Conditional Entropy and Data Processing: an Axiomatic Approach Based on Core-Concavity

Arthur Américo, MHR. Khouzani, Pasquale Malacaria
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
emails: {a.passosderezende, arman.khouzani, p.malacaria}@qmul.ac.uk .

***Abstract*—This work presents an axiomatization for entropy based on an extension of concavity called core-concavity. We show that core-concavity characterizes the largest class of functions for which the data-processing inequality holds, under the assumption that conditional entropy is defined as a generalized average. Also, under the same assumption, we show that data-processing and "conditioning reduces entropy" properties are equivalent. We prove several properties of core-concave functions, including generalization of perfect secrecy and of Fano's inequality. We also show that definitions of conditional entropy based on worst-case can be retrieved as limit cases of generalized averages. A connection between statistical decision making and this axiomatic approach is also presented.**

I. Introduction

Information theoretical entropy was introduced by Shannon using a set of three axioms [2]. Several variations and extensions of Shannon's entropy have also been introduced in an axiomatic way, the most well-known being Rényi's entropies [3]. While these works derive some specific entropy or family of entropies, this paper aims to characterize all entropies satisfying some desirable properties regarding their conditional form. Possibly the most celebrated property of entropy is that processing data can never increase information: the data-processing inequality (DPI). Another well-known property is that the uncertainty about X knowing Y is never higher than the uncertainty about X alone, i.e., "conditioning reduces entropy" (CRE). At a high level, we seek to answer "what is the largest class of entropies that satisfy DPI and/or CRE?".

This class is characterized by the axiomatization presented here. We define conditional entropy via a "generalized averaging" which we call *η-averaging* (EAVG). Assuming this form of conditional entropy, we show that an entropy satisfies DPI and CRE if and only if it is "core-concave" (CCV), i.e., it is an increasing transformation of a concave function [4].

The contributions of this paper are the following: In Section II we define core-concave entropies, and explain their relationship with quasi-concave and Schur-concave functions. In Section III, we introduce our axiomatic approach and prove the claim made above in Theorem 2. Section IV illustrates examples from the literature that are core-concave entropies.

Next, in Section V-A, we consider the "additional information increases entropy" property and establish that all symmetric and expansible core-concave entropies satisfy it. For such symmetric and expansible core-concave entropies, we then prove two important properties: first, a general "perfect secrecy" theorem in Section V-C; second, a generalization of the Fano's inequality in Section V-D, which provides bounds for such entropies in terms of the probability of error.

Subsequently, Section V-E shows that conditional entropies defined as a "worst-case scenario", as adopted in some contexts, can be recaptured as a limit construction of our generalized averages. Section V-F establishes a natural connection to the problem of statistical decision making, in which a statistical experiment is considered where each state of the nature can result in different possible observables. The decision maker then has to decide which state generated the given observable, where the decision has to minimize a loss function. We interpret this setting within our axiomatic framework.

Finally, in Section V-G, we investigate the relationship between core-concave entropies and channel ordering. Namely, "degradedness" ordering [5], also called "channel refinement" [6] and "matrix majorization" [7], is the order defined as follows: $K_2 \leq K_1 \iff K_2 = K_1 K$ for channels K_1, K_2, K, i.e., channel K_2 can be derived by a post-processing of channel K_1. We show that for channels $K_1 : \mathcal{X} \to \mathcal{Y}, K_2 : \mathcal{X} \to \mathcal{Z}$ channel K_2 is degraded from K_1 if and only if for all probability distributions over $\mathcal{X}$ and all core-concave entropies H, we have $H(X|Y) \leq H(X|Z)$.

A. Background and Related Literature

A portion of the results in this paper had appeared in the workshop proceedings [1].

The axiomatic approach to information measures is arguably as old as the field of information theory itself, as already in [2] Shannon proved his entropy to be the only one (up to a scaling factor) that satisfies some intuitive requirements. Much work in that direction has been undertaken in exploring sets of postulates that characterize Shannon entropy or its generalizations [3], [8], [9]. A review of these efforts is provided in [10].

The axiomatic approach of Section III-A is inspired by the work of Alvim *et al.* [11], which was concerned with an axiomatic treatment for the field of Quantitative Information Flow. Instead of the generalized averaging (EAVG), they confined their study to regular averaging, reaching a conclusion

Parts of this work have been presented in the proceeding in [1].

similar to our Theorem 2. It follows from their results that conditional entropies under regular averaging respect DPI and CRE if and only if they are concave. However, their framework was not directly suitable to define or analyze conditional forms of fairly common entropies, such as the Rényi families. By relaxing the regular averaging to EAVG, we are able to expand the scope of their results, encompassing those entropies.

One of the motivations for this work was the competing definitions of conditional Rényi entropies in the literature, as surveyed in [12] and later in [13]. In these works, properties such as CRE and DPI were proved or disproved in a case by case basis for each candidate conditional form. In [12], Teixeira *et al.* investigated the first two definitions in Table I. In addition to those two, Iwamoto and Shikata [13] also studied definitions (1a) and (1b). Theorems 2 and 3 simplify the verification of both CRE and DPI for each form of Rényi entropies and each α. Moreover, the results in Section V can be instantiated to obtain information-theoretic results for each form of conditional Rényi entropies.

The notion of preorders over channels has also been extensively studied in the literature, both within Information Theory and outside it. In the setting of comparison of experiments, the degradedness order has been used to establish whether an experiment is more informative than another: Theorem 6 can be traced back to Blackwell's Theorem [14]. In our setting, the partial proof of Sherman [15], which only concerns experiments with finite outcomes, suffices. A simple proof of the Theorem can also be found in [16]. Our proof of Theorem 6 is inspired by Dahl [7]. Dahl's work is concerned with a preorder over all real matrices, which coincides with the degradedness ordering when these matrices are row-stochastic.

B. Notations & Conventions

Throughout the paper, X, Y, $Z, \ldots$ represent discrete random variables (abbreviated as r.v.) with alphabets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}, \ldots$. We assume that the elements of each alphabet are indexed, e.g., denoting by $x_1, \ldots, x_{|\mathcal{X}|}$ the elements of $\mathcal{X}$, and so on. Given $x_i \in \mathcal{X}$, we write $p(x_i)$ or p_i to mean $Pr\{X = x_i\}$, and use p to refer to the distribution. We may specify the r.v. with a subscript, e.g., write $p_X(x)$, if it is not clear from the context.

Let $\Delta_n \subset \mathbb{R}^n$ be the $(n-1)$-dimensional probability simplex. Given a probability distribution p over $\{x_1, \ldots, x_n\}$ we overload the notation and use p to refer to its probability vector $(p_1, \ldots, p_n) \in \Delta_n$. Given a function F over Δ_n and a random variable X with distribution $p = (p_1, \ldots, p_n)$, we will use $F(X)$, $F(p_1, \ldots, p_n)$ and $F(p)$ interchangeably.

A channel K is a row stochastic matrix with rows indexed by $\mathcal{X}$ and columns indexed by $\mathcal{Y}$. The value $K(y|x)$ is equal to $p(y|x) = Pr\{Y = y|X = x\}$, i.e., the conditional probability that y is produced by the channel K when x is the input value. The notation $K : \mathcal{X} \to \mathcal{Y}$ means that the channel K has $\mathcal{X}$ and $\mathcal{Y}$ as its input and output alphabets.

II. Core-concavity and Basic Properties

Entropy is classically motivated as a measure of the "uncertainty" associated with a distribution. For example, if the entropy function is symmetric, it should attain its minimum on a *point distribution* like $(1, 0, \ldots, 0)$, and its maximum on the uniform distribution $(1/n, \ldots, 1/n)$.

A further basic property for an entropy function can be derived by studying the following example: Consider a coin toss with "head" and "tails" probabilities of α and $1-\alpha$ and a random variable X. If the outcome of the coin flip is "head", the distribution that X is drawn from is p^1, and if it is "tails", it is p^2, for some $p^1, p^2 \in \Delta_n$. If the outcome of the coin flip is not observed, the distribution of X will be $p^3 = \alpha p^1 + (1-\alpha)p^2$. One would expect that a measure of uncertainty F to be lower in the scenario that the outcome of the coin-flip is observed compared to when it is not; that is:

$$\alpha F(p^1) + (1-\alpha)F(p^2) \leq F\left(\alpha p^1 + (1-\alpha)p^2\right).$$

In other words, F should be concave.

However, and crucial to this work, we can make the observation that applying an increasing function η to both sides of the above inequality preserves it, i.e.,:

$$\eta\left(\alpha F(p^1) + (1-\alpha)F(p^2)\right) \leq \eta\left(F\left(\alpha p^1 + (1-\alpha)p^2\right)\right).$$

Hence the above inequality, using the pair (η, F), can be justified as a measure of uncertainty in the same way as F is. Crucially, the pair (η, F) describes strictly more functions than just concave functions. The pair (η, F) defines a *core-concave* entropy [4]:

Definition 1. *A "core-concave entropy" $H = (\eta, F)$ is a pair such that:*

1) *F is a real-valued function over Δ_n that is continuous and concave;*
2) *η is a continuous and strictly increasing real valued function defined over the image of F.*

Let $\mathcal{H}$ denote the set of all core-concave entropies. Given $H = (\eta, F) \in \mathcal{H}$, we define $H(p) = \eta(F(p))$. If X is a discrete random variable with distribution $p \in \Delta_n$ we may use $H(X)$ to refer to $H(p)$.

Notice that so far we do not require H to be "expansible", that is, $H(p_1, \ldots, p_n)$ may be different from $H(p_1, \ldots, p_n, 0)$ and so on. There are indeed core-concave entropies that are not expansible, for example, the function $F(p) = -\sum_i 2^{p_i}$ is concave for each Δ_n but not expansible.

A downside of not assuming expansibility is that it would make little sense to compare entropies of distributions that have different dimensions. Expansibility, along with symmetry, will be added at a later stage (Section V) when considering the "additional information increases entropy" property.

The definition of core-concavity can be readily seen to capture:

- Shannon entropy: $H_1(p) = -\sum_i p_i \log p_i$, with $\eta(t) = t$ and $F(p) = -\sum_i p_i \log p_i$.
- Min-entropy: $H_\infty(p) = -\log(\max_i p_i)$, with $\eta(t) = -\log(-t)$, and $F(p) = -\max_i p_i$.
- Guessing entropy $H_G(p) = \sum_i i p_{[i]}$ (where $p_{[1]}, \ldots, p_{[n]}$ is a non-increasing rearrangement of p), with $\eta(t) = t$ and $F(p) = \sum i p_{[i]}$.

Core-concave entropies also encompass a more general family of entropies referred to as Sharma-Mittal [17] defined as follows:

$$H_{\alpha,\beta}(p) = \frac{1}{\beta-1}\left(1-(\|p\|_\alpha^\alpha)^{\frac{1-\beta}{1-\alpha}}\right), \quad \alpha \geq 0,\ \alpha,\beta \neq 1.$$

This family generalizes Rényi entropies by $H_{\alpha,\beta\to 1}(p)$, Shannon by $H_{\alpha\to 1,\beta\to 1}(p)$, and Havrda-Tsallis entropies [18], [19] as $H_{\alpha,\alpha}(p) = \frac{1}{1-\alpha}\left(1-\|p\|_\alpha^\alpha\right)$. Core-concavity of Sharma-Mittal family $H_{\alpha,\beta}(p)$ (and hence any of its subfamilies) can be seen by taking:

$$\eta(t) = \frac{1}{\beta-1}\left(1-t^{\frac{1-\beta}{1-\alpha}}\right),\ F(p) = \|p\|_\alpha^\alpha \qquad 0<\alpha<1,$$
$$\eta(t) = \frac{1}{\beta-1}\left(1-(-t)^{\frac{1-\beta}{1-\alpha}}\right),\ F(p) = -\|p\|_\alpha^\alpha \qquad 1<\alpha.$$

A. Core/Quasi/Schur-Concavity

Besides being a generalizing framework for entropy measures, core-concavity can be seen as a property of real valued functions over Δ_n. Given $f:\Delta_n\to\mathbb{R}$, we say that f is a *core-concave function* if there is a $(\eta,F)\in\mathcal{H}$ such that $f(p)=\eta(F(p))$. In particular, the unconditional form of a core-concave entropy is a core-concave function. In the following, we shed some light on the relation of the core-concave property with the related notions of quasi-concavity and Schur-concavity.

By definition, a real valued function ϕ over some convex subset of $\mathbb{R}^n$ is *quasi-concave* if for all $\alpha\in\mathbb{R}$, the set $\{x|\phi(x)\leq\alpha\}$ is a convex set. Equivalently, it must be that for all $\lambda\in[0,1]$, and all x, y in the domain of ϕ we have: $\phi(\lambda x+(1-\lambda)y)\geq\min\{\phi(x),\phi(y)\}$. Note that trivially, any concave function is also both core-concave and quasi-concave. We also have:

Proposition 1. *Any core-concave function is also quasi-concave.*

Proof: The statement of the proposition follows from the following two facts:

1) All concave functions are quasi-concave;
2) If F is quasi-concave and η is increasing then $\eta\circ F$ is quasi-concave. ∎

Given $p,q\in\Delta_n$, p *majorizes* q if for all $k\leq n$, $\sum p_{[i]}\geq\sum q_{[i]}$, where $(p_{[1]},\ldots,p_{[n]})$ and $(q_{[1]},\ldots,q_{[n]})$ are non increasing rearrangements of p and q. A function ϕ is said to be *Schur-concave* iff $\phi(p)\leq\phi(q)$ whenever p majorizes q. As any symmetric quasi-concave function is Schur-concave [20, Chapter 3.C], an immediate consequence of Proposition 1 is the following:

Corollary 1. *Any symmetric core-concave function is Schur-concave.*

Note, however, that we are not assuming symmetry in this paper until Section V.

B. Conditional Entropy

Definition 2. *Given a core-concave entropy $H=(\eta,F)$, we define its "conditional" form as:*

$$H(X|Y) = \eta\left(\sum_{y\in\mathcal{Y}^+} p(y)F(X|y)\right),$$

where $\mathcal{Y}^+$ is the support of Y and $F(X|y)$ is shorthand for $F(p_{X|y})$.

Critically, note that in terms of the (unconditional) entropy, the above is equivalent to:

$$H(X|Y) = \eta\left(\sum_{y\in\mathcal{Y}^+} p(y)\eta^{-1}\left(H(X|y)\right)\right),$$

which is a generalization of the expected value of $H(X|y)$ with respect to p_Y.

From the definitions above, it is also possible to define, for each $H\in\mathcal{H}$, a quantity $I_H(X;Y)=H(X)-H(X|Y)$, analogous to the classical mutual information. Moreover, following Csiszár [10], a core-concave divergence can also be defined, at least for core-concave entropies where F is such that $F(p)=\sum_i f(p_i)$ for some convex function f (this is the "postulate of Sum property" in [10]). Then one can define a (η,F) divergence as:

$$D_{(\eta,F)}(p\,\|\,q) = -\eta\left(-\sum_i q_i f(p_i/q_i)\right).$$

C. Choices of η and F

When describing an entropy within our framework the choices of η and F are, in general, not unique. For example, one could equally define Shannon entropy in this framework by choosing $\eta(x)=2x$ and $F(p)=-\frac{1}{2}\sum_i p_i\log p_i$. While the choices of η and F are immaterial to the values of the entropy $H(p)$, they can radically change its conditional form. Consider, for instance, the Rényi entropies:

$$H_\alpha(p) = \frac{\alpha}{1-\alpha}\log\|p\|_\alpha,$$

which, for $\alpha>1$, can be recovered by choosing either $\eta(x)=(\alpha/1-\alpha)\log(-x)$ and $F(p)=-\|p\|_\alpha$, or $\eta^*(x)=(1/1-\alpha)\log(-x)$ and $F^*(p)=-\|p\|_\alpha^\alpha$. These choices induce, respectively, the following two conditional forms:

$$H_\alpha(X|Y) = \frac{\alpha}{1-\alpha}\log\sum_{y\in\mathcal{Y}^+} p(y)\left\|p_{X|y}\right\|_\alpha, \tag{1a}$$

$$H_\alpha^*(X|Y) = \frac{1}{1-\alpha}\log\sum_{y\in\mathcal{Y}^+} p(y)\left\|p_{X|y}\right\|_\alpha^\alpha. \tag{1b}$$

Indeed, both forms have been proposed in the literature [21], [22], and they do not coincide. In particular, only (1a) coincides with conditional min-entropy as $\alpha\to\infty$, that is, $\lim_{\alpha\to\infty}H_\alpha(X|Y)=H_\infty(X|Y)$ (see e.g. [13]).

Nevertheless, the conditional form of an entropy uniquely identifies the choice of η and F up to a linear transformation:

Theorem 1. *Let $H^1=(\eta_1,F_1)$ and $H^2=(\eta_2,F_2)$ be core-concave entropies. If $H^1(X|Y)=H^2(X|Y)$ for all r.v. X,Y,*

then $\eta_2(x) = \eta_1(ax + b)$ *and* $F_2(p) = (1/a)F_1(p) - b/a$ *for some* $a, b \in \mathbb{R}$.

Conversely, if there are $a, b \in \mathbb{R}$ *such that* $\eta_2(x) = \eta_1(ax + b)$ *and* $F_2(p) = (1/a)F_1(p) - b/a$, *then* $H^1(X|Y) = H^2(X|Y)$ *for all r.v.* X, Y.

Proof: The converse follows directly from Definition 2.

For the direct implication, let $H^1 = (\eta_1, F_1)$, $H^2 = (\eta_2, F_2) \in \mathcal{H}$ such that for all r.v. X, Y, $\eta_1(\sum_y p(y)F_1(X|y)) = \eta_2(\sum_y p(y)F_2(X|y))$. Since the images of η_1 and η_2 coincide, and η_1 is strictly increasing, the function $\phi = \eta_1^{-1} \circ \eta_2$ is well defined. Thus, for all probability vectors $p \in \Delta_n$, we have $F_1(p) = \phi(F_2(p))$.

Now, let $p^1, p^2 \in \Delta_n$, $\mathcal{X} = \{x_1, \ldots, x_n\}$, $\mathcal{Y} = \{y_1, y_2\}$, and define the joint distribution:[1]

$$p_{(X,Y)}(x, y) = \begin{cases} tp^1(x) & \text{if } y = y_1, \\ (1-t)p^2(x) & \text{if } y = y_2. \end{cases} \tag{2}$$

From the definition of $p_{(X,Y)}$, one can easily check that the marginal distributions are given by $p_X = tp^1 + (1-t)p^2$ and $p_Y = (t, 1-t)$, and that $p_{X|y_1} = p^1$ and $p_{X|y_2} = p^2$. Thus, from the assumption that $\eta_1(\sum_y p(y)F_1(X|y)) = \eta_2(\sum_y p(y)F_2(X|y))$, we obtain:

$$\begin{aligned} \phi(tF_2(p^1) + (1-t)F_2(p^2)) &= tF_1(p^1) + (1-t)F_1(p^2) \\ &= t\phi(F_2(p^1)) + (1-t)\phi(F_2(p^2)). \end{aligned}$$

Note that the above equation holds for all choices of $p^1, p^2 \in \Delta_n$ and $t \in [0, 1]$. Now, since F is a continuous function over the compact metric space Δ_n, it attains its maximum and minimum. Let $w = \min_{p \in \Delta_n} F_2(p)$ and $z = \max_{p \in \Delta_n} F_2(p)$. Thus, the range of F_2 is $[w, z]$. Let $r \in [w, z]$. Substituting $t = \frac{r-w}{z-w}$ in the above equation, we obtain:

$$\begin{aligned} \phi(r) &= \frac{r-w}{z-w}\phi(z) + \left(1 - \frac{r-w}{z-w}\right)\phi(w) \\ &= \left(\frac{\phi(z) - \phi(w)}{z-w}\right) r + \frac{z\phi(w) - w\phi(z)}{z-w}. \end{aligned}$$

Hence, $\phi(r) = ar + b$, for $a = \frac{\phi(z)-\phi(w)}{z-w}$ and $b = \frac{z\phi(w)-w\phi(z)}{z-w}$. Therefore, $\eta_2(r) = \eta_1(\phi(r)) = \eta_1(ar + b)$, and $F_1(p) = \phi(F_2(p)) = aF_2(p) + b$. ∎

III. Axioms

Let $H = (\eta, F)$ be a pair such that $F : \Delta_n \to \mathbb{R}$ and η is a strictly increasing real-valued function defined on the image of F, and define $H(X) = \eta(F(X))$. Note that every real-valued function $H : \Delta_n \to \mathbb{R}$ can be constructed in this way for any η, by taking $F = \eta^{-1} \circ H$.

Given the pair $H = (\eta, F)$, we associate with it a conditional form $H(X|Y)$, which is the extension of H to a distribution over distributions. In particular, $(X|Y)$ represents the family of random variables $(X|y)_{y \in \mathcal{Y}}$ where $(X|y)$ is the random variable with distribution $p_{X|y}(x) = p_{(X,Y)}(x, y)/p_Y(y)$. The family of distributions associated to the family of random variables $(X|Y)$ is hence $\{p_{X|y} : y \in \mathcal{Y}\}$. Given a channel $K : \mathcal{X} \to \mathcal{Y}$ and a distribution over $\mathcal{X}$, the family of distributions $\{p_{X|y} : y \in \mathcal{Y}\}$ can be computed using Bayes' rule. This construction is similar to the *hyper-distributions* as defined in [11], [23].

Definition 3. *For any* $n > 0$, *a pair* $H = (\eta, F)$, $F : \Delta_n \to \mathbb{R}$ *as defined above, respects:*

CCV ***(core-concavity)****: if* H *is core-concave (as described in Definition 1).*

EAVG ***(η-averaging)****: if given r.v.* X,Y, *its conditional form* $H(X|Y)$ *is defined as:*

$$H(X|Y) = \eta\left(\sum_{y \in \mathcal{Y}^+} p(y)F(X|y)\right).$$

CRE ***(conditioning reduces entropy):*** *if for all r.v.* X,Y, $H(X|Y) \leq H(X)$, *with equality holding if* X *and* Y *are independent.*

DPI ***(data-processing inequality):*** *if, for all r.v.* X,Y,Z *such that* $X \to Y \to Z$ *(i.e., if* X *and* Z *are conditionally independent given* Y*),*

$$H(X|Y) \leq H(X|Z).$$

In all axioms, we assume $|\mathcal{X}| = n$, *and that* $\mathcal{Y}, \mathcal{Z}$ *are finite, nonempty sets.*

A. Relations Between the Axioms

This section proves the results illustrated in Figure 1. These axioms were inspired by those in [11]. However the axioms therein fail to capture most entropies, e.g. all Rényi entropies with $\alpha > 1$, which are not concave [24, Theorem 1].

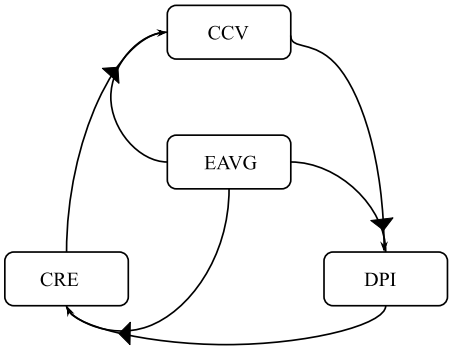

Fig. 1. The implications graph among the axioms.

We begin by noting that the second part of CRE is a straightforward consequence of EAVG:

Lemma 1. *If* EAVG *holds for* $H = (\eta, F)$, *then* $H(X|Y) = H(X)$ *for all independent r.v.* X, Y.

[1] This joint distribution is the same one used in [11, Proposition 16], defined in terms of a marginal π^3 and a channel C^*.

Proof: Assume X and Y are independent r.v.'s. Then:

$$H(X|Y) = \eta\left(\sum_{y\in\mathcal{Y}^+} p(y)F(X|y)\right)$$
$$= \eta\left(\sum_{y\in\mathcal{Y}^+} p(y)F(X)\right) = \eta\circ F(X) = H(X). \qquad \blacksquare$$

For the sake of brevity, the implications between the axioms in the following results are stated using the acronyms in Definition 3. Hence, for example, the statement (EAVG and DPI) $\Rightarrow$ CRE means that if a pair $H = (\eta, F)$ satisfies EAVG and DPI, then it satisfies CRE.

Proposition 2. *(EAVG and DPI) $\Rightarrow$ CRE.*

Proof: Let Z be a random variable over the singleton set $\mathcal{Z} = z_1$, so that $p(z_1) = 1$. Consider the composition $X \to Y \to Z$. Then:

$$H(X|Y) \;\leq\; H(X|Z) \;=\; H(X).$$

The inequality follows from DPIand the equality follows from Lemma 1 by noting that Z is independent of X given Y. $\blacksquare$

Proposition 3. *(EAVG and CRE) $\Rightarrow$ CCV.*

Proof: Let X_1, X_2 be random variables with distributions $p^1, p^2 \in \Delta_n$, $t \in [0,1]$ and define random variables X, Y with the range and joint distribution as in (2). Then, the marginal and conditional distributions obtained are as in the proof of Theorem 1, and EAVG implies:

$$H(X|Y) \;=\; \eta\left(tF(p^1) + (1-t)F(p^2)\right).$$

On the other hand, $H(X)$ can be written as $\eta\left(F\left(tp^1 + (1-t)p^2\right)\right)$. From CRE, we have $H(X|Y) \leq H(X)$. Hence:

$$\eta(tF(p^1) + (1-t)F(p^2)) \leq \eta\left(F\left(tp^1 + (1-t)p^2\right)\right).$$

Since η is strictly increasing, the above yields concavity of F:

$$tF(p^1) + (1-t)F(p^2) \leq F\left(tp^1 + (1-t)p^2\right).$$

Hence $H = (\eta, F)$ is core-concave. $\blacksquare$

Proposition 4. *(EAVG and CCV) $\Rightarrow$ DPI.*

Proof: Assume $X \to Y \to Z$. Since $p(y) = \sum_{z\in\mathcal{Z}^+} p(z)p(y|z)$, we can write:

$$\sum_{y\in\mathcal{Y}^+} p(y)F\left(X|y\right) = \sum_{y\in\mathcal{Y}^+}\left(\sum_{z\in\mathcal{Z}^+} p(z)p(y|z)\right)F\left(X|y\right)$$
$$=_{(a)} \sum_{y\in\mathcal{Y}^+, z\in\mathcal{Z}^+} p(z)p(y|z)F\left(X|y,z\right)$$
$$\leq \sum_{z\in\mathcal{Z}^+} p(z)F\left(\sum_{y\in\mathcal{Y}^+} p(y|z)p_{X|y,z}\right)$$
$$=_{(b)} \sum_{z\in\mathcal{Z}^+} p(z)F\left(X|z\right).$$

Equality (a) follows because $X \to Y \to Z$ implies $p_{X|y} = p_{X|y,z}$. Next, Jensen's inequality is applied for concave F, noting that $p(y|z)$ for $y \in \mathcal{Y}^+$ constitute convex coefficients. Equality (b) uses:

$$\sum_{y\in\mathcal{Y}^+} p(y|z)p_{X|y,z} = p_{X|z}.$$

The proposition follows by applying $\eta(\cdot)$ to the steps, noting that η preserves the inequality since it is increasing. $\blacksquare$

This completes all the relations between axioms in Figure 1.

An important consequence of the axioms is the following:

Theorem 2. *Given EAVG, the properties of CRE, CCV and DPI become equivalent. That is:*

$$EAVG \Rightarrow (CRE \Leftrightarrow CCV \Leftrightarrow DPI).$$

Proof: Split the theorem in the following two statements:

1) EAVG $\Rightarrow$ (CRE $\Leftrightarrow$ CCV),
2) EAVG $\Rightarrow$ (CCV $\Leftrightarrow$ DPI).

For (1), the proof of the "only if" part is Proposition 3, and the "if" part is proved by Proposition 4 combined with Proposition 2.

For (2), the proof of the "only if" part is Proposition 4, and the "if" part is proved by Proposition 2 combined with Proposition 3. $\blacksquare$

The definition of EAVG might seem somewhat peculiar. However, it is equivalent to applying a generalized definition of mean as per [25, Section 6.20], and taking $H(X|Y)$ as this generalized mean of the values of $H(X|y)$ weighted by p_Y.

IV. Applications to Entropy definitions in the literature

The results of Section III can be applied to conditional entropies defined in the literature. The work of verifying whether a given conditional form satisfies CRE and DPI, which in many works in the literature is often done in a case-by-case manner (see, e.g., [12], [13]) can be, in most cases, replaced by a direct application of Theorem 2. An application of Theorem 2 to possible conditional Rényi entropies suggested in the literature is presented in Table I.

Another application of Theorem 2 is as a "recipe" for defining well-behaved conditional entropies – that is, conditional entropies that satisfy the properties DPI and CRE. From the discussion on Section II, for example, one might define a conditional version of the Sharma-Mittal entropy as:

$$H_{\alpha,\beta}(X|Y) = \frac{1}{\beta-1}\left(1 - \left(\sum_y p(y)\left\|p_{X|y}\right\|_\alpha^\alpha\right)^{\frac{1-\beta}{1-\alpha}}\right).$$

Since this conditional version satisfies EAVG and Sharma-Mittal entropies are core-concave, it follows that DPI and CRE hold for this definition.

As another example, consider the (h,Φ)-entropies [26], which are a generalization of the f-entropies as defined in (1.4) in [27]. They are defined by a strictly increasing function h and a concave Φ, as follows

$$H_{h,\Phi}(X) = h\left(\sum_x \Phi(p(x))\right).$$

TABLE I
CHARACTERISTICS OF RÉNYI ENTROPY MEASURES DIRECTLY OBTAINABLE BY THE RESULTS IN SECTION III

Unconditional form	Conditional form $H(X\|Y)$	$\eta(r)$	$F(p)$	EAVG	CCV	DPI and CRE
$H_\alpha(X) = \frac{\alpha}{1-\alpha}\log\|p\|_\alpha$	$\sum_y p(y)H_\alpha(X\|y)$	r	$H_\alpha(p)$	Yes	iff $\alpha \leq 1$	iff $\alpha \leq 1$
	$\frac{1}{1-\alpha}\log\left(\frac{\sum_{x,y}p(x,y)^\alpha}{\sum_y p(y)^\alpha}\right)$	-	-	No	No	-
	$\frac{\alpha}{1-\alpha}\log\left(\sum_y p(y)\|p_{X\|y}\|_\alpha\right)$	$\frac{\alpha}{1-\alpha}\log(-r)$	$-\|p\|_\alpha$	Yes	Yes	Yes
	$\frac{1}{1-\alpha}\log\left(\sum_y p(y)\|p_{X\|y}\|_\alpha^\alpha\right)$	$\frac{1}{1-\alpha}\log(-r)$	$-\|p\|_\alpha^\alpha$	Yes	Yes	Yes
	$-\log\left(\sum_y p(y)\|p_{X\|y}\|_\alpha^{\frac{\alpha}{\alpha-1}}\right)$	$-\log(-r)$	$-\|p\|_\alpha^{\frac{\alpha}{\alpha-1}}$	Yes	Yes	Yes

These entropies are easily seen to be core-concave, by taking $\eta = h$ and $F(X) = \sum_x \Phi(p(x))$. From Theorem 2, the naive way of defining a conditional (h,Φ)-entropy as $\sum_y p(y)H_{h,\Phi}(X|y)$ would, in general, fail to satisfy DPI and CRE. Instead, our results suggest the definition

$$H_{h,\Phi}(X|Y) = h\left(\sum_y p(y)\sum_x \Phi\left(p_{X|y}(x)\right)\right).$$

The following generalized entropy, introduced by Arimoto [27], depends on a function $f : (0,1] \to \mathbb{R}_{\geq 0}$ of class C^1 (i.e., functions whose first derivative is continuous) such that $f(1) = 0$:

$$H_f(X) = \inf_{\mathbf{q}\in\Delta_n} \sum_{i=1}^n p(x_i)f(q_i).$$

As shown in [27], this entropy is continuous and concave, and therefore core-concave (with η as the identity function). Hence, Theorem 2 ensures that the conditional form given by:

$$H_f(X|Y) = \sum_y p(y)H_f(X|y)$$

satisfies DPI and CRE.

V. ADDITIONAL PROPERTIES

In this section, some classical information-theoretic results are generalized to the core-concave framework. For many of these results, the formalism developed above is not sufficiently expressible, as one needs to compare the entropy measure of probability distributions of different dimensions. To address this shortcoming, we introduce the following definition:

Definition 4. *A collection $\{H^i\}_{i\in\mathbb{N}}$ of core concave entropies is "expansible" if, for each $i \in \mathbb{N}$,*

1) *H^i is a core-concave entropy over Δ_i;*
2) *for all $(p_1,\ldots,p_i) \in \Delta_i$, $H^i(p_1,\ldots,p_i) = H^{i+1}(p_1,\ldots,p_i,0)$.*

Another property that will be necessary for the results of this section is *symmetry*:

Definition 5. *A core concave entropy $H = (\eta, F)$ over Δ_n is "symmetric" if for all $(p_1,\ldots,p_n) \in \Delta_n$ and all bijective functions $\phi : \{1,\ldots,n\} \to \{1,\ldots,n\}$, we have $F(p_1,\ldots,p_n) = F(p_{\phi(1)},\ldots,p_{\phi(n)})$. We say that a collection $\{H^i\}_{i\in\mathbb{N}}$ is symmetric if all its elements are symmetric.*

Generally, we say that a collection $\{H^i\}_{i\in\mathbb{N}}$ satisfies a property if all its elements satisfy that property.

All the results in Sections V-A, V-B, V-C and V-D regard collections of core-concave entropies. Thus, for brevity, we will refer to the collection $\{H^i\}_{i\in\mathbb{N}}$ simply by H.

A. AIE (additional information increases entropy)

A common requirement for an entropy function is that "additional information increases entropy", i.e., $\forall X,Y,\ H(X) \leq H(X,Y)$ (see e.g. [13]). It is easy to prove that this property, and its conditional extension, are a consequence of symmetry and expansibility for core-concave entropies:

Proposition 5. *Let H be symmetric and expansible. Then:*

1) $\forall X,Y\ H(X) \leq H(X,Y)$,
2) $\forall X,Y,Z\ H(X|Z) \leq H(X,Y|Z)$.

Proof: Recall that that a core-concave entropy which is symmetric and expansible is also Schur-concave (Corollary 1).

For (1), note that p_X (the distribution of X) majorizes $p_{(X,Y)}$ (the distribution of (X,Y)). Hence by Schur-concavity of H, $H(p_X) \leq H(p_{(X,Y)})$, i.e. $H(X) \leq H(X,Y)$.

For (2), the argument is the same noticing that for all z, $p_{X|z}$ majorizes $p_{(X,Y)|z}$. ∎

B. Subadditivity

It is known that subadditivity, that is

$$H(X,Y) \leq H(X) + H(Y)$$

does not hold for most Rényi entropies. For example, consider the joint distribution over (X,Y), with $\mathcal{X} = \{x_1,x_2\}$ and $\mathcal{Y} = \{y_1,y_2\}$ given by

$$p_{(X,Y)}(x_1,y_1) = 0, \qquad p_{(X,Y)}(x_1,y_2) = 1/4,$$
$$p_{(X,Y)}(x_2,y_1) = 1/4, \qquad p_{(X,Y)}(x_2,y_2) = 1/2.$$

This distribution has marginals $p_X = p_Y = (1/4, 3/4)$, and one can check that $H(X,Y) > H(X) + H(Y)$ for H being any Rényi entropy with $\alpha > 1.61$.

Consequently subadditivity does not hold for core-concave entropies. However, we have the following (tight) inequality:

Proposition 6. *Let H be symmetric and expansible. Then $H(X,Y) \leq H(\tilde{p})$ where $\tilde{p}$ is the following $|\mathcal{X}||\mathcal{Y}|$-sized distribution:*

$$\tilde{p}(x,y) = p_X(x)/|\mathcal{Y}|, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

Proof: Without loss of generality, arrange $p_{(X,Y)}$ as a $|\mathcal{X}||\mathcal{Y}|$-sized vector where the first $|\mathcal{Y}|$ elements are $p(x_1, y_1)$ to $p(x_1, y_{|\mathcal{Y}|})$, the second $|\mathcal{Y}|$ elements are $p(x_2, y_1)$ to $p(x_2, y_{|\mathcal{Y}|})$, and so on. Thanks to the Schur-concavity of H, the result follows if we show $p_{(X,Y)}$ majorizes $\tilde{p}$. For this, it is sufficient to find a doubly stochastic matrix D (of size $|\mathcal{X}||\mathcal{Y}|$-by-$|\mathcal{X}||\mathcal{Y}|$) such that $\tilde{p} = Dp_{(X,Y)}$ (see e.g. [28, Theorem 2.1]). The following is such a D:

$$D = \begin{pmatrix} \overline{Y} & 0 & \cdots & 0 \\ 0 & \overline{Y} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \overline{Y} \end{pmatrix}$$

where each zero represents a $|\mathcal{Y}|$-by-$|\mathcal{Y}|$ matrix will all zero entries, and $\overline{Y}$ is a $|\mathcal{Y}|$-by-$|\mathcal{Y}|$ matrix with all entries equal to $1/|\mathcal{Y}|$. It is straightforward to check that D is indeed doubly stochastic and $\tilde{p} = Dp_{(X,Y)}$. ∎

For Rényi entropies, this proposition takes a more concise form sometimes referred to as "weak subadditivity" and was proved for quantum systems in [29]. An alternative proof can be obtained as a corollary of Proposition 6.

Corollary 2. *Let H_α denote the Rényi entropy with parameter α. We have:*

$$H_\alpha(X,Y) \leq H_\alpha(X) + \log|\mathcal{Y}|.$$

Proof: Applying Prop. 6 to Rényi entropy, we get:

$$\begin{aligned} H_\alpha(X,Y) &\leq H_\alpha(\tilde{p}) \\ &= \frac{1}{1-\alpha}\log\left(\sum_x\sum_y (p(x)/|\mathcal{Y}|)^\alpha\right) \\ &= \frac{1}{1-\alpha}\log\left(\sum_x |\mathcal{Y}|(p(x)/|\mathcal{Y}|)^\alpha\right) \\ &= \frac{1}{1-\alpha}\left(\log\left(\sum_x p(x)^\alpha\right) + \log\left(|\mathcal{Y}|^{1-\alpha}\right)\right) \\ &= H_\alpha(X) + \log|\mathcal{Y}|. \end{aligned}$$ ∎

C. Perfect Secrecy

Shannon's perfect secrecy theorem can be generalized to all expansible and symmetric core-concave entropies, by extending an argument from [13]. As in [13] we assume a symmetric encryption scheme E that satisfies (1) "perfect secrecy" i.e. plaintext and ciphertext are interpreted as independent random variables, and (2) "perfect correctness" that is the encryption scheme makes no decryption errors, i.e. for all keys k, ciphertext c, plaintexts m:

$$p(m,k|c) = \begin{cases} 0 & \text{if } E(k,c) \neq m, \\ p(k|c) & \text{if } E(k,c) = m. \end{cases}$$

Proposition 7. *Let H be symmetric and expansible, and let K, M, C be the r.v. associated respectively with keys, plaintexts, and ciphertexts in a symmetric encryption system satisfying perfect secrecy and perfect correctness. Then the following holds: $H(M) \leq H(K)$.*

Proof: We have:

$$H(M) = H(M|C) \leq_a H(M,K|C) =_b H(K|C) \leq_c H(K),$$

where $H(M) = H(M|C)$ because M and C are independent (because of perfect secrecy), (a) is Proposition 5-2, (b) is because the encryption makes no decryption errors (perfect correctness), and (c) is CRE. ∎

D. Bounds in terms of probability of error

Symmetric core-concave entropies can be bounded in terms of the probability of error. These bounds generalize Fano's inequality. Let H be symmetric and expansible and X a r.v. with distribution $(p_1, \ldots, p_n)$. Then the probability of error e is defined as $e = 1 - \max_{x\in\mathcal{X}} p(x)$. Given a family of random variables $(X|Y)$, the average-probability of error $\hat{e}$ is defined as the average probability of all errors in the family, i.e., $\hat{e} = \sum_y p(y)e_y$ where $e_y = 1 - \max_{x\in\mathcal{X}} p_{X|y}(x)$.

Proposition 8. *(Fano's generalization)*

1) $H(X) \leq H\left(1-e, \frac{e}{n-1}, \ldots, \frac{e}{n-1}\right)$.
2) $H(X|Y) \leq H\left(1-\hat{e}, \frac{\hat{e}}{n-1}, \ldots, \frac{\hat{e}}{n-1}\right)$.

Proof: For part (1): A symmetric core-concave entropy is Schur concave and the distribution $(p_1, \ldots, p_n)$ majorizes $\left(1-e, \frac{e}{n-1}, \ldots, \frac{e}{n-1}\right)$;. This argument was originally used in Vajda and Vašek work [9] for Schur concave entropies.

For part (2): Let e_y be the associated probability of error to the distribution $(X|y)$. Following part (1):

$$F(X|y) \leq F\left(1-e_y, \frac{e_y}{n-1}, \ldots, \frac{e_y}{n-1}\right).$$

Hence:

$$\begin{aligned} \sum_y p(y)F(X|y) &\leq \sum_y p(y)F\left(1-e_y, \frac{e_y}{n-1}, \ldots, \frac{e_y}{n-1}\right) \\ &\leq_{(a)} F\left(\sum_y p(y)(1-e_y), \frac{\sum_y p(y)e_y}{n-1}, \ldots, \frac{\sum_y p(y)e_y}{n-1}\right) \\ &= F\left(1-\hat{e}, \frac{\hat{e}}{n-1}, \ldots, \frac{\hat{e}}{n-1}\right). \end{aligned}$$

where inequality (a) is by concavity of F. Part (2) now follows by applying the (increasing) function η to the above. ∎

Proposition 8 is a generalization of the Fano's inequality:

$$H(X|Y) \leq H(\hat{e}, 1-\hat{e}) + \hat{e}\log(n-1).$$

This is because when H is the Shannon entropy, we have:

$$\begin{aligned} H(X|Y) &\leq H(1-\hat{e}, \frac{\hat{e}}{n-1}, \ldots, \frac{\hat{e}}{n-1}) \\ &= (1-\hat{e})\log(\frac{1}{1-\hat{e}}) + (n-1)\frac{\hat{e}}{n-1}\log(\frac{n-1}{\hat{e}}) \\ &= H(\hat{e}, 1-\hat{e}) + \hat{e}\log(n-1). \end{aligned}$$

E. Recovering MIN as a Limit Case of CCV and EAVG

We now consider a definition of conditional entropy based on the "worst-case" scenario, by taking the conditional entropy to be the minimum value of the entropy over the posterior distributions. That is, given an entropy measure $H_M : \Delta_n \to \mathbb{R}$, the conditional has the form

$$H_M(X|Y) = \min_{y\in\mathcal{Y}^+} H_M(X|y).$$

This form has also been studied in [11], in which results similar to the ones of the last section are derived. Given an (possibly not core-concave) entropy H we say that it satisfies:

- **QCV** **(quasi-concavity)**: if H, as a function over Δ_n, is quasi-concave.
- **MIN** **(minimum)**: if given r.v. X,Y, its conditional form $H(X|Y)$ is defined as:

$$H(X|Y) = \min_{y\in\mathcal{Y}^+} H(X|y).$$

The next result, similar to Theorem 2, is a rewording of the results in [11]:

Theorem 3. *MIN* $\Rightarrow$ (*QCV* $\Leftrightarrow$ *DPI* $\Leftrightarrow$ *CRE*).

A straightforward result from Theorem 3 is the following:

Proposition 9. *Let H be a symmetric and expansible entropy that satisfies MIN and QCV. Then, it satisfies the conclusions of Propositions 5 and 7.*

Proof: From Theorem 3, such an entropy satisfies DPI and CRE. Moreover, a symmetric quasi-concave function is Schur-concave [20, Section 3.C]. The proof is then almost identical to the ones on Propositions 5 and 7. ∎

In general, entropies that satisfy MIN do not satisfy EAVG, and therefore, are not encompassed by the core-concave framework. However, if there is $H = (\eta, F) \in \mathcal{H}$ such that $H_M(X) = H(X)$ (that is, if the unconditional form of H_M coincides with that of a core-concave entropy), then $H_M(X|Y)$ can be retrieved as a limit of a sequence $\{(\eta_i, F_i)\}_{i\in\mathbb{N}}$ in $\mathcal{H}$. Before establishing this claim, we need an auxiliary result:

Lemma 2. *Given $V : \Delta_n \to \mathbb{R}_{\geq 0}$, we have:*

$$\lim_{\beta\to\infty} \left(\sum_{y\in\mathcal{Y}^+} p(y) \left(V(X|y)\right)^\beta \right)^{1/\beta} = \max_{y\in\mathcal{Y}^+} V(X|y).$$

Proof: Fix $\beta > 0$ and choose y^* such that $y^* \in \arg\max_{y\in\mathcal{Y}^+} (V(X|y))^\beta$. Then, because $\sum_{y\in\mathcal{Y}^+} p(y) = 1$,

$$\sum_{y\in\mathcal{Y}^+} p(y) \left(V(X|y)\right)^\beta \geq p(y^*) \left(V(X|y^*)\right)^\beta, \qquad \text{and}$$
$$\sum_{y\in\mathcal{Y}^+} p(y) \left(V(X|y)\right)^\beta \leq \max_{y\in\mathcal{Y}^+} \left(V(X|y)\right)^\beta.$$

For $\beta > 0$, the function $f(t) = t^{1/\beta}$ is increasing in $\mathbb{R}_{\geq 0}$. Its application to the terms of the inequalities above yields

$$\left(\sum_{y\in\mathcal{Y}^+} p(y)(V(X|y))^\beta \right)^{1/\beta} \geq (p(y^*))^{1/\beta} V(X|y^*), \qquad \text{and}$$
$$\left(\sum_{y\in\mathcal{Y}^+} p(y)(V(X|y))^\beta \right)^{1/\beta} \leq \max_{y\in\mathcal{Y}^+} V(X|y).$$

Since the function $f(t) = t^\beta$ is strictly increasing, $y^* \in \arg\max_{y\in\mathcal{Y}^+} V(X|y)$. Therefore,

$$\lim_{\beta\to\infty} (p(y^*))^{1/\beta} V(X|y^*) = V(X|y^*) \qquad \text{and}$$
$$\max_{y\in\mathcal{Y}^+} V(X|y) = V(X|y^*)$$

and the claim follows from the sandwich theorem. ∎

Theorem 4. *Let $H_M : \Delta_n \to \mathbb{R}$ be an entropy measure associated with a conditional form as $H_M(X|Y) = \min_{y\in\mathcal{Y}^+} H_M(X|y)$. If there is $(\eta, F) \in \mathcal{H}$ such that $H_M(p) = \eta(F(p))$, then there is a sequence $\{H^i = (\eta_i, F_i)\}_{i\in\mathbb{N}}$ in $\mathcal{H}$ such that:*

$$H_M(X|Y) = \lim_{i\to\infty} H^i(X|Y) = \lim_{i\to\infty} \eta_i\Big(\sum_{y\in\mathcal{Y}^+} p(y) F_i(X|y) \Big).$$

Proof: Suppose that $H_M(p) = \eta(F(p))$ for some $(\eta, F) \in \mathcal{H}$. Without loss of generality, we can assume F to be nonpositive, by taking $a = 1$ and $b = \max_{p\in\Delta_n} F(p)$ in Theorem 1. Now, define $\eta_i(x) = \eta(-(-x)^{1/i})$ and $F_i(p) = -(-F(p))^i$. We first show that $(\eta_i, F_i) \in \mathcal{H}$. Note that for all $i > 0$, η_i is increasing, since it is the composition of η and the function $x \mapsto -(-x)^{1/i}$, which is increasing for $x \in \mathbb{R}_{\leq 0}$. Moreover, for all $p^1, p^2 \in \Delta_n$, $\beta \in [0, 1]$,

$$\begin{aligned}(-F(\beta p^1 + (1-\beta)p^1))^i &\leq (\beta(-F(p^1)) + (1-\beta)(-F(p^2)))^i \\ &\leq \beta(-F(p^1))^i + (1-\beta)(-F(p^2))^i,\end{aligned}$$

where the inequalities follow from F being concave and $x \mapsto x^i$ increasing. Thus, $(-F)^i$ is convex, and each F_i is concave. This proves that $(\eta_i, F_i) \in \mathcal{H}$.

Now, we have:

$$\begin{aligned}\lim_{i\to\infty} H^i(X|Y) &= \lim_{i\to\infty} \eta_i \left(\sum_{y\in\mathcal{Y}^+} p(y) F_i(X|y) \right) \\ &= \lim_{i\to\infty} \eta \left(- \left(- \left(\sum_{y\in\mathcal{Y}^+} p(y) F_i(X|y) \right) \right)^{1/i} \right) \\ &= \eta \left(\lim_{i\to\infty} - \left(- \left(\sum_{y\in\mathcal{Y}^+} p(y) \left(-(-F(X|y))^i\right) \right) \right)^{1/i} \right) \\ &= \eta \left(- \lim_{i\to\infty} \left(\sum_{y\in\mathcal{Y}^+} p(y)(-F(X|y))^i \right)^{1/i} \right) \\ &=_{(a)} \eta \left(- \max_{y\in\mathcal{Y}^+} (-F(X|y)) \right)\end{aligned}$$

$$= \eta\left(\min_{y\in\mathcal{Y}^+} F(X|y)\right) =_{(b)} \min_{y\in\mathcal{Y}^+} H_M(X|y) = H_M(X|Y).$$

where the equality (a) follows from Lemma 2, substituting $-F$ for V, and equality (b) because η is increasing. ∎

Note that, in the statement of Theorem 4, H_M and H are different core-concave entropies. Although H_M coincides with H in the unconditional form, their conditional forms do not.

F. Connection to statistical decision making

A particular case of the axiomatization presented in this work is when η is the identity function. In this case, core-concavity reduces to concavity, η-averaging to simple averaging, and the results in Section III reduce to those in [11].

This particular case is of interest because it captures the problem of experiment comparison and statistical decision making. In this setting a channel is seen as a *statistical experiment*, where the rows of the channel corresponds to the possible values of an unknown parameter (the state of nature) given by the r.v. X, and the columns corresponds to the possible observations given the experiment, represented by the r.v. Y. A statistical decision problem is to choose an element on a set $\mathcal{A} = \{a_1, \ldots, a_k\}$ of possible *actions* based on the observations of the experiment such that a certain loss function is minimized.

To formalize the setting, and following [7], consider a probabilistic decision function δ where $\delta(y, a)$ is the probability of taking action a if y has been observed. Let's also define the loss function $L : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that $L(x, a)$ is the loss when x is the true state and a is the decision. The risk in a statistical experiment K (i.e., a channel $K : X \to Y$) when the true state of nature is x, the decision function is δ, and the loss function is L is defined then as:

$$R_K(x, \delta, L) = \sum_{y\in\mathcal{Y}} K(y|x) \sum_{a\in\mathcal{A}} L(x, a)\delta(y, a).$$

Taking the expectation w.r.t. the state of nature, the "expected risk" (denoted by ER) in a statistical experiment K when the decision function is δ and the loss is L can be taken as:

$$ER_K(\delta, L) = \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} K(y|x) \sum_{a\in\mathcal{A}} L(x, a)\delta(y, a).$$

When making a decision, one is usually interested in choosing δ that minimizes $ER_K(\delta, L)$. The minimum expected risk of K under loss function L, denoted by $\mu R_K(L)$, is thus:

$$\mu R_K(L) = \min_{\delta} ER_K(\delta, L).$$

With a bit of manipulation, we obtain:

$$\begin{aligned}\mu R_K(L) &= \min_{\delta} \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} K(y|x) \sum_{a\in\mathcal{A}} L(x, a)\delta(y, a)\\ &= \min_{\delta} \sum_{y\in\mathcal{Y}} p(y) \sum_{x\in\mathcal{X}, a\in\mathcal{A}} L(x, a)\delta(y, a)p_{X|y}(x)\\ &= \sum_{y} p(y) \min_{a\in\mathcal{A}} \sum_{x} L(x, a)p_{X|y}(x).\end{aligned}$$

The last inequality follows from observing that the minimum can always be attained by a deterministic δ, i.e., one that uniquely maps each output to a possible state.

Notice that the function $p \to \min_a \sum_x L(x, a)p(x)$ is concave. Therefore, letting:

$$H_R = \left(x \mapsto x,\ p \to \min_a \sum_x L(x, a)p(x)\right),$$

we see that $H_R \in \mathcal{H}$, and $H_R(Y|X) = \mu R_K(L)$. In the following, we will drop μ for brevity.

The value $H_R(X)$ can be interpreted simply as the minimum expected risk of K, when K is a null channel $\mathbf{0}$ i.e. all rows in K are identical. The channel $\mathbf{0}$ corresponds to a *non-informative experiment*, that is, a statistical experiment where any observation gives no information about the true state. From Theorem 2, some interesting properties regarding statistical decisions can be derived from the fact that $H_R(X) \in \mathcal{H}$.

- From CRE, we have that the minimum expected risk of any experiment K is never greater than that of a non-informative experiment: $\forall K,\ R_K(L) \leq R_{\mathbf{0}}(L)$.
- From DPI, whenever an experiment can be emulated from the observable of another experiment, the minimum expected risk of the former will never be greater than that of the latter. More formally, given $K_1 : \mathcal{X} \to \mathcal{Y}$ and $K_2 : \mathcal{X} \to \mathcal{Z}$, if

$$\exists W : \mathcal{Y} \to \mathcal{Z};\quad K_2(z|x) = \sum_{y\in\mathcal{Y}} K_1(y|x)W(z|y), \quad (3)$$

  then $R_{K_1}(L) \leq R_{K_2}(L)$.

An important remark to make is that concavity and risk functions in fact coincide, that is, any concave function F can be written as $\min_{a\in\mathcal{A}} \sum_x L(x, a)p(x)$ for some loss function L, if we allow $\mathcal{A}$ to be infinite. This is a known result in the statistics community [11], [30]. At a high level the proof of this fact is based on the following two observations:

- All concave functions can be written as an infimum of a family of linear functions;
- Given a loss function L, $p \mapsto \sum_x L(x, a)p(x)$ is a linear function for each $a \in \mathcal{A}$.

In the next section we further study the relation (3), and what is known in the literature as the Blackwell Theorem, which can be summarized by stating that if $R_{K_1}(L) \leq R_{K_2}(L)$ holds for all choices of L, then (3) also holds.

Blackwell's result was motivated by comparison of statistical experiments. It turns out that a similar conclusion can be derived when the quantification is taken over the set of core-concave entropies, instead of the loss functions L. To this end, we dedicate the next section to a derivation of Blackwell's theorem using the framework developed in this paper.

G. Blackwell Theorem for Core-concave entropies

Throughout this section, let $K_1 : \mathcal{X} \to \mathcal{Y}$ and $K_2 : \mathcal{X} \to \mathcal{Z}$ share an input X and produce outputs Y and Z, respectively.

Channel K_2 is *degraded from* K_1 [5], written as $K_1 \geq_{\mathrm{d}} K_2$, if exists a channel $R : \mathcal{Y} \to \mathcal{Z}$ such that $K_2 = K_1 R$, i.e.,

$$K_2(z|x) = \sum_{y\in\mathcal{Y}} K_1(y|x)R(z|y) \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}.$$

It follows immediately from Theorem 2, by DPI, that whenever $K_1 \geq_{\mathrm{d}} K_2$, then:

$$\forall H \in \mathcal{H}, p_X;\ H(X|Y) \leq H(X|Z). \tag{4}$$

Whenever channels K_1, K_2 respect (4), we write $K_1 \geq_{\mathcal{H}} K_2$.

The converse statement, i.e., that $K_1 \geq_{\mathcal{H}} K_2 \Rightarrow K_1 \geq_{\mathrm{d}} K_2$ is a result that can be traced back to a theorem by Blackwell on comparison of experiments [14], [31]. It also relates to more recent results in [7], [23]. Here, we prove the theorem using Dahl's work on matrix majorization [7]. We start with the following definition which provides a useful characterization of conditional entropy.

Definition 6. *Given a continuous and concave function $F : \Delta_n \to \mathbb{R}$, we define G_F, for all $\mathbf{q} = \{q_1, \ldots, q_n\} \in \mathbb{R}^n_{\geq 0}$, as*

$$G_F(\mathbf{q}) = \left(\sum_{i=1}^{n} q_i\right) F\left(\frac{q_1}{\sum_{j=1}^n q_j}, \ldots, \frac{q_n}{\sum_{j=1}^n q_j}\right)$$

if $\mathbf{q}$ is not the null vector, and $G_F(0, \ldots, 0) = 0$.[2] In short, using the 1-norm notation, we have $G_F(\mathbf{q}) = \|\mathbf{q}\|_1 F\left(\frac{\mathbf{q}}{\|\mathbf{q}\|_1}\right)$.

Given $H = (\eta, F) \in \mathcal{H}$, it is possible to define the conditional entropy $H(X|Y)$ in terms of the functions G_F and of the joint probability distribution $p(x, y)$ as the following:

$$H(X|Y) = \eta\left(\sum_{y \in \mathcal{Y}^+} G_F(p(x_1, y), \ldots p(x_n, y))\right). \tag{5}$$

Definition 7. *A* cone $V \subset \mathbb{R}^n$ *is a set such that for all $\lambda \in \mathbb{R}_{>0}$, $\mathbf{q} \in V \implies \lambda\mathbf{q} \in V$. A function $\phi : V \to \mathbb{R}$ over a cone $V \subset \mathbb{R}^n$ is* positively homogeneous *if, for all $\lambda \in \mathbb{R}_{>0}$ and $\mathbf{q} \in V$, $\phi(\lambda\mathbf{q}) = \lambda\phi(\mathbf{q})$.*

Proposition 10. *A function $\phi : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ is continuous, positively homogeneous and concave if, and only if, it coincides with G_F for some concave function F.*

Proof: First, we prove that G_F is continuous, positively homogeneous and concave. It is immediate to see that G_F is continuous and positively homogeneous. To see that G_F is concave, consider the affine map:

$$g_1(q_1, \ldots, q_n) = \left(q_1, \ldots, q_n, \sum_i q_i\right),$$

and the perspective function $g_2(\mathbf{q}, t) = tF(\mathbf{q}/t)$, which is concave since F is concave [32, Chapter 3.2.6]. Then, for all non zero $\mathbf{q} \in \mathbb{R}^n_{\geq 0}$, $G_F(\mathbf{q}) = g_2(g_1(\mathbf{q}))$. Therefore, being the composition of an affine map and a concave function, G_F is concave [32, Chapter 3.2.2].

Conversely, suppose $\phi : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ is continuous, positively homogeneous and concave. Then we can write ϕ as a G_F by taking $F = \phi|_{\Delta_n}$, where $\phi|_{\Delta_n}$ the restriction of ϕ to Δ_n. To see that, notice that $\phi(\frac{\mathbf{q}}{\|\mathbf{q}\|_1}) = \frac{1}{\|\mathbf{q}\|_1}\phi(\mathbf{q})$, and hence:

$$G_{\phi|_{\Delta_n}}(\mathbf{q}) = \|\mathbf{q}\|_1 \phi\left(\frac{\mathbf{q}}{\|\mathbf{q}\|_1}\right) = \phi(\mathbf{q}). \qquad \blacksquare$$

[2]Notice that, as F is continuous over a compact set (and thus, bounded), $\lim_{\epsilon \to 0^+} G_F(\epsilon\mathbf{q}) = 0 = G_F(0, \ldots, 0)$, for all $\mathbf{q} \in \mathbb{R}^n_{\geq 0}$.

The proof of the main result relies on the following ([7]):

Theorem 5. *Let A and B be real-valued matrices with m rows, and denote by A^i, B^i their i-th column. There is a row stochastic matrix R such that $AR = B$ if and only if for all positively homogeneous convex functions $\phi : \mathbb{R}^m \to \mathbb{R}$,*

$$\sum_i \phi(A^i) \geq \sum_j \phi(B^j).$$

Theorem 6 (Blackwell Theorem for core-concavity)**.** *We have:*

$$K_1 \geq_{\mathrm{d}} K_2 \Leftrightarrow K_1 \geq_{\mathcal{H}} K_2,$$

that is, given channels $K_1 : \mathcal{X} \to \mathcal{Y}$ and $K_2 : \mathcal{X} \to \mathcal{Z}$, $K_2 = K_1 R$ for some channel $R : \mathcal{Y} \to \mathcal{Z}$ if, and only if, for all distributions p over X, $H(X|Y) \leq H(X|Z)$ for all core-concave entropies H.

Proof: The forward implication is the "data processing" inequality and is proved in Proposition 4.

For the reverse implication, suppose that $H(X|Y) \leq H(X|Z)$ for a full support p and all core-concave H. Let $\mathrm{diag}(p)$ be the matrix with the values of p in the diagonal and 0 elsewhere, and let $A = \mathrm{diag}(p)K_1$, $B = \mathrm{diag}(p)K_2$ – that is, A and B are the matrices of the joint distributions obtained from p and K_1, K_2. Then, for all choices of F, we have

$$\sum_{y \in \mathcal{Y}^+} G_F(A^y) \leq \sum_{z \in \mathcal{Z}^+} G_F(B^z).$$

Notice that all positively homogeneous functions over $\mathbb{R}^n$ are also positively homogeneous over $\mathbb{R}^n_{\geq 0}$, and that a convex function over $\mathbb{R}^n$ is continuous. Therefore, from Theorem 5 and Proposition 10, there is a row stochastic (channel) matrix R such that $AR = B$. As p has full support, $\mathrm{diag}(p)$ is non-singular, and therefore $K_1 R = \mathrm{diag}(p)^{-1}AR = \mathrm{diag}(p)^{-1}B = K_2$. ∎

Conclusion

In this work, we formalized a general form of conditional entropy by introducing the notion of "core-concavity" in an axiomatic framework. Using these axioms, we showed that core-concavity characterizes the largest class of functions for which the data-processing inequality holds, under the assumption that conditional entropy is defined as a generalized average. We also showed that given core-concavity, the data-processing inequality and "conditioning reduces entropy" are equivalent. Several other well known properties of Shannon entropy are generalized in this framework, including perfect secrecy and Fano's inequality. A natural connection to the problem of statistical decision making was also developed. Finally, we established that core-concavity completely characterizes the "degradedness" ordering.

A potential future direction of this axiomatic work is to study divergence-based definitions of entropy within the core-concave framework. For instance, [33], [34] develop improved lower and upper bounds on data-processing inequalities and generalized Fano's inequality based on higher order properties of the divergence functions. It would be interesting to study these bounds within the core-concave framework and their possible applications to computer security and privacy.

REFERENCES

[1] A. Américo, M. Khouzani, and P. Malacaria, "Core-concavity, Gain Functions and Axioms for Information Leakage," in *The Art of Modelling Computational Systems: A Journey from Logic and Concurrency to Security and Privacy*. Springer Nature Switzerland AG, 2019.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 625–56, 1948.

[3] A. Rényi, "On Measures of Entropy and Information," in *Proc. 4th Berkeley Symposium on Mathematics, Statistics, and Probability*, 1961, pp. 547–561.

[4] M. Khouzani and P. Malacaria, "Generalized Entropies and Metric-Invariant Optimal Countermeasures for Information Leakage Under Symmetric Constraints," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 888–901, 2018.

[5] T. M. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, 1972.

[6] M. S. Alvim, K. Chatzikokolakis, C. Palamidessi, and G. Smith, "Measuring information leakage using generalized gain functions," in *Proc. IEEE 25th Computer Security Foundations Symposium (CSF)*, 2012, pp. 265–279.

[7] G. Dahl, "Matrix majorization," *Linear Algebra and its Applications*, vol. 288, pp. 53 – 73, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0024379598101751

[8] D. K. Faddeev, "On the concept of entropy of a finite probabilistic scheme," *Uspekhi Matematicheskikh Nauk*, vol. 11, no. 1, pp. 227–231, 1956.

[9] I. Vajda and K. Vašek, "Majorization, concave entropies, and comparison of experiments." *Problems of Control and Information Theory*, vol. 14, no. 2, pp. 105–116, 1985.

[10] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–273, 2008. [Online]. Available: https://www.mdpi.com/1099-4300/10/3/261

[11] M. S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith, "An axiomatization of information flow measures," *Theoretical Computer Science*, vol. 777, pp. 32 – 54, 2019, in memory of Maurice Nivat, a founding father of Theoretical Computer Science - Part I. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304397518306376

[12] A. Teixeira, A. Matos, and L. Antunes, "Conditional rényi entropies," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4273–4277, July 2012.

[13] M. Iwamoto and J. Shikata, "Information theoretic security for encryption based on conditional rényi entropies," in *Information Theoretic Security*, C. Padró, Ed. Cham: Springer International Publishing, 2014, pp. 103–121.

[14] D. Blackwell, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, vol. 24, no. 2, pp. 265–272, 1953. [Online]. Available: http://www.jstor.org/stable/2236332

[15] S. Sherman, "On a theorem of hardy, littlewood, polya, and blackwell," *Proceedings of the National Academy of Sciences*, vol. 37, no. 12, pp. 826–831, 1951. [Online]. Available: https://www.pnas.org/content/37/12/826

[16] M. Leshno and Y. Spector, "An elementary proof of blackwell's theorem," *Mathematical Social Sciences*, vol. 25, no. 1, pp. 95 – 98, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0165489692900284

[17] B. Sharma and D. Mittal, "New non-additive measures of entropy for discrete probability distributions," *Journal of Mathematical Science (Soc. Math. Sci., Calcutta, India*, vol. 10, pp. 28–40, 1975.

[18] J. Havrda and F. Charvát, "Quantification method of classification processes. concept of structural α-entropy," *Kybernetika*, vol. 3, no. 1, pp. 30–35, 1967.

[19] C. Tsallis, "Possible generalization of boltzmann-gibbs statistics," *Journal of statistical physics*, vol. 52, no. 1-2, pp. 479–487, 1988.

[20] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: theory of majorization and its applications*. Mathematics In Science And Engineering, Academic Press, 1979, vol. 143.

[21] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," *Topics in information theory*, 1977.

[22] M. Hayashi, "Exponential decreasing rate of leaked information in universal random privacy amplification," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3989–4001, 2011.

[23] A. McIver, C. Morgan, G. Smith, B. Espinoza, and L. Meinicke, "Abstract channels and their robust information-leakage ordering," in *Proc. 3rd Int. Conf. Principles of Security and Trust (POST)*, ser. LNCS, vol. 8414. Springer, 2014, pp. 83–102.

[24] M. Ben-Bassat and J. Raviv, "Renyi's entropy and the probability of error," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 324–331, May 1978.

[25] G. Hardy, J. Littlewood, K. M. R. Collection, G. Pólya, D. Littlewood, and G. Pólya, *Inequalities*, ser. Cambridge Mathematical Library. Cambridge University Press, 1952. [Online]. Available: https://books.google.co.uk/books?id=t1RCSP8YKt8C

[26] M. Salicru, M. Menendez, D. Morales, and L. Pardo, "Asymptotic distribution of (h, ϕ)-entropies," *Communications in Statistics - Theory and Methods*, vol. 22, no. 7, pp. 2015–2031, 1993.

[27] S. Arimoto, "Information-theoretical considerations on estimation problems," *Information and Control*, vol. 19, no. 3, pp. 181 – 194, 1971. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0019995871900659

[28] B. C. Arnold, *Majorization and the Lorenz order: A brief introduction*. Springer Science & Business Media, 2012, vol. 43.

[29] W. van Dam and P. Hayden, "Renyi-entropic bounds on quantum communication," *arXiv preprint quant-ph/0204093*, 2002.

[30] P. D. Grünwald, A. P. Dawid *et al.*, "Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory," *the Annals of Statistics*, vol. 32, no. 4, pp. 1367–1433, 2004.

[31] D. Blackwell, "The comparison of experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, Ed. Berkeley: Univ. of California Press, 1951, pp. 93–102.

[32] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[33] I. Sason and S. Verdú, "Arimoto–rényi conditional entropy and bayesian m -ary hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 4–25, 2018.

[34] I. Sason, "On data-processing and majorization inequalities for f-divergences with applications," *Entropy*, vol. 21, no. 10, 2019.

Arthur Américo received his bachelor's degree in Physics and master's degree in Computer Science from Universidade Federal de Minas Gerais, in Belo Horizonte, Brazil. He is currently a Ph.D. student in Computer Science at Queen Mary University of London, under the supervision of Prof. Malacaria and Dr. Khouzani. His research focuses on mathematical tools for quantifying information leakage in security systems.

MHR Khouzani received his Ph.D. in Electrical and Systems Engineering in 2011 from University of Pennsylvania. He is currently a Lecturer in the EECS department of Queen Mary University of London (QMUL). Dr. Khouzani's research has been in the area of information security.

Pasquale Malacaria received his Laurea in Philosophy from "La Sapienza" University in Rome and his PhD in "Logique et fondements de l'Informatique" from the University of Paris VII in France. His work focuses on information theory, game theory, verification and their applications to computer security. He is a Professor of Computer Science at Queen Mary University of London. He has been an EPSRC advanced research fellow, is a recipient of the Alonzo Church award 2017 and of the Facebook Faculty awards 2015.