Hypercontractivity and Information Theory

Adway Girish
LTHI, IC, EPFL

***Abstract*—The notion of hypercontractivity has been well-studied as a mathematical tool during the last five decades. Numerous powerful applications have been discovered in quantum physics, Fourier analysis, theoretical computer science, and even probability theory, but surprisingly, remains underutilized in information theory. The goal here is to study and document the progress that has been made so far. Starting with a simple question about the size of "decoding sets" as in channel coding, we see that answering a more general question instead leads to unexpected connections with various information measures. We also look at some applications of hypercontractivity in solving information-theoretic problems, and comment on what improvements can be made.**

***Index Terms*—hypercontractivity, Markov operators, information measures, contraction coefficients, maximal correlation, Boolean functions**

I. Introduction[1]

INFORMATION theorists, particularly in the style of Csiszár and Körner [1], frequently make use of "decoding sets" to prove achievability and converse results in coding theorems. These decoding sets are simply the sets $B \subseteq \mathcal{Y}$ such that for some $A \subseteq \mathcal{X}$, the probability $\mathsf{W}(B \mid x) \geq \lambda$ for all $x \in A$, for some large enough $\lambda \in (0,1)$. In the channel coding problem, these represent the sets that any input $x \in A$ has a high probability to get sent to, when transmitted over the channel W. For a given B, define $g_\lambda(B)$ to be the set of all $x \in \mathcal{X}$ such that $\mathsf{W}(B \mid x) \geq \lambda$, i.e., the set of input messages that are likely to end up in the same output set. The goal of channel coding is then to have B as small as possible while still allowing the resolution of $\mathcal{X}$ into disjoint sets of the form $g_\lambda(B)$. More generally, we can pose the following problem: *for $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$, where each (X_i, Y_i) is an independent, identically distributed (i.i.d.) copy of (X, Y), and sets $A \subseteq \mathcal{X}^n$, $B \subseteq \mathcal{Y}^n$, under the requirement that $W(B \mid A) \geq \lambda$, what is the relation between $\mathbb{P}(Y^n \in B)$ and $\mathbb{P}(X^n \in A)$?*

Ahlswede and Gács [2] give us a lower bound for $\mathbb{P}(Y^n \in B)$ in terms of $\mathbb{P}(X^n \in A)$, by showing that there exist positive numbers $r < 1$ and p such that

$$\mathbb{P}(Y^n \in B) \geq \mathsf{W}(B \mid A)^p \, \mathbb{P}(X^n \in A)^r,$$

under the assumption that the distribution of (X_i, Y_i) is indecomposable, i.e., there do not exist nontrivial sets A and B such that $\mathbb{P}(X^n \in A$ if and only if $Y^n \in B) = 1$.

By rearranging terms, the above inequality is equivalent to

$$\mathbb{P}(X^n \in A, \, Y^n \in B) \leq \mathbb{P}(X^n \in A)^{1-\frac{r}{p}} \, \mathbb{P}(Y^n \in B)^{\frac{1}{p}}.$$

For $r \geq 1$, this is always true, as it is simply a special case of Hölder's inequality[2] applied to the p-norm $\|X\|_p = \mathbb{E}[|X|^p]^{\frac{1}{p}}$, and the fact that the p-norm is nondecreasing in p. Thus, Ahlswede and Gács [2] study the more general question: *do there also exist some numbers $r < 1$ and $p > 0$ such that*

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

for all bounded functions f and g, where p' is the Hölder conjugate of p, and $q = pr < p$? This is easily seen to be equivalent to the following: *do there exist $r < 1$ and $p > 0$ such that*

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

for all bounded functions g, where $q = pr < p$? Going one step further and defining the Markov operator T, which takes bounded functions on $\mathcal{Y}$ to bounded functions on $\mathcal{X}$, as

$$(Tg)(x) = \mathbb{E}[g(Y) \mid X = x], \tag{1}$$

the above is also equivalent to $\|(Tg)(X)\|_p \leq \|g(Y)\|_{pr}$. This is always true when $r = 1$, i.e., $\|(Tg)(X)\|_p \leq \|g(Y)\|_p$ for all g, and hence T "contracts" its arguments (as expected with bounded operators). However, when $r < 1$, the right-hand side decreases, but we still want this stronger inequality to hold—hence the term "hyper–contractivity".

This write-up is centered on the three background papers [2], [3], [4]. Beginning as above, Ahlswede and Gács [2] go on to study the above problem in detail, defining relevant "hypercontractivity parameters" (Section II-C), obtaining estimates and discovering several properties for the smallest such r (Section III). Among these properties, we see the unexpected appearance of various information measures in the characterization of these parameters; this connection was studied and generalized by Chandra Nair [3]. Anantharam et al. [4] then use some properties of hypercontractivity to study a more natural information-theoretic problem—on the mutual information between Boolean functions [5]. We look at this, along with other applications, in Section IV. The goal of this write-up is to examine how information theory can benefit from having mathematical tools such as hypercontractivity, while simultaneously contributing to the development of the theory of hypercontractivity itself.

II. Preliminaries

In this section, we establish the notation that is used throughout this write-up (we follow standard conventions, but this is for completeness), and then define the relevant hypercontractivity terms.

[1] The (standard) notation used is explained in Section II.

[2] $\|f(X)g(Y)\|_1 \leq \|f(X)\|_p \|g(Y)\|_{p'}$ for all $p, p' > 1$ with $\frac{1}{p} + \frac{1}{p'} = 1$; such p, p' are called Hölder conjugates of each other.

A. Notation

Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets, and (X, Y) be a pair of random variables on $\mathcal{X} \times \mathcal{Y}$, with the joint probability mass function (pmf) μ_{XY}, which we denote as $(X, Y) \sim \mu_{XY}$. Let μ_X and μ_Y be the marginal pmfs of X and Y respectively. We assume an underlying probability measure $\mathbb{P}$, and sometimes say $\mathbb{P}(X \in A)$ to mean $\sum_{x \in A} \mu_X(x)$. Further, let $\mathsf{W}_{Y|X}$ or simply W denote the conditional pmf of Y given X, i.e., $\mathsf{W}(y \mid x) = \frac{\mu_{XY}(x,y)}{\mu_X(x)}$, called the *channel* from X to Y. Given any distribution μ_X and a channel W, we define the corresponding "channel output" distribution μ_Y as the marginal distribution induced by the channel, i.e., $\mu_Y(y) = \sum_{x \in \mathcal{X}} \mu_X(x) \, \mathsf{W}(y \mid x)$.

We also abuse notation in the following standard manner: for $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$, where each $(X_i, Y_i) \sim \mu_{XY}$ i.i.d., we also use $\mu_X(A) = \mathbb{P}(X^n \in A)$, $\mathsf{W}(y^n \mid x^n) = \mathsf{W}^n(y^n \mid x^n) = \mathbb{P}(Y^n = y^n \mid X^n = x^n)$, and $\mathsf{W}(B \mid A) = \mathbb{P}(Y^n \in B \mid X^n \in A)$, for $A \subseteq \mathcal{X}^n$ and $B \subseteq \mathcal{Y}^n$.

The expectation of any function f of a random variable X is given by $\mathbb{E}_{\mu_X}[f(X)] \triangleq \sum_{x \in \mathcal{X}} \mu_X(x) \, f(x)$; we simply say $\mathbb{E}[f(X)]$ when the distribution is clear. The indicator function of a set A is defined as follows: $\mathbb{1}_A(x) = 1$ when $x \in A$ and 0 otherwise. All logarithms are taken to the base 2.

B. Information measures

Given two distributions ν_X and μ_X on $\mathcal{X}$, we say that ν_X is *absolutely continuous* w.r.t. μ_X, if $\nu_X(x) = 0$ for some $x \in \mathcal{X}$ implies that $\mu_X(x) = 0$; this is denoted as $\nu_X \ll \mu_X$. We then define the *KL-divergence* from μ_X to ν_X as

$$D_{\mathsf{KL}}(\nu_X \,||\, \mu_X) \triangleq \sum_{x \in \mathcal{X}} \nu_X(x) \log \frac{\nu_X(x)}{\mu_X(x)}.$$

For any pair of random variables $(X, Y) \sim \mu_{XY}$, we define the *mutual information* between them as $I(X; Y) \triangleq D_{\mathsf{KL}}(\mu_{XY} \,||\, \mu_X \mu_Y)$, and the *entropy* of X as $H(X) \triangleq I(X; X)$. In particular, we require the entropy of $\{0, 1\}$-valued random variables, so for X with $\mu_X(0) = p = 1 - \mu_X(1)$, define

$$h_2(p) \triangleq H(X) = -p \log p - (1 - p) \log(1 - p). \tag{2}$$

Also of interest to us is the *maximal correlation* between X and Y, given by

$$\rho_{\mathsf{m}}(X; Y) \triangleq \sup_{(f(X), g(Y)) \in \mathcal{S}} \mathbb{E}[f(X) g(Y)],$$

where $\mathcal{S}$ is the collection of pairs of random variables $(f(X), g(Y))$ such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f(X)^2] = \mathbb{E}[g(Y)^2] = 1$.

C. Hypercontractivity parameters

Given the random variable $X \sim \mu_X$, for $p > 1$, we define the p-norm of any function f defined on X as

$$\|f(X)\|_p \triangleq \mathbb{E}[|f(X)|^p]^{\frac{1}{p}} = \left(\sum_{x \in \mathcal{X}} \mu_X(x) \, |f(x)|^p \right)^{\frac{1}{p}}.$$

It follows from the (strict) convexity of the mapping $t \mapsto t^\alpha$ for $\alpha > 1$, that the p-norm is nondecreasing in p (and is strictly increasing if there exist $x_1 \neq x_2$ with $\mu_X(x_1), \mu_X(x_2) > 0$ such that $f(x_1) \neq f(x_2)$).

Following on from the discussion in Section I, we now define some hypercontractivity parameters. We say that the pair of random variables (X, Y) is (p, q)*-hypercontractive* for $1 \leq q \leq p < \infty$ if, for every bounded function g, we have

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

or equivalently, $\|(Tg)(X)\|_p \leq \|g(Y)\|_q$, where T is the Markov operator defined in (1). For $p \geq 1$, we define

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

and $r_p(X; Y) = \frac{q_p(X;Y)}{p}$, i.e., $r_p(X; Y)$ is the smallest r such that $\|(Tg)(X)\|_p \leq \|g(Y)\|_{pr}$ for some fixed $p \geq 1$. We will also be interested in the quantity $r^*(X; Y) = \inf_{p \geq 1} r_p(X; Y)$. When the random variables involved are clear from context (or unimportant), we simply write r_p and r^*.

III. Properties of Hypercontractivity Parameters

We now study the quantities defined in Section II-C in more detail, and see that in spite of their rather abstract definition, they satisfy some useful properties, and that there is a deep connection with several seemingly unrelated quantities. Full proofs of most results are lengthy, so we do not describe them here—they can be found in the original papers.

We start with some simple observations about r_p and r^* which follow from the p-norm being increasing in p. Before doing so, we require the following definition (which was given informally in Section I). A pair of random variables (X, Y) is said to be *decomposable* if there exist sets $A \subseteq \mathcal{X}, B \subseteq \mathcal{Y}$ such that $0 < \mathbb{P}(X \in A), \mathbb{P}(Y \in B) < 1$, and $\mathbb{P}(X \in A, Y \in B) + \mathbb{P}(X \in A^c, Y \in B^c) = 1$, i.e., with probability 1, (X, Y) is either in $A \times B$ or in $A^c \times B^c$. The pair (X, Y) is *indecomposable* if it is not decomposable. Note that this is more general than the strong requirement of positivity, i.e., $\mu_{XY}(x, y) > 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Indeed, even if (X, Y) were not indecomposable, we could split (X, Y) into some number of indecomposable blocks and study them separately.

Property 1 (Monotonicity [2])**.** *The quantity r_p is nonincreasing in p, and $r_1 = 1$. This implies that $r^* = \inf_{p \geq 1} r_p = \lim_{p \to \infty} r_p$. Further, if (X, Y) is indecomposable, we also have that $r_p(X; Y)$ is strictly decreasing in p.*

Property 2 (Tensorization [2])**.** *For independent (but not necessarily identically distributed) pairs of random variables $\{(X_i, Y_i)\}_{i=1}^n$,*

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

In particular, if they all also happen to have the same distribution as (X, Y), then $r_p(X^n; Y^n) = r_p(X; Y)$.

Property 3 (Relation to maximal correlation [2]). *The parameter $r_p(X;Y)$ can be lower bounded in terms of $\rho_{\mathsf{m}}(X;Y)$, as*

$$r_p(X;Y) \geq p^{-1} + (1-p^{-1})\rho_{\mathsf{m}}(X;Y)^2.$$

Letting $p \to \infty$, we have $r^*(X;Y) \geq \rho_{\mathsf{m}}(X;Y)^2$. On the other hand, if X and Y are uncorrelated, this reduces to $r_p(X;Y) \geq p^{-1}$; when X and Y are also independent, this holds with equality, i.e., $r_p(X;Y) = p^{-1}$.

Property 4 (Appearance of information measures). *For $(X,Y) \sim \mu_{XY}$, Ahlswede and Gács [2] obtain a surprising equivalent characterization of r^* in terms of information measures, as*

$$r^*(X;Y) = \sup_{\nu_X : \substack{\nu_X \neq \mu_X, \\ \nu_X \ll \mu_X}} \frac{D_{\mathsf{KL}}(\nu_Y \,||\, \mu_Y)}{D_{\mathsf{KL}}(\nu_X \,||\, \mu_X)}, \tag{3}$$

where ν_Y is the marginal distribution induced on Y by ν_X through the same conditional distribution W. Nair [3] generalizes this to any finite $p \geq 1$, as

$$r_p(X;Y) = \\ \sup_{\nu_{XY} : \substack{\nu_{XY} \neq \mu_{XY}, \\ \nu_{XY} \ll \mu_{XY}}} \frac{D_{\mathsf{KL}}(\nu_Y \,||\, \mu_Y)}{pD_{\mathsf{KL}}(\nu_{XY} \,||\, \mu_{XY}) - (p-1)D_{\mathsf{KL}}(\nu_X \,||\, \mu_X)}.$$

Similarly, Anantharam et al. [6] show another equivalent characterization of r^*,

$$r^*(X;Y) = \sup_{U : \substack{U - X - Y, \\ I(U;X) > 0}} \frac{I(U;Y)}{I(U;X)}, \tag{4}$$

which Nair [3] extends to any finite $p \geq 1$, as

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

where $\mathcal{P}_\mu = \{\nu_{UXY} : \mathcal{U} \times \mathcal{X} \times \mathcal{Y} \to [0,1], |\mathcal{U}| \leq |\mathcal{X}||\mathcal{Y}|, \sum_{u \in \mathcal{U}} \nu_{UXY}(u,\cdot,\cdot) = \mu_{XY}\}$ is the set of distributions of (U,X,Y) such that the marginal of (X,Y) is μ_{XY}.

The above characterization of the hypercontractivity parameters is remarkable, because it shows a connection with information measures which appears seemingly out of nowhere. Also surprising is that the proofs of the characterizations for any finite $p \geq 1$ due to Nair [3] are simpler and more elegant than the proofs of (3) and (4) in the limiting case. Among other things, this characterization makes the following property almost trivial, since it follows immediately from the convexity of $D_{\mathsf{KL}}(\cdot \,||\, \cdot)$.

Property 5 (Convexity [4]). *Given a random variable X with a fixed distribution μ_X, let the channel W, generating the random variable $Y \sim \mu_Y$, be varying. Then $r^*(X;Y)$ is convex in W for a fixed μ_X.*

While the above characterizations show us some connections between information-theoretic quantities and hypercontractivity parameters, it is unclear how to compute r_p or r^* for a given (X,Y). Property 6 gives a geometric interpretation and computable characterization of r_p (and hence r^*) as the smallest λ at which some function of λ matches its lower convex envelope[3].

Property 6 (Geometric characterization). *Let $\mathsf{K}[f](x)$ denote the lower convex envelope of the function f evaluated at the point x. Anantharam et al. [6] show that for any pair of random variables (X,Y), where $X \sim \mu_X$ and $\mathsf{W}_{Y|X} = \mathsf{W}$,*

$$r^*(X;Y) = \inf\{\lambda : \mathsf{K}[t_\lambda^{\mathsf{W}}](\mu_X) = t_\lambda^{\mathsf{W}}(\mu_X)\},$$

where the function t_λ^{W} is defined on the set of distributions on X as

$$t_\lambda^{\mathsf{W}}(\mu) = H(Y) - \lambda H(X),$$

with $X \sim \mu$ and Y generated from X by the conditional distribution $\mathsf{W}_{Y|X} = \mathsf{W}$. Once again, Nair [3] generalizes this to finite $p \geq 1$, as

$$r_p(X;Y) = \inf\{\lambda : \mathsf{K}[t_{p,\lambda}^{\mathsf{W}}](\mu_{XY}) = t_{p,\lambda}^{\mathsf{W}}(\mu_{XY})\},$$

where the function $t_{p,\lambda}^{\mathsf{W}}$ is defined on the set of distributions on (X,Y) as

$$t_{p,\lambda}^{\mathsf{W}}(\mu) = H(Y) - \lambda H(X) - p\lambda H(Y \mid X).$$

Property 7 provides a neat dual relation between the maximizing distributions in the characterization (3) for binary-valued random variables, which can be used to simplify the computation of r^*.

Property 7 (A duality result [4]). *Let $(X,Y) \sim \mu_{XY}$ be a pair of binary-valued random variables. Let ν_X^* be the maximizing distribution in the characterization (3) of $r^*(X;Y)$ (this is well-defined since the supremum is over a finite space). Define the random variables (U,V), with $U \sim \nu_X^*$ and $\mathsf{W}_{V|U} = \mathsf{W}_{Y|X}$. Then, μ_X is also the maximizing distribution in (3) for $r^*(U;V)$, and $r^*(X;Y) = r^*(U;V)$. Further, the line segment connecting the curve $\mathbb{P}(X=1) \mapsto H(Y) - \lambda H(X)$ at the points $\mu_X(1)$ and $\nu_X^*(1)$ exactly matches the lower convex envelope of the same curve.*

IV. Applications to Information Theory

In this section, we make use of properties from Section III to try and answer some information-theoretic problems. It is worth noting that each of the problems considered is of a different flavour and requires different techniques to be studied, but hypercontractivity allows us to, if not solve, at least view the problems from a unified perspective.

A. Probabilities of decoding sets

We start by returning to the example discussed in Section I to motivate the study of hypercontractivity. As seen previously, Ahlswede and Gács [2] show the following result.

Theorem 1. *Given an i.i.d. sequence of random variables $\{(X_i,Y_i)\}_{i=1}^n$ such that (X_i,Y_i) is indecomposable, for any*

[3] The lower convex envelope $\mathsf{K}[f]$ of a function f is given by $\mathsf{K}[f](x) = \sup\{g(x) : g \text{ is convex and } g \leq f \text{ on the entire domain of } f\}$.

sets $A \subseteq \mathcal{X}^n, B \subseteq \mathcal{Y}^n$, there exist positive numbers $r < 1$ and p such that

$$\mathbb{P}(Y^n \in B) \geq \mathsf{W}(B \mid A)^p \, \mathbb{P}(X^n \in A)^r.$$

Proof: Fix some $p > 1$, and let p' be its Hölder conjugate. Let T^n be the Markov operator associated with the channel W (actually W^n), given by

$$(T^n g)(x^n) = \sum_{y^n \in \mathcal{Y}^n} g(y^n) \mathsf{W}(y^n \mid x^n),$$

for any bounded function g. By Hölder's inequality, we have

$$\begin{aligned}
\mathbb{P}(X^n \in A, Y^n \in B) &= \sum_{x^n \in \mathcal{X}^n} \mu_{X^n}(x^n) \mathbb{1}_A(x^n) \mathsf{W}(B \mid x^n) \\
&= \mathbb{E}[|\mathbb{1}_A(X^n)(T^n \, \mathbb{1}_B)(X^n)|] \\
&\leq \mathbb{E}[\mathbb{1}_A(X^n)^{p'}]^{\frac{1}{p'}} \, \mathbb{E}[(T^n \, \mathbb{1}_B)(X^n)^p]^{\frac{1}{p}} \\
&= \mathbb{P}(X^n \in A)^{\frac{1}{p'}} \, \|(T^n \, \mathbb{1}_B)(X^n)\|_p \\
&\leq \mathbb{P}(X^n \in A)^{\frac{1}{p'}} \, \|\mathbb{1}_B(X^n)\|_{pr_p} \\
&= \mathbb{P}(X^n \in A)^{1-\frac{1}{p}} \, \mathbb{P}(Y^n \in B)^{\frac{1}{pr_p}},
\end{aligned}$$

where we write $r_p = r_p(X_i; Y_i) = r_p(X^n; Y^n)$ by the tensorization property. Rearranging the above, we have

$$\begin{aligned}
\mathbb{P}(Y^n \in B) &\geq \mathsf{W}(B \mid A)^{pr_p} \, \mathbb{P}(X^n \in A)^{r_p} \\
&\geq \mathsf{W}(B \mid A)^p \, \mathbb{P}(X^n \in A)^{r_p}.
\end{aligned}$$

Since (X_i, Y_i) is indecomposable, we have that r_p is strictly decreasing in p. Together with $r_1 = 1$, we have that $r_p < 1$ for any $p > 1$, which completes the proof. ■

This result has a neat information-theoretic interpretation. An idea that goes all the way back to Shannon [7] is to think of the occurrence of an event with probability p as giving us $-\log p$ units of "information". Suppose we consider sets A, B such that $\mathsf{W}(B \mid A) \geq \lambda$, some constant independent of n (as is the case when B is a decoding set of A). If we know that $X^n \in A$, i.e., our "information" about X^n is $-\log \mathbb{P}(X^n \in A)$, then our "information" about Y^n is nearly $-r_p \log \mathbb{P}(X^n \in A)$, i.e., a constant $r_p < 1$ times lesser (this interpretation is "more correct" as $n \to \infty$, with the assumption that $\mathbb{P}(Y^n \in B)$ goes to zero exponentially in n).

B. Mutual information between Boolean functions

The data processing inequality [8] in information theory states that for any random variables $X \leftrightarrow Y \leftrightarrow Z$ forming a Markov chain, $I(X; Z) \leq I(X; Y)$, i.e., further processing can only decrease the statistical dependence as measured by the mutual information. This seemingly simple observation is fundamental in information theory and has been used to prove several classical converse results. As a direct consequence of the characterization in (4), it is easy to see that the hypercontractivity parameter $r^*(X; Y)$ also satisfies a data processing inequality, i.e., for $X \leftrightarrow Y \leftrightarrow Z$, $r^*(X; Z) \leq r^*(X; Y)$.

Additionally, when X and Y are independent, $r^*(X; Y) = \lim_{p \to \infty} r_p = \lim_{p \to \infty} \frac{1}{p} = 0$, just like $I(X; Y)$. These similarities seem to suggest that r^* (or some appropriate monotone function of it) can be used as a proxy to measure the dependence between random variables, just like mutual information or maximal correlation (which also satisfies a data processing inequality). Indeed, another connection comes from Property 5, which tells us that $r^*(X; Y)$ is convex in W for a fixed μ_X; this is also the case with $I(X; Y)$.

We now look at an application with this idea in mind. A DSBS(α) (doubly symmetric binary source with parameter α) is a pair of random variables (X, Y) such that X and Y are both i.i.d. Bernoulli$\left(\frac{1}{2}\right)$, and $\mathbb{P}(X = 0 \mid Y = 1) = \mathbb{P}(X = 1 \mid Y = 0) = \frac{\alpha}{2}$. Kumar and Courtade [5] pose the following obvious-looking conjecture which (in their words) is "the simplest, nontrivial embodiment of Boolean functions in an information-theoretic context". Given the sequence $\{(X_i, Y_i)\}_{i=1}^n$ drawn i.i.d. from a DSBS(α), for any Boolean function $b : \{0,1\}^n \to \{0,1\}$, the conjecture is that $I(b(X^n); Y^n) \leq I(X_i; Y^n) = 1 - h_2(\alpha)$, for any i in $1, \ldots, n$. While this seems trivial at first glance, a proof is not known yet. Even a numerical proof is not possible, since n is arbitrary.

We can make some (marginal) progress using hypercontractivity. From the characterization in (4), we have that $I(U; Y) \leq r^*(X; Y) \, I(U; X)$ for any $U - X - Y$ forming a Markov Chain. Ahlswede and Gács [2] calculate that for a DSBS(α), $r^*(X_i; Y_i) = (1 - 2\alpha)^2$, so by the tensorization property, $r^*(X^n; Y^n) = (1 - 2\alpha)^2$. Let b be any Boolean function, then taking $U = b(X^n)$, we have

$$\begin{aligned}
I(b(X^n); Y^n) &\leq r^*(X^n; Y^n) \, I(b(X^n); X^n) \\
&= (1 - 2\alpha)^2 \, H(b(X^n)) \leq (1 - 2\alpha)^2.
\end{aligned}$$

For all $\alpha \in [0, 1]$, $(1 - 2\alpha^2) \geq 1 - h_2(\alpha)$, but equality holds at $\alpha = 0, 1, \frac{1}{2}$, which seems to suggest some kind of local tightness in the "low-noise" or "high-noise" regimes, but even this fails since the first derivatives are unequal at $\alpha = 0, 1$ and the second derivatives are unequal at $\alpha = \frac{1}{2}$ (both of their first derivatives are zero at $\alpha = \frac{1}{2}$).

Kumar and Courtade [5] also propose a weaker form of the conjecture, namely that $I(b_1(X^n); b_2(Y^n)) \leq 1 - h_2(\alpha)$ for all Boolean functions b_1, b_2. Instead, Anantharam et al. [4] propose the following stronger form of this second conjecture, in terms of the hypercontractivity parameter r^*.

Conjecture 1. *For any pair of binary-valued random variables (W, Z),*

$$I(W; Z) \leq 1 - h_2\left(\frac{1 - \sqrt{r^*(W; Z)}}{2}\right) \tag{5}$$

Supposing for a moment that this conjecture is true, we have, for $\{(X_i, Y_i)\}_{i=1}^n$ i.i.d. as a DSBS(α),

$$\begin{aligned}
I(b_1(X^n); b_2(Y^n)) &\leq 1 - h_2\left(\frac{1 - \sqrt{r^*(b_1(X^n); b_2(Y^n))}}{2}\right) \\
&\leq 1 - h_2\left(\frac{1 - \sqrt{r^*(X^n; Y^n)}}{2}\right) \\
&= 1 - h_2\left(\frac{1 - \sqrt{r^*(X_i; Y_i)}}{2}\right) \\
&= 1 - h_2\left(\frac{1 - \sqrt{(1 - 2\alpha)^2}}{2}\right)
\end{aligned}$$

$$= 1 - h_2(\alpha).$$

In essence, rather than simply considering r^* (which gives the loose constant $(1-2\alpha)^2$ instead of $1-h_2(\alpha)$) as the mutual information proxy, we look to use $f(r^*) = 1 - h_2\left(\frac{1-\sqrt{r^*}}{2}\right)$, which is monotonically increasing and satisfies $f(0) = 0$, $f(1) = 1$.

Anantharam et al. [4] analytically prove Conjecture 1 only under some conditions on the distribution of (W, Z), which includes, in particular, when the channel between them is symmetric, i.e., $\mathbb{P}(Z = 1 \mid W = 0) = \mathbb{P}(Z = 0 \mid W = 1)$. Note, however, that this is not enough to prove the weaker form of the Kumar-Courtade conjecture [5], since the functions b_1 and b_2 could be arbitrary, destroying the symmetry of $(b_1(X^n), b_2(Y^n))$.

Nonetheless, Conjecture 1 can be verified numerically, since unlike the Kumar-Courtade conjecture, there is no arbitrary n to deal with. This gives further reason to believe that the weaker conjecture involving two Boolean functions is true. Indeed, a proof for this weaker conjecture has been discovered since, using Fourier-analytic techniques [9].

C. Strictness of strong data processing inequalities

The data processing inequality can be stated in more generality than that in Section IV-B: given any fixed channel W from X to Y, for any distributions $\mu_X \neq \nu_X$ on X, $D_{\mathsf{KL}}(\nu_Y \,||\, \mu_Y) \leq D_{\mathsf{KL}}(\nu_X \,||\, \mu_X)$. In many cases, however, a stronger result holds, namely that there exists some constant $\eta_{\mathsf{KL}}(\mathsf{W}) < 1$ such that $D_{\mathsf{KL}}(\nu_Y \,||\, \mu_Y) \leq \eta_{\mathsf{KL}}(\mathsf{W}) D_{\mathsf{KL}}(\nu_X \,||\, \mu_X)$, where the smallest such $\eta_{\mathsf{KL}}(\mathsf{W})$ is called the contraction coefficient [10] of the channel W; such a result is called a *strong data processing inequality*. We may also consider input-dependent contraction coefficients by fixing some μ_X together with the channel W, and letting $\eta_{\mathsf{KL}}(\mu_X, \mathsf{W}) < 1$ be the smallest value such that

$$D_{\mathsf{KL}}(\nu_Y \,||\, \mu_Y) \leq \eta_{\mathsf{KL}}(\mu_X, \mathsf{W}) D_{\mathsf{KL}}(\nu_X \,||\, \mu_X)$$

for all $\nu_X \neq \mu_X$. One notices immediately from the characterization in (3) that $\eta_{\mathsf{KL}}(\mu_X, \mathsf{W}) = r^*(X; Y)$. A natural question to ask is whether this inequality holds with equality for some ν_X, i.e., whether the coefficient $\eta_{\mathsf{KL}}(\mu_X, \mathsf{W})$ is achievable or is only a limiting value. Ahlswede and Gács [2] provide an exact characterization of the distributions for which equality holds.

Theorem 2. *The supremum in (3) is attained by some distribution ν_X if and only if $r^*(X;Y) > \rho_{\mathsf{m}}^2(X;Y)$, or equivalently (since $r^*(X;Y) \geq \rho_{\mathsf{m}}^2(X;Y)$ in general), $r^*(X;Y) = \rho_{\mathsf{m}}^2(X;Y)$ if and only if*

$$\eta_{\mathsf{KL}}(\mu_X, \mathsf{W}) = r^*(X;Y) > \frac{D_{\mathsf{KL}}(\nu_Y \,||\, \nu_Y)}{D_{\mathsf{KL}}(\nu_X \,||\, \nu_X)}$$

for all $\nu_X \ll \mu_X$.

D. Connections to learning theory

The Information Bottleneck (IB) problem was introduced by Tishby et al. [11] as an information-theoretic approach to learning. In particular, we consider a pair of correlated random variables $(X, Y) \sim \mu_{XY}$, where Y is some target variable that we wish to study, and X is an observation that depends on Y. The goal is to find a representation $T(X)$ (that may be randomized) which characterizes the trade-off between T depending too much on X itself and not having enough dependence on Y. Formally, the problem is

$$\inf_{\substack{T: \, T-X-Y \\ I(Y;T) \geq \alpha}} I(X;T) \equiv \sup_{\substack{T: \, T-X-Y \\ I(X;T) \leq \alpha}} I(Y;T),$$

and the relation with hypercontractivity is apparent, since the optimization in the characterization (4) of r^* has a similar structure. This optimization problem was studied several years before it was proposed in [11], by Witsenhausen and Wyner [12], who obtained a geometric solution involving the lower convex envelopes of certain curves, similar to that of r^* and r_p in Property 6.

V. Future Plans and Conclusion

While there is undoubtedly a clear connection between hypercontractivity and information theory, as evidenced by various useful characterizations and interpretations of hypercontractivity parameters in terms of information-theoretic quantities, there have been only a small number of applications of hypercontractivity with conclusive results. One major bottleneck is the lack of a closed form expression for r_p or r^*, making it difficult to actually put the properties to use. As a result, we are constrained to rely exclusively on manipulations that give us equivalent characterizations. Conversely, this also means that making some progress on obtaining closed form expressions for these parameters would almost simultaneously provide answers to several seemingly unrelated questions that are unified through the lens of hypercontractivity.

Thus, we are confident that hypercontractivity has the potential to shed light on several areas of information theory, particularly with the increasing interest from the rapidly growing learning community. Even in cases where hypercontractivity may not directly give us results, as in the attempts at the Kumar-Courtade conjecture in Section IV-B, it is worth noting that a solution was given by Fourier analysis, which has benefited from decades of study using hypercontractive techniques. Similarly, in a less technical sense, we believe that hypercontractivity can (and should) be used as an effective tool to supplement the study of information theory, and the cross-pollination of such evidently related mathematical areas is sure to bear fruit in due time.

References

[1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Cambridge University Press, 2011.

[2] R. Ahlswede and P. Gacs, “Spreading of sets in product spaces and hypercontraction of the Markov operator,” *The Annals of Probability*, 1976.

[3] C. Nair, “Equivalent formulations of hypercontractivity using information measures,” in *Proc. International Zurich Seminar on Communications*, 2014.

[4] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, “On hypercontractivity and the mutual information between boolean functions,” in *Proc. Allerton Conference on Communication, Control, and Computing*, 2013.

[5] T. A. Courtade and G. R. Kumar, "Which boolean functions maximize mutual information on noisy inputs?" *IEEE Transactions on Information Theory*, 2014.

[6] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *CoRR*, 2013.

[7] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, 1948.

[8] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[9] G. Pichler, G. Matz, and P. Piantanida, "A tight upper bound on the mutual information of two boolean functions," in *2016 IEEE Information Theory Workshop (ITW)*, 2016.

[10] A. Makur, "Information contraction and decomposition," Ph.D. dissertation, Massachusetts Institute of Technology, 2019.

[11] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Allerton Conference on Communication, Control and Computing*, 1999.

[12] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Transactions on Information Theory*, 1975.