Information Divergences and the Curious Case of the Binary Alphabet

Jiantao Jiao*, Thomas Courtade†, Albert No*, Kartik Venkat*, and Tsachy Weissman*
*Department of Electrical Engineering, Stanford University; †EECS Department, University of California, Berkeley
Email: {jiantao, albertno, kvenkat, tsachy}@stanford.edu, courtade@eecs.berkeley.edu

***Abstract*—Four problems related to information divergence measures defined on finite alphabets are considered. In three of the cases we consider, we illustrate a contrast which arises between the binary-alphabet and larger-alphabet settings. This is surprising in some instances, since characterizations for the larger-alphabet settings do not generalize their binary-alphabet counterparts. For example, we show that f-divergences are not the unique decomposable divergences on binary alphabets that satisfy the data processing inequality, despite contrary claims in the literature.**

I. Introduction

Divergence measures play a central role in information theory and other branches of mathematics. Many special classes of divergences, such as Bregman divergences, f-divergences, and Kullback-Liebler-type f-distance measures, enjoy various properties which make them particularly useful in problems related to learning, clustering, inference, optimization, and quantization, to name a few. In this paper, we investigate the relationships between these three classes of divergences, each of which will be defined formally in due course, and the subclasses of divergences which satisfy desirable properties such as monotonicity with respect to data processing. Roughly speaking, we address the following four questions:

QUESTION 1: If a decomposable divergence satisfies the data processing inequality, must it be an f-divergence?

QUESTION 2: Is Kullback-Leibler (KL) divergence the unique KL-type f-distance measure which satisfies the data processing inequality?

QUESTION 3: Is KL divergence the unique Bregman divergence which is invariant to statistically sufficient transformations of the data?

QUESTION 4: Is KL divergence the unique Bregman divergence which is also an f-divergence?

Of the above four questions, only QUESTION 4 has an affirmative answer (despite indications to the contrary having appeared in the literature; a point which we address further in Section III-A). However, this assertion is slightly deceptive. Indeed, if the alphabet size n is at least 3, then all four questions can be answered in the affirmative. Thus, counterexamples only arise in the binary setting when $n = 2$.

This is perhaps unexpected. Intuitively, the data processing inequality is not a stringent requirement. In this sense, the answers to the above series of questions imply that the class of "interesting" divergence measures can be very small when $n \geq 3$ (e.g., restricted to the class of f-divergences, or a multiple of KL divergence). However, in the binary alphabet setting, the class of "interesting" divergence measures is strikingly rich, a point which will be emphasized in our results. In many ways, this richness is surprising since the binary alphabet is usually viewed as the simplest setting one can consider. In particular, we would expect that the class of interesting divergence measures corresponding to binary alphabets would be less rich than the counterpart class for larger alphabets. However, as we will see, the opposite is true.

The observation of this dichotomy between binary and larger alphabets is not without precedent. For example, Fischer proved the following in his 1972 paper [1].

Theorem 1 *Suppose $n \geq 3$. If, and only if, f satisfies*

$$\sum_{k=1}^{n} p_k f(p_k) \leq \sum_{k=1}^{n} p_k f(q_k) \tag{1}$$

for all probability distributions $P = (p_1, p_2, \ldots, p_n), Q = (q_1, q_2, \ldots, q_n)$, then it is of the form

$$f(p) = c \log p + b \quad \text{for all } p \in (0,1), \tag{2}$$

where b and $c \leq 0$ are constants.

As implied by his supposition that $n \geq 3$ in Theorem 1, Fischer observed and appreciated the distinction between the binary and larger alphabet settings when considering so-called *Shannon-type* inequalities of the form (1). Indeed, in the same paper [1], Fischer gave the following result:

Theorem 2 *The functions of the form*

$$f(q) = \int \frac{G(q)}{q} dq, \quad q \in (0,1), \tag{3}$$

with G arbitrary, nonpositive, and satisfying $G(1-q) = G(q)$ for $q \in (0,1)$, are the only absolutely continuous functions[1] on $(0,1)$ satisfying (1) when $n = 2$.

Only in the special case where G is taken to be constant in (3), do we find that f is of the form (2). We direct the interested reader to [2, Chap. 4] for a detailed discussion.

In part, the present paper was inspired and motivated by

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

[1]There also exist functions f on $(0,1)$ which are not absolutely continuous and satisfy (1). See [1].

Theorems 1 and 2, and the distinction they draw between binary and larger alphabets. Indeed, the answers to questions QUESTION 1 – QUESTION 3 are in the same spirit as Fischer's results. For instance, our answer to question QUESTION 2 demonstrates that the functional inequality (1) *and* a data processing requirement are still not enough to demand f take the form (2) when $n = 2$. To wit, we prove an analog of Theorem 2 for this setting (see Section III-B).

This paper is organized as follows. In Section II, we recall several important classes of divergence measures and define what it means for a divergence measure to satisfy the data processing, sufficiency, or separability properties. In Section III, we investigate each of the questions posed above and state our main results. The Appendices contain all proofs.

II. PRELIMINARIES: DIVERGENCES, DATA PROCESSING, AND SUFFICIENCY PROPERTIES

Let $\bar{\mathbb{R}} \triangleq [-\infty, +\infty]$ denote the extended real line. Throughout this paper, we only consider finite alphabets. To this end, let $\mathcal{X} = \{1, 2, \ldots, n\}$ denote the alphabet, which is of size n, and let $\Gamma_n = \{(p_1, p_2, \ldots, p_n) : \sum_{i=1}^n p_i = 1, p_i \geq 0, i = 1, 2, \ldots, n\}$ be the set of probability measures on $\mathcal{X}$, with $\Gamma_n^+ = \{(p_1, p_2, \ldots, p_n) : \sum_{i=1}^n p_i = 1, p_i > 0, i = 1, 2, \ldots, n\}$ denoting its relative interior.

We refer to a nonnegative function $D : \Gamma_n \times \Gamma_n \to \bar{\mathbb{R}}$ simply as a divergence function (or, divergence measure). Of course, essentially all common divergences – including Bregman and f-divergences, which are defined shortly – fall into this general class. In this paper, we are primarily interested in divergence measures which satisfy either of two properties: the *data processing property* and the *sufficiency property*.

In the course of defining these properties, we will consider (possibly stochastic) transformations $P_{Y|X} : X \mapsto Y$, where $Y \in \mathcal{X}$. That is, $P_{Y|X}$ is a Markov kernel with source and target both equal to $\mathcal{X}$ (equipped with the discrete σ-algebra). If $X \sim P_X$, we will write $P_X \to P_{Y|X} \to P_Y$ to denote that P_Y is the marginal distribution of Y generated by passing X through the channel $P_{Y|X}$. That is, $P_Y(\cdot) \triangleq \sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(\cdot|x)$.

Now, we are in a position to formally define the *data processing property*.

Definition 1 (Data Processing) *A divergence function D satisfies the data processing property if, for all $P_X, Q_X \in \Gamma_n$, we have*

$$D\left(P_X; Q_X\right) \geq D\left(P_Y; Q_Y\right) \tag{4}$$

for any transformation $P_{Y|X} : X \mapsto Y$, where P_Y and Q_Y are defined via $P_X \to P_{Y|X} \to P_Y$ and $Q_X \to P_{Y|X} \to Q_Y$, respectively.

A weaker version of the data processing inequality is the *sufficiency property*. For two arbitrary distributions P, Q, we define a joint distribution P_{XZ}, $Z \in \{1, 2\}$, such that

$$P_{X|1} = P, \quad P_{X|2} = Q. \tag{5}$$

A transformation $P_{Y|X} : X \mapsto Y$ is said to be a *sufficient transformation of X for Z* if Y is a sufficient statistic of X for Z. We remind the reader that Y is a sufficient statistic of X for Z if the following two Markov chains hold:

$$Z - X - Y \qquad\qquad Z - Y - X. \tag{6}$$

Definition 2 (Sufficiency) *A divergence function D satisfies the sufficiency property if, for all $P_{X|1}, P_{X|2} \in \Gamma_n$ and $Z \in \{1, 2\}$, we have*

$$D(P_{X|1}; P_{X|2}) \geq D\left(P_{Y|1}; P_{Y|2}\right) \tag{7}$$

for any sufficient transformation $P_{Y|X} : X \mapsto Y$ of X for Z, where $P_{Y|z}$ is defined by $P_{X|z} \to P_{Y|X} \to P_{Y|z}$ for $z \in \{1, 2\}$.

Clearly the sufficiency property is weaker than the data processing property because our attention is restricted to only those (possibly stochastic) transformations $P_{Y|X}$ for which Y is a sufficient statistic of X for Z. Given the definition of a sufficient statistic, we note that the inequality in (7) can be replaced with equality to yield an equivalent definition.

Henceforth, we will simply say that a divergence function $D(\cdot;\cdot)$ satisfies DATA PROCESSING when it satisfies the data processing property. Similarly, we say that a divergence function $D(\cdot;\cdot)$ satisfies SUFFICIENCY when it satisfies the sufficiency property.

Remark 1 *In defining both* DATA PROCESSING *and* SUFFICIENCY*, we have required that $Y \in \mathcal{X}$. This is necessary since the divergence function $D(\cdot;\cdot)$ is only defined on $\Gamma_n \times \Gamma_n$.*

We make one more definition following [3].

Definition 3 (Decomposibility) *A divergence function D is said to be decomposable (or, separable) if there exists a bivariate function $\delta(u, v) : [0, 1]^2 \to \bar{\mathbb{R}}$ such that*

$$D(P; Q) = \sum_{i=1}^n \delta(p_i, q_i) \tag{8}$$

for all $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$ in Γ_n.

Having defined divergences in general, we will now recall three important classes of divergences which will be of interest to us: Bregman divergences, f-divergences, and KL-type divergences.

A. Bregman Divergences

Let $G(P) : \Gamma_n \to \mathbb{R}$ be a convex function defined on Γ_n, differentiable on Γ_n^+. For two probability measures $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$ in Γ_n, the Bregman divergence generated by G is defined by

$$D^G(P; Q) \triangleq G(P) - G(Q) - \langle \nabla G(Q), P - Q \rangle, \tag{9}$$

where $\langle \nabla G(Q), P - Q \rangle$ denotes the standard inner product between $\nabla G(Q)$ and $(P - Q)$ (interpreted as vectors in $\mathbb{R}^n$). Note that Bregman divergences can also be defined over domains other than Γ_n.

B. f-Divergences

Csiszár [4], and independently Ali and Silvey [5], introduced the notion of f-divergences, which take the form

$$D_f(P;Q) \triangleq \sum_{i=1}^{n} p_i f\left(\frac{q_i}{p_i}\right), \tag{10}$$

where f is a convex function with $f(1) = 0$. Examples of f-divergences include Kullback–Leibler divergence, Hellinger distance, and total variation distance. All f-divergences are decomposable by definition. Many important properties of f-divergences can be found in [6] and references therein.

C. Kullback-Leibler-type f-distance measures

A Kullback-Leibler-type f-distance measure (or, KL-type f-distance measure) [7] takes the form

$$L(P;Q) = \sum_{k=1}^{n} p_k\Big(f(q_k) - f(p_k)\Big) \geq 0. \tag{11}$$

If a particular divergence $L(P;Q)$ is defined by (11) for a given f, we say that f generates $L(P;Q)$. Theorems 1 and 2 characterize all permissible functions f which generate KL-type f-distance measures. As with f-divergences, KL-type f-distance measures are decomposable by definition.

III. Main Results

In this section, we address each of the questions posed in the introduction.

A. Question 1: *Are f-Divergences the unique decomposable divergences which satisfy* Data Processing*?*

Recall that a decomposable divergence D takes the form given in (8). Theorem 1 in [3] asserts that any decomposable divergence which satisfies Data Processing must be an f-divergence. However, the proof of [3, Theorem 1] only works when $n \geq 3$, a fact which apparently went unnoticed[2]. The proof of the same claim in [9, Appendix] suffers from a similar flaw and also fails when $n = 2$. Of course, knowing that the assertion holds for $n \geq 3$, it is natural to expect that it also must hold for $n = 2$. As it turns out, this is not true. In fact, counterexamples exist in great abundance.

To this end, take any f-divergence $D_f(P;Q)$ and let $k : \mathbb{R} \to \mathbb{R}$ be an arbitrary nondecreasing function, such that $k(0) = 0$. Since all f-divergences satisfy Data Processing (cf. [6] and references therein), the divergence function $\tilde{D}(P;Q) \triangleq k\left(D_f(P;Q)\right)$ must also satisfy Data Processing. [3] It turns out that the divergence function $\tilde{D}(P;Q)$ is also decomposable in the binary case, which follows immediately from decomposability of f-divergences and the following lemma.

[2]The propagation of this error has led to other claims which are incorrect in the binary alphabet setting (cf. [8, Theorem 2]).

[3]However, divergences of the form $k\left(D_f(P;Q)\right)$ do not constitute all of the decomposable divergences satisfying Data Processing on binary alphabet. A simple counterexample would be $k_1(D_{f_1}(P;Q)) + k_2(D_{f_2}(P;Q))$, where k_1, k_2 are two different non-decreasing functions satisfying $k_i(0) = 0$, $i = 1, 2$.

Lemma 1 *A divergence function D on the binary alphabet is decomposable if and only if*

$$D((p, 1-p);(q, 1-q)) = D((1-p, p);(1-q, q)). \tag{12}$$

Therefore, if $\tilde{D}(P;Q)$ is not itself an f-divergence, we can conclude that $\tilde{D}(P;Q)$ constitutes a counterexample to [3, Theorem 1] for the binary case. Indeed, $\tilde{D}(P;Q)$ is generally not an f-divergence. For example, if $f(x) = |x-1|$, the generated f-divergence in the binary case reduces to

$$D_f\Big((p, 1-p);(q, 1-q)\Big) = 2|p-q|. \tag{13}$$

Letting $k(x) = x^2$, we have

$$\tilde{D}(P;Q) = 4(p-q)^2 = \delta(p,q) + \delta(1-p, 1-q), \tag{14}$$

where $\delta(p,q) = 2(p-q)^2$. Since $\tilde{D}(P;Q) = 4(p-q)^2$ is a Bregman divergence, we will see later in Theorem 5 that it cannot also be an f-divergence because it is not proportional to KL-divergence[4]. Thus, the answer to Question 1 is negative for the case $n = 2$. As mentioned above, [3, Theorem 1] implies the answer is affirmative when $n \geq 3$.

B. Question 2: *Is KL divergence the only KL-type f-distance measure which satisfies* Data Processing*?*

Recall from Section II that a KL-type f-distance measure takes the form given in (11). As alluded to in the introduction, there is a dichotomy between KL-type f-distance measures on binary alphabets, and those on larger alphabets. In particular, we have the following result:

Theorem 3 *If $L(P;Q)$ is a Kullback-Leibler-type f-distance measure which satisfies* Data Processing*, then*

1) *If $n \geq 3$, $L(P;Q)$ is equal to KL divergence up to a nonnegative multiplicative factor;*
2) *If $n = 2$ and the function $f(x)$ that generates $L(P;Q)$ is continuously differentiable, then $f(x)$ is of the form*

$$f(x) = \int \frac{G(x)}{x} dx, \quad \text{for } x \in (0,1), \tag{15}$$

where $G(x)$ satisfies the following properties:

   a) *$xG(x) = (x-1)h(x)$ for $x \in (0, 1/2]$ and some nonnegative, nondecreasing continuous function $h(x)$.*
   b) *$G(x) = G(1-x)$ for $x \in [1/2, 1)$.*

Conversely, any nonnegative, non-decreasing continuous function $h(x)$ generates a KL-type divergence in the manner described above which satisfies Data Processing.

To illustrate the last claim of Theorem 3, take for example $h(x) = x^2, x \in [0, 1/2]$. In this case, we obtain

$$f(x) = \frac{1}{2}x^2 - x + C, \quad \forall x \in [0,1], \tag{16}$$

[4]The differentiability hypotheses of Theorem 5 are satisfied in this case due to the smoothness of $4(p-q)^2$ and convexity of the function f defining an f-divergence.

where C is a constant. Letting $P = (p, 1-p), Q = (q, 1-q)$, and plugging (16) into (11), we obtain the KL-type divergence $L(P;Q) = (p-q)^2/2$, which satisfies DATA PROCESSING, but certainly is not proportional to KL divergence. Thus, the answer to question QUESTION 2 is negative.

At this point it is instructive to compare with the discussion on question QUESTION 1. In Section III-A, we showed that a divergence which is decomposable and satisfies DATA PROCESSING is not necessarily an f-divergence when the alphabet is binary (contrary to the assertion of [3, Theorem 1]). From the above example, we see that the much stronger hypothesis – that a divergence is a Kullback-Leibler-type f-distance measure which satisfies DATA PROCESSING – still does not necessitate an f-divergence in the binary setting.

C. QUESTION 3: *Is KL divergence the unique Bregman divergence which satisfies* SUFFICIENCY*?*

In this section, we investigate whether KL divergence is the unique Bregman divergence that satisfies SUFFICIENCY. Again, the answer to this is affirmative for $n \geq 3$, but negative in the binary case. This is captured by the following theorem.

Theorem 4 *If $D^G(P;Q)$ is a Bregman divergence which satisfies* SUFFICIENCY *and*

1) *$n \geq 3$, then $D^G(P;Q)$ is equal to the KL divergence up to a nonnegative multiplicative factor;*
2) *$n = 2$, then $D^G(P;Q)$ can be any Bregman divergence generated by a symmetric bivariate convex function $G(P)$ defined on Γ_2.*

The first part of Theorem 4 is possibly surprising, since we do not assume the Bregman divergence $D^G(P;Q)$ to be decomposable a priori. In an informal statement immediately following Theorem 2 in [8], a claim similar to first part of our Theorem 4 was proposed. However, we are the first to give a complete proof of this result, as no proof was previously known [10].

We have already seen an example of a Bregman divergence which satisfies DATA PROCESSING (and therefore SUFFICIENCY) in our previous examples. Letting $P = (p, 1-p), Q = (q, 1-q)$ and defining $G(P) = p^2 + (1-p)^2$ generates the Bregman divergence

$$D^G(P;Q) = 2(p-q)^2. \tag{17}$$

The second part of Theorem 4 characterizes all Bregman divergences on binary alphabets which satisfy SUFFICIENCY as being in precise correspondence with the set of symmetric bivariate convex functions defined on Γ_2.

D. QUESTION 4: *Is KL divergence the unique Bregman divergence which is also an f-divergence?*

We conclude our investigation by asking whether Kullback–Leibler divergence is the unique divergence which is both a Bregman divergence and an f-divergence. The first result of this kind was proved in [11] in the context of linear inverse problems requiring an alphabet size of $n \geq 5$, and is hard to extract as an independent result. Amari has also addressed question QUESTION 4 in [9], in which he studied divergences defined on the space of nonnegative measures. Amari showed that α-divergences (see [9] for a definition) are the unique divergences lying in the intersection of *decomposable* Bregman divergences and f-divergences defined on the space of positive measures. Using methods from information geometry, Amari showed that the uniqueness of α-divergences implies that KL divergence is the unique divergence lying in the intersection of decomposable Bregman divergences and f-divergences defined on the space of probability measures. Here, we give an elementary proof of this result without restricting our attention to decomposable Bregman divergences. Our main result in this context is the following:

Theorem 5 *Suppose $D(P;Q)$ is both a Bregman divergence and an f-divergence for some $n \geq 2$. If $f'(x)$, and $f''(1)$ exist, then $D(P;Q)$ is equal to KL divergence up to a nonnegative multiplicative factor.*

Remark 2 *In the following appendices, we provide very brief sketches of all the proofs due to space constraints. The interested reader is referred to an extended version [12] for complete details.*

APPENDIX A
PROOF OF THEOREM 3

It suffices to consider $n = 2$, since the first part of Theorem 3 follows from Theorem 1. In this case, any transform can be represented by the following channel:

$$P_{Y|X}(1|1) = \alpha \quad P_{Y|X}(2|1) = 1-\alpha; \tag{18}$$
$$P_{Y|X}(1|2) = \beta \quad P_{Y|X}(2|2) = 1-\beta, \tag{19}$$

where $\alpha, \beta \in [0,1]$. Define $\tilde{p} = p\alpha + \beta(1-p), \tilde{q} = q\alpha + \beta(1-q)$. The data processing property asserts that

$$\begin{aligned} &p\left(f(q) - f(p)\right) + (1-p)\left(f(1-q) - f(1-p)\right) \\ &\quad \geq \tilde{p}\left(f(\tilde{q}) - f(\tilde{p})\right) + (1-\tilde{p})\left(f(1-\tilde{q}) - f(1-\tilde{p})\right) \end{aligned} \tag{20}$$

for all $p, q, \alpha, \beta \in [0,1]$. Theorem 2 implies that $f'(p) = G(p)/p, G(p) \leq 0, G(p) = G(1-p)$ for $p \in (0,1)$.

Fixing α, β, p, we know LHS minus RHS of (20) achieves its minimum (i.e., zero) at $q = p, \forall \alpha, \beta$. Hence, $\exists \delta > 0$ for which we can take derivatives w.r.t. q on both sides of (20), plug in $f'(p) = G(p)/p, G(p) = G(1-p)$, and conclude that

$$\frac{G(q)}{G(\tilde{q})} \geq \frac{(\alpha-\beta)^2 q(1-q)}{\tilde{q}(1-\tilde{q})} \quad \text{for all } q \in (p, p+\delta). \tag{21}$$

Since (21) does not depend on p, it must hold for all α, β, q. Eliminating α, β, and exploiting symmetry of $G(q)$, we obtain

$$G(q)\frac{q}{1-q} \leq G(\tilde{q})\frac{\tilde{q}}{1-\tilde{q}}, \ 0 \leq q \leq 1/2, \ 0 \leq \tilde{q} \leq q. \tag{22}$$

Our claims follow from this inequality.

APPENDIX B
PROOF OF THEOREM 4

The roadmap for case $n \geq 3$ is as follows. First we show SUFFICIENCY implies the convex function $G(P)$ has the form

$$G(P) = (p_1 + p_2)U\left(\frac{p_1}{p_1 + p_2}\right) + E(p_1 + p_2), \tag{23}$$

where $U(\cdot; p_4, \ldots, p_n)$ and $E(\cdot; p_4, \ldots, p_n)$ are two univariate functions with parameters $p_4, \ldots, p_n$. Then, we show the only symmetric function $G(P)$ taking this form is the negative Shannon entropy up to nonnegative multiplicative factor by appealing to the so-called generalized *fundamental equation of information theory* (cf. [13]).

Parametrize $P_\lambda^{(t)}$ as follows: $P_\lambda^{(t)} = (\lambda t, \lambda(1-t), r - \lambda, p_4, \ldots, p_n)$, where $r \triangleq 1 - \sum_{i\geq 4} p_i, t \in [0,1]$. For $0 < \lambda_1 < \lambda_2$, taking $P_{X|1} \triangleq P_{\lambda_1}^{(t)}, P_{X|2} \triangleq P_{\lambda_2}^{(t)}$, we have

$$D^G(P_{X|1}; P_{X|2}) = G(P_{\lambda_1}^{(t)}) - G(P_{\lambda_2}^{(t)}) - \langle \nabla G(P_{\lambda_2}^{(t)}), P_{\lambda_1}^{(t)} - P_{\lambda_2}^{(t)} \rangle$$

Note that the following transform is sufficient for Z.

$$Y(X) = \begin{cases} 1 & X \in \{1,2\} \\ X & \text{otherwise} \end{cases} \tag{24}$$

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

$$R(\lambda, t) \triangleq R(\lambda, t; p_4, p_5, \ldots, p_n) \triangleq G(P_\lambda^{(t)}). \tag{25}$$

Fixing λ_2, choose functions $U(t), F(t)$ in the way such that

$$\tilde{R}(\lambda_2, t) = \tilde{R}(\lambda_2, 1), \tag{26}$$

$$\langle \nabla \tilde{R}(\lambda_2, t), P_{\lambda_1}^{(t)} - P_{\lambda_2}^{(t)} \rangle = \langle \nabla \tilde{R}(\lambda_2, 1), P_{\lambda_1}^{(1)} - P_{\lambda_2}^{(1)} \rangle \tag{27}$$

where $\tilde{R}(\lambda, t) \triangleq R(\lambda, t) - \lambda U(t) - F(t)$. The fact that $\tilde{R}(\lambda, t)$ generates the same Bregman divergence as $R(\lambda, t)$ implies that there exists a function $E : [0,1] \to \mathbb{R}$, such that $\tilde{R}(\lambda, t) = E(\lambda)$. Hence, $R(\lambda, t) = F(t) + \lambda U(t) + E(\lambda)$.

Taking $p_1 = xa, p_2 = x(1-a), a \in [0,1], p_3 = 1 - \left(\sum_{i\geq 4} p_i\right) - x$ and letting $x \downarrow 0$, convexity of G and the Gale–Klee–Rockafellar Theorem [14] imply that $F(\cdot)$ must be constant, else the limit as $x \downarrow 0$ depends on how we approach the simplex boundary. This proves (23).

Considering a permutation $Y = \pi(X)$, and taking $P_{X|1} = P, P_{X|2} = N = (1/n, 1/n, \ldots, 1/n)$, it follows from the sufficiency property that $G(P) - \langle \nabla G(N), P \rangle$ is a symmetric function. Re-defining $G(P)$ by $G(P) - \langle \nabla G(N), P \rangle$, using the symmetry of $G(P)$, we know for $r = p_1 + p_2 + p_3$,

$$\begin{aligned} &(r - p_3) U\left(\frac{p_1}{r - p_3}\right) + E(r - p_3) \\ &\quad = (r - p_1) U\left(\frac{p_3}{r - p_1}\right) + E(r - p_1). \end{aligned} \tag{28}$$

This functional equation has a known solution [13, Cor. 10.7a]; consequently, we can show that $G(P) \propto \sum_{i=1}^n p_i \ln p_i$.

The $n = 2$ case follows from a permutation transform and SUFFICIENCY.

APPENDIX C
PROOF OF THEOREM 5

Setting $p_i = q_i = 0, i \geq 3$, and denoting p_1 by p, q_1 by q, $G(p, 1-p, 0, \ldots, 0)$ by $h(p)$, we have

$$h(p) - h(q) - h'(q)(p - q) = pf\left(\frac{q}{p}\right) + (1-p) f\left(\frac{1-q}{1-p}\right).$$

Defining $g(x) = xf(1/x)$, we know $g'(x), g''(1)$ exists, and

$$h(p) - h(q) - h'(q)(p - q) = qg\left(\frac{p}{q}\right) + (1-q) g\left(\frac{1-p}{1-q}\right).$$

Taking derivatives w.r.t. p on both sides, parametrizing $p = at, q = t, a \neq 1$, and letting $t \downarrow 0$, we obtain

$$\lim_{t \downarrow 0} (h'(at) - h'(t)) = g'(a) - g'(1). \tag{29}$$

We have

$$g'(x) = g'(1) + \lim_{t\downarrow 0} (h'(xt) - h'(t)) \tag{30}$$

$$= g'(1) + \lim_{t \downarrow 0} \left(\sum_{i=1}^n h'(x^{\frac{n-i+1}{n}} t) - h'(x^{\frac{n-i}{n}} t) \right) \tag{31}$$

$$= g'(1) + \sum_{i=1}^n \lim_{t\downarrow 0} (h'(x^{\frac{n-i+1}{n}} t) - h'(x^{\frac{n-i}{n}} t)) \tag{32}$$

$$= g'(1) + \sum_{i=1}^n \left(g'(x^{1/n}) - g'(1) \right) \tag{33}$$

$$= n g'(x^{1/n}) - (n-1) g'(1). \tag{34}$$

Expanding $g'(x)$ around $x = 1$ using Taylor series on the RHS and taking $n \to \infty$, we have $g(x) = g''(1) x \ln x + C(x - 1)$. (The authors thank Jingbo Liu for a suggestion contributing to this proof.)

REFERENCES

[1] P. Fischer, "On the inequality $\sum p_i f(p_i) \geq \sum p_i f(q_i)$," *Metrika*, vol. 18, pp. 199–208, 1972.

[2] J. Aczél and Z. Daróczy, "On measures of information and their characterizations," *New York*, 1975.

[3] M. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data processing theorem of information theory," *Information Theory, IEEE Transactions on*, vol. 43, no. 4, pp. 1288–1293, 1997.

[4] I. Csiszár *et al.*, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[5] S. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142, 1966.

[6] M. Basseville, "Divergence measures for statistical data processing," 2010.

[7] P. Kannappan and P. Sahoo, "Kullback-leibler type distance measures between probability distributions," *J. Math. Phys. Sci*, vol. 26, pp. 443–454, 1992.

[8] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*. IEEE, 2007, pp. 566–570.

[9] S.-I. Amari, "α-divergence is unique, belonging to both f-divergence and bregman divergence classes," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4925–4931, 2009.

[10] P. Harremoës, Private communication with the authors, Dec. 2013.

[11] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The annals of statistics*, pp. 2032–2066, 1991.

[12] J. Jiao, T. Courtade, A. No, K. Venkat, and T. Weissman, "Information measures: the curious case of the binary alphabet," *arXiv preprint*, 2014.

[13] P. Kannappan, *Functional equations and inequalities with applications*. Springer, 2009.

[14] 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.