Brascamp-Lieb Inequality and Its Reverse: An Information Theoretic View

Jingbo Liu*, Thomas A. Courtade†, Paul Cuff* and Sergio Verdú*

*Department of Electrical Engineering, Princeton University
†Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
Email: {jingbo,cuff,verdu}@princeton.edu, courtade@eecs.berkeley.edu

***Abstract*—We generalize a result by Carlen and Cordero-Erausquin on the equivalence between the Brascamp-Lieb inequality and the subadditivity of relative entropy by allowing for random transformations (a broadcast channel). This leads to a unified perspective on several functional inequalities that have been gaining popularity in the context of proving impossibility results. We demonstrate that the information theoretic dual of the Brascamp-Lieb inequality is a convenient setting for proving properties such as data processing, tensorization, convexity and Gaussian optimality. Consequences of the latter include an extension of the Brascamp-Lieb inequality allowing for Gaussian random transformations, the determination of the multivariate Wyner common information for Gaussian sources, and a multivariate version of Nelson's hypercontractivity theorem. Finally we present an information theoretic characterization of a reverse Brascamp-Lieb inequality involving a random transformation (a multiple access channel).**

I. Introduction

The Brascamp-Lieb (BL) inequality [1][2][3] in functional analysis concerns the optimality of Gaussian functions in a certain class of integral inequalities. To be concrete, consider an inequality of the following general form, which we shall call a *Brascamp-Lieb like* (BLL) inequality:

$$\mathbb{E}\left[\prod_{j=1}^{m} f_j(B_j(X))\right] \le D \prod_{j=1}^{m} \|f_j\|_{\frac{1}{c_j}}, \tag{1}$$

where the expectation is with respect to $X \sim Q$, for each $j \in \{1, \dots, m\}$, $c_j \in (0, \infty)$, $B_j \colon \mathcal{X} \to \mathcal{Y}_j$ is measurable, $f_j \colon \mathcal{Y}_j \to \mathbb{R}$ is nonnegative measurable, and

$$\|f_j\|_{\frac{1}{c_j}} := \left(\int f_j^{1/c_j} \mathrm{d}R_{Y_j}\right)^{c_j}$$

for some R_{Y_j}. Conventionally, the BL inequality is the Gaussian case of (1) where Q is Gaussian (or the Lebesgue measure, with the expectations replaced by integrals) and $(B_j)_{j=1}^m$ are linear projections. In this setting, Brascamp and Lieb [1] showed that (1) holds if and only if it holds for all centered Gaussian functions $(f_j)_{j=1}^m$. Generalizing a result in [1], Lieb [2] extended the validity of the result to arbitrary surjective linear maps (B_j). Lieb's proof used a rotational invariance property of Gaussian random variables. Given the fundamental nature of Lieb's result and its far-reaching consequences, alternative proof methods have attracted wide interest; see [3, Remark 1.10] for the history and references.

Motivated by the quest for an alternative proof of Lieb's result using the superadditivity of Fisher information, Carlen and Cordero-Erausquin proved a duality between the BLL inequality in (1) and a super-additivity of relative entropy. We remind the reader that the BLL inequality is more general than the setting initially considered by Brascamp and Lieb [1] since the random variables need not be Gaussian and (B_j) need not be linear.

In this paper, we extend the duality result of Carlen and Cordero-Erausquin, by allowing (B_j) to be non-deterministic random transformations (that is, there is a broadcast channel). This is motivated by two considerations. First, random transformations are natural and essential for many information theoretic applications; see for example [4]. Second, the result subsumes the equivalent formulation of the strong data processing inequality (in addition to that of hypercontractivity and Loomis-Whitney inequality/Shearer's lemma, which is already contained in Carlen and Cordero-Erausquin's result). These inequalities have recently attracted significant attention in information theory [5][6][7][8], theoretical computer science [9][10] and statistics [11][12]. Previous proofs of their equivalent formulations have been discovered independently and sometimes rely on the finiteness of the alphabet. In contrast, the present approach is based on the nonnegativity of relative entropy (which corresponds to the Donsker-Varadhan formula used by Carlen and Cordero-Erausquin) and holds for general alphabets.

In the same vein as Brascamp and Lieb's original result that Gaussian kernels have Gaussian maximizers, we establish Gaussian optimality in several information theoretic optimization problems related to the dual form of the BL inequality. Roughly speaking, our approach is based on the fact that two independent random variables are both Gaussian if their sum is independent of their difference, a fact which was also used in establishing Gaussian extremality in information theory by Geng-Nair [13] [14]. It is worth noting that similar techniques have appeared previously in the literature on the Brascamp-Lieb inequality: Lieb used a rotational invariance argument in [2], and it was observed that convolution preserves the extremizers of Brascamp-Lieb inequality [15, Lemma 2]. However, as keenly noted in [13], working with the information theoretic counterparts offers certain advantages, partly because of the similarity between the proof techniques with certain converses in data transmission. Implications of Gaussian optimality are discussed in Section IV-B.

Finally, we provide an information theoretic formulation of Barthe's reverse Brascamp-Lieb inequality (RBL) [16]. In fact, we shall consider a generalization of RBL involving a multiple access channel (MAC) - pairing nicely with the broadcast channel in the forward inequality. From this formulation it is seen that the mysterious "sup" operation in RBL disappears when the MAC is a bijective mapping, which is the special case of reverse hypercontractivity considered by Kamath [17]. Moreover, the strong data processing inequality is also a

This work was supported in part by NSF Grants CCF-1528132, CCF-0939370 (Center for Science of Information), CCF-1319299, CCF-1319304, CCF-1350595 and AFOSR FA9550-15-1-0180.

978-1-5090-1805-5/16/$31.00 ©2016 IEEE

special case of generalized RBL when the MAC is a point to point channel. Another direction of generalizing the reverse hypercontractivity, where the stochastic map is still a bijective mapping but the coefficients (c_j) are allowed to be negative, has been recently considered by Beigi and Nair (see [18]). To our knowledge, their result does not imply ours, or vice versa.

Omitted proofs are given in [19].

II. A General Duality Result

Given two probability measures $P \ll Q$ on $\mathcal{X}$, define the *relative information* as the log Radon-Nikodym derivative

$$\imath_{P\|Q}(x) := \log \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \tag{2}$$

The relative entropy is defined as

$$D(P\|Q) := \mathbb{E}\left[\imath_{P\|Q}(X)\right] \tag{3}$$

where $X \sim P$, if $P \ll Q$, and infinity otherwise.

Theorem 1. *Fix Q_X, positive integer m, $d \in \mathbb{R}$, and $Q_{Y_j|X}$, measure R_{Y_j} on $\mathcal{Y}_j$, $c_j \in (0,\infty)$ for each $j \in \{1,\dots,m\}$. Let $(X, Y_j) \sim Q_X Q_{Y_j|X}$. Then the following statements are equivalent:*

1) *For any non-negative measurable functions $f_j \colon \mathcal{Y}_j \to \mathbb{R}$,*

$$\mathbb{E}\left[\exp\left(\sum_{j=1}^{m} \mathbb{E}[\log f_j(Y_j)|X] - d\right)\right] \le \prod_{j=1}^{m} \|f_j\|_{\frac{1}{c_j}}. \tag{4}$$

2) *For $P_X \ll Q_X$ and $P_X \to Q_{Y_j|X} \to P_{Y_j}$,*

$$D(P_X\|Q_X) + d \ge \sum_{j=1}^{m} c_j D(P_{Y_j}\|R_{Y_j}). \tag{5}$$

The special case where $Q_{Y_j|X}$, $j \in \{1,\dots,m\}$ are deterministic is established in [20, Theorem 2.1]. We refer to (4) as a *generalized Brascamp-Lieb like* (GBLL) inequality.

Proof Sketch: The key idea of the proof is to define certain auxiliary distributions. Later we will reveal a nice symmetry with the auxiliary distributions used in the proof of the reverse inequality.

1)⇒2) Define an auxiliary measure S_X via

$$\imath_{S_X\|Q_X}(x) := -d_0 + \sum_{j=1}^{m} c_j \mathbb{E}[\imath_{P_{Y_j}\|R_{Y_j}}(Y_j)|X = x]$$

where d_0 is a normalization constant, and $f_j \leftarrow \left(\frac{\mathrm{d}P_{Y_j}}{\mathrm{d}R_{Y_j}}\right)^{c_j}$. Then 2) follows from 1) and the nonnegativity of $D(P_X\|S_X)$.

2)⇒1) Define P_X and S_{Y_j} through

$$\imath_{P_X\|Q_X}(x) = -d - d_0 + \mathbb{E}\left[\sum_{j=1}^{m} \log f_j(Y_j) \,\middle|\, X = x\right] \tag{6}$$

$$\imath_{S_{Y_j}\|R_{Y_j}}(y_j) := \frac{1}{c_j}\log f_j(y_j) - d_j, \tag{7}$$

where d_j's are normalization constants. Then 2) follows from 1) and the nonnegativity of $D(P_{Y_j}\|S_{Y_j})$. ■

III. Notable Special Cases of Theorem 1

Not only does Theorem 1 admit a very simple proof but it unifies the equivalent formulations of several functional inequalities and information theoretic inequalities, and the approach applies to general alphabets, in contrast to some previous methods requiring finite alphabets (cf. [21], [22]).

A. Variational Formula for Rényi Divergence

Suppose R, Q and T are probability measures on $(\mathcal{X}, \mathscr{F})$, $R, Q \ll T$, $\alpha \in (0,1) \cup (1,\infty)$. Define $D_\alpha(Q\|R)$ as

$$\frac{1}{\alpha-1}\log\left(\mathbb{E}\left[\exp\left(\alpha \imath_{Q\|T}(\bar{X}) + (1-\alpha)\imath_{R\|T}(\bar{X})\right)\right]\right) \tag{8}$$

where $\bar{X} \sim T$. [23] showed that (8) equals the supremum of

$$\frac{\alpha}{\alpha-1}\log \mathbb{E}[\exp((\alpha-1)g(\hat{X}))] - \log \mathbb{E}[\exp(\alpha g(X))] \tag{9}$$

over bounded nonnegative measurable g such that (9) is well-defined, where $X \sim R$ and $\hat{X} \sim Q$. For $\alpha \in (1,\infty)$, by setting $\exp(g(\cdot))$ as an indicator function of a set, one recovers the logarithmic probability comparison bound (LPCB) [24], which is useful in the error exponent analysis.

Now, a simple proof of (9) for $\alpha \in (1,\infty)$ can be obtained from Theorem 1 by setting $m \leftarrow 1$, $Y_1 = X$, $c \leftarrow \frac{\alpha-1}{\alpha}$ and $d = \frac{\alpha-1}{\alpha} D(P\|R)$. Note that (5) is then reduced to

$$D(P\|Q) + \frac{\alpha-1}{\alpha} D_\alpha(Q\|R) \ge \frac{\alpha-1}{\alpha} D(P\|R) \tag{10}$$

which is well-known and can be easily shown using the nonnegativity of relative entropy. Meanwhile, setting $f \leftarrow \exp((\alpha-1)g)$ in (4), we see (9) is less than or equal to $D_\alpha(Q\|R)$. The equality is achieved when

$$\frac{\mathrm{d}Q}{\mathrm{d}R}(x) = \frac{\exp(\alpha g(x))}{\mathbb{E}[\exp(\alpha g(X))]}. \tag{11}$$

B. Strong Data Processing Constant

A strong data processing inequality (SDPI) is an inequality of the form

$$D(P_X\|Q_X) \ge c D(P_Y\|Q_Y), \quad \text{for all } P_X \ll Q_X \tag{12}$$

where $P_X \to Q_{Y|X} \to P_Y$, and we have fixed $Q_{XY} = Q_X Q_{Y|X}$ [21][25][26]. The conventional data processing inequality corresponds to $c = 1$, so SDPI's generally specify $c > 1$. The study of the largest constant c for (12) to hold can be traced back to Ahlswede and Gács [21], who showed its equivalence to

$$\mathbb{E}[\exp(\mathbb{E}[\log f(Y)|X])] \le \|f\|_{\frac{1}{c}}, \quad \text{for all } f \ge 0. \tag{13}$$

The proof in [21, Theorem 5], which is based on a connection between SDPI and hypercontractivity, relies heavily on the finiteness of the alphabet, and is quite technical even in that case. From Theorem 1, however, it is straightforward to check that such equivalence holds for general alphabets.

C. Hypercontractivity

The BLL inequality also encompasses

$$\mathbb{E}\left[f_1(Y_1) f_2(Y_2)\right] \le \|f_1\|_{p_1}\|f_2\|_{p_2} \tag{14}$$

where $p_1, p_2 \in [0,\infty)$. Using the method of types/typicality, it is shown in [22] that (14) is equivalent to

$$D(P_{Y_1Y_2}\|Q_{Y_1Y_2}) \ge \tfrac{1}{p_1} D(P_{Y_1}\|Q_{Y_1}) + \tfrac{1}{p_2} D(P_{Y_2}\|Q_{Y_2}) \tag{15}$$

for all $P_{Y_1Y_2} \ll Q_{Y_1Y_2}$ in the case of finite alphabets. The proof of Theorem 1 based on nonnegativity of relative entropy establishes this equivalence for general alphabets, and in particular allows one to prove Nelson's inequality for Gaussian hypercontractivity from (15); see the end of Section IV-B.

D. Loomis-Whitney Inequality and Shearer's Lemma

The combinatorial Loomis-Whitney inequality [27, Theorem 2] can be recovered from the following integral inequality: let μ be the counting measure on $\mathcal{A}^m$, then

$$\int_{\mathcal{A}^m} \prod_{j=1}^{m} f_j(\pi_j(x))\mathrm{d}\mu(x) \le \prod_{j=1}^{m} \|f_j\|_{m-1} \tag{16}$$

for all nonnegative f_j's, where where π_j is the projection operator deleting the j-th coordinate and the norm on the right is with respect to the counting measure on $\mathcal{A}^{m-1}$. This is an extension of the BLL inequality to counting measures[1], and by Theorem 1 is equivalent to the entropy inequality known as Shearer's Lemma [28]

$$H(X_1,\dots,X_m) \le \sum_{j=1}^{m} \frac{1}{m-1} H(X_1^{j-1}, X_{j+1}^{m}). \tag{17}$$

IV. Applications of the Information Theoretic Formulation

A. Data Processing, Tensorization and Convexity

The information theoretic formulation in Theorem 1 leads to simple proofs of basic and important properties of BLL inequalities. Assuming $R_{Y_j} = Q_{Y_j}$ for simplicity, one has [19]:

- *Data processing*: if $(Q_X, (Q_{Y_j|X}), d, c^m)$ is such that (5) holds, then by data processing for the relative entropy, $(Q_X, (Q_{Z_j|X}), d, c^m)$ also holds for any $(Q_{Z_j|Y_j})$, where $Q_{Y_j|X} \to Q_{Z_j|Y_j} \to Q_{Z_j|X}$. A similar property holds for processing the input.
- *Tensorization*: if $(Q_X^i, (Q_{Y_j|X}^i), d^i, c^m)$, $i = 1, 2$ satisfies (5), then $(Q_X^1 \times Q_X^2, (Q_{Y_j|X}^1 \times Q_{Y_j|X}^2), d^1 + d^2, c^m)$ satisfies (5). The proof is similar to standard converse proofs in information theory.
- *Convexity*: if $(Q_X^i, (Q_{Y_j|X}^i), d^i, (c^i{}_j))$ $i = 1, 2$ satisfies (5), then for any $\theta \in (0, 1)$, $(Q_X^\theta, (Q_{Y_j|X}^\theta), d^\theta, (c^\theta{}_j))$ also satisfies (5) where

  $$d^\theta := (1-\theta)d^0 + \theta d^1, \tag{18}$$

  $$c_j^\theta := (1-\theta)c_j^0 + \theta c_j^1, \quad \forall j \in \{1,\dots,m\}. \tag{19}$$

  This is equivalent to the Riesz-Thorin interpolation theorem in functional analysis in the case of non-negative kernels. The proof in the information theoretic setting is much simpler because the c_j's only affect the right side of (5) as linear coefficients.

B. Gaussian Optimality

A less direct application is found in establishing Gaussian optimality for several information theoretic inequalities related to the BL inequalities. Toward this end, assume for the remainder of the section that $\mathcal{X}, \mathcal{Y}_1, \dots, \mathcal{Y}_m$ are Euclidean spaces of dimensions $n, n_1, \dots, n_m$, respectively, and $Q_{\mathbf{X}}$ and $(Q_{\mathbf{Y}_j|\mathbf{X}})$ are Gaussian. To be precise about the notions of Gaussian optimality, we adopt terminology from [3]:

- *Extremisability*: sup/inf is finitely attained.
- *Gaussian extremisability*: sup/inf is finitely attained by Gaussian functions/distributions.
- *Gaussian exhaustibility*: the value of sup/inf does not change when the arguments are restricted to Gaussian functions/distributions.

Most of the time, we prove Gaussian exhaustibility in general, while the more restrictive property of Gaussian extremisability is shown imposing non-degeneracy assumptions.

Fix $\mathbf{M} \succeq 0$, positive constants c_j, and Gaussian random transformations $Q_{\mathbf{Y}_j|\mathbf{X}}$ for $j \in \{1,\dots,m\}$. Define

$$F(P_{\mathbf{X}U}) := -h(\mathbf{X}|U) + \sum_{j=1}^{m} c_j h(\mathbf{Y}_j|U) + c_0 \operatorname{Tr}[\mathbf{M}\mathbf{\Sigma}_{\mathbf{X}|U}], \tag{20}$$

where $\mathbf{X} \sim P_{\mathbf{X}}$, $P_{\mathbf{X}} \to Q_{\mathbf{Y}_j|\mathbf{X}} \to P_{\mathbf{Y}_j}$ and $\mathbf{\Sigma}_{\mathbf{X}|U} := \mathbb{E}[\operatorname{Cov}(\mathbf{X}|U)]$.

Definition 1. *We say $(Q_{\mathbf{Y}_1|\mathbf{X}}, \dots, Q_{\mathbf{Y}_m|\mathbf{X}})$ is* non-degenerate *if each $Q_{\mathbf{Y}_j|\mathbf{X}=\mathbf{0}}$ is a n_j-dimensional Gaussian distribution with invertible covariance matrix.*

We have an *extremisability* result under a covariance constraint:

Theorem 2. *If $(Q_{\mathbf{Y}_1|\mathbf{X}}, \dots, Q_{\mathbf{Y}_m|\mathbf{X}})$ are non-degenerate, then $\inf_{P_{\mathbf{X}U}}\{F(P_{\mathbf{X}U}) \colon \mathbf{\Sigma}_{\mathbf{X}|U} \preceq \mathbf{\Sigma}\}$ is finite and is attained by a Gaussian $\mathbf{X}$ and constant U.*

Proof Sketch: Assume that both $P_{\mathbf{X}^{(1)}U^{(1)}}$ and $P_{\mathbf{X}^{(2)}U^{(2)}}$ are minimizers of (20) subject to $\mathbf{\Sigma}_{\mathbf{X}|U} \preceq \mathbf{\Sigma}$ (see [19] for the proof of the existence of minimizer). Let

$$(U^{(1)}, \mathbf{X}^{(1)}, \mathbf{Y}_1^{(1)}, \dots, \mathbf{X}_m^{(1)}) \sim P_{\mathbf{X}^{(1)}U^{(1)}} Q_{\mathbf{Y}_1|\mathbf{X}} \cdots Q_{\mathbf{Y}_m|\mathbf{X}}$$

$$(U^{(2)}, \mathbf{X}^{(2)}, \mathbf{Y}_1^{(2)}, \dots, \mathbf{X}_m^{(2)}) \sim P_{\mathbf{X}^{(2)}U^{(2)}} Q_{\mathbf{Y}_1|\mathbf{X}} \cdots Q_{\mathbf{Y}_m|\mathbf{X}}$$

be mutually independent. Define $\mathbf{X}^{\pm} = \frac{1}{\sqrt{2}}\left(\mathbf{X}^{(1)} \pm \mathbf{X}^{(2)}\right)$. Define $\mathbf{Y}_j^+$ and $\mathbf{Y}_j^-$ similarly for $j = 1, \dots, m$, and put $\hat{U} = (U^{(1)}, U^{(2)})$. We now observe that

1) Due to the Gaussian nature of $Q_{\mathbf{Y}_j|\mathbf{X}}$, $\mathbf{Y}_j^+ | \{\mathbf{X}^+ = \mathbf{x}^+, \mathbf{X}^- = \mathbf{x}^-, \hat{U} = u\} \sim Q_{\mathbf{Y}_j|\mathbf{X}=\mathbf{x}^+}$ is independent of $\mathbf{x}^-$. Thus $\mathbf{Y}_j^+ | \{\mathbf{X}^+ = \mathbf{x}, \hat{U} = u\} \sim Q_{\mathbf{Y}_j|\mathbf{X}=\mathbf{x}}$ as well. Similarly, $\mathbf{Y}_j^- | \{\mathbf{X}^- = \mathbf{x}, \hat{U} = u\} \sim Q_{\mathbf{Y}_j|\mathbf{X}=\mathbf{x}}$.
2) $\mathbf{\Sigma}_{\mathbf{X}^+|\hat{U}}, \mathbf{\Sigma}_{\mathbf{X}^-|\mathbf{X}^+\hat{U}} \preceq \mathbf{\Sigma}_{\mathbf{X}^-|\hat{U}}$ so both $P_{\mathbf{X}^+,\hat{U}}$ and $P_{\mathbf{X}^-,\hat{U}\mathbf{X}^+}$ satisfy the covariance constraint.

Using steps similar to the conventional converse proofs in information theory, one can show that

$$\sum_{i=1}^{2} F(P_{\mathbf{X}^{(i)}U^{(i)}}) \ge F(P_{\mathbf{X}^+,\hat{U}}) + F(P_{\mathbf{X}^-,\hat{U}\mathbf{X}^+}). \tag{21}$$

But both $P_{\mathbf{X}^+,\hat{U}}$ and $P_{\mathbf{X}^-,\hat{U}\mathbf{X}^+}$ are candidate optimizers of (20) subject to the given covariance constraint whereas $P_{U^{(i)}\mathbf{X}^{(i)}}$ are the optimizers by assumption ($i = 1, 2$), so

$$\max_i F(P_{\mathbf{X}^{(i)}U^{(i)}}) \le \min\{F(P_{\mathbf{X}^+,\hat{U}}), F(P_{\mathbf{X}^-,\hat{U}\mathbf{X}^+})\}, \tag{22}$$

which combined with (21) implies that $F(\cdot)$ has the same value at $P_{\mathbf{X}^{(1)}U^{(1)}}$, $P_{\mathbf{X}^{(2)}U^{(2)}}$, $P_{\mathbf{X}^+,\hat{U}}$ and $P_{\mathbf{X}^-,\hat{U}\mathbf{X}^+}$. We now need the following basic observation to conclude that *each term* in the linear combination in the definition of $F(\cdot)$ is also equal under those four distributions.

Lemma 3. *Let p and q be real-valued functions on an arbitrary set $\mathcal{D}$. If $f(t) := \min_{x\in\mathcal{D}}\{p(x) + tq(x)\}$ is attained for all $t \in (0, \infty)$, then for almost all t, $f'(t)$ exists and equals*

$$q(x^\star) \quad \forall x^\star \in \arg\min_{x\in\mathcal{D}}\{p(x) + tq(x)\}. \tag{23}$$

In particular, for all such t, $q(x^\star) = q(\tilde{x}^\star)$ and $p(x^\star) = p(\tilde{x}^\star)$ for all $x^\star, \tilde{x}^\star \in \arg\min_{x\in\mathcal{D}}\{p(x) + tq(x)\}$.

[1] Theorem 1 allows obvious extensions to nonnegative σ-finite measures.

Geometrically $f(\cdot)$ can be viewed as the negative of the Legendre-Fenchel transformation[2] of $\mathcal{S} := \{(-q(x), p(x))\}_{x\in\mathcal{D}}$. Hence $f(\cdot)$ is concave, and the left and the right derivatives are determined by the two extreme points of the intersection between $\mathcal{S}$ and the supporting hyperplane. See [19] for a complete proof of Lemma 3.

By Lemma 3 and symmetry, for almost all $(c_0, \dots, c_m)$,

$$\begin{aligned} & h(\mathbf{X}^+|\hat{U}) = h(\mathbf{X}^-|\mathbf{X}^+, \hat{U}) = h(\mathbf{X}^+|\mathbf{X}^-, \hat{U}) \\ \Longrightarrow & I(\mathbf{X}^+; \mathbf{X}^-|\hat{U}) = 0. \end{aligned} \tag{24}$$

The proof is completed by a strengthening of the Skitovic-Darmois characterization of normal distributions [13]:

Lemma 4. *If $\mathbf{A}_1$ and $\mathbf{A}_2$ are mutually independent random vectors such that $\mathbf{A}_1+\mathbf{A}_2$ is independent of $\mathbf{A}_1-\mathbf{A}_2$, then $\mathbf{A}_1$ and $\mathbf{A}_2$ are normally distributed with identical covariances.*

■

Remark 1. New ingredients added to the Geng-Nair approach [13][14] for establishing Gaussian optimality include:

- Lemma 3, that is, by differentiating with respect to the linear coefficients, we can conveniently obtain information theoretic identities which helps us to conclude the conditional independence of $\mathbf{X}^+$ and $\mathbf{X}^-$ quickly.[3] For small m, in principle, this may be avoided by exhaustively enumerating the expansions of two-letter quantities (e.g. as done in [14]), but that approach becomes increasingly complicated and unstructured as m increases.
- A semicontinuity property is used in the proof of the existence of the minimizer (not discussed here, but see [19]). The continuity of differential entropy argument in [13][14] does not apply here for a sequence of weakly convergent $\mathbf{X}_n$, since their densities are not regularized by convolving with the Gaussian density.

If we do not impose the non-degenerate assumption and the regularization $\mathbf{\Sigma}_{\mathbf{X}|U} \preceq \mathbf{\Sigma}$, it is possible that the optimization in Theorem 2 is nonfinite and/or not attained by any $P_{U\mathbf{X}}$. In this case, we can prove that the optimization is *exhausted* by Gaussian distributions, by taking the limit in Theorem 2 as the variance of the additive noise converges to zero. To state the result conveniently, for any $P_{\mathbf{X}}$, define

$$F_0(P_{\mathbf{X}}) := -h(\mathbf{X}) + \sum_{j=1}^m c_j h(\mathbf{Y}_j) + c_0 \operatorname{Tr}[\mathbf{M}\mathbf{\Sigma}_{\mathbf{X}}], \tag{25}$$

where $(\mathbf{X}, \mathbf{Y}_j) \sim P_{\mathbf{X}} Q_{\mathbf{Y}_j|\mathbf{X}}$.

Theorem 5. *In the general (possibly degenerate) case, For any given positive semidefinite $\mathbf{\Sigma}$,*

$$\inf_{P_{\mathbf{X}U}, \mathbf{\Sigma}_{\mathbf{X}|U} \preceq \mathbf{\Sigma}} F(P_{\mathbf{X}U}) = \inf_{P_{\mathbf{X}} \text{ Gaussian}, \mathbf{\Sigma}_{\mathbf{X}|U} \preceq \mathbf{\Sigma}} F_0(P_{\mathbf{X}}). \tag{26}$$

The same holds when the covariance constraint is dropped.

Note that Theorem 5 reduces an infinite dimensional optimization problem to a finite dimensional one. From Theorem 2 and Theorem 5 one easily obtains Gaussian optimality results in a related optimization problem involving mutual information; see [19] for details. We close the section by mentioning several implications of the Gaussian optimality results:

- Extension of BL to Gaussian transformations: when $Q_{\mathbf{X}}$ and $(Q_{\mathbf{Y}_j|\mathbf{X}})$ are Gaussian, (4) holds if and only if it holds for all Gaussian functions (f_j).
- Multivariate Gaussian hypercontractivity: we say an m-tuple of random variables $(X_1, \dots, X_m) \sim Q_{X^m}$ is $(p_1, \dots, p_m)$-hypercontractive for $p_j \in [1, \infty]$ if
$$\mathbb{E}\left[\prod_{j=1}^m f_j(X_j)\right] \le \prod_{j=1}^m \|f_j(X_j)\|_{p_j} \tag{27}$$
for all bounded measurable (f_j). Suppose $Q_{X^m} = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a positive semidefinite matrix whose diagonal values are all 1. By choosing an appropriate $\mathbf{M}$ in (25), one sees that Theorem 5 continues to hold if the differential entropies are replaced by relative entropies with Gaussian reference measures. Then by the dual formulation of hypercontractivity, (p_j) is hypercontractive if and only if a certain matrix inequality is satisfied, which can be simplified [19] to the following condition:
$$\mathbf{P} \succcurlyeq \mathbf{\Sigma}, \tag{28}$$
that is, $\mathbf{P} - \mathbf{\Sigma}$ is positive semidefinite, where $\mathbf{P}$ is a diagonal matrix with (p_j) on its diagonal. The $m = 2$ case is Nelson's hypercontractivity theorem.
- The rate region for certain network information theory problems can be solved by finite dimensional matrix optimizations. For example, the multivariate Wyner common information [30] of m Gaussian scalar random variables $X_1, \dots, X_m$ with covariance matrix $\mathbf{\Sigma} \succ \mathbf{0}$ is given by
$$\frac{1}{2} \inf_{\mathbf{\Lambda}} \log \frac{|\mathbf{\Sigma}|}{|\mathbf{\Lambda}|} \tag{29}$$
where the infimum is over all diagonal matrices $\mathbf{\Lambda}$ satisfying $\mathbf{\Lambda} \preceq \mathbf{\Sigma}$. Previously, an estimation theoretic argument [30, Corollary 1] only establishes the Gaussian optimality of the auxiliary provided that $\mathbf{\Sigma}$ satisfies certain conditions. Other examples include one communicator common randomness generation and omniscient helper key generation [7].

V. Duality Result for the Reverse Inequality

In this section we give a dual to Theorem 1. Here we state and sketch proof for finite $\mathcal{X}^m$ - an assumption used in showing a "splitting" property of relative information. Extension to more general alphabets is treated in [19].

Theorem 6. *Fix $Q_{Y|X^m}$, (Q_{X_j}), R_Y and $d \in \mathbb{R}$. Assume $|\mathcal{X}^m| < \infty$ and $Q_Y \ll \mu$ where $\prod_j Q_{X_j} \to Q_{Y|X^m} \to Q_Y$. The following two statements are equivalent:*

1) *For any $f_j\colon \mathcal{X}_j \to [0, +\infty)$, if $F\colon \mathcal{Y} \to [0, +\infty)$ is such that $\mathbb{E}[\log F(Y)|X^m = x^m] \ge \sum_j c_j \log f_j(x_j)$, $\prod_j Q_{X_j}$-almost surely, then*
$$\int F \mathrm{d}R_Y \ge \exp(d) \prod_j \left(\int f_j \mathrm{d}Q_{X_j}\right)^{c_j}. \tag{30}$$
2) *For any (P_{X_j}), there exists a coupling P_{X^m} such that*
$$D(P_Y \| R_Y) + d \le \sum_j c_j D(P_{X_j} \| Q_{X_j}) \tag{31}$$
where $P_{X^m} \to Q_{Y|X^m} \to P_Y$.

[2] An extension of the Legendre-Fenchel transformation of a convex set.

[3] The idea of differentiating the coefficients has been used in [29].

Notice that compared to Theorem 1, the inequality signs in (30) and (31) are reversed, and the computation of the best constant d involves an extra optimization (over F or P_{X^m}).

Proof Sketch: We consider $d = 0$ only as the general case can be handled similarly. 1)⇒2) This is the nontrivial direction which uses the finiteness of $|\mathcal{X}^m|$. Given (P_{X_j}), suppose P_{X^m} is a coupling that minimizes $D(P_Y\|R_Y)$ (which exists because $D(P_Y\|R_Y)$ is lower semicontinuous in P_{X^m}). It is a standard exercise using KKT conditions to show an important splitting property: there exist $g_j\colon \mathcal{X}_j \to \mathbb{R}$ such that

$$\mathbb{E}[\imath_{P_Y\|R_Y}(Y)|X^m = x^m] \ge \sum_j c_j g_j(x_j) \tag{32}$$

for all x^m, and the equality holds P_{X^m}-almost surely.[4] The latter claim follows by applying complementary slackness to the nonnegativity constraint on P_{X^m}. Now define S_{X_j} by

$$\imath_{S_{X_j}\|Q_{X_j}}(x_j) = -d_j + g_j(x_j) \tag{33}$$

where d_j's are normalization constants. Applying 1) with $F \leftarrow \frac{\mathrm{d}P_Y}{\mathrm{d}R_Y}$, $f_j \leftarrow \exp(g_j)$, and using the fact that $D(P_{X_j}\|S_{X_j}) \ge 0$, we obtain 2) upon rearranging.

2)⇒1) Given F, (f_j), define S_Y, (P_{X_j}) by

$$\imath_{S_Y\|R_Y}(y) = -d_0 + \log F(y); \tag{34}$$

$$\imath_{P_{X_j}\|Q_{X_j}}(x_j) = -d_j + \log f_j(x_j), \tag{35}$$

where $d_0,\dots,d_m$ are normalization constants. The assumption $Q_Y \ll \mu$ guarantees that S_Y and (P_{X_j}) are well-defined except for the trivial case where F and some f_j are zero almost surely. Then choose the P_Y such that (31) holds. Finally 1) follows from 2) and the nonnegativity of $D(P_Y\|S_Y)$. ■

Remark 2. Once the finiteness assumption on $\mathcal{X}^m$ is dropped (see [19]), we can recover Barthe's formulation of the reverse Brascamp-Lieb inequality [16] from (30): when $Q_{Y|X^m}$ is deterministic given by $\phi\colon \mathcal{X}^m \to \mathcal{Y}$, then the first statement in Theorem 6 holds if and only if it holds for

$$F(y) := \sup_{x^m\colon \phi(x^m)=y} \prod_j f_j^{c_j}(x_j), \quad \forall y. \tag{36}$$

The RBL is recovered by letting $\phi(\cdot)$ be a linear function.

Remark 3. Both the forward and reverse duality theorems recover the strong data processing inequality when $m = 1$ (that is, the MAC or the broadcast channel is a point-to-point channel). No meaningful version of reverse strong data processing is immediately apparent; the naive candidate which simply reverses the inequality sign doesn't even tensorize [32].

References

[1] H. J. Brascamp and E. H. Lieb, "Best constants in Young's inequality, its converse, and its generalization to more than three functions," *Advances in Mathematics*, vol. 20, no. 2, pp. 151–173, 1976.

[2] E. H. Lieb, "Gaussian kernels have only Gaussian maximizers," *Inventiones Mathematicae*, vol. 102, no. 1, pp. 179–208, 1990.

[3] J. Bennett, A. Carbery, M. Christ, and T. Tao, "The Brascamp–Lieb inequalities: finiteness, structure and extremals," *Geometric and Functional Analysis*, vol. 17, no. 5, pp. 1343–1415, 2008.

[4] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú, "Smoothing Brascamp-Lieb inequalities and strong converses for CR generation," in *Proc. of IEEE International Symposium on Information Theory*, July 2016, Barcelona, Spain.

[5] T. Courtade, "Outer bounds for multiterminal source coding via a strong data processing inequality," in *Proc. of IEEE International Symposium on Information Theory*, pp. 559–563, July 2013, Istanbul, Turkey.

[6] Y. Polyanskiy and Y. Wu, "A note on the strong data-processing inequalities in Bayesian networks," *http://arxiv.org/pdf/1508.06025v1.pdf.*

[7] J. Liu, P. Cuff, and S. Verdú, "Secret key generation with one communicator and a one-shot converse via hypercontractivity," in *Proc. of 2015 IEEE International Symposium on Information Theory*, pp. 710–714, June 2015, Hong Kong, China.

[8] A. Xu and M. Raginsky, "Converses for distributed estimation via strong data processing inequalities," in *Proc. of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China*, July 2015.

[9] A. Ganor, G. Kol, and R. Raz, "Exponential separation of information and communication," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 176–185, 2014.

[10] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," *arXiv preprint arXiv:1506.07216*, 2015.

[11] E. Mossel, R. O'Donnell, and K. Oleszkiewicz, "Noise stability of functions with low influences: Invariance and optimality," *Annals of Mathematics*, vol. 171, no. 1, pp. 295–341, 2010.

[12] M. J. John C Duchi and M. J. Wainwright, "Local privacy and statistical minimax rates," in *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 429–438, 2013.

[13] Y. Geng and C. Nair, "The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2087–2104, April, 2014.

[14] C. Nair, "An extremal inequality related to hypercontractivity of Gaussian random variables,"

[15] F. Barthe, "Optimal Young's inequality and its converse: a simple proof," *Geometric and Functional Analysis*, vol. 8, no. 2, pp. 234–242, 1998.

[16] F. Barthe, "On a reverse form of the Brascamp-Lieb inequality," *Inventiones Mathematicae*, vol. 134, no. 2, pp. 335–361, (see also arXiv:math/9705210 [math.FA]), 1998.

[17] S. Kamath, "Reverse hypercontractivity using information measures," in *Proc. of the 53rd Annual Allerton Conference on Communications, Control and Computing*, pp. 627–633, Sept. 30-Oct. 2, 2015, UIUC, Illinois.

[18] C. Nair, "Tensorization: information theory and hypercontractivity." http://www.ee.iitb.ac.in/bits/wp-content/uploads/2016/01/BITS.pdf. based on work by S. Beigi and C. Nair in 2015, presented at Bombay Information Theory Seminar 2016.

[19] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú, "Information theoretic perspectives on Brascamp-Lieb inequality and its reverse," *draft*.

[20] E. A. Carlen and D. Cordero-Erausquin, "Subadditivity of the entropy and its relation to Brascamp–Lieb type inequalities," *Geometric and Functional Analysis*, vol. 19, no. 2, pp. 373–405, 2009.

[21] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *The Annals of Probability*, vol. 4, pp. 925–939, 1976.

[22] C. Nair, "Equivalent formulations of hypercontractivity using information measures," *International Zurich Seminar*, Feb. 2014.

[23] R. Atar, K. Chowdhary, and P. Dupuis, "Robust bounds on risk-sensitive functionals via Rényi divergence," *SIAM J. Uncertainty Quant.*, vol. 3, pp. 18–33, 2015.

[24] R. Atar and N. Merhav, "Information-theoretic applications of the logarithmic probability comparison bound," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5366–5386, Oct. 2015.

[25] I. Csiszar and J. Körner, *Information theory: coding theorems for discrete memoryless systems (Second edition).* Cambridge University Press, 2011.

[26] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *http://arxiv.org/pdf/1304.6133v1.pdf.*

[27] L. H. Loomis and H. Whitney, "An inequality related to the isoperimetric inequality," *Bull. Amer. Math. Soc.*, vol. 55, pp. 961–962, 1949.

[28] P. F. F. Chung, R. Graham and J. Shearer, "Some intersection theorems for ordered sets and graphs," *J. Combinatorial Theory Series A*, vol. 43, no. 1, pp. 23–37, 1986.

[29] T. Courtade and J. Jiao, "An extremal inequality for long Markov chains," in *Proc. of the 52rd Annual Allerton Conference on Communications, Control and Computing*, pp. 763–770, Oct. 1-3, 2014, UIUC, Illinois.

[30] G. Xu, W. Liu, and B. Chen, "Wyner's common information: Generalizations and a new lossy source coding interpretation," *arXiv preprint arXiv:1301.2237*, 2013.

[31] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, pp. 146–158, 1975.

[32] J. Liu, P. Cuff, and S. Verdú, "Key capacity for product sources with application to stationary Gaussian processes," *IEEE Transactions on Information Theory*, vol. 62, pp. 984–1005, Feb. 2016.

[4] This property is related (but not equivalent) to a property of I-projections onto linear sets discussed by Csiszár [31, Section 3]; see [19] for a discussion.