Some Properties of Rényi-like Information Measures

Ganesh Ajjanagadde,Yury Polyanskiy

***Abstract*—Rényi entropy, Rényi divergence, and α-mutual information are all generalizations of their classical counterparts which have been useful in information theory and its applications. A number of their interesting properties have been established in the literature, and although they are similar in many respects to their classical counterparts, they also suffer from some serious disadvantages. One such property is the lack of a nice single-letterization for these quantities. We compute capacities defined with respect to α-mutual information for some simple channels, and illustrate the lack of a single-letterization. We then examine a potential application of Rényi entropy to the binary adder MAC (multiple access channel). In spite of not improving existing converse bounds due to a lack of single-letterization for Rényi entropy, we establish some bounds that generalize classical entropy sub-additivity (a form of single-letterization) in various directions. These can be viewed as a non-standard form of single-letterization.**

***Index Terms*—Rényi entropy, α-mutual information, strong converse, multiple access channel**

I. Introduction

In [1], through an alternative set of axioms that a notion of entropy must satisfy, Alfred Rényi generalized the classical Shannon entropy as a measure of randomness of a probability distribution:

Definition 1. *For a discrete random variable X on alphabet $\mathcal{A}$, the Rényi entropy of order $\alpha \in [0, \infty]$ is given by*

$$H_\alpha(X) = \begin{cases} \log |\{a \in \mathcal{A} : P_X(a) > 0\}| & \alpha = 0 \\ \frac{1}{1-\alpha} \log \left(\sum_{a \in \mathcal{A}} P_X(a)^\alpha \right) & \alpha \in (0,1) \cup (1, \infty) \\ H(X) & \alpha = 1 \\ \min_{a \in \mathcal{A}} \ -\log P_X(a) & \alpha = \infty \end{cases} \tag{1}$$

where the Shannon entropy $H(X) = \mathbb{E}[-\log P_X(x)]$.

Note that in the above definition, the behaviors at $\alpha \in \{0, 1, \infty\}$ are justified from continuity. In the sequel, for simplicity we shall omit these special cases unless they deserve particular attention. Also, throughout this paper all random variables are discrete unless otherwise noted.

Rényi also generalized the classical KL (Kullback-Leibler) divergence, which results in a similar mathematical form:

Definition 2. *For two distributions P and Q, the Rényi divergence of order $\alpha \geq 0$ is given by*

$$D_\alpha(P||Q) = \frac{1}{1-\alpha} \log \left(\sum_{a \in \mathcal{A}} P(a)^\alpha Q(a)^{1-\alpha} \right). \tag{2}$$

This definition may be used in the natural way to define conditional Rényi divergences as well.

A number of proposals were made over the years regarding an analogous generalization of mutual information [2], [3], [4]. However, here the choice is less clear, and [5] provides a summary of the various proposals. We follow [5] in our choice to settle on a default definition of α-mutual information (Arimoto's proposal):

Definition 3. *Let $P_X \to P_{Y|X} \to P_Y$, where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. The α-mutual information for $\alpha > 0$ is:*

$$I_\alpha(X;Y) = \min_{Q_Y} D_\alpha \left(P_{Y|X} || Q_Y | P_X \right)$$

A more explicit (non-variational) form is:

$$I_\alpha(X;Y) = \frac{\alpha}{\alpha - 1} \log \sum_{y \in \mathcal{Y}} \left(\sum_{x \in \mathcal{X}} P_X(x) P_{Y|X=x}^\alpha(y) \right)^{\frac{1}{\alpha}} \tag{3}$$

Using α-mutual information, we may naturally define (analogously to the classical case) α-capacity for a fixed channel/kernel $P_{Y|X}$ as:

Definition 4.

$$C_\alpha \left(P_{Y|X} \right) = \sup_{P_X} I_\alpha(X;Y)$$

where $P_X \to P_{Y|X} \to P_Y$ and $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$.

From the explicit formula for α-mutual information, it is clear that for a identity kernel $P_{Y|X}$,

$$I_\alpha(X;X) = H_{\frac{1}{\alpha}}(X) \tag{4}$$

The importance of this equation will be clear later, as it links the failure of single-letterization for I_α and H_α together.

We also note that for the purposes of determining capacity, since $\log()$ is an increasing function, it suffices to maximize (minimize) the following expression for $\alpha > 1$ ($\alpha < 1$) respectively:

$$J_\alpha(X;Y) = \sum_{y \in \mathcal{Y}} \left(\sum_{x \in \mathcal{X}} P_X(x) P_{Y|X=x}^\alpha(y) \right)^{\frac{1}{\alpha}} \tag{5}$$

II. Computation of Simple Single-letter Capacities

In this section, we compute the α-capacity for simple single-letter channels. In [5], it is mentioned that the α-mutual information is a concave function of P_X for $\alpha \geq 1$, and that a monotone transformation from $I_\alpha(X;Y)$ to $\frac{1}{\alpha-1} J_\alpha(X;Y)$ makes the function concave in P_X for $\alpha > 0$. This makes the α-capacity determination a convex optimization problem for

$\alpha > 0$. As such, for the simple single-letter channels (BSC and BEC) considered here, the problem is easily solved for all α. In order to present the results, we first recall the ideas of majorization/smoothing and Karamata's inequality:

Definition 5. *Let $\vec{p}, \vec{q}$ denote two vectors in $\mathbb{R}^d$. We say that $\vec{p}$ majorizes $\vec{q}$ (symbolically $\vec{p} \succ \vec{q}$) iff:*

$$\begin{aligned} p_{(1)} &\geq q_{(1)} \\ \sum_{i=1}^{k} p_{(i)} &\geq \sum_{i=1}^{k} q_{(i)} \quad (\forall\, 2 \leq k \leq d-1) \\ \sum_{i=1}^{d} p_{(i)} &= \sum_{i=1}^{d} q_{(i)} \end{aligned}$$

where $p_{(i)}$ denotes the i-th order statistic of $\vec{p}$ in descending order.

A nice, useful, equivalent characterization of majorization is the following result (smoothing) [6]:

Theorem 1. *Let $\vec{p}, \vec{q}$ denote two vectors in $\mathbb{R}^d$. Then $\vec{p} \succ \vec{q}$ iff there exists a consecutive sequence of "Robin-Hood" transformations from $\vec{p}$ to $\vec{q}$. A Robin-Hood transformation of $\vec{p}$ consists of picking two distinct indices i, j, and replacing p_i by $p_i - \epsilon$, and p_j by $p_j + \epsilon$, where $0 \leq \epsilon \leq \frac{|p_i - p_j|}{2}$. Here $p_i > p_j$ without loss of generality.*

The utility of defining majorization lies in Karamata's inequality:

Theorem 2. *Let f denote a real-valued convex function defined on some interval I of $\mathbb{R}$. Let $\vec{p} \succ \vec{q}$. Then, $\sum_{i=1}^{d} f(p_i) \geq \sum_{i=1}^{d} f(q_i)$. An analogous inequality holds (with the sign flipped) for concave functions.*

We collect the results for the BSC(δ) and BEC(δ) in the following theorem. Here, δ denotes the probability of bit-flip and probability of erasure respectively:

Theorem 3. $\forall \alpha \geq 0$,

For the BSC(δ),

$$C_\alpha(X;Y) = \delta_\alpha\left(\delta || \frac{1}{2}\right), \tag{6}$$

where $\delta_\alpha(p||q) = \frac{1}{\alpha - 1}\log\left(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\right)$ is the binary Rényi divergence. Equality is achieved only when P_X is uniform on $\{0,1\}$.

For the BEC(δ),

$$C_\alpha(X;Y) = \frac{\alpha}{\alpha - 1}\log\left(\delta + (1-\delta)2^{\frac{\alpha-1}{\alpha}}\right). \tag{7}$$

Equality is achieved only when P_X is uniform on $\{0,1\}$.

Proof. Let $P_X = (1-p, p)$ on $(0,1)$ respectively. For the BSC,

$$\begin{aligned} J_\alpha(X;Y) = &(p\delta^\alpha + (1-p)(1-\delta)^\alpha)^{\frac{1}{\alpha}} \\ &+ ((1-p)\delta^\alpha + p(1-\delta)^\alpha)^{\frac{1}{\alpha}}. \end{aligned}$$

Observe that the sum of the two terms that are being raised to $\frac{1}{\alpha}$ is $\delta^\alpha + (1-\delta)^{1-\alpha}$, which is independent of p. Thus, we may smooth $(1-p, p)$ by $(0.5, 0.5)$, since for $\alpha < 1$, $\frac{1}{\alpha} > 1$, for $\alpha > 1$, $\frac{1}{\alpha} < 1$, and x^a is concave (convex) for $a < 1$ ($a > 1$) respectively. This proves 6.

For the BEC,

$$J_\alpha(X;Y) = ((1-p)(1-\delta)^\alpha)^{\frac{1}{\alpha}} + (p(1-\delta)^\alpha)^{\frac{1}{\alpha}} + \delta.$$

Like the above, the sum of the first two terms that are being rasied to $\frac{1}{\alpha}$ is independent of p. Thus, we may smooth $(1-p, p)$ by $(0.5, 0.5)$ to get the desired result. This proves 7. □

We also remark here that the single-letter Z channel capacity does not have a nice closed form, and that the capacity achieving input distribution for this channel varies with α, unlike the BSC and BEC cases.

A more non-trivial application of smoothing is in solving the cost-constrained BSC capacity for a single-channel use:

Theorem 4. $\forall \alpha \geq 0$,

$$\begin{aligned} C_\alpha(X;Y) = \frac{\alpha}{\alpha - 1}\log\Big(&(P\delta^\alpha + (1-P)(1-\delta)^\alpha)^{\frac{1}{\alpha}} \\ &+ ((1-P)\delta^\alpha + P(1-\delta)^\alpha)^{\frac{1}{\alpha}}\Big). \end{aligned}$$

for the BSC(δ) and cost constraint $\mathbb{E}[X] \leq P$, where (without loss of generality) $P \leq 0.5$.

Proof. We repeat the proof as in the unconstrained case. Here, we may smooth till $(1-P, P)$ and no more. More formally, $(1-P, P)$ is the only extremal point with respect to the partial order of majorization in this single-letter, cost-constrained case. □

III. Generalization to Multiple Uses of Channel

In [5], it is mentioned that one can single-letterize capacity in the case of discrete memoryless channels. This gives the result that the (unconstrained) capacity of n uses of the channel is nC, where C denotes the single-letter (i.e single use) capacity. However, this result does not extend to the case where one places constraints on the input distributions, e.g Hamming weight constraints on n uses of the BSC. An interesting idea [7] allows one to single-letterize for memoryless channels satisfying the following two constraints:

1) $I_\alpha\left(X^2; Y^2\right) \leq I_\alpha(X_1;Y_1) + I_\alpha(X_2;Y_2)$ (X^2 denotes the joint (X_1, X_2)), i.e single-letterization for α-mutual information
2) concavity in P_X of $I_\alpha(X;Y)$.

However, these conditions can't be simultaneously met (for $\alpha \neq 1$). The second condition is met for all $\alpha \geq 1$ as noted earlier. The first condition is not true for general channels, though for specific channels it is possible that it could hold. The reason for this is as follows. Consider the identity channel. Then, by 4, the first condition becomes

$$H_{\frac{1}{\alpha}}\left(X^2\right) \leq H_{\frac{1}{\alpha}}(X_1) + H_{\frac{1}{\alpha}}(X_2).$$

Unfortunately, this subadditivity for Rényi entropy is not true. In fact, simple modifications designed to obtain some sort of

single-letterization for Rényi entropy are bound to fail due to a result in [8], which states that one can fix $H_\alpha(X)$ and $H_\alpha(Y)$ and make $H_\alpha(X,Y)$ go to ∞ by increasing the alphabet size. This result is true for all orders of $\alpha \neq 1$.

We also illustrate the lack of nice single-letterization with a more non-trivial example of the BSC with cost constraints (multi-letter case). First, we introduce some notation. Let $H(x,r)$ denote the Hamming ball of radius r centered at the bit-vector x, i.e it is the set of all bit-vectors y of the same length as x such that $|x-y| \leq r$, where $|x|$ denotes the Hamming weight of x, and $|x-y|$ denotes the Hamming distance between bit-vectors x and y. We also introduce the non-standard notation of "Hamming shell" $S(x,r)$ to denote the set of bit-vectors y such that $|x-y| = r$. Then, we have the following theorem illustrating the lack of a nice single-letterization for the multi-letter, cost-constrained BSC:

Theorem 5. *Consider a BSC(δ), and cost constraint $\mathbb{E}[|X^n|] \leq nP$. Then, $\forall \alpha \geq 0$, the capacity achieving input distribution (caid) ${P_{X^n}}^*$ must be "spherically symmetric", in the sense that it should have a uniform distribution conditioned on $\mathbb{1}\{S(0,r)\}$ for each $0 \leq r \leq n$. Moreover, $\forall \alpha \neq 1$, the capacity achieving input distribution is not a product distribution.*

Proof. Observe that the BSC(δ) has permutation symmetry, in the sense that $I_\alpha(X^n;Y^n) = I_\alpha(PX^n;PY^n)$, where P denotes an arbitrary permutation of $(1,2,\ldots,n)$. Now, $\mathbb{E}[|X^n|] \leq nP \Leftrightarrow \mathbb{E}[|PX^n|] \leq nP$. Thus, by the concavity of $\frac{1}{\alpha-1}J_\alpha(X;Y)$, we may then average over all permutations of X^n to conclude that $I_\alpha(X^{n*};Y^{n*}) \geq I_\alpha(X^n;Y^n)$ for the "spherically symmetric" ${P_{X^n}}^*$ obtained by averaging over permutations of P_{X^n}. Now suppose that the capacity achieving distribution is a product distribution. Observe that the only "spherically symmetric" input distribution which is also a product distribution is an i.i.d Bernoulli ensemble. Furthermore, since I_α is additive over product distributions, we may invoke Theorem 4 to conclude that X^n are distributed i.i.d $Ber(P)$. However, take $n=2$. It is easily checked that $\forall \alpha \neq 1$, there exist input distributions which result in greater mutual information than the above. These counterexamples for $n=2$ may be then combined with product distributions over the remaining $n-2$ letters to generate counterexamples $\forall n \geq 2$. □

IV. Applications of the Rényi Information Measures

In this section, we explore a possible application of Rényi information measures to tightening the best-known converse bound for the binary adder MAC. The binary adder MAC is defined as follows. $\mathcal{A} = \mathcal{B} = \{0,1\}$ are the alphabets of the two users. $Y = A+B$ is the channel output, and is in $\{0,1,2\}$. By using results in [9] and [5], one can tighten best known converse [10] second order term from $O(\sqrt{n \log n})$ to $O(\sqrt{n})$ if the following conjecture is true:

$$H_\alpha(Y^n) \leq nH_\alpha(Y^*) \quad \forall \alpha \in (0,1), n \geq 1 \tag{8}$$

This follows from general converse bounds established in [5] and [9], by taking the order of the Rényi entropy $\alpha = 1 - O\left(\frac{1}{\sqrt{n}}\right)$. Note that the $\alpha \leq 1$ assumption is necessary for the conjecture. For $\alpha > 1$ and $n = 2$, consider the family $\vec{p} = \{x, 0.5-x, 0.5-x, x\}$ and $\vec{q} = \{0.5-x, x, x, 0.5-x\}$ for some $x \in [0, 0.5]$. Then for $\alpha > 1$, it may be easily checked that $x = 0.25$ results in a local minimum of the joint Rényi entropy, and so 8 is false for $\alpha > 1$ and $n = 2$. Note that by looking at product distributions for $n > 2$ and using the above counterexample for $n = 2$, one generates counterexamples for all $n > 2$ as well.

Unfortunately, 8 remains open. The best result we have managed so far is that the conjecture is true for $n = 1$ and for $n = 2, \alpha \leq 0.5$. Computer simulations indicate that the conjecture is likely true till $n = 5$.

For $n = 1$, 8 is easy to verify via the following argument. Let $P_A = \{p, 1-p\}$ and $P_B = \{1-q, q\}$. A simple manipulation of the resulting derivative expressions shows that $p = q$ for a local optimum. On $p = q$, the Rényi entropy becomes an expression in a single variable, whose optimum is easily checked to be at $p = q = 0.5$. Of course, the necessary second derivative tests and boundary checks need to be completed as well, which we omit.

For $n = 2, \alpha \leq 0.5$, the proof is technical and tedious and is hence in an appendix. Note that this result is very pessimistic from another perspective as well. The whole purpose of 8 is to establish a tighter converse bound. As one can see in the argument above, one is really interested in $\alpha = 1 - O(\frac{1}{\sqrt{n}})$, while 0.5 is much less than 1. Not much can be said about this, since $n = 2$ is too small to conclude anything about asymptotic behavior.

V. Approaches Towards Single-letterization

As can be seen from above discussion, one of the key difficulties with Rényi-like measures of information is the lack of a nice single-letterization. This is to a large extent responsible for the incomplete understanding of product channels (i.e n uses) in terms of these measures, as well as the failed attempt at tightening a converse bound for the binary adder MAC. Shannon entropy and the related classical measures of information offer a much cleaner picture in that respect, since not only do the variational quantities such as capacity single-letterize, but also the measures themselves in a very simple fashion. In light of this, in spite of the negative result on single-letterization of Rényi entropy [8], it is somewhat surprising that one can still furnish some interesting bounds on Rényi entropy:

Theorem 6. *Let $|\mathcal{X}| = M, |\mathcal{Y}| = N$. Then:*

$$H_\alpha(P_{XY}) \leq H_\alpha(P_X) + \max_{1 \leq i \leq M} H_\alpha\left(P_{Y|X=i}\right) \quad (\forall \alpha \in (0,1)). \tag{9}$$

$$H_\alpha(P_{XY}) \leq \log\left(\sum_{i=1}^{M} \exp\left(H_\alpha\left(P_{Y|X=i}\right)\right)\right) \quad (\forall \alpha \in (0,1)). \tag{10}$$

$$H_\alpha(P_{XY}) \le t(H_\alpha(Q_X) + H_\alpha(R_Y)) + (1-t)\left[\frac{1}{1-\alpha}\log\left(\sum_{i,j} P_{ij}^\alpha \left(\frac{P_{ij}}{Q_i R_j}\right)^{\frac{t\alpha}{1-t}}\right)\right] \quad (11)$$

$$(\forall \alpha \in (0,1), t \in (0,1), Q_X, R_Y).$$

In the special case as $t \to 1$ in 11, we get:

$$H_\alpha(P_{XY}) \le H_\alpha(Q_X) + H_\alpha(R_Y) + \frac{\alpha}{1-\alpha} D_\infty(P_{XY}||Q_X R_Y). \quad (12)$$

We also have:

$$H_\alpha(P_{XY}) \le H_\alpha(P_X) + H_{\frac{1}{\alpha}}(P_{Y_\alpha}) \, (\forall \alpha \ge 0), \quad (13)$$

where

$$\mathbb{P}[Y_\alpha = y] = \sum_x P_{XY}(x,y)^\alpha \exp[(\alpha-1)H_\alpha(X,Y)].$$

More generally, for any $0 < u < v < \infty$, we have:

$$\frac{1-u}{u} H_u(X,Y) + \frac{v-1}{v} H_v(X,Y) \le \left(\frac{1}{u} - \frac{1}{v}\right)(H_{\frac{u}{v}}(X_v) + H_{\frac{v}{u}}(Y_u)). \quad (14)$$

In particular, for $\alpha < 1$, we have:

$$H_\alpha(X,Y) \le (1+\alpha) H_{\alpha^2}(X,Y) - H_\alpha(X,Y) \le H_\alpha(X_\alpha) + H_{\frac{1}{\alpha}}(Y_{\alpha^2}). \quad (15)$$

Proof.

$$\sum_{i,j} P_{i,j}^\alpha = \sum_i P_i^\alpha \sum_j P_{j|i}^\alpha \le \left(\max_{1 \le i \le M} \sum_j P_{j|i}^\alpha\right) \sum_i P_i^\alpha.$$

Taking logarithms and dividing everything by $1-\alpha$, we get 9. Proof of 10 follows from Hölder's inequality with conjugate exponents $\frac{1}{\alpha}$, $\frac{1}{1-\alpha}$ applied to the two sequences consisting of the marginal P_X and conditional $P_{Y|X}$ respectively. Proof of 11 follows from Hölder's inequality applied to the product of two sequences consisting of tilted joint distribution P_{XY}, tilted product distribution $Q_X R_Y$, and with a degree of freedom that enables one to pick $t \in (0,1)$. The special case 12 follows by considering the limiting behavior of the RHS as $t \to 1$. Moving to the other class of bounds, the general statement 14 with arbitrary u, v follows from Minkowski's inequality for mixed $l_p - l_q$ norms applied to the marginals of the tilted joint distribution. The remaining statements 13 and 15 are various specializations of the general statement. For instance, setting $v = \alpha, u = \alpha^2$ yields 15. □

IDEA: For a noiseless MAC, Tsallis entropy single-letterization yields the following converse bound:

$$M^{\lambda-1} \ge \frac{\left[1 + n\left(\sum_i {p_i}^\lambda - 1\right)\right]^\lambda - \epsilon^\lambda}{(1-\epsilon)^\lambda} \quad \forall \lambda \in (0,1) \quad (16)$$

where p_i denotes the single-letter channel output distribution. One nice feature of this is seen when one Taylor-expands $\sum_i {p_i}^x$ about $x = 1$: zeroth order term cancels the 1, first order term yields entropy, second order term something involving varentropy, etc. As long as powers don't mess things up, I think this should yield a nice, strong converse bound. Note that this bound uses the tighter inequalities for Rényi divergence in your paper with Verdu.

VI. Conclusion

Although [5] presents many analogs of classical results that hold for Rényi-like measures of information, some critical aspects such as single-letterization lack the same structure. At some level, this paper thus tempers the optimism surrounding these Rényi-like measures of information, since it may be argued that single-letterization of some sort is the "heart" of information theory as applied in the traditional context of communication. However, some of the successful applications of Rényi entropy in particular have been in non-traditional information theory applications, such as DNA sequencing [11], target tracking [12], and guessing [13]. The successes above demonstrate clearly that Rényi-like measures of information are useful. Perhaps a study of these measures in application domains apart from communication could yield more insight into what sort of "information" do these measures capture.

At a more technical level, we feel that the bounds established in this paper provide a useful first step in understanding how one can single-letterize Rényi-like measures of information. In future work, we plan to examine these bounds more carefully.

Appendix A
Proof for $n = 2, \alpha \le 0.5$

Below we present a proof for $n = 2, \alpha \le 0.5$. It was initially hoped that the same ideas could be used for all $\alpha \le 1$, but we ran into difficulties. For the rest of this proof, we let $P_{A^2} = \{p_0, p_1, p_2, p_3\}$ and likewise $P_{B^2} = \{q_0, q_1, q_2, q_3\}$. The labelling is based on base two representation ($0 \to 00, 1 \to 01, 2 \to 10, 3 \to 11$). Then, our conjecture is:

$$(p_0q_0)^\alpha + (p_1q_1)^\alpha + (p_2q_2)^\alpha + (p_3q_3)^\alpha + (p_0q_1 + p_1q_0)^\alpha + (p_0q_2 + p_2q_0)^\alpha + (p_1q_3 + p_3q_1)^\alpha + (p_2q_3 + p_3q_2)^\alpha + (p_0q_3 + p_1q_2 + p_2q_1 + p_3q_0)^\alpha \le \frac{1}{4^\alpha} + \frac{4}{8^\alpha} + \frac{4}{16^\alpha} \quad (17)$$

The idea is essentially "trading mass". We "equalize" (p_0, p_3) and (q_0, q_3) simultaneously. By "equalizing", we mean that we replace a pair (a, b) by $\left(\frac{a+b}{2}, \frac{a+b}{2}\right)$. We claim that this increases the left hand side of the inequality. Observe that $(p_0q_1 + p_1q_0) + (p_1q_3 + p_3q_1)$ is invariant under this operation. Likewise, $(p_0q_2 + p_2q_0) + (p_2q_3 + p_3q_2)$ is also invariant under this operation. Moreover, it is clear that the terms $p_0q_1 + p_1q_0$ and $p_1q_3 + p_3q_1$ have been "equalized". Similarly, $p_0q_2 + p_2q_0$ and $p_2q_3 + p_3q_2$ have been "equalized" as well. Thus, we have increased the sum of the corresponding four terms of 17. The terms p_1q_1 and p_2q_2 are unaffected by this operation. Thus, for the above claim, it suffices to check that the sum of the remaining three terms in 17 has not decreased. For this, observe that if (p_0, p_3) and (q_0, q_3) are "opposite

sorted" (terminology that is used regarding rearrangements of sequences), we have a successful majorization:

$\left(p_1q_2 + p_2q_1 + 2\frac{p_0+p_3}{2}\frac{q_0+q_3}{2} \geq \frac{p_0+p_3}{2}\frac{q_0+q_3}{2} \geq \frac{p_0+p_3}{2}\frac{q_0+q_3}{2}\right)$ is majorized by

$(p_1q_2 + p_2q_1 + p_0q_3 + p_3q_0 \geq p_0q_0?p_3q_3)$,

where '?' denotes an intermediate inequality that is not needed to have a definite direction for the majorization to hold.

In the case of similar sorting, by the rearrangement inequality, it follows that the term $p_0q_3 + p_1q_2 + p_2q_1 + p_3q_0$ has not decreased. Thus, it suffices to check that $(p_0q_0)^\alpha + (p_3q_3)^\alpha$ has increased. For this, we depend critically on $\alpha \leq 0.5$.

By normalizing, we may assume that $p_0 + p_3 = q_0 + q_3 = 2$ without loss of generality, since the desired inequality is homogenous. Thus, it suffices to check that for any variables w, x, y, z such that $w+z = x+y = 2$, and $\alpha \leq 0.5$, we have:

$$(wx)^\alpha + (yz)^\alpha \leq \sqrt{(w^{2\alpha} + z^{2\alpha})(x^{2\alpha} + y^{2\alpha})}$$
$$\leq \sqrt{(w+z)(x+y)} \leq 2,$$

using $\alpha \leq 0.5$ as desired.

Thus, we have the claim that the "0-3" equalization can't decrease the left hand side when $\alpha \leq 0.5$. By symmetry, we may follow the "0-3" equalization by a "1-2" equalization to further not decrease the left hand side. Thus, for 8 (under $\alpha \leq 0.5$), it suffices to prove 8 for for all choices of $0 \leq p, q \leq 0.5$, where A^2 takes the p.m.f $(p, 0.5-p, 0.5-p, p)$, and B^2 takes the p.m.f $(0.5-q, q, q, 0.5-q)$.

Let $\vec{y}$ denote the channel output distribution for the above A^2 and B^2. Then,

$\vec{y} = (2p(0.5-q)+2q(0.5-p), pq+(0.5-p)(0.5-q), pq+(0.5-p)(0.5-q), pq+(0.5-p)(0.5-q), pq+(0.5-p)(0.5-q), p(0.5-q), p(0.5-q), q(0.5-p), q(0.5-p))$.

Suppose $2p(0.5-q)+2q(0.5-p) \leq 0.25$. Then, the equation $4r(0.5-r) = 2p(0.5-q)+2q(0.5-p)$ has a solution in $0 \leq r \leq 0.5$. In this case, consider $A^{*2} = (r, 0.5-r, 0.5-r, r)$ and $B^{*2} = (0.5-r, r, r, 0.5-r)$ respectively. Then, the channel output

$\vec{y^*} = (4r(0.5-r), r^2+(0.5-r)^2, r^2+(0.5-r)^2, r^2+(0.5-r)^2, r^2+(0.5-r)^2, r^2+(0.5-r)^2, r(0.5-r), r(0.5-r), r(0.5-r), r(0.5-r))$.

This effectively "matches" the desired $\vec{y}$. More precisely, it is clear that the first term is the same in both vectors; $r(0.5-r)$ is the average of the sum of the last 4 terms of $\vec{y}$; and hence in fact the first five terms are the same in both vectors.

Putting these claims together, we have that $\vec{y^*}$ is majorized by $\vec{y}$ in this case.

Now suppose $2p(0.5-q) + 2q(0.5-p) > 0.25$. Then, the equation $2r^2 + 2(0.5-r)^2 = 2p(0.5-q) + 2q(0.5-p)$ has a solution in $0 \leq r \leq 0.5$. In this case, consider $A^{**2} = (r, 0.5-r, 0.5-r, r)$ and $B^{**2} = (r, 0.5-r, 0.5-r, r)$ respectively. Then, the channel output

$\vec{y^{**}} = (2r^2+2(0.5-r)^2, 2r(0.5-r), 2r(0.5-r), 2r(0.5-r), 2r(0.5-r), r^2, r^2, (0.5-r)^2, (0.5-r)^2)$.

We claim that this effectively "matches" the desired $\vec{y}$. The first term is the same in both vectors by the choice of r. Also, the sum of the last four terms is the same in both vectors. Thus, the first five terms are the same in both vectors.

We now claim that $|r^2 - (0.5-r)^2| \leq |p(0.5-q) - q(0.5-p)|$.

This claim shows that $(r^2, (0.5-r)^2)$ can be obtained by "trading mass" between $(p(0.5-q), q(0.5-p))$.

We now prove the claim.

For ease of dealing with absolute values, we assume without loss of generality that $r \geq 0.25$ and $p \geq q$.

Using the fact that r and $0.5-r$ are the two roots of a quadratic $x^2 + (0.5-x)^2 = p(0.5-q) + q(0.5-p)$, and simplifying, we see that this is equivalent to:

$\frac{\sqrt{4p+4q-16pq-1}}{2} \leq p-q$.

Squaring both sides, it suffices to show that:

$4p^2 + 4q^2 - 4p - 4q + 8pq + 1 \geq 0$, or equivalently,

$(2p + 2q - 1)^2 \geq 0$, which is clearly true.

Collecting all these claims, we see that $\vec{y^{**}}$ is majorized by $\vec{y}$ in this case.

Altogether, we have now reduced the task to proving two single variable inequalities, one corresponding to $\vec{y^*}$, and the other corresponding to $\vec{y^{**}}$ parametrized by the variable r.

These inequalities are easy to establish by derivative tests.

Thus, we have resolved the conjecture when $n = 2, \alpha \leq 0.5$.

ACKNOWLEDGMENT

Ganesh Ajjanagadde would like to thank Anuran Makur for stimulating discussions.

REFERENCES

[1] A. Rényi, "On measures of entropy and information," in *Fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1961, pp. 547–561.

[2] I. Csiszár, "Generalized cutoff rates and renyi's information measures," *Information Theory, IEEE Transactions on*, vol. 41, no. 1, pp. 26–34, 1995.

[3] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.

[4] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," *Topics in Information Theory, Proc. Coll. Math. Soc. János Bolyai*, pp. 41–52, 1975.

[5] S. Verdú, "α-mutual information," in *Information Theory and Applications Workshop*, 2015.

[6] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of majorization and its applications: Theory of majorization and its applications*. Springer Science & Business Media, 2010.

[7] Y. Polyanskiy and Y. Wu, "Lecture Notes on Information Theory," http://people.lids.mit.edu/yp/homepage/data/itlectures_v3.pdf, 2015, [Online; accessed 03-April-2015].

[8] M. Kovacevic, I. Stanojevic, and V. Senk, "On the entropy of couplings," *arXiv preprint arXiv:1303.3235*, 2013.

[9] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and rényi divergence," in *48th Annual Allerton Conference on Communication, Control, and Computing*, October 2010, pp. 1327–1333.

[10] R. Ahlswede, "An elementary proof of the strong converse theorem for the multiple-access channel," *J. Combinatorics, Information and System Sciences*, vol. 7, no. 3, 1982.

[11] A. S. Motahari, G. Bresler, and D. N. Tse, "Information theory of dna shotgun sequencing," *Information Theory, IEEE Transactions on*, vol. 59, no. 10, pp. 6273–6289, 2013.

[12] C. Kreucher, K. Kastella, and A. O. Hero, "Multi-target sensor management using alpha-divergence measures," in *Information Processing in Sensor Networks*. Springer, 2003, pp. 209–222.

[13] T. van Erven and P. Harremoës, "Rényi divergence and majorization," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1335–1339.