EQUIVALENT FORMULATIONS OF HYPERCONTRACTIVITY USING INFORMATION MEASURES

CHANDRA NAIR[1 †]

[1]*Dept. of Information Engineering,*
The Chinese University of Hong Kong

Abstract

We derive alternate characterizations for the hypercontractive region of a pair of random variables using information measures.

1. Introduction

A pair of random variables (X, Y) defined on some probability space $(\Omega, \mathcal{F}, \mu)$, is said to be (p, q)-hypercontractive for $1 \leq q \leq p < \infty$ if the inequality

$$\| \mathrm{E}[g(Y)|X]\|_p \leq \|g(Y)\|_q$$

holds for every bounded measurable function $g(Y)$.

The following is another well-known equivalent definition (equivalence being a direct application of Hölder's inequality). A pair of random variables (X, Y) defined on some probability space $(\Omega, \mathcal{F}, \mu)$, is said to be (p, q)-hypercontractive for $1 \leq q \leq p < \infty$ if the inequality

$$\mathrm{E}[f(X)g(Y)] \leq \|f(X)\|_{p'}\|g(Y)\|_q$$

holds for every pair of bounded measurable functions $f(X), g(Y)$. Here $p' = \frac{p}{p-1}$ denotes the Hölder conjugate of p.

For any $p \geq 1$ one can define

$$q_p(X; Y) = \inf\{q : (X, Y) \text{ is } (p, q)\text{-hypercontractive}\}.$$

Define the ratio $r_p(X; Y) = \frac{q_p(X;Y)}{p}$.

Hypercontractive inequalities have found a variety of applications in quantum physics [3], theoretical computer science [4], analysis [6], and in information theory [1, 2]. In this talk we present the following alternate characterizations of $r_p(X; Y)$ using information measures.

A very useful property of the hypercontractive parameter $r_p(X; Y)$ is the so-called *tensorization property* which states: if $\{(X_i, Y_i)\}_{i=1}^n$ are independent random variables, then

$$r_p(X^n; Y^n) = \max_{i=1}^{n} r_p(X_i; Y_i).$$

For the purpose of this manuscript, let us assume that the random variables X, Y take values in a finite alphabet space $\mathcal{X} \times \mathcal{Y}$. Thus we can talk about a probability mass function $\mu_{XY}(x, y)$ and the induced marginal distributions $\mu_X(x), \mu_Y(y)$. We assume, without loss of generality, that $\mu_X(x) > 0, \forall x \in \mathcal{X}$ and $\mu_Y(y) > 0, \forall y \in \mathcal{Y}$. We will omit the subscripts on the measures when there is no ambuguity.

† Corresponding author: Chandra Nair, email: chandra.nair@gmail.com

Consider two measures ν and μ on a space, say $\mathcal{X}$. If the measure ν is absolutely continuous with respect to the measure μ, then we denote it as $\nu \ll \mu$. Let $D(\nu(x)\|\mu(x)) = \sum_{x\in\mathcal{X}} \nu(x)\log_2 \frac{\nu(x)}{\mu(x)}$ be the relative entropy between the two measures $\nu(x)$ and $\mu(x)$, when $\nu \ll \mu$.

For $p \geq 1$, define

$$k_p(X;Y) := \sup_{\substack{\nu(X,Y)\ll\mu(X,Y)\\ \nu(X,Y)\neq\mu(X,Y)}} \frac{D(\nu(y)\|\mu(y))}{pD(\nu(x,y)\|D(\mu(x,y)) - (p-1)D(\nu(x)\|D(\mu(x))}.$$

Let $\mathcal{P}_\mu = \{\nu_{UXY}(u,x,y) : u \in \mathcal{U}, |\mathcal{U}| < \infty, \sum_{u\in\mathcal{U}} \nu(u,x,y) = \mu(x,y)\ \forall x \in \mathcal{X}, y \in \mathcal{Y}\}$ be the collection of measures induced by random variables (U,X,Y) such that the marginal law of (X,Y) is consistent with $\mu(X,Y)$. Here U is a random variable taking finitely many values and is often referred to as *an auxiliary random variable* in multiuser information theory. Later we will see that we can restrict the size of $|\mathcal{U}|$ to $|\mathcal{X}||\mathcal{Y}|$.

For $p \geq 1$ define

$$u_p(X;Y) = \sup_{\substack{\nu_{UXY}\in\mathcal{P}_\mu\\ I(U;XY)>0}} \frac{I(U;Y)}{pI(U;XY) - (p-1)I(U;X)}.$$

Theorem 1. *The following equivalence holds:*

$$r_p(X;Y) = k_p(X;Y) = u_p(X;Y).$$

Further the common value is also equal to

$$\inf\{\lambda : H(Y) - \lambda pH(X,Y) + \lambda(p-1)H(X) = \mathsf{K}[H(Y) - \lambda pH(X,Y) + \lambda(p-1)H(X)]_{\mu_{XY}}\},$$

where $\mathsf{K}[f]_x$ represents the lower convex envelope of the function f evaluated at x.

Remark: The above result generalizes the equivalence results in both [1] and [2] which deal with the limiting case $p \to \infty$.

2. Proof of the main result

The equality in Theorem 1 will be established by showing a sequence of inequalities. In particular, we will show that

(i) $r_p(X;Y) \leq k_p(X;Y)$

(ii) $k_p(X;Y) \leq u_p(X;Y)$

(iii) $u_p(X;Y) \leq r_p(X;Y)$.

Proof of $r_p(X;Y) \leq k_p(X;Y)$: Given $p > 1$, and $\epsilon > 0$ arbitrary, let non-negative functions $f(X)$ and $g(Y)$ satisfy

$$E(f(X)g(Y)) > \|f(X)\|_{p'}\|g(Y)\|_{(r_p-\epsilon)p}. \tag{1}$$

W.l.o.g. assume that $\|f(X)\|_{p'} = \|g(Y)\|_{(r_p-\epsilon)p} = 1$.

Define $f(x)^{p'} = h(x)$ and $g(y)^{(r_p-\epsilon)p} = j(y)$. We have

$$\sum_x \mu(x)h(x) = \sum_y \mu(y)j(y) = 1.$$

Since $\mathrm{E}(f(X)g(Y)) > 1$, let $C < 1$ be such that

$$\sum_{x,y} C\mu(x,y)h(x)^{\frac{1}{p'}} j(y)^{\frac{1}{(r_p-\epsilon)p}} = 1.$$

Define

$$\nu(x,y) := C\mu(x,y)h(x)^{\frac{1}{p'}} j(y)^{\frac{1}{(r_p-\epsilon)p}}.$$

One easily verifies that $\nu_{XY} \ll \mu_{XY}$ and $\nu_{XY} \neq \mu_{XY}$.

Now observe that

$$\begin{aligned}
&pD(\nu(x,y)\|\mu(x,y)) - (p-1)D(\nu(x)\|\mu(x)) \\
&\quad = p\log C + \frac{p}{p'}\sum_x \nu(x)\log h(x) + \frac{1}{(r_p-\epsilon)}\sum_y \nu(y)\log j(y) - (p-1)\sum_x \nu(x)\log\frac{\nu(x)}{\mu(x)} \\
&\quad = p\log C + (p-1)\sum_x \nu(x)\log\frac{\mu(x)h(x)}{\nu(x)} + \frac{1}{(r_p-\epsilon)}\sum_y \nu(y)\log\frac{\mu(y)j(y)}{\nu(y)} \\
&\qquad + \frac{1}{(r_p-\epsilon)}\sum_y \nu(y)\log\frac{\nu(y)}{\mu(y)} \\
&\quad \le \frac{1}{(r_p-\epsilon)}\sum_y \nu(y)\log\frac{\nu(y)}{\mu(y)} = \frac{1}{(r_p-\epsilon)}D(\nu(y)\|\mu(y)).
\end{aligned}$$

Here the last inequality follows since $C < 1, \sum_x \mu(x)h(x) = 1, \sum_y \mu(y)j(y) = 1$. Since $\epsilon > 0$ is arbitrary we are done. □

Proof of $k_p(X;Y) \le u_p(X;Y)$: The argument below is identical to the argument in [2] which in turn is motivated by similar arguments in [5].

Let $\delta \in (0, k_p(X;Y))$ be arbitrary. Let $\nu_{XY} \ll \mu_{XY}, \nu_{XY} \neq \mu_{XY}$ be any distribution satisfying

$$\frac{D(\nu(y)\|\mu(y))}{pD(\nu(x,y)\|D(\mu(x,y)) - (p-1)D(\nu(x)\|D(\mu(x))} > k_p(X;Y) - \delta.$$

Let $\mathcal{U}_\epsilon := \{1,2\}$. Fix a sufficiently small[1] $\epsilon > 0$ and define a triple (U_ϵ, X, Y) according to:

- $\mathrm{P}(U_\epsilon = 1) = \epsilon$; Conditional distribution of $(X,Y)|(U_\epsilon = 1) = \nu_{XY}$,
- $\mathrm{P}(U_\epsilon = 2) = 1-\epsilon$; Conditional distribution of $(X,Y)|(U_\epsilon = 2) = \mu_{XY} + \frac{\epsilon}{1-\epsilon}(\mu_{XY} - \nu_{XY}) = \frac{1}{1-\epsilon}\mu_{XY} - \frac{\epsilon}{1-\epsilon}\nu_{XY}$.

Note that for any $\epsilon > 0$ the distribution of (U_ϵ, X, Y) belongs to $\mathcal{P}_\mu$.

For any $0 < \lambda < k_p(X;Y) - \delta$ define the function

$$g(\epsilon) := I(U_\epsilon;Y) - \lambda(pI(U_\epsilon;X,Y) - (p-1)I(U_\epsilon;X)).$$

Elementary calculations yield that

$$\frac{dg(\epsilon)}{d\epsilon}\Big|_{\epsilon\downarrow 0} = D(\nu(y)\|\mu(y)) - \lambda(pD(\nu(x,y)\|\mu(x,y)) - (p-1)D(\nu(x)\|\mu(x)) > 0,$$

where the last inequality is because of the choice of $\nu(X,Y)$ and as $0 < \lambda < k_p(X;Y) - \delta$. Since $g(0) = 0$ this implies that for some $\epsilon' > 0$ we have $I(U'_\epsilon;Y) - \lambda(pI(U'_\epsilon;X,Y) - (p-1)I(U'_\epsilon;X)) > 0$. Note also that $I(U'_\epsilon;X,Y) > 0$ since $\nu_{XY} \neq \mu_{XY}$.

This implies that

$$\sup_{\substack{\nu_{UXY}\in\mathcal{P}_\mu \\ I(U;XY)>0}} \frac{I(U;Y)}{pI(U;XY) - (p-1)I(U;X)} \ge \frac{I(U_{\epsilon'};Y)}{pI(U'_\epsilon;X,Y) - (p-1)I(U'_\epsilon;X)} > \lambda.$$

Since the above holds for all $\lambda < k_p(X;Y) - \delta$ we have

$$u_p(X;Y) = \sup_{\substack{\nu_{UXY}\in\mathcal{P}_\mu \\ I(U;XY)>0}} \frac{I(U;Y)}{pI(U;XY) - (p-1)I(U;X)} \ge k_p(X;Y) - \delta.$$

Finally, since $\delta > 0$ is arbitrary, we are done. □

[1] For instance $\epsilon < \min_{(x,y):\nu(x,y)>0} \frac{\mu(x,y)}{\nu(x,y)}$ and $\epsilon > 1 - \max_{(x,y)} \mu(x,y)$ suffices. Observe that such an ϵ exists if X and Y are not constants.

Proof of $u_p(X;Y) \leq r_p(X;Y)$: This proof uses the tensorization property of $r_p(X;Y)$ as well as the notion of typical sequences often employed in multi-terminal information theory.

Consider any $(U,X,Y) \sim \nu_{UXY} \in \mathcal{P}_\mu$. Consider (U^n,X^n,Y^n) distributed according to $\prod_i \nu_{UXY}(u_i,x_i,y_i)$, i.e. the components are independent and identically distributed.

Pick a single u^n such that

$$\{|\{i : u_i = u\}| - n\nu_U(u) \leq |\mathcal{U}|\}.$$

For instance, let $\mathcal{U} = \{1,2,\ldots,m\}$ and set the first $\lceil n\nu_U(1)\rceil$ entries of u^n to be 1, then next $\lceil n\nu_U(2)\rceil$ entries of u^n to be 2, and so on. At the end one would have at least

$$n - \sum_{i=1}^{m-1} \lceil n\nu_U(i)\rceil \geq n - (m-1) - n\sum_{i=1}^{m-1}\nu_U(i) = n\nu_U(m) - (m-1)$$

entries taking the final value m.

Define two sets according to

$$\mathcal{A} = \{x^n : \left||\{i : (u_i,x_i) = (u,x)\}| - n\nu_{UX}(u,x)\right| \leq \sqrt{n}\log(n)\nu_{UX}(u,x) \text{ for all } (u,x)\}$$

and

$$\mathcal{B} = \{y^n : \left||\{i : (u_i,y_i) = (u,y)\}| - n\nu_{UY}(u,y)\right| \leq \sqrt{n}\log(n)\nu_{UY}(u,y) \text{ for all } (u,y)\}.$$

In the language used in network information theory, these are the sets of sequences x^n and y^n respectively that are *jointly typical* with the u^n sequence chosen earlier.

Note that for any set $\mathcal{A}$ and $\mathcal{B}$ we have (Lemma 1 in [1])

$$\mathrm{P}(X^n \in \mathcal{A}, Y^n \in \mathcal{B}) = \mathrm{E}(1_{\mathcal{A}}\,\mathrm{E}(1_{\mathcal{B}}|X^n)) \leq \mathrm{P}(\mathcal{A})^{1-\frac{1}{p}} \|\,\mathrm{E}(1_{\mathcal{B}}|X^n)\|_p \leq \mathrm{P}(\mathcal{A})^{1-\frac{1}{p}}\,\mathrm{P}(B)^{\frac{1}{r_p p}}. \tag{2}$$

The first inequality follows from Hölder and the second one by the definition and tensorization property of $r_p(X;Y)$ (which implies $r_p(X^n;Y^n) = r_p(X;Y)$ when $\{X_i,Y_i\}$ are i.i.d. according to (X,Y)).

It is known (by a simple counting argument) that $\frac{1}{n}\log_2 \mathrm{P}(\mathcal{A}) \to -I(U;X)$ and $\frac{1}{n}\log_2 \mathrm{P}(\mathcal{B}) \to -I(U;Y)$ as $n \to \infty$.

Define

$$C = \{(x^n,y^n) : \left||\{i : (u_i,x_i,y_i) = (u,x,y)\}| - n\nu_{UXY}(u,x,y)\right| \leq \sqrt{n}\log(n)\nu_{UXY}(u,x,y) \text{ for all } (u,x,y)\}.$$

Clearly if $(x^n,y^n) \in C$ then $x^n \in \mathcal{A}$ and $y^n \in \mathcal{B}$. Thus $\mathrm{P}(C) \leq \mathrm{P}(X^n \in \mathcal{A}, Y^n \in \mathcal{B})$. A counting argument again shows that $\frac{1}{n}\log_2 \mathrm{P}(C) \to -I(U;XY)$.

Since we have

$$\mathrm{P}(C) \leq \mathrm{P}(X^n \in \mathcal{A}, Y^n \in \mathcal{B}) \leq \mathrm{P}(\mathcal{A})^{1-\frac{1}{p}}\,\mathrm{P}(B)^{\frac{1}{r_p p}}$$

by taking logarithms and dividing by n and letting n go to infinity, we obtain that

$$-I(U;XY) \leq -\left(1-\frac{1}{p}\right)I(U;X) - \frac{1}{r_p p}I(U;Y).$$

This implies that

$$r_p(pI(U;XY) - (p-1)I(U;X)) \geq I(U;Y)$$

for every $\nu_{UXY} \in \mathcal{P}_\mu$. When $I(U;XY) > 0$ we see that $pI(U;XY) - (p-1)I(U;X) > 0$ implying

$$r_p \geq \sup_{\substack{\nu_{UXY}\in\mathcal{P}_\mu \\ I(U;XY)>0}} \frac{I(U;Y)}{pI(U;XY) - (p-1)I(U;X)} = u_p(X;Y),$$

as desired. □

The remaining part of the proof is to show that the common value is also given by

$$\inf\{\lambda : H(Y) - \lambda pH(X,Y) + \lambda(p-1)H(X) = \mathsf{K}[H(Y) - \lambda pH(X,Y) + \lambda(p-1)H(X)]_{\mu_{XY}}\}. \tag{3}$$

It is an easy exercise to observe that

$$\mathsf{K}[H(Y) - \lambda p H(X,Y) + \lambda(p-1)H(X)]_{\mu_{XY}} = \inf_{\nu_{UXY} \in \mathcal{P}_\mu} H(Y|U) - \lambda p H(X,Y|U) + \lambda(p-1)H(X|U).$$

Thus if the equality

$$H(Y) - \lambda p H(X,Y) + \lambda(p-1)H(X) = \mathsf{K}[H(Y) - \lambda p H(X,Y) + \lambda(p-1)H(X)]_{\mu_{XY}}$$

holds for some λ then for every $\nu_{UXY} \in \mathcal{P}_\mu$

$$H(Y|U) - \lambda p H(X,Y|U) + \lambda(p-1)H(X|U) \geq H(Y) - \lambda p H(X,Y) + \lambda(p-1)H(X).$$

Rearrangement yields

$$\lambda(pI(U;XY) - (p-1)I(U;X)) \geq I(U;Y),$$

implying

$$\lambda \geq \sup_{\substack{\nu_{UXY} \in \mathcal{P}_\mu \\ I(U;XY)>0}} \frac{I(U;Y)}{pI(U;XY) - (p-1)I(U;X)} = u_p(X;Y).$$

Let $\lambda_p(X;Y)$ denote the infimum of λ satisfying (3). Then we have $\lambda_p(X;Y) \geq u_p(X;Y)$.

On the other hand, for any $\epsilon > 0$ there must exist a $\nu_{UXY} \in \mathcal{P}_\mu$ such that

$$\begin{aligned} &H(Y|U) - (\lambda_p(X;Y) - \epsilon)pH(X,Y|U) + (\lambda_p(X;Y) - \epsilon)(p-1)H(X|U) \\ &\quad < H(Y) - (\lambda_p(X;Y) - \epsilon)pH(X,Y) + (\lambda_p(X;Y) - \epsilon)(p-1)H(X). \end{aligned}$$

Re-arrangement yields

$$(\lambda_p(X;Y) - \epsilon)(pI(U;XY) - (p-1)I(U;X)) < I(U;Y).$$

For such a ν_{UXY} clearly $I(U;XY) > 0$, hence

$$\lambda_p(X;Y) - \epsilon < \frac{I(U;Y)}{pI(U;XY) - (p-1)I(U;X)} \leq u_p(X;Y).$$

Taking $\epsilon \to 0$ yields the desired equality that $\lambda_p(X;Y) \leq u_p(X;Y)$ completing the proof of the equivalence. □

The last part of this section is to show that in the above calculations one can restrict to random variables U that take at most $|\mathcal{X}||\mathcal{Y}|$ distinct values. (These are also standard arguments in network information theory.)

For any distribution ν_{XY} on $\mathcal{X} \times \mathcal{Y}$ define the function

$$f(\nu) = H(Y) - \lambda p H(X,Y) + \lambda(p-1)H(X)$$

where the entropies are evaluated at the distribution ν_{XY}.

Observe that (by Caratheodory-Fenchel-Bunt theorem) the lower convex envelope of $f(\nu)$ and the distribution μ_{XY} denoted earlier as

$$\mathsf{K}[H(Y) - \lambda p H(X,Y) + \lambda(p-1)H(X)]_{\mu_{XY}}$$

can be computed as a convex combination of at most $|\mathcal{X}||\mathcal{Y}|$ distributions $\nu_{XY}^{(i)}$, $i = 1, .., |\mathcal{X}||\mathcal{Y}|$. Consider $\nu_{XY}^{(i)}$ to be the conditional distribution of (X,Y) when $U = i$ and thus observe that we have

$$\mathsf{K}[H(Y) - \lambda p H(X,Y) + \lambda(p-1)H(X)]_{\mu_{XY}} = \inf_{\substack{\nu_{UXY} \in \mathcal{P}_\mu \\ |\mathcal{U}| \leq |\mathcal{X}||\mathcal{Y}|}} H(Y|U) - \lambda p H(X,Y|U) + \lambda(p-1)H(X|U).$$

Remark 1. This remark is for those unfamiliar with the cardinality bounding arguments in network information theory. To apply Caratheodory-Fenchel-Bunt theorem one considers the continuous mapping from ν_{XY} to $\mathcal{S} \subset \mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$ where the first $|\mathcal{X}||\mathcal{Y}| - 1$ co-ordinates represent the values of $\nu_{XY}(i,j)$ (except the entry $i = \mathcal{X}$, $j = \mathcal{Y}$, whose value is forced by the remaining entries since ν_{XY} is a probability vector) and the last co-ordinate is the value $f(\nu)$. One is interested in obtaining a particular point in the convex hull of $\mathcal{S}$ using as few points from $\mathcal{S}$ as possible. Since $\mathcal{S}$ is connected it suffices to use $|\mathcal{X}||\mathcal{Y}|$ distinct points. This improvement over Caratheodory is due to Fenchel and later generalized by Bunt. □

Historical background and Acknowledgements

This work is directly motivated by the work of the author with Venkat Anantharam, Amin Gohari, and Sudeep Kamath in [2] which deals with the equivalent version of Theorem 1 when $p \to \infty$. Trying to generalize this result to finite p, the author had obtained that both $u_p(X;Y)$ and $k_p(X;Y)$ were lower bounds to $r_p(X;Y)$. Numerical experiments conducted by Sudeep Kamath suggested that these bounds were tight, prompting the author to complete the proof of the equivalence. The author wishes to thank Venkat Anantharam, Amin Gohari, and Sudeep Kamath for various stimulating discussions relating to the connections between hypercontractivity parameters and information measures.

This work was supported by an AoE grant (Project No. AoE/E-02/08) and two GRF grants (Project Nos. 415810 and 415612) from the UGC of the Hong Kong Special Administrative Region, China.

References

[1] Rudolf Ahlswede and Peter Gács, Spreading of sets in product spaces and hypercontraction of the markov operator, The Annals of Probability (1976), 925–939.

[2] Venkat Anantharam, Amin Aminzadeh Gohari, Sudeep Kamath, and Chandra Nair, On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover, CoRR **abs/1304.6133** (2013).

[3] E. Brian Davies, Leonard Gross, and Barry Simon, Hypercontractivity: a bibliographic review, Ideas and methods in quantum and statistical physics (Oslo, 1988), Cambridge Univ. Press, Cambridge, 1992, pp. 370–389. MR 1190534 (93g:47052)

[4] Jeff Kahn, G. Kalai, and Nathan Linial, The influence of variables on boolean functions, Foundations of Computer Science, 1988., 29th Annual Symposium on, 1988, pp. 68–80.

[5] J Körner and K Marton, Comparison of two noisy channels, Topics in Inform. Theory(ed. by I. Csiszar and P.Elias), Keszthely, Hungary (August, 1975), 411–423.

[6] Michel Talagrand, On russo's approximate zero-one law, The Annals of Probability **22** (1994), no. 3, pp. 1576–1587 (English).