On Hypercontractivity and a Data Processing Inequality

Venkat Anantharam
EECS Department,
University of California,
Berkeley, CA, USA
ananth@eecs.berkeley.edu

Amin Gohari
ISSL Lab, EE Department,
Sharif University of Technology,
Tehran, Iran
aminzadeh@sharif.edu

Sudeep Kamath
ECE Department,
University of California,
San Diego, CA, USA
sukamath@ucsd.edu

Chandra Nair
IE Department,
The Chinese University
of Hong Kong
chandra@ie.cuhk.edu.hk

***Abstract*—In this paper we provide the correct tight constant to a data-processing inequality claimed by Erkip and Cover. The correct constant turns out to be a particular hypercontractivity parameter of (X, Y), rather than their squared maximal correlation. We also provide alternate geometric characterizations for both maximal correlation as well as the hypercontractivity parameter that characterizes the data-processing inequality.**

I. Introduction

Given a pair of random variables[1] (X, Y), the *data-processing inequality* states that whenever $U \to X \to Y$ form a Markov chain, we have

$$I(U; Y) \leq I(U; X).$$

A natural question to ask is the following: what is the smallest r such that the inequality

$$I(U; Y) \leq r I(U; X)$$

holds for every U whenever $U \to X \to Y$ is Markov. Erkip and Cover [3] claimed that the smallest possible r is $\rho_m^2(X; Y)$, the squared *maximal correlation* between X and Y. We show that this result is incorrect and we establish that the right constant is related to a particular hypercontractivity parameter of (X, Y).

A. Definitions and preliminaries

Definition 1. *For any real-valued random variable W with finite support, and any real number $\mathsf{p} \geq 1$, define $\|W\|_{\mathsf{p}} := (\mathrm{E}\,|W|^{\mathsf{p}})^{\frac{1}{\mathsf{p}}}$.*

Definition 2. *Given random variables X and Y, the Hirschfeld-Gebelein-Rényi maximal correlation of (X, Y) is defined as follows:*

$$\rho_m(X; Y) := \sup \mathrm{E}[f(X)g(Y)], \tag{1}$$

where the supremum is taken over all functions f, g such that

$$\mathrm{E}\, f(X) = \mathrm{E}\, g(Y) = 0, \text{ and } \mathrm{E}\, f^2(X), \mathrm{E}\, g^2(Y) \leq 1.$$

[1] Throughout this paper, random variables (X, Y) take values in $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}|, \mathcal{Y}| < \infty$. Further we assume that $\mathbb{P}(X = x) > 0\ \forall x \in \mathcal{X}$ and $\mathbb{P}(Y = y) > 0\ \forall y \in \mathcal{Y}$.

Definition 3. *A pair of random variables (X, Y) is said to be (p, q)-hypercontractive for $1 \leq \mathsf{q} \leq \mathsf{p} < \infty$ if the inequality*

$$\|\,\mathrm{E}(g(Y)|X)\|_{\mathsf{p}} \leq \|g(Y)\|_{\mathsf{q}}$$

holds for all functions $g(Y)$.

Definition 4. *For $\mathsf{p} \geq 1$ define $\mathsf{q}_{\mathsf{p}}^*(X; Y)$ as*

$$\inf\{\mathsf{q} : \mathsf{q} \geq 1, (X, Y) \textit{ is } (\mathsf{p}, \mathsf{q}) - \textit{hypercontractive}\}.$$

For $\mathsf{p} \geq 1$, we define the following quantity:

$$r_{\mathsf{p}}(X; Y) := \frac{\mathsf{q}_{\mathsf{p}}^*(X; Y)}{\mathsf{p}}.$$

An important property of the hypercontractivity parameter $r_{\mathsf{p}}(X; Y)$ is the so-called *tensorization property*. It is known [1] that if (X_1, Y_1) is independent of (X_2, Y_2), then

$$r_{\mathsf{p}}(X_1, X_2; Y_1, Y_2) = \max\{r_{\mathsf{p}}(X_1; Y_1), r_{\mathsf{p}}(X_2; Y_2)\}.$$

The following theorem summarizes some results about the quantity $r_{\mathsf{p}}(X; Y)$.

Theorem 1 ([1] Theorem 3a,3b)**.** *The following statements hold:*

(i) $r_{\mathsf{p}}(X; Y)$ *is non-increasing in* p.
(ii) $r_{\mathsf{p}}(X; Y) \geq \rho_m^2(X; Y) + \frac{1 - \rho_m^2(X;Y)}{\mathsf{p}}$ *for all* $\mathsf{p} \geq 1$.

Denote

$$r_\infty(X; Y) := \inf_{\mathsf{p} \geq 1} r_{\mathsf{p}}(X; Y). \tag{2}$$

Remark 1. *It is clear from Theorem 1 that*

$$r_\infty(X; Y) = \lim_{\mathsf{p} \to \infty} r_{\mathsf{p}}(X; Y) \geq \rho_m^2(X; Y). \tag{3}$$

Let $\nu_X(x)$ and $\mu_X(x)$ be probability distributions on the same finite set. We use $D_{KL}(\nu_X \| \mu_X)$ to denote the relative entropy or the Kullback-Liebler divergence between $\nu_X(x)$ and $\mu_X(x)$, i.e.

$$D_{KL}(\nu_X \| \mu_X) := \sum_x \nu_X(x) \log \frac{\nu_X(x)}{\mu_X(x)}\ .$$

Given a pair of random variables $(X, Y) \sim \mu(x, y)$ where $\mu(x, y)$ denotes their probability mass function, let $\mu(y|x)$ be the channel from X to Y induced by $\mu(x, y)$. We consider X to be the input and Y to be the output of the channel. For any input $\nu_X(x)$, let $\nu_Y^\mu(y) = \sum_x \nu_X(x)\mu(y|x)$ denote the induced output distribution by the channel $\mu(y|x)$ when the input distribution is ν_X. Define

$$d_*(X;Y) := \sup_{\nu_X \neq \mu_X} \frac{D_{KL}(\nu_Y^\mu \| \mu_Y)}{D_{KL}(\nu_X \| \mu_X)}. \tag{4}$$

Finally define the main quantity of interest:

$$m_*(X;Y) := \sup_{U:\ U-X-Y, I(U;X)>0} \frac{I(U;Y)}{I(U;X)}. \tag{5}$$

Here again we think of X as the input and Y as the output of the channel characterized by $\mu(y|x)$.

Remark 2. *Witsenhausen and Wyner [13] consider the trade-off between possible values of $I(U;X)$ and $I(U;Y)$. The tensorization property of $m_*(X;Y)$ can be inferred from their results.*

B. Summary of results

We provide a proof of the following equivalence result:

Theorem 2. *Given a pair of random variables (X, Y) and the quantities $r_\infty(X;Y)$, $d_*(X;Y)$, and $m_*(X;Y)$ as defined in the previous section, we have*

$$r_\infty(X;Y) = d_*(X;Y) = m_*(X;Y).$$

Remark 3. *The equality $r_\infty(X;Y) = d_*(X;Y)$ was established in [1, Theorem 5a], and only the second equality $d_*(X;Y) = m_*(X;Y)$ is new here. However we will prove a three way equivalence in this paper as opposed to only establishing $d_*(X;Y) = m_*(X;Y)$.*

Remark 4. *In [3, Theorem 8] it was claimed that the following inequality holds:*

$$I(U;Y) \leq \rho_m^2(X;Y)I(U;X),\ \forall\ U - X - Y.$$

It turns out that this inequality is incorrect. Indeed from Theorem 2 and Remark 1 it is immediate that $m_*(X;Y) \geq \rho_m^2(X;Y)$. We will prove later in the paper that the inequality is strict in general by providing an explicit example. The strictness of the inequality $r_\infty(X;Y) \geq \rho_m^2(X;Y)$ in general is known from [1, Theorem 9b]. The strictness of the inequality $m_*(X;Y) \geq \rho_m^2(X;Y)$ would also alternately follow from our Theorem 2.

In this paper we will also provide alternate geometric characterizations of both $\rho_m^2(X;Y)$ and $r_\infty(X;Y)$. Fix a channel $\mu_{Y|X}(y|x)$, fix $\lambda \in [0, 1]$, and consider the function[2] of the probability distribution of X denoted by $t_\lambda(X)$ which is defined by

$$t_\lambda(X) := H(Y) - \lambda H(X).$$

[2]We abuse notation when we write $t_\lambda(X)$. We really wish to think of t_λ as a function of the probability distribution of X.

We will show in Theorem 3 that $\rho_m^2(X, Y)$ is the smallest λ such that $t_\lambda(X)$ has a positive semidefinite Hessian at $\mu(x)$ and $r_\infty(X;Y)$ is the smallest λ such that $t_\lambda(X)$ matches its lower convex envelope, denoted by $\mathcal{K}[t_\lambda](X)$, at $\mu_X(x)$.

C. Organization of the paper

In Section II, we will prove Theorem 2. In Section III we will establish the alternate geometric characterizations for $\rho_m^2(X;Y)$ and $r_\infty(X;Y)$. The paper we uploaded on arXiv [15] presents the key results from a slightly different perspective.

II. Proof of Theorem 2

Remark: The proof below is stitched together using a judicious borrowing of arguments from [1], [14], and standard techniques. In the authors' opinion, the three way equivalence argument elucidates the proof of the equivalence $r_\infty(X;Y) = d_*(X;Y)$ (in [1]). Further this proof idea allows for a natural generalization and provides an alternate (new) characterization of $r_{\mathsf{p}}(X;Y), \mathsf{p} \geq 1$ [18].

If X and Y are independent, then it is easy to see that $r_{\mathsf{p}}(X;Y) = \frac{1}{\mathsf{p}}\ \forall \mathsf{p} \geq 1$; hence $r_\infty(X;Y) = 0$. It is also easy to see that $d_*(X;Y) = m_*(X;Y) = 0$. The theorem clearly holds in this case.

We will assume then that X and Y are not independent. By choosing $f(x) = 1_{x \in A} - P(X \in A), g(y) = 1_{y \in B} - P(Y \in B)$ in (1) for appropriate sets A, B, we can obtain $\rho_m(X;Y) > 0$. From (3), we get $r_\infty(X;Y) > 0$.

The proof will follow from the following sequence of implications that we will establish.

(a) $r_\infty(X;Y) \leq d_*(X;Y)$,
(b) $d_*(X;Y) \leq m_*(X,Y)$,
(c) $m_*(X,Y) \leq r_\infty(X,Y)$.

Proof of (a): Writing r_{p}, r_∞ for $r_{\mathsf{p}}(X;Y), r_\infty(X;Y)$ respectively, from the definition we have the following inequality:

$$\| \mathrm{E}(g(Y)|X)\|_{\mathsf{p}} \leq \|g(Y)\|_{r_{\mathsf{p}}\mathsf{p}},$$

for all $g(Y) \geq 0$.

Define $h(y) := g(y)^{r_{\mathsf{p}}\mathsf{p}}$, and note that

$$\|g(Y)\|_{r_{\mathsf{p}}\mathsf{p}}^{r_{\mathsf{p}}\mathsf{p}} = \mathrm{E}(h(Y)).$$

We also obtain that for all $\mathsf{p} \geq 1$

$$\| \mathrm{E}(g(Y)|X)\|_{\mathsf{p}}^{r_{\mathsf{p}}\mathsf{p}} = \left(\sum_x \mu(x) \left(\sum_y \mu(y|x) h(y)^{\frac{1}{r_{\mathsf{p}}\mathsf{p}}} \right)^{\mathsf{p}} \right)^{r_{\mathsf{p}}}.$$

Using the well-known limit $\lim_{r \downarrow 0} ||W||_r = \exp(\mathrm{E} \log |W|)$, we get that as $\mathsf{p} \to \infty$,

$$\left(\sum_y \mu(y|x) h(y)^{\frac{1}{r_{\mathsf{p}}\mathsf{p}}} \right)^{\mathsf{p}} \to \prod_y h(y)^{\frac{\mu(y|x)}{r_\infty}}.$$

Thus we have for all $h(Y) > 0$

$$\left(\sum_x \mu(x) \prod_y h(y)^{\frac{\mu(y|x)}{r_\infty}} \right)^{r_\infty} \leq \mathrm{E}(h(Y)). \tag{6}$$

From the definition of $r_\infty(X;Y)$ and the continuity and strict monotonicity of $\|\cdot\|_r$ in r for non-constant random variables, it is immediate that it is the smallest such number for which the above inequality holds for all $h(Y) > 0$.

Therefore for any $\epsilon > 0$ there exists an $h_\epsilon(Y)$ such that

$$\left(\sum_x \mu(x) \prod_y h_\epsilon(y)^{\frac{\mu(y|x)}{r_\infty - \epsilon}}\right)^{r_\infty - \epsilon} > \mathrm{E}(h_\epsilon(Y)).$$

Further w.l.o.g. assume that

$$\mathrm{E}(h_\epsilon(Y)) = 1.$$

Define a probability distribution (see [1, (5.11)])

$$\nu(x) = C\mu(x) \prod_y h_\epsilon(y)^{\frac{\mu(y|x)}{r_\infty - \epsilon}},$$

where $C < 1$ is a normalizing constant. Now, note that

$$\nu(x) \log \frac{\nu(x)}{C\mu(x)} = \nu(x) \sum_y \frac{\mu(y|x)}{r_\infty - \epsilon} \log h_\epsilon(y)$$

$$\implies \quad \sum_x \nu(x) \log \frac{\nu(x)}{\mu(x)} = \sum_y \frac{\nu^\mu(y)}{r_\infty - \epsilon} \log h_\epsilon(y) + \log C.$$

Finally observe that since $C < 1$

$$\begin{aligned}
&\sum_x \nu(x) \log \frac{\nu(x)}{\mu(x)} \\
&< \sum_y \frac{\nu^\mu(y)}{r_\infty - \epsilon} \log h_\epsilon(y) \\
&= \sum_y \frac{\nu^\mu(y)}{r_\infty - \epsilon} \log \frac{\nu^\mu(y)}{\mu(y)} + \sum_y \frac{\nu^\mu(y)}{r_\infty - \epsilon} \log \frac{\mu(y) h_\epsilon(y)}{\nu^\mu(y)} \\
&\leq \sum_y \frac{\nu^\mu(y)}{r_\infty - \epsilon} \log \frac{\nu^\mu(y)}{\mu(y)} + \frac{1}{r_\infty - \epsilon} \log(\sum_y \mu(y) h_\epsilon(y)) \\
&= \sum_y \frac{\nu^\mu(y)}{r_\infty - \epsilon} \log \frac{\nu^\mu(y)}{\mu(y)},
\end{aligned}$$

where the last equality follows since $\mathrm{E}(h_\epsilon(Y)) = 1$.

Thus $r_\infty(X;Y) - \epsilon < \frac{D(\nu^\mu(y)||\mu(y))}{D(\nu(x)||\mu(x))} \leq d_*(X;Y)$. Since $\epsilon > 0$ is arbitrary we are done.

Proof of (b): Let $\mathcal{U}_\epsilon := \{1, 2\}$. Fix a sufficiently small $\epsilon > 0$ and define U_ϵ satisfying $U_\epsilon - X - Y$ by

- $P(U_\epsilon = 1) = \epsilon, P(X = x | U_\epsilon = 1) = \nu_\delta(x)$,
- $P(U_\epsilon = 2) = 1 - \epsilon, P(X = x | U_\epsilon = 2) = \mu(x) + \frac{\epsilon}{1-\epsilon}(\mu(x) - \nu_\delta(x)) = \frac{1}{1-\epsilon}\mu(x) - \frac{\epsilon}{1-\epsilon}\nu_\delta(x)$,

where $\nu_\delta(x) \neq \mu(x)$ is a probability distribution satisfying $\frac{D\left(\nu^\mu_\delta(y)\|\mu(y)\right)}{D\left(\mu_\delta(x)\|\mu(x)\right)} > d_*(X;Y) - \delta > 0$. For sufficiently small $\epsilon > 0$, we have that $\frac{1}{1-\epsilon}\mu(x) - \frac{\epsilon}{1-\epsilon}\nu_\delta(x)$ is a probability distribution (as $\mu(x)$ was assumed to have full support). Note that

$$\begin{aligned}
&P(U_\epsilon = 1) P(X = x | U_\epsilon = 1) + \\
&\quad P(U_\epsilon = 2) P(X = x | U_\epsilon = 2) = \mu(x) \ \forall x \in \mathcal{X},
\end{aligned}$$

so that this specified chain $U_\epsilon - X - Y$ has the correct marginal distribution for (X, Y).

For any $0 < \theta < d^*(X;Y) - \delta$ define the function

$$g(\epsilon) := I(U_\epsilon; Y) - \theta I(U_\epsilon; X).$$

We have

$$\begin{aligned}
\frac{dg(\epsilon)}{d\epsilon} &= -\frac{d}{d\epsilon}\left(\epsilon H(\nu^\mu_\delta(y)) + (1-\epsilon) H\left(\frac{1}{1-\epsilon}\mu(y) - \frac{\epsilon}{1-\epsilon}\nu^\mu_\delta(y)\right)\right) \\
&+ \theta \frac{d}{d\epsilon}\left(\epsilon H(\nu_\delta(x)) + (1-\epsilon) H\left(\frac{1}{1-\epsilon}\mu(x) - \frac{\epsilon}{1-\epsilon}\nu_\delta(x)\right)\right) \\
&= -H(\nu^\mu_\delta(y)) + H\left(\frac{\mu(y) - \epsilon\nu^\mu_\delta(y)}{1-\epsilon}\right) + \theta H(\nu_\delta(x)) \\
&- \theta H\left(\frac{\mu(x) - \epsilon\nu_\delta(x)}{1-\epsilon}\right) - \sum_y \frac{\nu^\mu_\delta(y) - \mu(y)}{1-\epsilon} \log\left(\frac{\mu(y) - \epsilon\nu^\mu_\delta(y)}{1-\epsilon}\right) \\
&+ \theta \sum_x \frac{\nu_\delta(x) - \mu(x)}{1-\epsilon} \log\left(\frac{\mu(x) - \epsilon\nu_\delta(x)}{1-\epsilon}\right).
\end{aligned}$$

Thus

$$\frac{dg(\epsilon)}{d\epsilon}\Big|_{\epsilon=0} = D\big(\nu^\mu_\delta(y)\|\mu(y)\big) - \theta D\big(\nu_\delta(x)\|\mu(x)\big) > 0,$$

where the last inequality is because $0 < \theta < d^*(X;Y) - \delta$ and $\frac{D\left(\nu^\mu_\delta(y)\|\mu(y)\right)}{D\left(\nu_\delta(x)\|\mu(x)\right)} > d_*(X;Y) - \delta$. Since $g(0) = 0$ this implies that for some $\epsilon' > 0$ we have $I(U_{\epsilon'}; Y) - \theta I(U_{\epsilon'}; X) > 0$ or that

$$m_*(X;Y) = \sup_{U:\ U - X - Y, I(U;Y) > 0} \frac{I(U;Y)}{I(U;X)} \geq \frac{I(U_{\epsilon'}; Y)}{I(U_{\epsilon'}; X)} > \theta.$$

Since the above holds for all $\theta < d_*(X;Y) - \delta$ we have

$$m_*(X;Y) \geq d_*(X;Y) - \delta.$$

Finally, since $\delta > 0$ is arbitrary, we let $\delta \to 0$, and we are done.

Proof of (c): This part uses standard typicality arguments in information theory and our definition of (and notation for) ϵ-typical sets are borrowed from [16].

For any $U \to X \to Y$ let $(U^n, X^n, Y^n) \sim \prod_i \mu(u_i, x_i, y_i)$. Pick a single $u^n \in \mathcal{T}^{(n)}_\epsilon(U)$. For some $\epsilon_1 > \epsilon$ let $\mathcal{A}_n = \{x^n : (u^n, x^n) \in \mathcal{T}^{(n)}_{\epsilon_1}(U, X)\}$ and for $\epsilon_2 > \epsilon_1$ let $\mathcal{B}_n = \{y^n : (u^n, y^n) \in \mathcal{T}^{(n)}_{\epsilon_2}(U, Y)\}$. Note that

$$\begin{aligned}
&\mathrm{P}(X^n \in \mathcal{A}_n, Y^n \in \mathcal{B}_n) \\
=& \mathrm{E}[1_{X^n \in \mathcal{A}_n} \mathrm{E}(1_{Y^n \in \mathcal{B}_n} | X^n)] \\
\leq& \mathrm{E}\, \|1_{X^n \in \mathcal{A}_n}\|_{\frac{\mathsf{p}}{\mathsf{p}-1}} \|\mathrm{E}(1_{Y^n \in \mathcal{B}_n} | X^n)\|_{\mathsf{p}} \\
\leq& \mathrm{E}\, \|1_{X^n \in \mathcal{A}_n}\|_{\frac{\mathsf{p}}{\mathsf{p}-1}} \mathrm{E}\, \|1_{Y^n \in \mathcal{B}_n}\|_{r_{\mathsf{p}}\mathsf{p}} \\
=& P(X^n \in \mathcal{A}_n)^{1 - \frac{1}{\mathsf{p}}} \mathrm{P}(Y^n \in \mathcal{B}_n)^{\frac{1}{r_{\mathsf{p}}\mathsf{p}}},
\end{aligned}$$

where we write r_{p} to denote $r_{\mathsf{p}}(X;Y)$ for convenience. The first inequality follows from Hölder's inequality and the second inequality from the definition and tensorization property of $r_{\mathsf{p}}(X;Y)$.

Standard calculations tell us that $\frac{1}{n}\log_2 \mathrm{P}(X^n \in \mathcal{A}_n) \to -I(U;X)$ and $\frac{1}{n}\log_2 \mathrm{P}(Y^n \in \mathcal{B}_n) \to -I(U;Y)$ as $n \to \infty$. From the law of large numbers we know that $\mathrm{P}(Y^n \in \mathcal{B}_n | X^n \in \mathcal{A}_n) \to 1$ as $n \to \infty$. Therefore, taking the logarithm on both sides, dividing by n and letting $n \to \infty$ we obtain that

$$-I(U;X) \le -\left(1-\frac{1}{\mathsf{p}}\right) I(U;X) - \frac{1}{r_{\mathsf{p}}\mathsf{p}} I(U;Y).$$

Rearranging we obtain $r_{\mathsf{p}}(X;Y) \ge \frac{I(U;Y)}{I(U;X)}$ and since this is true for any U, by taking the supremum on the right hand side we obtain $r_{\mathsf{p}}(X;Y) \ge m_*(X;Y)$. Since the right hand side does not depend on p, we take $\mathsf{p} \to \infty$ to obtain $r_\infty(X;Y) \ge m_*(X;Y)$. This completes the proof of Theorem 2.

III. A GEOMETRIC CHARACTERIZATION OF $\rho_m^2(X;Y)$ AND $m^*(X;Y)$

Let $\mu(y|x)$ be the channel transition probability from X to Y induced by a joint distribution $\mu(x,y)$. For a fixed channel $\mu(y|x)$, consider a function of the input distribution $\nu(x)$,

$$t_\lambda(X) := H(Y) - \lambda H(X),$$

where λ is a constant in $[0,1]$. Observe that the function is concave when $\lambda = 0$ and convex when $\lambda = 1$.[3]

We write $\mathcal{K}[t_\lambda](X)$ for the lower convex envelope of $t_\lambda(X)$.

Proposition 1. *If $\mathcal{K}[t_\lambda](X) = t_\lambda(X)$ at $\nu(x)$ for some λ then $\mathcal{K}[t_{\lambda_1}](X) = t_{\lambda_1}(X)$ at $\nu(x)$ for all $\lambda_1 \ge \lambda$.*

Proof. If $\mathcal{K}[t_\lambda](X) = t_\lambda(X)$ at $\nu(x)$ for some λ, then note that for any $\lambda_1 \ge \lambda$

$$\begin{aligned} & t_{\lambda_1}(X) = t_\lambda(X) - (\lambda_1 - \lambda)H(X) \\ \implies \quad & \mathcal{K}[t_{\lambda_1}](X) \ge \mathcal{K}[t_\lambda](X) - (\lambda_1 - \lambda)H(X). \end{aligned}$$

Here the inequality comes since $\mathcal{K}[f+g] \ge \mathcal{K}[f] + \mathcal{K}[g]$; note that $-(\lambda_1 - \lambda)H(X)$ is convex. Therefore at $\nu(x)$ we will have that

$$\begin{aligned} t_{\lambda_1}(X) &\ge \mathcal{K}[t_{\lambda_1}](X) \ge \mathcal{K}[t_\lambda](X) - (\lambda_1 - \lambda)H(X) \\ &= t_\lambda(X) - (\lambda_1 - \lambda)H(X) = t_{\lambda_1}(X), \end{aligned}$$

establishing the proposition. □

The following theorem gives a geometric interpretation of $\rho_m^2(X;Y)$ and $m_*(X;Y)$ in terms of the behaviour of the function $t_\lambda(X)$.

Theorem 3. *Let $(X,Y) \sim \mu(x,y)$. The following statements hold:*

1) *$\rho_m^2(X;Y)$ is the minimum value of λ such that the function $t_\lambda(X)$ has a positive semidefinite Hessian at $\mu(x)$.*
2) *$m_*(X;Y)$ is the minimum value of λ such that the function $t_\lambda(X)$ touches its lower convex envelope at $\mu(x)$, i.e. such that $\mathcal{K}[t_\lambda](X) = t_\lambda(X)$ at $\mu(x)$.*

Proof of 1): The claim is straightforward when X, Y are independent. When X, Y are not independent, Rényi's characterization of the maximal correlation [19] states that

$$\rho_m^2(X;Y) = \sup_{f(X):\mathrm{E}\, f(X)=0, \mathrm{E}[f^2(X)]=1} \mathrm{E}[\mathrm{E}[f(X)|Y]^2].$$

Take an arbitrary multiplicative perturbation of the form $\mu_\epsilon(x) = \mu(x)(1+\epsilon f(x))$. For μ_ϵ to stay a valid perturbation we need $\mathrm{E}[f(X)] = 0$. Furthermore we can normalize f by assuming that $\mathrm{E}[f^2(X)] = 1$. The second derivative in ϵ of $H(Y) - \lambda H(X)$ is equal to [7]

$$-\mathrm{E}[\mathrm{E}[f(X)|Y]^2] + \lambda\,\mathrm{E}[f^2(X)] = -\mathrm{E}[\mathrm{E}[f(X)|Y]^2] + \lambda\ ,$$

which is non-negative as long as $\lambda \ge \mathrm{E}[\mathrm{E}[f(X)|Y]^2]$. Thus the minimum value λ^* such that the second derivative is non-negative for all local perturbations is

$$\lambda^* = \sup_{f(X):\mathrm{E}\, f(X)=0, \mathrm{E}[f^2(X)]=1} \mathrm{E}[\mathrm{E}[f(X)|Y]^2] = \rho_m^2(X;Y).$$

■

Proof of 2): Consider the minimum value of λ, say $\lambda^\dagger$, such that the function $t_\lambda(X)$ touches its lower convex envelope at $\mu(x)$. Thus, we are looking for the minimum λ such that for $(X,Y) \sim \mu(x,y)$ we have[4]

$$H(Y) - \lambda H(X) \le H(Y|U) - \lambda H(X|U), \quad \forall\ U : U - X - Y.$$

Equivalently we require the minimum λ such that,

$$\lambda \ge \frac{I(U;Y)}{I(U;X)}, \quad \forall\ U : U - X - Y \text{ with } I(U;X) > 0.$$

Thus,

$$\lambda^\dagger = \sup_{U:\ U-X-Y, I(U;X)>0} \frac{I(U;Y)}{I(U;X)} = m_*(X;Y). \qquad ■$$

Remark 5. *Since $t_\lambda(X) = \mathcal{K}[t_\lambda](X)$ at $\mu(x)$ implies that the Hessian of $t_\lambda(X)$ at $\mu(x)$ is positive semidefinite, we have*

$$m_*(X;Y) \ge \rho_m^2(X;Y).$$

A. Counterexample to the Erkip-Cover data processing inequality

In [3, Theorem 8], Erkip and Cover claimed that

$$I(U;Y) \le \rho_m^2(X;Y) I(U;X)$$

holds whenever $U - X - Y$ form a Markov chain. Furthermore they claimed that, $\rho_m^2(X;Y)$ is the minimum such constant, i.e.

$$m_*(X;Y) = \rho_m^2(X;Y). \tag{7}$$

We will first provide a counterexample to these claims and then point out a gap in their argument.

1) Counterexample to (7): Let X be a binary random variable with $p(X=0) = \frac{1}{2}$. Define $p(x,y)$ by passing X through the asymmetric erasure channel given in Fig. 1. When either X or Y is binary then it is known [20] that

$$\rho_m^2(X;Y) = \left[\sum_{x,y} \frac{p(x,y)^2}{p(x)p(y)}\right] - 1.$$

We then have $\rho_m^2(X;Y) = 0.6$. Suppose we construct U satisfying $U - X - Y$ such that $U|\{X=0\} \sim \mathrm{Ber}(0.1), U|\{X=1\} \sim \mathrm{Ber}(0.4)$. Then $I(U;Y) = 0.055770...$ and $I(U;X) =$

[3]This convexity at $\lambda = 1$ follows from the fact that for any $U - X - Y$ we have $I(U;X) \ge I(U;Y)$ or equivalently $H(Y) - H(X) \le H(Y|U) - H(X|U)$.

[4]Note that if U is independent of X, i.e. $I(U;X) = 0$ then the above inequality is always true.

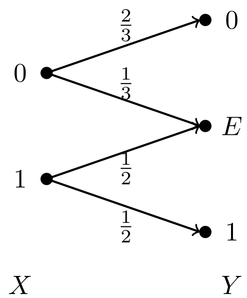

Fig. 1. An asymmetric erasure channel.

0.09130..., so that $\frac{I(Y;U)}{I(X;U)} = 0.6108... > 0.6 = \rho_m^2(X;Y)$, and this contradicts (7). It can be shown in a reasonably straightforward manner, using our characterization in Theorem 3, that $m_*(X;Y) = \frac{1}{2}\log_2\left(\frac{12}{5}\right) = 0.631517...$ for this pair of random variables (X,Y).

One can show that if a measure $\mu(x,y)$ is drawn uniformly at random from the set of all probability measures on pairs of binary random variables, then with probability one we have $m_*(X;Y) > \rho_m^2(X;Y)$.

The error of the Erkip-Cover proof lies in their use of a Taylor's series expansion. Consider the expansion in the left column of page 1037 of their paper [3], where they use their equation (16) to expand around $p(\tilde{v})$. It is possible that $p(\tilde{v})$ is zero for some $\tilde{v}$ and this causes an error as the derivative in this direction is infinity and the Taylor's series expansion is no longer valid. As our counterexample shows, this seems to be a significant but subtle error that cannot be worked around.

Some of the works that use this incorrect result of [3], such as [21], are affected by this error. A claim similar to that of [3], which appears in [9], is also false.[5]

IV. Conclusion

In this paper we showed the equivalence between the optimal constant in the data-processing inequality and a hypercontractivity parameter connecting random variables X and Y. This corrects an incorrect claim due to Erkip and Cover [3]. We also presented a new geometric characterization of the maximal correlation and of this hypercontractivity parameter.

Acknowledgements

S. Kamath and V. Anantharam gratefully acknowledge research support from the ARO MURI grant W911NF-08-1-0233, "Tools for the Analysis and Design of Complex Multi-Scale Networks", from the NSF grant CNS-0910702, and from the NSF Science & Technology Center grant CCF-0939370, "Science of Information". The work of Chandra Nair was partially supported by the following: an area of excellence grant (Project No. AoE/E-02/08) and two GRF grants (Project Nos. 415810 and 415612) from the University Grants Committee of the Hong Kong Special Administrative Region, China. The work of Amin Gohari was partially supported by Iranian National Science Foundation (INSF) network information theory chair.

[5]This paper studies the ratio $\frac{I(U;Y)}{I(U;X)}$ when $I(U;X)$ is very small. However, as pointed out in [3], the supremum of $\frac{I(U;Y)}{I(U;X)}$ occurs when $I(U;X) \to 0$. So the problem studied by [9] is the same as that of [3].

References

[1] R. Ahlswede and P. Gács "Spreading of Sets in Product Spaces and Hypercontraction of the Markov Operator," Annals of Probability, vol. 4, pp. 925–939, Dec. 1976.

[2] S. Beigi, "A New Quantum Data Processing Inequality," arXiv: 1210.1689, 2013.

[3] E. Erkip and T. Cover, "The efficiency of investment information," IEEE Transactions On Information Theory, vol. 44, no. 3, pp. 1026-1040, May 1998.

[4] P. Gács and J. Körner, "Common information is far less than mutual information," Problems of Control and Information Theory, Vol. 2, No. 2, pp. 149 -162, 1973.

[5] H. Gebelein, "Das statistische Problem der Korrelation als Variations- und Eigenwert-problem und sein Zusammenhang mit der Ausgleichungsrechnung," Zeitschrift für angew. Math. und Mech. 21, pp. 364-379, 1941.

[6] Y. Geng and C. Nair, "The capacity region of the two-receiver vector gaussian broadcast channel with private and common messages," arXiv: 1202.0097, 2012.

[7] A. Gohari and V. Anantharam, "Evaluation of Marton's Inner Bound for the General Broadcast Channel," IEEE Transactions On Information Theory, vol. 58, no. 2, pp. 608 -619, Feb 2012.

[8] H. O. Hirschfeld, "A connection between correlation and contingency," Proc. Cambridge Philosophical Soc. 31, pp 520-524, 1935.

[9] S.-L. Huang and L. Zheng, "Linear Information Coupling Problems", Procedings of the IEEE Symposium On Information Theory (ISIT), pp.1029-1033, 2012.

[10] S. Kamath and V. Anantharam, "Non-interactive Simulation of Joint Distributions: The Hirschfeld-Gebelein-Rényi Maximal Correlation and the Hypercontractivity Ribbon," Proceedings of the 50th Annual Allerton Conference on Communications, Control and Computing, pp. 1057-1064, October 2012, Monticello, Illinois.

[11] P. Delgosha and S. Beigi, " Impossibility of Local State Transformation via Hypercontractivity", arXiv: 1307.2747, 2013.

[12] W. Kang and S. Ulukus, "A New Data Processing Inequality and Its Applications in Distributed Source and Channel Coding," IEEE Transactions On Information Theory, vol. 57, no. 1, pp. 56-69, Jan 2011.

[13] H. Witsenhausen and A. Wyner, "A Conditional Entropy Bound for a Pair of Discrete Random Variables," IEEE Transactions On Information Theory, vol. 21, no. 5, pp. 493-501, September 1975.

[14] Janos Körner and Katalin Marton, "Comparison of two noisy channels," Topics in Inform. Theory(ed. by I. Csiszar and P.Elias), Keszthely, Hungary, pp 411-423, Aug. 1975.

[15] V. Anantharam, A. Gohari, S. Kamath and C. Nair, "On Maximal Correlation, Hypercontractivity, and the Data Processing Inequality studied by Erkip and Cover", arXiv: 1304.6133, 2013.

[16] A. El Gamal and Y.-H. Kim, "Network Information Theory", Cambridge University Press, Cambridge, U.K., 2012.

[17] E. Mossel, K. Oleszkiewicz and A. Sen, "On Reverse Hypercontractivity", Geometric and Functional Analysis, pp. 1-36, 2013.

[18] C. Nair, " Equivalent formulations of Hypercontractivity using Information Measures", Presented at IZS workshop, Zurich, Feb. 2014.

[19] A. Rényi, "On measures of dependence,"Acta Math. Hung., vol. 10, pp. 441-451, 1959.

[20] H.S. Witsenhausen, "On sequences of pairs of dependent random variables," SIAM Journal on Applied Mathematics, vol. 28, no. 1, pp. 100-113, January 1975.

[21] L. Zhao and Y.-K. Chia, "The efficiency of common randomness generation", 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 944 - 950, 2011.