathcal{Y}} \mathbb{E}_P[\ell(X, y)], \tag{14}$$

and its details could be found in [11]. In the statistics literature, the quantity $-V(P)$ is usually called the *generalized entropy* or *Bayes envelope*, and we refer to Dawid [28] for details.

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

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

$$C(\ell, P_{XZ}) = \left(\sum_z P_Z(z) G(P_{X|Z}) \right) - G(P_X), \tag{15}$$

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

$$G(P) = V(P) + \langle c, P \rangle, \tag{16}$$

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.

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

A. The case $n \geq 3$

It suffices to consider the constrained case when $Z \in \{1, 2\}$, and show that the Data Processing Axiom in this case mandates the usage of logarithmic loss.

Define $\alpha \triangleq P_Z(1)$. Take $P^{(t)}_{\lambda_1}, P^{(t)}_{\lambda_2}$ to be two probability vectors on $\mathcal{X}$ parametrized in the following way:

$$P^{(t)}_{\lambda_1} = (\lambda_1 t, \lambda_1(1-t), r - \lambda_1, p_4, \ldots, p_n) \tag{17}$$

$$P^{(t)}_{\lambda_2} = (\lambda_2 t, \lambda_2(1-t), r - \lambda_2, p_4, \ldots, p_n), \tag{18}$$

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

Taking $P_{X|1} \triangleq P^{(t)}_{\lambda_1}, P_{X|2} \triangleq P^{(t)}_{\lambda_2}$, we have

$$\begin{aligned} &C(\ell, P_{XZ}) \\ &= \alpha V(P^{(t)}_{\lambda_1}) + (1-\alpha) V(P^{(t)}_{\lambda_2}) - V(\alpha P^{(t)}_{\lambda_1} + (1-\alpha) P^{(t)}_{\lambda_2}). \end{aligned}$$

Note that for any $\alpha, t, \lambda_1, \lambda_2$, the following transformation is a sufficient statistic for Z.

$$T(X) = \begin{cases} x_1 & X \in \{x_1, x_2\} \\ X & \text{otherwise} \end{cases} \tag{19}$$

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

$$\begin{aligned} &\alpha V(P^{(t)}_{\lambda_1}) + (1-\alpha) V(P^{(t)}_{\lambda_2}) - V(\alpha P^{(t)}_{\lambda_1} + (1-\alpha) P^{(t)}_{\lambda_2}) \\ &= \alpha V(P^{(1)}_{\lambda_1}) + (1-\alpha) V(P^{(1)}_{\lambda_2}) \\ &\quad - V(\alpha P^{(1)}_{\lambda_1} + (1-\alpha) P^{(1)}_{\lambda_2}). \end{aligned} \tag{20}$$

Fixing $p_4, p_5, \ldots, p_n$, we define the function

$$R(\lambda, t; p_4, p_5, \ldots, p_n) \triangleq V(P^{(t)}_{\lambda}), \tag{21}$$

and further denote $R(\lambda, t; p_4, p_5, \ldots, p_n)$ by $R(\lambda, t)$ for simplicity.

Note that if we define $\tilde{R}(\lambda, t) \triangleq R(\lambda, t) - \lambda U(t) - F(t)$, where $U(t), F(t)$ are arbitrary real-valued functions, for all λ_1, λ_2, t we have

$$\begin{aligned} &\alpha R(\lambda_1, t) + (1-\alpha) R(\lambda_2, t) - R(\alpha\lambda_1 + (1-\alpha)\lambda_2, t) \\ &\quad = \alpha \tilde{R}(\lambda_1, t) + (1-\alpha) \tilde{R}(\lambda_2, t) - \tilde{R}(\alpha\lambda_1 + (1-\alpha)\lambda_2, t), \end{aligned}$$

which, recalling (20), implies that

$$\begin{aligned} &\alpha \tilde{R}(\lambda_1, t) + (1-\alpha) \tilde{R}(\lambda_2, t) - \tilde{R}(\alpha\lambda_1 + (1-\alpha)\lambda_2, t) \\ &= \alpha \tilde{R}(\lambda_1, 1) + (1-\alpha) \tilde{R}(\lambda_2, 1) \\ &\quad - \tilde{R}(\alpha\lambda_1 + (1-\alpha)\lambda_2, 1). \end{aligned} \tag{22}$$

Taking $\lambda_1 = 0$ and $\lambda_2 = r = 1 - \sum_{i \geq 4} p_i$, we can choose the functions $U(t), F(t)$ in a way such that

$$\tilde{R}(0, t) = A(p_4, \ldots, p_n), \quad \tilde{R}(r, t) = B(p_4, \ldots, p_n), \tag{23}$$

where A, B are some functions of $(p_4, \ldots, p_n)$.

Plugging (23) into (22), we know that

$$\begin{aligned} &\alpha A(p_4, \ldots, p_n) + (1-\alpha) B(p_4, \ldots, p_n) - \tilde{R}((1-\alpha)r, t) \\ &= \alpha A(p_4, \ldots, p_n) + (1-\alpha) B(p_4, \ldots, p_n) - \tilde{R}((1-\alpha)r, 1), \end{aligned}$$

which implies that

$$\tilde{R}((1-\alpha)r, t) = \tilde{R}((1-\alpha)r, 1), \ \forall \alpha \in [0, 1], t \in [0, 1]. \tag{24}$$

In other words, there exists a function $E : [0, 1] \to \mathbb{R}$, such that $\tilde{R}(\lambda, t) = E(\lambda)$. Therefore, expressing λ, t in terms of p_1, p_2, and taking $p_1 = xa$, $p_2 = x(1-a)$, $a \in [0, 1]$, $p_3 = 1 - (\sum_{i \geq 4} p_i) - x$ yields the expression

$$V(P) = E(x) + xU(a) + F(a), \tag{25}$$

where $P = (p_1, p_2, \ldots, p_n)$. Now, we recall the following lemma.

Lemma 3 (Gale-Klee-Rockafellar [29])**.** *If D is boundedly polyhedral and ϕ is a convex function on the relative interior of D which is bounded on bounded sets, then ϕ has a unique continuous covex extension on D.*

Lemma 3 implies that we must have $F \equiv \mathsf{const}$ in order to extend $V(P)$ to a continuous function on Γ_n, else the limit of $V(P)$ as $x \to 0$ in (25) depends on how we approach the boundary. Without loss of generality, we take $F \equiv 0$.

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

$$(r-p_3)U\left(\frac{p_1}{r-p_3}\right) + E(r-p_3) + (c_3-c_1)p_3$$
$$= (r-p_1)U\left(\frac{p_3}{r-p_1}\right) + E(r-p_1) + (c_3-c_1)p_1. \quad (26)$$

Lemma 4 ([30]). *Any measurable solution of*

$$f(x)+(1-x)g\left(\frac{y}{1-x}\right) = h(y)+(1-y)k\left(\frac{x}{1-y}\right), \quad (27)$$

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

$$f(x) = aH_2(x) + b_1 x + d, \quad (28)$$
$$g(y) = aH_2(y) + b_2 y + b_1 - b_4, \quad (29)$$
$$h(x) = aH_2(x) + b_3 x + b_1 + b_2 - b_3 - b_4 + d, \quad (30)$$
$$k(y) = aH_2(y) + b_4 y + b_3 - b_2, \quad (31)$$

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.

After significant algebraic manipulation (which is omitted due to space constraints), it follows from Lemma 4 and (26) that

$$G(P) = A\sum_{i=1}^{n} p_i \ln p_i + \mathsf{const}, \quad (32)$$

where A is a positive constant. Plugging (32) into Lemma 2 completes the proof for $n \geq 3$.

B. The case $n = 2$

The 'if' part of Theorem 2 follows from Lemma 2. Savage's representation of proper scoring rules [6] gives the 'only if' direction. In particular, 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

$$\ell_G(x,Q) \triangleq \langle G'(Q), Q\rangle - G(Q) - G'_x(Q), \quad (33)$$

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., [6] for details). The loss function $\ell_G(x, Q)$ also satisfies

$$P \in \inf_{Q\in\Gamma_n} \mathbb{E}_P[\ell_G(X,Q)]. \quad (34)$$

Substituting loss function $\ell_G(x, Q)$ into (2) defines a valid $C(\ell, P_{XZ})$.

References

[1] I. J. Good, "Rational decisions," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 107–114, 1952.

[2] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[3] T. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *Information Theory, IEEE Transactions on*, vol. 60, no. 1, pp. 740–761, 2014.

[4] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[5] A. P. Dawid, "Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design," Department of Statistical Science, University College London. http://www.ucl.ac.uk/Stats/research/abs94.html, Tech. Rep. 139, 1998.

[6] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[7] E. Erkip and T. M. Cover, "The efficiency of investment information," *Information Theory, IEEE Transactions on*, vol. 44, no. 3, pp. 1026–1040, 1998.

[8] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–273, 2008.

[9] J. Aczél, B. Forte, and C. Ng, "Why the shannon and hartley entropies are'natural'," *Advances in Applied Probability*, pp. 131–146, 1974.

[10] J. Jiao, T. Courtade, A. No, K. Venkat, and T. Weissman, "Information measures: the curious case of the binary alphabet," *submitted to IEEE Transactions on Information Theory*.

[11] J. Jiao, T. Courtade, K. Venkat, and T. Weissman, "Justification of logarithmic loss via the benefit of side information," *arXiv preprint arXiv:1403.4679*, 2014.

[12] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[13] ——, "Testing for causality: a personal viewpoint," *Journal of Economic Dynamics and control*, vol. 2, pp. 329–352, 1980.

[14] C. A. Sims, "Money, income, and causality," *The American Economic Review*, vol. 62, no. 4, pp. 540–552, 1972.

[15] Y. Hosoya, "On the granger condition for non-causality," *Econometrica*, vol. 45, no. 7, pp. 1735–36, 1977.

[16] G. Chamberlain, "The general equivalence of granger and sims causality," *Econometrica: Journal of the Econometric Soc.*, pp. 569–581, 1982.

[17] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American Statistical Association*, vol. 77, no. 378, pp. 304–313, 1982.

[18] J. F. Geweke, "Measures of conditional linear dependence and feedback between time series," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 907–915, 1984.

[19] C. Gourieroux, A. Monfort, and E. Renault, "Kullback causality measures," *Annales d'Economie et de Statistique*, pp. 369–410, 1987.

[20] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for gaussian variables," *Physical review letters*, vol. 103, no. 23, p. 238701, 2009.

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

[22] H. Marko, "The bidirectional communication theory–a generalization of information theory," *IEEE Trans. Comm.*, vol. 21, pp. 1345–1351, 1973.

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

[24] H. H. Permuter, Y.-H. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *Information Theory, IEEE Transactions on*, vol. 57, no. 6, pp. 3248–3259, 2011.

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

[26] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[27] G. Kramer, *Directed Information for Channels with Feedback*. Konstanz: Hartung-Gorre Verlag, 1998, Dr. sc. thchn. Dissertation, Swiss Federal Institute of Technology (ETH) Zurich.

[28] A. P. Dawid, "The geometry of proper scoring rules," *Annals of the Institute of Statistical Mathematics*, vol. 59, no. 1, pp. 77–93, 2007.

[29] D. Gale, V. Klee, and R. Rockafellar, "Convex functions on convex polytopes," *Proceedings of the American Mathematical Society*, vol. 19, no. 4, pp. 867–873, 1968.

[30] P. Kannappan and C. Ng, "Measurable solutions of functional equations related to information theory," *Proceedings of the American Mathematical Society*, pp. 303–310, 1973.