Rényi Divergence and Majorization

Tim van Erven
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
E-mail: Tim.van.Erven@cwi.nl

Peter Harremoës
Copenhagen Business College
Copenhagen, Denmark
E-mail: harremoes@ieee.org

***Abstract*—Rényi divergence is related to Rényi entropy much like information divergence (also called Kullback-Leibler divergence or relative entropy) is related to Shannon's entropy, and comes up in many settings. It was introduced by Rényi as a measure of information that satisfies almost the same axioms as information divergence.**

We review the most important properties of Rényi divergence, including its relation to some other distances. We show how Rényi divergence appears when the theory of majorization is generalized from the finite to the continuous setting. Finally, Rényi divergence plays a role in analyzing the number of binary questions required to guess the values of a sequence of random variables.

I. Introduction

Since Shannon's introduction of his entropy function various other similar measures of uncertainty or information have been introduced. Most of these have found no applications and some have found applications only in quite special cases. An exception is formed by Rényi entropy and Rényi divergence, which pop up again and again. They are far from being as well understood as Shannon entropy and Shannon divergence, and do not have as simple an interpretation. Erdal Arikan observed that the discrete version of Rényi entropy is related to so-called guessing moments [1].

In this short note we shall first review the most important properties of Rényi divergence in Section II. In Section III we give a very brief introduction to Markov ordering and its relation to majorization. Then in Sections IV, and V we relate Rényi divergence to the theory of majorization. And finally, in Section VI we will show that, like its entropy counterpart, Rényi divergence is related to guessing moments.

II. Rényi Divergence

Let P and Q be probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$, and let p and q be their densities with respect to a common σ-finite dominating measure μ. Then for any $0 < \alpha < \infty$ except $\alpha = 1$, the *Rényi divergence* D_α of *order* α of P from Q is defined as

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu, \tag{1}$$

with the conventions that $p^\alpha q^{1-\alpha} = 0$ if $p = q = 0$, even for $\alpha < 0$ and $\alpha > 1$, and that $x/0 = \infty$ for $x > 0$. Continuity considerations lead to the following extensions for $\alpha \in \{0, 1\}$:

$$D_0(P\|Q) = \lim_{\alpha \downarrow 0} D_\alpha(P\|Q) = -\log Q(p > 0),$$
$$D_1(P\|Q) = \lim_{\alpha \uparrow 1} D_\alpha(P\|Q) = D(P\|Q),$$

where $D(P\|Q) = \int p \log p/q \, \mathrm{d}\mu$ (with the conventions that $0 \log 0/x = 0$ and $x \log x/0 = \infty$ if $x > 0$) denotes the *information divergence*, which is also known as Kullback-Leibler divergence or relative entropy. For $\alpha > 0$, it was introduced by Rényi [2], who provided an axiomatic characterization in terms of "intuitively evident postulates". An operational characterizations of Rényi divergence via coding has been described [3].

We will first review some of the basic properties of D_α. Whenever these properties can easily be derived from known results, we will point to the relevant literature. For other properties, space requirements limit us to only hint at their proofs. A longer version of this paper with full proofs will be published elsewhere, and will include results for negative values of the order α.

Let us start by noting that, for finite orders $0 < \alpha \neq 1$, D_α is a continuous, strictly increasing function of the power divergence

$$d_\alpha(P, Q) = \frac{\int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu - 1}{\alpha - 1}.$$

As d_α are *f-divergences*, we may derive properties for D_α from general properties of f-divergences [4].

In particular, Rényi divergence satisfies the *data processing inequality*

$$D_\alpha(P_{|\mathcal{G}} \| Q_{|\mathcal{G}}) \leq D_\alpha(P\|Q)$$

for any σ-subalgebra $\mathcal{G} \subseteq \mathcal{F}$, where $P_{|\mathcal{G}}$ and $Q_{|\mathcal{G}}$ denote the restrictions of P and Q to $\mathcal{G}$. As a special case, taking $\mathcal{G} = \{0, \mathcal{X}\}$ to be the trivial algebra, we find that

$$D_\alpha(P\|Q) \geq 0.$$

$D_\alpha(P\|Q) = 0$ if and only if $P = Q$. Taking $\mathcal{G} = \sigma(\mathcal{P})$ to be the σ-algebra generated by a finite partition $\mathcal{P}$ of $\mathcal{X}$, the data processing inequality implies that discretizing $\mathcal{X}$ can only decrease D_α. However, because of the following property, which carries over from f-divergences, D_α may be approximated arbitrarily well by such finite partitions:

$$D_\alpha(P\|Q) = \sup_{\mathcal{P}} D_\alpha(P_{|\sigma(\mathcal{P})} \| Q_{|\sigma(\mathcal{P})}) \qquad (\alpha > 0), \tag{2}$$

where the supremum is over all finite partitions $\mathcal{P}$ of $\mathcal{X}$. This characterization also shows that we have found the right generalization of Rényi's definition for finite $\mathcal{X}$.

Using the dominated convergence theorem it can be shown that:

Theorem 1: D_α is continuous in α on

$$A = \{\alpha \mid 0 \leq \alpha \leq 1 \text{ or } D_\alpha(P\|Q) < \infty\}.$$

D_α is also nondecreasing in α, and on A it is constant if and only if q/p is constant P-a.s.

The fact that D_α is nondecreasing, together with Equation (2), implies that $\lim_{\alpha\uparrow 1} D_\alpha = D$, as asserted in our definition of D_1: for finite $\mathcal{X}$, this can be verified directly using l'Hôpital's rule. Therefore

$$\begin{aligned}\lim_{\alpha\uparrow 1} D_\alpha(P\|Q) &= \sup_{\alpha<1}\sup_{\mathcal{P}} D_\alpha(P_{|\sigma(\mathcal{P})}\|Q_{|\sigma(\mathcal{P})}) \\ &= \sup_{\mathcal{P}}\sup_{\alpha<1} D_\alpha(P_{|\sigma(\mathcal{P})}\|Q_{|\sigma(\mathcal{P})}) = D(P\|Q). \quad (3)\end{aligned}$$

The assertion that $\lim_{\alpha\downarrow 0} D_\alpha = -\log Q(p > 0)$ is verified differently, using the dominated convergence theorem and the observation that $\lim_{\alpha\downarrow 0} p^\alpha q^{1-\alpha}$ equals q if $p > 0$ and 0 otherwise. Rényi divergence may be extended to $\alpha = \infty$ by letting α tend to ∞. Then, for finite $\mathcal{X}$,

$$D_\infty(P\|Q) = \log \max_{x\in\mathcal{X}} \frac{P(x)}{Q(x)},$$

and by an interchanging of suprema similar to (3) we find that

$$D_\infty(P\|Q) = \log \sup_{A\in\mathcal{F}} \frac{P(A)}{Q(A)} = \log \operatorname*{ess\,sup}_{x\in\mathcal{X}} \frac{\mathrm{d}P}{\mathrm{d}Q}(x)$$

in general. Consequently, $D_\infty(Q\|P)$ (note the reversal of P and Q) is a one-to-one function of the separation distance $s(P,Q) = \max_x(1 - P(x)/Q(x))$, defined only for countable $\mathcal{X}$, which has been used to obtain bounds on the rate of convergence to the stationary distribution for certain Markov chains [5], [6].

Equation 2 implies that there exists a sequence $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of σ-algebras generated by finite partitions such that

$$\lim_{n\to\infty} D_\alpha(P_{|\mathcal{F}_n}\|Q_{|\mathcal{F}_n}) = D_\alpha(P\|Q).$$

By the connection to f-divergences, such a convergence result holds for any *increasing* sequence of σ-algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_\infty = \sigma\left(\bigcup_{n=1}^\infty \mathcal{F}_n\right) \subseteq \mathcal{F}$:

$$\lim_{n\to\infty} D_\alpha(P_{|\mathcal{F}_n}\|Q_{|\mathcal{F}_n}) = D_\alpha(P_{|\mathcal{F}_\infty}\|Q_{|\mathcal{F}_\infty}) \qquad (\alpha > 0) \tag{4}$$

[4, Theorem 15]. By a suitable choice of $\mathcal{F}_n$ this result extends *additivity* for any distributions $P_1, P_2, \ldots$ and $Q_1, Q_2, \ldots$,

$$\sum_{n=1}^N D_\alpha(P_n\|Q_n) = D_\alpha(P_1 \times \cdots \times P_N\|Q_1 \times \cdots \times Q_N),$$

from any finite N (for which it is easy to prove) to $N = \infty$ (if $\alpha > 0$). For $\alpha = 0$ additivity only holds for finite N. By a direct proof we can also prove the counterpart to (4) for decreasing sequences of σ-algebras $\mathcal{F} \supseteq \mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \cdots \supseteq \mathcal{F}_\infty = \bigcap_{n=1}^\infty \mathcal{F}_n$ (for finite α) under the condition that the divergence is finite.

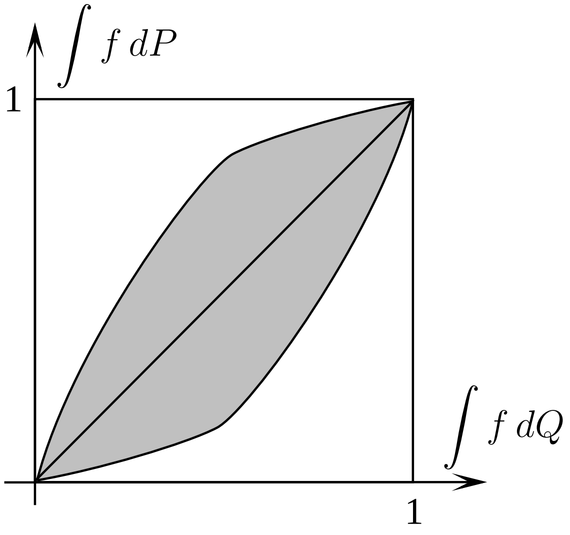

Fig. 1. Example of a Lorenz diagram.

Let

$$H^2(P,Q) = \int (p^{1/2} - q^{1/2})^2 \,\mathrm{d}\mu = 2 - 2d_{1/2}(P,Q)$$

denote the squared *Hellinger distance*, and let

$$\chi^2(P,Q) = \int \frac{(p-q)^2}{q}\mathrm{d}\mu = d_2(P,Q) - 1$$

denote the χ^2*-distance* [5]. We see that

$$D_{1/2}(P\|Q) = -2\log(1 - H^2(P,Q)/2)$$

and $D_2(P\|Q) = \log(1 + \chi^2(P,Q))$. Hence by $\log x \leq x - 1$

$$\begin{aligned}H^2(P,Q) \leq D_{1/2}(P\|Q) \leq D(P\|Q)& \\ \leq D_2(P\|Q) \leq \chi^2(P,Q).&\end{aligned}$$

III. Majorization, Markov ordering and Lorenz diagrams

The general theory of majorization is now a well established mathematical discipline [7]. The majorization lattice and its relation to discrete entropy was studied in [8] and later generalized in [9]. Recently a long article on this subject by Gorban, Gorban, and Judge has been accepted for publication [10]. We refer to these papers for a more complete discussion and further references. Here we shall relate the relative majorization lattice to Rényi divergence.

Definition 2: Let P and Q be measures on the same measurable set. The *Lorenz diagram* of (P,Q) is the range of

$$f \mapsto \left(\int f\ \mathrm{d}P, \int f\ \mathrm{d}Q\right),$$

where f is any measurable function with values in $[0,1]$.

If Q is the uniform distribution then the Lorenz diagram of (P_1,Q) is a subset of the Lorenz diagram of (P_2,Q) if and only if P_2 *majorizes* P_1.

Theorem 3: The Lorenz diagram of (P_1, Q) is a subset of the Lorenz diagram of (P_2, Q) if and only if there exists a Markov operator that transforms P_2 into P_1 and leaves Q invariant.

Definition 4: Let P_1, P_2 and Q be measures on the same measurable set $\mathcal{X}$. We write $P_2 \succeq_Q P_1$ if the Lorenz diagram of (P_1, Q) is a subset of the Lorenz diagram of (P_2, Q). If the Lorenz diagrams of (P_1, Q) and (P_2, Q) are equal, then we write $P_1 \simeq_Q P_2$.

This ordering that generalizes majorization will be celled the *Markov ordering* [10][1].

Theorem 5 ([9]): Let Q be a measure on a measurable set $\mathcal{X}$. If Q is a uniform distribution on a finite set or if Q has no atoms, then $M_+^1(\mathcal{X}) / \simeq_Q$ is a lattice, where $M_+^1(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$.

The Lorenz diagram is characterized by a lower bound curve that is convex and an upper bounding curve that is concave. Because of the symmetry around $(1/2, 1/2)$ the Lorenz diagram is completely determined by the lower bounding curve.

Definition 6: The *Lorenz curve* of (P, Q) is the convex envelope of the Lorenz diagram, i.e. the largest convex function such that all the points in the Lorenz diagram are at or above the curve.

Proposition 7 ([9]): Let P and Q be measures on the same measurable set $\mathcal{X}$. The Lorenz curve of (P, Q) is the convex envelop of the points $(P(A_t), Q(A_t))$ where A_t are events of the form $A_t = \left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P}{\mathrm{d}Q} \leq t\right\}$.

In statistics the sets $A_t = \left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P}{\mathrm{d}Q} \leq t\right\}$ play the role of acceptance sets related to the likelihood ratio test of ratio t. The proof of this proposition is therefore essentially the same as the proof of the Neyman-Pearson Lemma [9]. Note that for discrete measures there will only be finitely many different points of the form $(P(A_t), Q(A_t))$, and in that case the Lorenz curve is piecewise linear. For $t_1 < t_2$

$$\frac{P(A_{t_2}) - P(A_{t_1})}{Q(A_{t_2}) - Q(A_{t_1})} = \frac{P\left(\left\{x \mid t_1 < \frac{\mathrm{d}P}{\mathrm{d}Q} \leq t_2\right\}\right)}{Q\left(\left\{x \mid t_1 < \frac{\mathrm{d}P}{\mathrm{d}Q} \leq t_2\right\}\right)} \in \left]t_1, t_2\right],$$

so $(P(A_t), Q(A_t))$ gives a parametrization of the Lorenz curve in terms of its slope if it is differentiable.

Suppose Q is the counting measure on a finite set $\mathcal{X}$ of size n, and let $P_1 = (v_1, \ldots, v_n)$ be a discrete measure on $\mathcal{X}$. Then A_t is simply $\{i \mid v_i \leq t\}$. Let $P_2 = (w_1, \ldots, w_n)$ be another measure and let $B_t = \{i \mid w_i \leq t\}$. Then $P_1 \preceq P_2$ if and only if $P_1(A_{t_1}) \geq P_2(B_{t_2})$ whenever $Q(A_{t_1}) = Q(B_{t_2})$. Thus $P_1 \preceq P_2$ if and only if the Lorenz curve of (P_1, Q) is above the Lorenz curve of (P_2, Q).

If one of the conditions of Theorem 5 is fulfilled, then for each convex function f there exists a measure P such that f is the Lorenz curve of P. Thus $M_+^1(\mathcal{X}) / \simeq_Q$ can be identified with the set of Lorenz curves. Let P_1 and P_2 be measures and let L_1 and L_2 be their Lorenz curves. Then $P_1 \wedge P_2$ can be identified with the Lorenz curve $\max\{L_1, L_2\}$ and $P_1 \vee P_2$ can be identified with the Lorenz curve that is the convex envelop of $\min\{L_1, L_2\}$. In general this lattice is neither modular nor distributive [9].

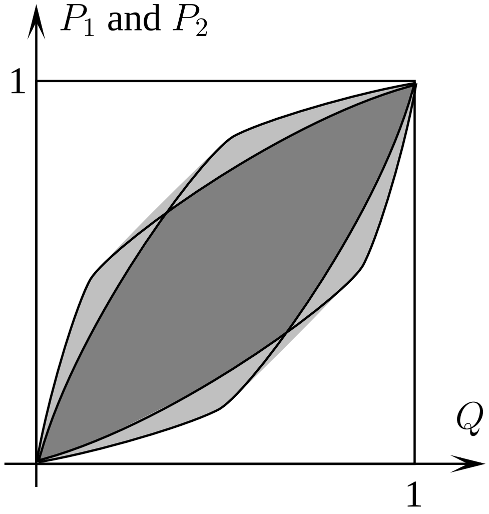

Fig. 2. The met and join of P_1 and P_2 have Lorenz diagrams that are the intersection (dark gray) and the convex hull of their union (light gray).

IV. Divergence, Convexity and Ordering

We will now consider properties of $D_\alpha(P\|Q)$ as we vary P and Q while keeping α fixed. Information divergence $D(P\|Q)$ is known to be jointly *convex* in the pair (P, Q) [11]. By an argument similar to the proof for D_1 in [11], this property generalizes to D_α for arbitrary order $0 \leq \alpha \leq 1$:

Theorem 8: For $0 \leq \alpha \leq 1$, $D_\alpha(P, Q)$ is jointly convex in the pair (P, Q).

Even though joint convexity does not generalize to $\alpha > 1$, we still have:

Theorem 9: For all α, $D_\alpha(P\|Q)$ is convex in Q.

The key step in proving the latter result for $\alpha > 1$ relies on Hölder's inequality.

Let P be absolutely continuous with respect to Q. If F denotes the curve that upper bounds the Lorenz diagram, then the Rényi divergence is given by

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int_0^1 (F'(t))^\alpha \ \mathrm{d}t.$$

Note that we can replace the upper bounding function by the lower bounding function (the Lorenz curve) without changing the integral.

Theorem 10: For $\alpha > 0$ the Rényi divergence $D_\alpha(P\|Q)$ is a increasing function of P on the lattice corresponding to Q.

Proof: Let F and G be concave functions on $[0, 1]$ such that $F \leq G$ and $F(0) = G(0) = 0$ and $F(1) = G(1) = 1$. Let $x \mapsto \Phi_x$ be a Markov kernel such that $x = \int y \ \mathrm{d}\Phi_x(y)$ for all $x \in [0, \infty[$. Then

$$\int G(y) \ \mathrm{d}\Phi_x(y) \leq G\left(\int y \ \mathrm{d}\Phi_x(y)\right) = G(x). \qquad (5)$$

[1] In [9] this ordering was called *relative majorization*.

Consider the set of all Markov kernels $x \mapsto \Phi_x$ such that $x = \int y \ \mathrm{d}\Phi_x(y)$ and $F(x) \leq \int G(y) \ \mathrm{d}\Phi_x(y)$ for all $x \in [0; \infty[$. This set is convex and contains an element such that $F(x) = \int G(y) \ \mathrm{d}\Phi_x(y)$. Then $F'(x) = \int G'(y) \ \mathrm{d}\Phi_x(y)$ and the theorem follows from Jensen's inequality. ■

Theorem 10 is essentially a noisy data processing inequality because the Markov kernel Φ_x in the proof essentially maps the measure corresponding to G into the measure corresponding to P. By adapting a proof from [9] is possible to prove the following theorem:

Theorem 11: Let P_1 and P_2 denote distributions that are absolutely continuous with respect to Q. If Markov ordering is taken with respect to Q then power divergence is sub-modular and super-additive, i.e.

$$d_\alpha(P_1, Q) + d_\alpha(P_2, Q) \geq d_\alpha(P_1 \wedge P_2, Q) + d_\alpha(P_1 \vee P_2, Q)$$

and

$$d_\alpha(P_1, Q) + d_\alpha(P_2, Q) \leq d_\alpha(P_1 \wedge P_2, Q).$$

Since power divergence is a function of Rényi divergence one can reformulate Theorem 11 in terms of Rényi divergence. Like Rényi divergence, the power divergence $d_\alpha(P, Q)$ tends to the information divergence $D(P\|Q)$ as $\alpha \uparrow 1$. This implies:

Corollary 12: Let P_1 and P_2 be distributions that are absolutely continuous with respect to Q. If the Markov ordering is taken with respect to Q then information divergence is sub-modular and super-additive, i.e.

$$\begin{aligned} D(P_1\|Q) + D(P_2\|Q) &\geq D(P_1 \wedge P_2\|Q) + D(P_1 \vee P_2\|Q) \\ D(P_1\|Q) + D(P_2\|Q) &\leq D(P_1 \wedge P_2\|Q). \end{aligned}$$

V. Continuity of Rényi divergence

The type of continuity of D_α in the pair (P, Q) turns out to depend on the topology and on α. We consider the *τ-topology*, in which convergence of P_n to P means that $P_n(A) \to P(A)$ for all $A \in \mathcal{F}$, and the *total variation topology* in which $P_n \to P$ if the variation distance between P_n and P goes to zero. In general the total variation topology is stronger than the τ-topology, but if $\mathcal{X}$ is countable, then the two topologies coincide.

Theorem 13: For any $\alpha > 0$, $D_\alpha(P\|Q)$ is a lower semi-continuous function of (P, Q) in the τ-topology.

Moreover:

Theorem 14: For $0 < \alpha < 1$, $D_\alpha(P\|Q)$ is a (uniformly) continuous function of (P, Q) in the total variation topology.

It remains to consider $\alpha = 0$. In this case:

Corollary 15: $D_0(P\|Q)$ is an upper semi-continuous function of (P, Q) in the total variation topology.

Using the Markov ordering we get more insight.

Theorem 16: If $\alpha \geq 1$ and $D_\alpha(\tilde{P}\|Q) < \infty$, then the Rényi divergence $D_\alpha(P\|Q)$ is continuous in P on the set $\left\{P \mid P \preceq_Q \tilde{P}\right\}$ when the set of probability measures is equipped with the topology of total variation.

Proof: If $P_n \to P$ in total variation for $n \to \infty$ then the Lorenz diagram of P_n tends to the Lorenz diagram of P in Hausdorff distance. Let $F, \tilde{F}$ and F_n denote the upper bounding functions for $P, \tilde{P}$ and P_n. Then for any $\varepsilon > 0$ eventually $F_n(t) \leq \min\left\{\tilde{F}(t), F(t+\varepsilon)\right\}$ for all $t \in [0, 1]$. Hence

$$\begin{aligned} &\limsup_{n\to\infty} D_\alpha(P_n\|Q) \\ &\quad \leq \frac{1}{\alpha - 1} \log \int_0^1 \left(\frac{d}{dt} \min\left\{\tilde{F}(t), F(t+\varepsilon)\right\}\right)^\alpha \ dt. \end{aligned}$$

This holds for all $\varepsilon > 0$ and, since the right-hand side tends to $\frac{1}{\alpha-1} \log \int_0^1 \left(\frac{d}{dt} F(t)\right)^\alpha \ dt = D_\alpha(P\|Q)$ for $\varepsilon \to 0$, the result follows. ■

VI. Guessing moments

Erdal Arikan observed that the discrete version of Rényi entropy is related to so-called guessing moments [1]. In this short note we shall see that Rényi divergences are also related to guessing moments.

Definition 17: Let P_1 and P_2 denote probability measures on $\mathcal{X}$. We say that P_1 is *a rearrangement* of P_2 if

$$Q\left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P_1}{\mathrm{d}Q}(x) \geq t\right\} = Q\left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P_2}{\mathrm{d}Q}(x) \geq t\right\}$$

for all $t \in \mathbb{R}$.

Definition 18: A guessing function in $\mathcal{X}$ is a function $g : \mathcal{X} \to \mathbb{R}$ such that $Q(\{x \mid g(x) \leq t\}) \leq t$ for $t \in [0, 1]$.

For a probability measure P on $\mathcal{X}$ with density $\frac{\mathrm{d}P}{\mathrm{d}Q}$ we are interested in bounds on the moments of guessing functions. For a guessing function g the ρ-th moment is given by

$$\|g\|_\rho = \left(\int_{\mathbb{R}^d} (g(x))^\rho \ \mathrm{d}P(x)\right)^{1/\rho}.$$

Definition 19: Let P be a probability measure on $\mathcal{X}$. For each Radon-Nikodým derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$, the *ranking function* r of $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is given by

$$r(x) = Q\left(\left\{y \mid \frac{\mathrm{d}P}{\mathrm{d}Q}(y) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right\}\right).$$

We note that if F is the distribution function of $\frac{\mathrm{d}P}{\mathrm{d}Q}$ then the ranking function is given by $r(x) = 1 - F(x)$. The ranking function is a guessing function.

$$\begin{aligned} &Q(\{x \mid r(x) \leq t\}) = \\ &\quad Q\left(\left\{x \mid Q\left(\left\{y \mid \frac{\mathrm{d}P}{\mathrm{d}Q}(y) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right\}\right) \leq t\right\}\right) \leq t. \end{aligned}$$

Note that $Q(\{x \mid r(x) \leq t\}) = t$ for all $t \in [0, 1]$ if and only if the distribution of the random variable $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is continuous.

Proposition 20: The ranking function is the guessing function that minimizes the ρ-th moment if $\rho > 0$ and maximizes the ρ-th moment if $\rho < 0$.

Guessing and ranking are closely related to majorization and the Markov ordering via the following proposition.

Proposition 21: Assume that P_1, P_2 and Q are probability measures on $\mathcal{X}$ and $P_1 \preceq_Q P_2$. Let r_1 and r_2 denote the ranking functions of P_1 and P_2. Then

$$\begin{aligned} \|r_1\|_\rho &\le \|r_2\|_\rho \text{ if } \rho > 0, \\ \|r_1\|_\rho &\ge \|r_2\|_\rho \text{ if } \rho < 0. \end{aligned}$$

Lemma 22: If $\alpha = \frac{1}{1+\rho} > 0$ then, for any probability measures P and Q,

$$-\log\left(\|r\|_\rho\right) \ge D_\alpha\left(P\|Q\right),$$

where the ρ-norm is calculated with respect to Q and r is the ranking function of $\frac{\mathrm{d}P}{\mathrm{d}Q}$.

Proof: We have

$$\begin{aligned} r\left(x\right) &= \int_{\frac{\mathrm{d}P}{\mathrm{d}Q}(y) \ge \frac{\mathrm{d}P}{\mathrm{d}Q}(x)} 1 \ \mathrm{d}Q(y) = \int_{\frac{\mathrm{d}P}{\mathrm{d}Q}(y) \ge \frac{\mathrm{d}P}{\mathrm{d}Q}(x)} 1^\alpha \ \mathrm{d}Q(y) \\ &\le \int_{\frac{\mathrm{d}P}{\mathrm{d}Q}(y) \ge \frac{\mathrm{d}P}{\mathrm{d}Q}(x)} \left(\frac{\frac{\mathrm{d}P}{\mathrm{d}Q}\left(y\right)}{\frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)}\right)^\alpha \ \mathrm{d}Q(y) \\ &\le \int \left(\frac{\frac{\mathrm{d}P}{\mathrm{d}Q}\left(y\right)}{\frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)}\right)^\alpha \ \mathrm{d}Q(y) = \frac{\int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(y\right)\right)^\alpha \ \mathrm{d}Q(y)}{\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)\right)^\alpha}. \end{aligned}$$

We get

$$\begin{aligned} E\left[r\left(X\right)^\rho\right] &\le \int \left(\frac{\int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(y\right)\right)^\alpha \ \mathrm{d}Q(y)}{\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)\right)^\alpha}\right)^\rho \frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right) \ \mathrm{d}Q(x) \\ &= \left(\int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(y\right)\right)^\alpha \ \mathrm{d}Q(y)\right)^\rho \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)\right)^{1-\alpha\rho} \ \mathrm{d}Q(x) \\ &= \left(\int \ \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)\right)^\alpha \ \mathrm{d}Q(x)\right)^{\frac{1}{\alpha}}. \end{aligned}$$

We raise to the power $1/\rho$ and take minus the logarithm and get

$$\begin{aligned} \log\left(E\left[r\left(X\right)^\rho\right]^{\frac{1}{\rho}}\right) &\le \log\left(\left(\int \ \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)\right)^\alpha \ \mathrm{d}Q(x)\right)^{\frac{1}{\alpha\rho}}\right) \\ &= \frac{1}{1-\alpha}\log\left(\int \ \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)\right)^\alpha \ \mathrm{d}Q(x)\right) = -D_\alpha\left(P\|Q\right). \end{aligned}$$

■

Using additivity of Rényi divergence and Lemma 22 we get the following theorem.

Theorem 23: If $\alpha = \frac{1}{1+\rho} > 0$ then for any i.i.d. sequence $X_1^n = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$ we have

$$-\frac{1}{n}\log\left(\|r\left(X_1^n\right)\|_\rho\right) \ge D_\alpha\left(P\|Q\right).$$

This bound is asymptotically tight as stated in the following theorem.

Theorem 24: If $\alpha = \frac{1}{1+\rho} > 0$ then for any i.i.d. sequence $X_1^n = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$ we have

$$\lim_{n\to\infty} -\frac{1}{n}\log\left(\|r\left(X_1^n\right)\|_\rho\right) = D_\alpha\left(P\|Q\right).$$

The result gives a new interpretation of Rényi divergence.

VII. Discussion

The results in this short paper are formulated under the assumption that the second argument Q in $D_\alpha\left(P\|Q\right)$ is a probability measure. Nevertheless many of the results still hold if Q is a more general positive measure. For instance many results on Rényi entropy are obtained when Q denotes the counting measure. Most of these results for Rényi entropy are well-known. Results for differential Rényi entropy are obtained when Q is the Lebesgue measure. For both Rényi entropy and differential Rényi entropy many results should first be formulated and proved for subsets of finite measure and then one should take a limit for an increasing sequence of subsets. In this sense our results on Rényi divergence are often more general than the results one will find in the literature.

We have related Rényi divergence to majorization and Markov ordering. An interesting related concept is catalytic majorization. It has been proved by M. Klimesh that one discrete distribution majorizes another distribution if and only if certain inequalities hold between their Rényi entropies [12]. A similar result is still to be proved for Rényi divergence.

VIII. Acknowledgments

We thank Christophe Vignat, Matthew Klimesh, and Erdal Arikan for useful discussions.

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

[1] E. Arikan, “An inequality on guessing and its application to sequential decoding,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 99–105, Jan. 1996.

[2] A. Rényi, “On measures of entropy and information,” in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 547–561, 1961.

[3] P. Harremoës, “Interpretations of Rényi entropies and divergences,” *Physica A: Statistical Mechanics and its Applications*, vol. 365, pp. 57–62, June 2006.

[4] F. Liese and I. Vajda, “On divergence and informations in statistics and information theory,” *IEEE Tranns. Inform. Theory*, vol. 52, pp. 4394 – 4412, Oct. 2006.

[5] A. L. Gibbs and F. E. Su, “On choosing and bounding probability metrics,” *International Statistical Review*, vol. 70, pp. 419–435, 2002.

[6] D. Aldous and P. Diaconis, “Strong uniform times and finite random walks,” *Advances in Applied Mathematics*, vol. 8, pp. 69–97, 1987.

[7] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*. New York: Academic Press, 1979.

[8] F. Cicalese and U. Vaccaro, “Supermodularity and subadditivity of the entropy on the majorization lattice,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 933–938, 2002.

[9] P. Harremoës, “A new look on majorization,” in *Proceedings ISITA 2004*, (Parma, Italy), pp. 1422–1425, Oct. 2004.

[10] A. N. Gorban, P. A. Gorban, and G. Judge, “The markov ordering approach,” *Entropy*, vol. 12, May 2010. To appear in a special Issue entitled ”Entropy in Model Reduction”.

[11] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[12] M. Klimesh, “Entropy measures and catalysis of bipartite quantum state transformations,” in *Proceedings 2004 IEEE International Symposium on Information Theory*, p. 357, June 27 - Luly 2 2004.