Unifying the Brascamp-Lieb Inequality and the Entropy Power Inequality

Venkat Anantharam*, Varun Jog †, and Chandra Nair‡

Abstract

The entropy power inequality (EPI) and the Brascamp-Lieb inequality (BLI) can be viewed as information inequalities concerning entropies of linear transformations of random variables. The EPI provides lower bounds for the entropy of linear transformations of random vectors with independent components. The BLI, on the other hand, provides upper bounds on the entropy of a random vector in terms of the entropies of its linear transformations. In this paper, we present a new entropy inequality that generalizes both the BLI and EPI by considering a variety of independence relations among the components of a random vector. Our main technical contribution is in the proof strategy that leverages the "doubling trick" to prove Gaussian optimality for certain entropy expressions under independence constraints.

1 Introduction

Information inequalities provide some of the most powerful mathematical tools in an information theorist's toolbox and are therefore a vital part of information theory. Inequalities such as the non-negativity of mutual information and the data processing inequality are so fundamental to information theory that they are inseparable from information-theoretic notation. These basic inequalities, combined with Fano's inequality, are powerful enough to yield the converse of Shannon's channel coding theorem. For harder problems in network information theory, it is necessary to develop more nuanced information inequalities. Not surprisingly, it is often the case that discovering new inequalities leads to breakthroughs in network information theory problems. Some examples of information inequalities that spurred such breakthroughs include the entropy power inequality [1, 2], numerous strengthened forms of the entropy power inequality [3, 4, 5], strong data processing inequalities [6], and inequalities that established certain continuity properties of entropy [7]. In this paper, we present a new class of information inequalities that unifies two fundamental inequalities in information theory: the entropy power inequality (EPI) and the Brascamp-Lieb inequality (BLI). This new formulation has potential applications in network information theory and beyond. In what follows, we provide a brief introduction to the EPI and the BLI and state our main results.

As notational conventions in what follows, $:=$ and $=:$ denote equality by definition depending on whether the expression being defined is on the left or on the right respectively, while, for an integer $n > 0$, $[n]$ denotes $\{1, \ldots, n\}$ and $I_{n\times n}$ denotes the $n \times n$ identity matrix. We use the notation $|A|$ for the determinant of a square matrix A. We use the term "entropy" as synonymous with "differential entropy" in this document. All vectors are assumed to be column vectors, and we will adopt the convention that if X is a $\mathbb{R}^k$-valued vector and Y is a $\mathbb{R}^l$-valued vector, then (X, Y) denotes the $\mathbb{R}^{k+l}$-valued vector that would normally be written as $(X^T, Y^T)^T$. Given a random

*Department of Electrical Engineering and Computer Sciences, UC Berkeley. Email: ananth@eecs.berkeley.edu
†Department of Electrical and Computer Engineering, UW - Madison. Email: vjog@wisc.edu
‡Department of Information Engineering Engineering, CUHK. Email: chandra@ie.cuhk.edu.hk

vector $(Z_1, \ldots, Z_n)$, we use the notation $Z_{a:b}$ to denote the random vector $(Z_a, Z_{a+1}, \ldots, Z_b)$, where $1 \leq a \leq b \leq n$. The notation $X \to U \to Y$ for random vectors X, U, and Y indicates that X and Y are conditionally independent given U.

Entropy power inequality: The EPI states that for any independent $\mathbb{R}^n$-valued random variables X and Y, the following inequality holds:

$$e^{\frac{2h(X+Y)}{n}} \geq e^{\frac{2h(X)}{n}} + e^{\frac{2h(Y)}{n}}. \tag{1}$$

Here, $h(\cdot)$ refers to the differential entropy function and all the differential entropies in equation (1) are assumed to exist. Equality holds if and only if X and Y are Gaussian random variables with proportional covariance matrices. The EPI was proposed by Shannon [1] and was first proved by Stam [8]. This proof was later simplified by Blachman [2]. A variety of simple and ingenious proofs have been discovered since, see [9].

The EPI has an equivalent statement, as discovered by Lieb [10]. This formulation states that for independent $\mathbb{R}^n$-valued random vectors X and Y, and $\lambda \in (0, 1)$, the following inequality holds:

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda h(X) + (1-\lambda)h(Y). \tag{2}$$

Equality holds in the above inequality if and only if X and Y are Gaussian random variables with identical covariance matrices. Note that $\sqrt{\lambda}X + \sqrt{1-\lambda}Y$ may be interpreted as a linear transformation of an $\mathbb{R}^{2n}$-valued random variable $Z := (X, Y)$ with some independence constraints on the components of Z, namely $X \perp\!\!\!\perp Y$. Another result along such lines is Zamir and Feder's EPI [4] for linear transformations of random vectors with independent components. This EPI has an equivalent formulation, discovered in [9, 11], that is analogous to Lieb's form of the EPI in equation (2): For an $\mathbb{R}^n$-valued random vector $X := (X_1, \ldots, X_n)$ with independent scalar components, and any $k \times n$ matrix A satisfying $AA^T = I_k$, Zamir and Feder's EPI states that

$$h(AX) \geq \sum_{j=1}^{n} \alpha_j^2 h(X_j), \tag{3}$$

where α_j^2 is the squared-norm of the j-th column of A; i.e., $\alpha_j^2 := \sum_{i=1}^{k} a_{ij}^2$.

Brascamp-Lieb inequality: The BLI [12] is actually a family of functional inequalities that lies, in some sense, at the intersection of information and functional inequalities. Many well-known and commonly used inequalities are special cases of the BLI, including Hölder's inequality, the Loomis-Whitney inequality, the Prékopa-Leindler inequality, and sharp forms of Young's convolution inequalities [13]. In Gardner's extensive survey [14], the author describes relationships between popular functional and information inequalities using a pyramid-like sketch, where inequalities at the top imply those below. The BLI and its reverse lie at the very apex of this inequality pyramid. A simple statement of the BLI is as follows:

Theorem 1 (Functional form of the BLI)**.** *Suppose that $a_1, \ldots, a_m$ are m vectors that span $\mathbb{R}^n$, $c_1, \ldots, c_m$ are positive numbers, and $f_1, \ldots, f_m$ are nonnegative integrable functions on $\mathbb{R}$. Define the function $\mathcal{F}$ via*

$$\mathcal{F}(f_1, \ldots, f_m) := \frac{\int_{\mathbb{R}^n} \prod_{j=1}^{m} f_j^{c_j}(a_j \cdot x)dx}{\prod_{j=1}^{m} \left(\int_{\mathbb{R}} f_j(t)dt\right)^{c_j}}.$$

Then the supremum of $\mathcal{F}$ over all nonnegative and integrable f_j is equal to the supremum of $\mathcal{F}$ when f_j are centered Gaussian functions, i.e., for all $j \in [m]$, we have $f_j(t) \propto e^{-t^2/b_j}$ for some $b_j > 0$.

The general form of the above result involves replacing the vectors a_j by matrices A_j, and the centered Gaussian functions by $f_j \propto e^{-x^T B_j x}$, for some positive definite matrices B_j with the appropriate dimensions [13]. Although our goal is to analyze this general form, it is sufficient to focus initially on the simpler form of the BLI to clearly explain our results. Surprisingly, a direct connection exists between the functional form of the BLI and a generalized subadditivity result for entropy. This link was first discovered in Carlen et al. [15], and has since led to newer proofs and generalizations of the original BLI [16, 17, 18, 19, 20]. The information-theoretic form of the BLI is the following:

Theorem 2 (Information-theoretic form of the BLI, Theorem 2.1 in [16])**.** *Let $a_1, \ldots, a_m$ and $c_1, \ldots, c_m$ be as in Theorem 1. For an $\mathbb{R}^n$-valued random variable X with a well-defined entropy (see Definition 1) and $\mathbb{E}\|X\|^2 < \infty$, define $f(X)$ as*

$$f(X) := h(X) - \sum_{j=1}^{m} c_j h(a_j \cdot X). \tag{4}$$

Then the supremum of f over all random variables X is equal to the supremum of f over all centered Gaussian random variables.

This information-theoretic form is completely equivalent to the functional form: For a fixed choice of the a_j and the c_j, the supremums in both problems have a direct relationship, and the cases of equality are also in correspondence [16, Theorem 2.1]. For this reason, we will only consider the information-theoretic form of the BLI in this paper. A defining feature of the BLI is that it reduces the infinite-dimensional optimization problem to a finite-dimensional optimization problem over the set of positive definite matrices. When the supremum in Theorem 2 is finite, random variables that achieve the supremum are called *extremizers*, and Gaussian random variables that achieve the supremum are called *Gaussian extremizers*. [1] The existence of extremizers or Gaussian extremizers and the finiteness of D are not addressed by Theorem 2, as stated above. However, this is well-understood in the literature [21, 16, 13].

Our contributions: The classical EPI and the EPI of Zamir and Feder are valid only under certain independence assumptions. To be precise, for an $\mathbb{R}^{2n}$-valued random vector Z, the EPI requires independence of $Z_{1:n}$ and $Z_{n+1:n}$ and considers the sum of these two vectors, whereas Zamir and Feder's EPI requires all the components to be independent and considers linear transformations of Z. It is natural to consider more general "mixed" independence constraints, for instance, independence of $Z_{1:k_1}, Z_{k_1+1:k_2}, \ldots, Z_{k_r+1:n}$ for suitable choices of k_i, and establish lower bounds on $h(AZ)$ for a matrix A. This is indeed a special case of the setting considered in our work.

Consider an $\mathbb{R}^n$-valued random vector $X := (X_1, \ldots, X_k)$, where $k \leq n$ and X_i are mutually independent $\mathbb{R}^{r_i}$-valued random variables. Note that $\sum_{i=1}^{k} r_i = n$. We consider the following function:

$$f(X) := \sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X), \tag{5}$$

for positive constants d_i and c_j where $i \in [k]$ and $j \in [m]$ for some $m \geq 1$, and surjective linear transformations A_j from $\mathbb{R}^n$ to $\mathbb{R}^{n_j}$. Just as in Theorem 2, our main result in Theorem 3 states that the supremum of $f(\cdot)$ over all random variables X satisfying the stated independence constraints

[1] In [13] a Gaussian extremizer is defined as a distribution that extremizes among the class of Gaussian distributions, but it turns out that this definition is identical to the one used here.

is the same as the supremum evaluated over centered Gaussian random variables. In Theorem 4, we identify necessary and sufficient conditions on n, k, m and the r_i, d_i, c_j, n_j and A_j, such that this supremum is finite. We show that the EPI, BLI, and Zamir and Feder's EPI easily follow from Theorem 3. Theorem 3 also provides a generalization of Zamir and Feder's result for certain kinds of dependent random variables.

Our main technical contribution is developing a new proof strategy that leverages the "doubling trick" in information theory appearing in [22]. The authors of [22] solved the problem of determining the capacity of a Gaussian vector broadcast channel, and invented the doubling trick as a part of their proof strategy to establish Gaussian optimality of certain information-theoretic expressions. This proof strategy has since been applied to a wide variety of problems [23, 24, 25, 26, 27, 5]. Although new to information theory, the doubling trick had previously been applied to functional inequalities: Lieb [28] used it to show that Gaussian kernels have Gaussian optimizers, whereas Carlen [29] used it to show Gaussian optimality in the log-Sobolev inequality. The proof strategy has been attributed to Ball [30]. Since many information-theoretic inequalities have equivalent functional formulations, it may seem that importing the doubling trick to information theory is just a matter of convenience – instead of working with functions and their norms, one chooses to work with random variables and their entropies. However, this is not the case: the doubling trick in information theory can be applied in conjunction with other information inequalities such as concavity of entropy, positivity of mutual information, and convexity of KL-divergence to yield entirely new results that may seem, at least at first sight, to be inaccessible from a functional viewpoint. Indeed, the proof strategy developed in this paper emphasizes this point of view.

Related work: The EPI may be thought of as a limiting special case of the BLI. Gardner [14] showed that the EPI follows from the sharp form of Young's inequality, which in turn is a special case of the BLI. This proof strategy is further clarified using a more geometric approach by Cordero-Erausquin and Ledoux [18]. The authors of [18] establish the EPI directly from Theorem 2 by carefully choosing the a_j and c_j as a function of a parameter ϵ that tends to 0 and yields the EPI in the limit. While these approaches do forge connections between the BLI and the EPI, they do not go much further. In particular, they do not suggest concrete approaches aimed at developing information inequalities for random vectors with more general independence properties.

Various information-theoretic analogues of hypercontractive inequalities and reverse Brascamp-Lieb inequalities in finite alphabet spaces have been studied in [19, 31]. A closely related work is that of Liu et al. [20], where a novel functional inequality called the forward-reverse Brascamp-Lieb inequality is formulated, and it is shown that there exists an analogous information-theoretic version of this inequality. Most relevant to us is the forward-reverse Brascamp-Lieb inequality with linear maps, where Liu et al. established Gaussian optimality using the doubling trick. Define a function F of the marginal densities of an $\mathbb{R}^n$-valued random variables X:

$$F(X_1, \ldots, X_n) := \inf_{\{Y | Y_i \stackrel{d}{=} X_i,\ i \in [n]\}} \sum_{i=1}^{n} d_i h(Y_i) - \sum_{j=1}^{m} c_j h(A_j Y). \tag{6}$$

Here, by $Y_i \stackrel{d}{=} X_i$ we mean that the distribution of Y_i is identical to that of X_i. Theorem 8 in [20] states that the supremum of F is obtained when each X_i is a centered Gaussian random variable, in which case the infimum in the definition in equation (6) is attained when the optimal coupling Y is a jointly Gaussian random vector. The expressions in equations (5) and (6) look very similar. The main difference is that equation (6) has an infimum over all possible couplings Y, whereas our definition in equation (5) enforces the unique coupling where components Y_i are mutually independent.

Structure of the paper: In Section 2, we introduce some preliminaries and set up the notation to be used in the rest of the paper. In Section 3 we state our main result in Theorem 3 and show that the EPI, BLI, and Zamir and Feder's EPI may be proved as special cases of this result. In Section 4, we prove Theorem 3. In Section 5, we establish necessary and sufficient conditions for the supremum of f in the expression in equation (5) to be finite. In Section 6, we provide a concrete example that demonstrates the utility of Theorem 3 in obtaining EPI-like results for dependent random variables. Finally, in Section 7 we conclude the paper and describe some open problems.

2 Preliminaries and notation

Definition 1. For $n > 0$, let X be an $\mathbb{R}^n$-valued random variable with density f_X that lies in the convex set of probability densities

$$\left\{ f \;\middle|\; \int_{\mathbb{R}^n} f(x) \log(1 + f(x)) dx < \infty \right\}. \tag{7}$$

Then we define the entropy of X as

$$h(X) := -\int_{\mathbb{R}^n} f_X(x) \log f_X(x) dx. \tag{8}$$

The entropy of a 0-dimensional random variable is defined to be 0.

Remark 2.1. *The integral in equation (7) is well-defined since the integrand is non-negative. The condition in equation (7) implies that the differential entropy integral in equation (8) is well-defined and lower-bounded away from $-\infty$. Also note that the condition in equation (7) is inherited by marginalization, i.e. if f satisfies the condition and g is a (multidimensional) marginal of f, then g also satisfies the condition.*

Definition 2 (BL datum)**.** For an integer $m > 0$, define an m-transformation as a triple

$$\mathbf{A} := (n, \{n_j\}_{j\in[m]}, \{A_j\}_{j\in[m]}),$$

where for each $j \in [m]$, $A_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$ is a surjective linear transformation, and $n_j \geq 0$. An m-exponent is defined as an m-tuple $\mathbf{c} = \{c_j\}_{j\in[m]}$, such that $c_j \geq 0$ for $j \in [m]$. A Brascamp-Lieb datum (BL datum) is defined as a pair $(\mathbf{A}, \mathbf{c})$ where $\mathbf{A}$ is an m-transformation and $\mathbf{c}$ is an m-exponent, for an integer $m > 0$.

Definition 3 (EPI datum)**.** For an integer $k > 0$, define a k-partition of n as $\mathbf{r} = \{r_i\}_{i\in[k]}$, such that $r_i > 0$ are integers and $\sum_{i\in[k]} r_i = n$. Let $\mathbf{d} = \{d_i\}_{i\in[k]}$ such that $d_i \geq 0$ for all i be a k-exponent. An EPI datum is a pair $(\mathbf{r}, \mathbf{d})$ where $\mathbf{r}$ is a k-partition and $\mathbf{d}$ is a k-exponent, for an integer $k > 0$.

Definition 4 (BL-EPI datum)**.** For an integer $n > 0$, a BL-EPI datum is defined as $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ where $(\mathbf{A}, \mathbf{c})$ is a BL datum for an integer $m > 0$, and $(\mathbf{r}, \mathbf{d})$ is an EPI datum for an integer $k > 0$.

Definition 5. Let $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ be a BL-EPI datum where $\mathbf{r}$ is a k-partition of n. Define $\mathcal{P}(\mathbf{r})$ to be the set of all $\mathbb{R}^n$-valued random variables $X := (X_1, X_2, \ldots, X_k)$ such that:

1. For $i \in [k]$, the random variables X_i take values in $\mathbb{R}^{r_i}$ and their densities satisfy the condition in equation (7);

2. $X_1, X_2, \ldots, X_k$ are independent;

3. $\mathbb{E}X = 0$ and $\mathbb{E}\,\|X\|_2^2 < \infty$;

Since entropy expressions are not affected by adding constants, the 0-mean assumption in Definition 5 may be made without loss of generality. Define $\mathcal{P}_g(\mathbf{r}) \subseteq \mathcal{P}(\mathbf{r})$ as the set of random variables X that satisfy the properties above, while, in addition, each X_i, $i \in [k]$ is Gaussian.

Remark 2.2. *Note that whether an $\mathbb{R}^n$-valued random variable X lies in $\mathcal{P}(\mathbf{r})$ or not is a property of its distribution. Note also that the finite variance assumption on random variables in $\mathcal{P}(\mathbf{r})$ implies that the entropies $h(X_i)$ for $i \in [k]$ and $h(A_jX)$ for $j \in [m]$ are bounded away from ∞. However, with only the variance assumption in place, it may happen that some of these entropies equal $-\infty$, which happens, for instance, when X is a constant. In this paper, we shall be dealing with differences of entropies of the form*

$$\sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X). \tag{9}$$

The condition in equation (7), together with the finite variance assumption, has the effect of ensuring that the absolute values of the differential entropies are finite, which ensures that the above difference is well-defined for $X \in \mathcal{P}(\mathbf{r})$. This is a technical assumption made for ease of presentation. In cases where the expression in equation (9) is not well-defined, we may redefine it to equal the limit

$$\limsup_{\delta \to 0_+} \sum_{i=1}^{k} d_i h(\tilde{X}_i) - \sum_{j=1}^{m} c_j h(A_j \tilde{X} + \sqrt{\delta} Z_j), \tag{10}$$

where $\tilde{X} := X + \sqrt{\delta} W$ for an standard normal W independent of X and the Z_j are standard normal random variables independent of (X, W). With this modification, our results continue to hold for random variables that satisfy all the conditions in Definition 5 except the condition in equation (7).

The following two concepts are required for Theorem 4.

Definition 6. Let $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ be a BL-EPI datum. Define a subspace $V \subseteq \mathbb{R}^n$ as being of $\mathbf{r}$-product form if V may be written as $V = V_1 \times V_2 \times \cdots \times V_k$ for subspaces $V_i \subseteq \mathbb{R}^{r_i}$, for $i \in [k]$.

Definition 7. Let $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ be a BL-EPI datum. An $\mathbf{r}$-product form subspace $V \subseteq \mathbb{R}^n$ is called a *critical subspace* if

$$\sum_{i=1}^{k} d_i \dim(V_i) = \sum_{j=1}^{m} c_j \dim(A_j V).$$

Definition 8. For a BL-EPI datum $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$, define $M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ as

$$M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d}) := \sup_{X \in \mathcal{P}(\mathbf{r})} \sum_{i=1}^{k} d_i h(X_i) - \sum_{j-1}^{m} c_j h(A_j X).$$

Similarly, define $M_g(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ as the above supremum taken over Gaussian inputs $X \in \mathcal{P}_g(\mathbf{r})$. When the BL-EPI datum is fixed, we shall omit the $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ argument and use the simplified notation M and M_g.

3 Main results

We are now in a position to state our main result:

Theorem 3 (Unified EPI and BLI)**.** *Let* $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ *be a BL-EPI datum. Recall the definition*

$$M_g := \sup_{Z \in \mathcal{P}_g(\mathbf{r})} \sum_{i=1}^{k} d_i h(Z_i) - \sum_{j=1}^{m} c_j h(A_j Z). \tag{11}$$

Then for any $X \in \mathcal{P}(\mathbf{r})$*, the following inequality holds:*

$$\sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X) \leq M_g. \tag{12}$$

Recall that in Definition 8 we introduced the quantity (with a simplified notation):

$$M := \sup_{X \in \mathcal{P}(\mathbf{r})} \sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X). \tag{13}$$

Naturally, we have $M \geq M_g$. Thus, if M_g is $+\infty$, then so is M. If $M_g < \infty$, then the above result implies $M \leq M_g$, and thus $M = M_g$. An equivalent way of stating the above result is asserting $M = M_g$.

Note that the theorem does not address the following points, which are worth investigating:

1. **Finiteness:** When is M_g (and therefore M) finite?
2. **Extremizability and Gaussian extremizability:** Assuming M is finite, when do extremizers exist for the supremum in equation (13), and when do Gaussian extremizers exist for the supremum in equation (12)? In particular, does extremizability imply Gaussian extremizability? (Clearly, the reverse implication is true because of Theorem 3.)
3. **Uniqueness of extremizers:** Assuming extremizers exists, are they unique in some appropriate sense?

The answers to all these questions will depend on the BL-EPI datum $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$. In this paper, we resolve the first question by identifying necessary and sufficient conditions on $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ that ensure finiteness of M and M_g. We do not address the latter two questions here. We show the following result:

Theorem 4. *For a BL-EPI datum* $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$*, we have* $M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d}) < \infty$ *if and only if the following conditions are satisfied:*

$$\sum_{i=1}^{k} d_i \dim(V_i) \leq \sum_{j=1}^{m} c_j \dim(A_j V) \quad \text{for all } \mathbf{r}\text{-product form } V, \quad \text{and} \tag{14}$$

$$\sum_{i=1}^{k} d_i r_i = \sum_{j=1}^{m} c_j n_j. \tag{15}$$

As we show below, Theorem 3 readily implies the EPI, BLI, and Zamir and Feder's EPI. For this reason, we choose to interpret the inequality in Theorem 3 as a unified version of the Brascamp-Lieb inequality and the entropy power inequality.

Entropy Power Inequality: We will prove the EPI in Lieb's form (2) using Theorem 3. Let X and Y be independent $\mathbb{R}^d$-valued random variables with zero means and bounded variances, and let $\lambda \in (0,1)$. The expression $\lambda h(X) + (1-\lambda)h(Y) - h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y)$ corresponds to $n = 2d$, $k = 2$, $r_1 = r_2 = d$, $d_1 = \lambda$, $d_2 = 1-\lambda$, $c_1 = 1$, and $A_1 = [\sqrt{\lambda}I_d, \sqrt{1-\lambda}I_d]$. Note that it is enough to prove $M_g = 0$ by explicit calculation. Consider Gaussian random variables $Z_1 \sim \mathcal{N}(0, \Sigma_1)$ and $Z_2 \sim \mathcal{N}(0, \Sigma_2)$. Plugging in the entropies of these Gaussian random variables and simplifying, we see that we need to evaluate the supremum

$$M_g = \sup_{\Sigma_1, \Sigma_2 \succeq 0} \lambda \log\det(\Sigma_1) + (1-\lambda)\log\det\Sigma_2 - \log\det(\lambda\Sigma_1 + (1-\lambda)\Sigma_2).$$

This supremum is seen to be 0 via the concavity of the log det function.

Brascamp-Lieb Inequality: When $k = 1$, $r_1 = n$, and $d_1 = 1$, we recover the setting of the Brascamp-Lieb inequality in its equivalent form of subadditivity of entropy:

$$h(X) \leq \sum_{j=1}^{m} c_j h(A_j X) + M_g, \tag{16}$$

for all $\mathbb{R}^n$-valued random variables X with $\mathbb{E}X = 0$ and $\mathbb{E}||X||_2^2 < \infty$.

Zamir and Feder's Inequality: Let A be a $k \times n$ matrix satisfying $AA^T = I_{k\times k}$. For $1 \leq j \leq n$, let the squared norm of the j-th column of A be denoted by α_j^2; i.e.,

$$\alpha_j^2 := \sum_{i=1}^{k} a_{ij}^2.$$

Just as we did for the EPI, it is enough to show that $M_g \leq 0$ by explicitly computing the supremum of $\sum_{j=1}^n \alpha_j^2 h(X_j) - h(AX)$ over Gaussian X. Let $\Lambda = \mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ be a positive definite matrix. Define a function F from the space of positive definite diagonal matrices to $\mathbb{R}$ as follows:

$$F(\Lambda) = \log|A\Lambda A^T| - \sum_{j=1}^{n} \alpha_j^2 \log\lambda_j.$$

If we show that $F(\Lambda) \geq 0$, then Theorem 3 will immediately imply Zamir and Feder's EPI for random vectors with independent components. Let $B := A\Lambda^{1/2}$, so that $A\Lambda A^T = BB^T$. Using the Cauchy-Binet formula for the determinant of BB^T, we obtain

$$|BB^T| = \sum_{1 \leq i_1 < \cdots < i_k \leq n} |B_{i_1 i_2 \ldots i_k}||B^T_{i_1 i_2 \ldots i_k}|,$$

where $B_{i_1 i_2 \ldots i_k}$ consists of the k columns of B corresponding to the indices $i_1, \ldots, i_k$. The right hand side of the above equality may be written explicitly as

$$\sum_{1 \leq i_1 < \cdots < i_k \leq n} \left(\prod_{j=1}^{k} \lambda_{i_j} \right) |A_{i_1 i_2 \ldots i_k}|^2.$$

Noting that $\sum_{1\le i_1<\cdots<i_k\le n}|A_{i_1i_2\ldots i_k}|^2 = |AA^T| = |I_k| = 1$ (again via the Cauchy-Binet formula), we may take logarithms and use Jensen's inequality to obtain

$$\log|A\Lambda A^T| \ge \sum_{1\le i_1<\cdots<i_k\le n}|A_{i_1i_2\ldots i_k}|^2 \log\left(\prod_{j=1}^{k}\lambda_{i_j}\right).$$

We now gather the coefficients of $\log\lambda_j$ for a fixed j. The coefficient of $\log\lambda_1$ is given by

$$\sum_{1=i_1<\cdots<i_k\le n}|A_{i_1i_2\ldots i_k}|^2 = 1-|A_{2,3,\ldots,n}A_{2,3,\ldots,n}^T| = 1-|I_n-A_1A_1^T| = \alpha_1^2.$$

Here, the first equality follows by using the Cauchy-Binet formula again, the second equality follows from the orthogonality of the rows of A, and the third equality is true because $|I_n-uu^T| = 1-\|u\|^2$ for any vector u. A similar calculation can be done to show that the coefficient of $\log\lambda_j$ is α_j^2 for all $1\le j\le n$, which completes the proof of $F(\Lambda)\ge 0$.

4 Proof of Theorem 3

Our proof strategy relies on the doubling trick [22] which was developed to solve optimization problems for random vectors under covariance constraints of the following form: $\sup_{\mathrm{Cov}(X)\preceq\Sigma} s(X)$. A rough sketch of the doubling trick proof strategy is outlined below:

- **Concave envelope:** Define the concave envelope of s, denoted by S, as the smallest concave function that pointwise dominates s. It can be seen that

$$S(X) = \sup_U s(X|U) = \sup_U \sum_{u\in\mathcal{U}} s(X|U=u)p_U(u),$$

  where the supremum is over finite auxiliary random variables U.

- **Subadditivity of S:** This step consists of defining S on the larger space of pairs of random variables (X,Y). A straightforward extension often exists for information-theoretic functions S. The subadditivity result shows that

$$S(X,Y)\le S(X)+S(Y).$$

  This is is the key step in the doubling trick, and is often the hardest step. The ingredients for establishing the subadditivity result developed here stems from the ideas to establish converses to coding theorems and outer bounds in network information theory. An argument with a flavor similar to the line of argument employed here can be found outlined in [32].

- **Optimizers of S:** In this step, we consider two i.i.d. copies of any optimizer X of S, say (X_1,X_2), and show that $(X_1+X_2)/\sqrt{2}$ and $(X_1-X_2)/\sqrt{2}$ are also optimizers. From here, we may use Gaussian characterization results [33] or the central limit theorem [22] to conclude that it is enough to consider only Gaussian optimizers.

- **Optimizers of s:** In this final step, we show that the optimal value for S is attained by a single Gaussian distribution; i.e., we may assume without loss of generality that $|\mathcal{U}|=1$, and thus this Gaussian also maximizes s.

As noted above, the most important step is to establish the subadditivity of S, for which we develop a new technique. The main idea is to exploit the chain rule for entropy in two separate ways. Given a random vector (X_1, X_2), we use the two expansions for the joint entropy $h(X_1, X_2)$:

(A) $h(X_1, X_2) = h(X_1) + h(X_2) - I(X_1; X_2)$,

(B) $h(X_1, X_2) = h(X_1|X_2) + h(X_2|X_1) + I(X_1; X_2)$.

To highlight the main ideas, we present a proof sketch of the subadditivity result for the EPI using our new technique.

4.1 Proving the EPI via subadditivity

Consider the function

$$s(X_1, Y_1) := \lambda h(X_1) + (1-\lambda)h(Y_1) - h(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1), \tag{17}$$

where $X_1 \perp\!\!\!\perp Y_1$. Define the lifting of s to the space of pairs of random variables by

$$s(X_{1:2}, Y_{1:2}) := \lambda h(X_{1:2}) + (1-\lambda)h(Y_{1:2}) - h(\sqrt{\lambda}X_{1:2} + \sqrt{1-\lambda}Y_{1:2}), \tag{18}$$

where $X_{1:2} \perp\!\!\!\perp Y_{1:2}$. Let $S(X_1, Y_1)$ and $S(X_{1:2}, Y_{1:2})$ be the respective concave envelopes of s and its lifting. We would like to show the subadditivity relation

$$S(X_{1:2}, Y_{1:2}) \leq S(X_1, Y_1) + S(X_2, Y_2). \tag{19}$$

Notice that

$$S(X_1, Y_1) = \sup_{X_1 \to U \to Y_1} s(X_1, Y_1|U), \tag{20}$$

and similarly for $S(X_{1:2}, Y_{1:2})$. For any auxiliary random variable U satisfying $X_{1:2} \to U \to Y_{1:2}$, applying expansion (A) to each entropy term in equation (18) (conditioned on U) yields

$$\begin{aligned}
s(X_{1:2}, Y_{1:2} \mid U) &= \lambda h(X_{1:2}|U) + (1-\lambda)h(Y_{1:2}|U) - h(\sqrt{\lambda}X_{1:2} + \sqrt{1-\lambda}Y_{1:2}|U) \\
&= \Big[\lambda h(X_1|U) + (1-\lambda)h(Y_1|U) - h(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1|U)\Big] \\
&+ \Big[\lambda h(X_2|U) + (1-\lambda)h(Y_2|U) - h(\sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2|U)\Big] \\
&+ \Big[-\lambda I(X_1; X_2|U) - (1-\lambda)I(Y_1; Y_2|U) + I(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1; \sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2|U)\Big].
\end{aligned} \tag{21}$$

For simplicity, call the terms in the brackets $T_1(U)$, $T_2(U)$, and $T_3(U)$ respectively, even though they actually depend on $p_{U|X_{1:2},Y_{1:2}}$. Observing that $X_i \to U \to Y_i$ for $i = 1, 2$, we may conclude $T_1(U) \leq S(X_1, Y_1)$ and $T_2(U) \leq S(X_2, Y_2)$. Substituting these inequalities, we arrive at

$$s(X_{1:2}, Y_{1:2}|U) \leq S(X_1, Y_1) + S(X_2, Y_2) + T_3(U). \tag{22}$$

We now expand the left hand side expression in equation (18) (conditioned on U) using expansion (B) for each entropy term:

$$\begin{aligned}
s(X_{1:2}, Y_{1:2}|U) &= \lambda h(X_{1:2}|U) + (1-\lambda)h(Y_{1:2}|U) - h(\sqrt{\lambda}X_{1:2} + \sqrt{1-\lambda}Y_{1:2}|U) \\
&= \Big[\lambda h(X_1|U, X_2) + (1-\lambda)h(Y_1|U, Y_2) - h(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1|U, \sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2)\Big] \\
&+ \Big[\lambda h(X_2|U, X_1) + (1-\lambda)h(Y_2|U, Y_1) - h(\sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2|U, \sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1)\Big] \\
&+ \Big[\lambda I(X_1; X_2|U) + (1-\lambda)I(Y_1; Y_2|U) - I(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1; \sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2|U)\Big].
\end{aligned} \tag{23}$$

For ease of notation, call the three terms $R_1(U)$, $R_2(U)$, and $R_3(U) = -T_3(U)$, even though they actually depend on $p_{U|X_{1:2},Y_{1:2}}$. Similar to inequality (22), we would like to upper bound $R_1(U)$ and $R_2(U)$ by $S(X_1, Y_1)$ and $S(X_2, Y_2)$ respectively. However, the conditioning for the entropy terms in each of the $R_i(U)$ is not the same, and we cannot directly conclude such a bound. Using the chain rule of mutual information and data-processing relations, we may make the conditioning in $R_1(U)$ and $R_2(U)$ uniform by introducing some extra mutual information terms:

$$\begin{aligned}
R_1(U) &= \Big[\lambda h(X_1|U, X_2) + (1-\lambda)h(Y_1|U, Y_2) - h(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1|U, \sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2)\Big] \\
&= \Big[\lambda h(X_1|U, X_2, Y_2) + (1-\lambda)h(Y_1|U, Y_2, X_2) - h(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1|U, X_2, Y_2)\Big] \\
&\quad - I(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1; X_2, Y_2|U, \sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2) \\
&=: \tilde{R}_1(U) - I_1(U),
\end{aligned}$$

where the notational conventions $\tilde{R}_1(U)$ and $I_1(U)$ are used even though the respective terms actually depend on $p_{U|X_{1:2},Y_{1:2}}$. The main step in the preceding equation is justified as follows. First, it it easy to check using the Markov relation $(X_1, X_2) \to U \to (Y_1, Y_2)$ that

$$h(X_1|U, X_2) = h(X_1|U, X_2, Y_2), \quad \text{and} \quad h(Y_1|U, Y_2) = h(Y_1|U, X_2, Y_2).$$

Also, we may verify that

$$\begin{aligned}
&h(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1|U, \sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2) \\
&= h(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1|U, X_2, Y_2) + I(\sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1; X_2, Y_2|U, \sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2).
\end{aligned}$$

Similar reasoning for $R_2(U)$ gives

$$\begin{aligned}
R_2(U) &= \Big[\lambda h(X_2|U, X_1) + (1-\lambda)h(Y_2|U, Y_1) - h(\sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2|U, \sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1)\Big] \\
&= \Big[\lambda h(X_2|U, X_1, Y_1) + (1-\lambda)h(Y_2|U, Y_1, X_1) - h(\sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2|U, X_1, Y_1)\Big] \\
&\quad - I(\sqrt{\lambda}X_2 + \sqrt{1-\lambda}Y_2; X_1, Y_1|U, \sqrt{\lambda}X_1 + \sqrt{1-\lambda}Y_1) \\
&=: \tilde{R}_2(U) - I_2(U),
\end{aligned}$$

where the notational conventions $\tilde{R}_2(U)$ and $I_2(U)$ are used even though the respective terms actually depend on $p_{U|X_{1:2},Y_{1:2}}$. Substituting the expressions for $R_1(U)$ and $R_2(U)$ in the expansion in equation (23), we arrive at

$$\begin{aligned}
s(X_{1:2}, Y_{1:2}|U) &= \tilde{R}_1(U) + \tilde{R}_2(U) - T_3(U) - I_1(U) - I_2(U) \\
&\overset{(a)}{\leq} S(X_1, Y_1) + S(X_2, Y_2) - T_3(U) - I_1(U) - I_2(U) \\
&\overset{(b)}{\leq} S(X_1, Y_1) + S(X_2, Y_2) - T_3(U). \qquad (24)
\end{aligned}$$

Here, in step (a) we used the Markov chains $X_1 \to (U, X_2, Y_2) \to Y_1$ and $X_2 \to (U, X_1, Y_1) \to Y_2$. Step (b) follows by noticing that $I_1(U)$ and $I_2(U)$ are non-negative, being mutual information expressions.

Inequalities (22) and (24) may now be used in tandem to conclude

$$s(X_{1:2}, Y_{1:2}|U) \leq S(X_1, Y_1) + S(X_2, Y_2). \qquad (25)$$

Taking the supremum over all auxiliary random variables U satisfying $X_{1:2} \to U \to Y_{1:2}$ leads to

$$S(X_{1:2}, Y_{1:2}) \leq S(X_1, Y_1) + S(X_2, Y_2). \tag{26}$$

Notice that the above proof not only gives us subadditivity, but also states that if there is equality in equation (25) for some optimal U^*, then $I_1(U^*) = I_2(U^*) = T_3(U^*) = 0$. This leads to several independence conditions that can be used establish Gaussian optimality. We do not sketch this part of the proof here.

In what follows, we develop this outline into a rigorous proof for a more general result. We do this in two stages: In Section 4.2 we establish the key subadditivity inequality and the independence relations that follow from the conditions for equality in that inequality, and in Section 4.3 we complete the proof of Theorem 3 by proving Gaussian optimality.

4.2 Subadditivity lemma

Let $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ be a BL-EPI datum. Let $X := (X_1, X_2, \ldots, X_k) \in \mathcal{P}(\mathbf{r})$, where $X_i \sim p_{X_i}$. A natural definition for $s(X)$ would be

$$s(X) := \sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X),$$

and one might then work with its concave envelope for the doubling trick. However, for ease of analysis, it is better to use Gaussian perturbed versions of s, with a small perturbation that tends to 0, and to work with the concave envelopes of such perturbed functions.

Definition 9. Let $W_i \sim \mathcal{N}(0, I_{r_i \times r_i}), i \in [k]$ be mutually independent standard normal random variables on $\mathbb{R}^{r_i}$, and let $W := (W_1, \ldots, W_k)$. For $j \in [m]$, define independent Gaussian random variables $Z_j \sim \mathcal{N}(0, I_{n_j \times n_j})$, and let $Z := (Z_1, Z_2, \ldots, Z_m)$. Assume that the random variables X, W and Z are mutually independent. For $\epsilon, \delta \geq 0$ define $s_{\epsilon,\delta} : \mathcal{P}(\mathbf{r}) \to \mathbb{R}$ as

$$s_{\epsilon,\delta}(X) := \sum_{i=1}^{k} d_i h(X_i + \sqrt{\delta} W_i) - \sum_{j=1}^{m} c_j h(A_j(X + \sqrt{\delta} W) + \sqrt{\epsilon} Z_j). \tag{27}$$

Let $S_{\epsilon,\delta}$ be the concave envelope of $s_{\epsilon,\delta}$; i.e., the smallest concave function that pointwise dominates $s_{\epsilon,\delta}$. Let U be an auxiliary random variable taking values in a finite set $\mathcal{U}$ such that we have $p_{X|U}(\cdot|U) \in \mathcal{P}(\mathbf{r})$. It is easy to see that the concave envelope has an equivalent definition in terms of such choices of U:

$$S_{\epsilon,\delta}(X) := \sup_{U \,:\, p_{X|U}(\cdot|U) \in \mathcal{P}(\mathbf{r})} \sum_{i=1}^{k} d_i h(X_i + \sqrt{\delta} W_i | U) - \sum_{j=1}^{m} c_j h(A_j(X + \sqrt{\delta} W) + \sqrt{\epsilon} Z_j \mid U), \tag{28}$$

where, on the right hand side of equation (28), we can assume that W, Z and (U, X) are mutually independent. For a particular choice of U, define

$$s_{\epsilon,\delta}(X \mid U) := \sum_{i=1}^{k} d_i h(X_i + \sqrt{\delta} W_i | U) - \sum_{j=1}^{m} c_j h(A_j(X + \sqrt{\delta} W) + \sqrt{\epsilon} Z_j | U). \tag{29}$$

Analogous to $\mathcal{P}(\mathbf{r})$, define $\mathcal{P}(2\mathbf{r})$ to be the set of random variables that take values in $\mathbb{R}^{2r_1} \times \cdots \times \mathbb{R}^{2r_k}$ and satisfy the conditions in Definition 5. More precisely, a random vector (X_1, X_2) is in

$\mathcal{P}(2\mathbf{r})$ if $X_1 := (X_{11}, \ldots, X_{k1})$ and $X_2 := (X_{12}, \ldots, X_{k2})$ are $\mathbb{R}^n$-valued random vectors such that the random vectors $(X_{i1}, X_{i2}) \in \mathbb{R}^{2r_i}$, $i \in [k]$ are mutually independent, satisfy the condition in equation (7), and condition 3 of Definition 5 holds for (X_1, X_2). Since the condition in equation (7) is inherited by marginalization, we have that if $(X_1, X_2) \in \mathcal{P}(2\mathbf{r})$ then $X_1 \in \mathcal{P}(\mathbf{r})$ and $X_2 \in \mathcal{P}(\mathbf{r})$.

We will need to define an extension of $S_{\epsilon,\delta}$ to the larger space $\mathcal{P}(2\mathbf{r})$. Consider a random vector $(X_1, X_2) \in \mathcal{P}(2\mathbf{r})$ as in the preceding paragraph. Define

$$\begin{aligned} s_{\epsilon,\delta}(X_1, X_2) := &\sum_{i=1}^{k} d_i h(X_{i1} + \sqrt{\delta} W_{i1}, X_{i2} + \sqrt{\delta} W_{i2}) \\ &- \sum_{j=1}^{m} c_j h(A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1}, A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2}), \end{aligned} \tag{30}$$

where (W_1, W_2, Z_1, Z_2) are mutually independent standard normal distributions of the appropriate dimensions that are independent of (X_1, X_2). The concave envelope of $s_{\epsilon,\delta}$ is defined as the smallest concave function that pointwise dominates $s_{\epsilon,\delta}$, and can be written as:

$$\begin{aligned} S_{\epsilon,\delta}(X_1, X_2) = &\sup_{U \,:\, p_{X_1 X_2|U}(\cdot,\cdot|U) \in \mathcal{P}(2\mathbf{r})} \sum_{i=1}^{k} d_i h(X_{i1} + \sqrt{\delta} W_{i1}, X_{i2} + \sqrt{\delta} W_{i2} \mid U) \\ &- \sum_{j=1}^{m} c_j h(A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1}, A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2} \mid U), \end{aligned} \tag{31}$$

where W_1, W_2, Z_1, Z_2 and (U, X_1, X_2) are mutually independent, with U taking values in finite sets $\mathcal{U}$ and $p_{X_1,X_2|U}(\cdot,\cdot|U) \in \mathcal{P}(2\mathbf{r})$. Figure 1 illustrates the relations between the random variables via a graphical model.

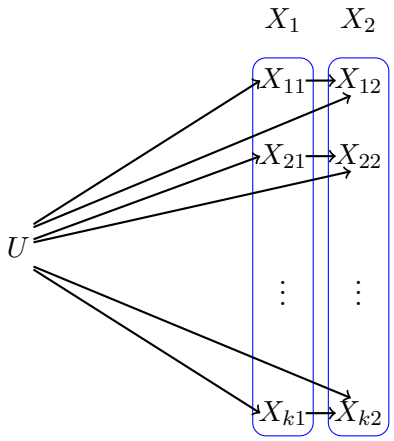

Figure 1: Illustration of the Markov relationship

Lemma 4.1 (Subadditivity lemma)**.** *The function S is subadditive; i.e., if $(X_1, X_2) \in \mathcal{P}(2\mathbf{r})$ then*

$$S_{\epsilon,\delta}(X_1, X_2) \leq S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2). \tag{32}$$

Corollary 4.1. *The function S tensorizes; i.e., if $X_1, X_2 \in \mathcal{P}(\mathbf{r})$ and if $X_1 \perp\!\!\!\perp X_2$, then*

$$S_{\epsilon,\delta}(X_1, X_2) = S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2). \tag{33}$$

Proof of Lemma 4.1. Let U be an auxiliary random variable taking values in a finite set $\mathcal{U}$, such that $p_{X_1,X_2|U}(\cdot,\cdot|U) \in \mathcal{P}(2\mathbf{r})$. Consider the following expansion, which comes from applying expansion (A) term by term:

$$\begin{aligned} s_{\epsilon,\delta}(X_1, X_2 \mid U) = &\left[\sum_{i=1}^{k} d_i h(X_{i1} + \sqrt{\delta}W_{i1}|U) - \sum_{j=1}^{m} c_j h(A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}|U)\right] \\ &+ \left[\sum_{i=1}^{k} d_i h(X_{i2} + \sqrt{\delta}W_{i2}|U) - \sum_{j=1}^{m} c_j h(A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}|U)\right] \\ &+ \Bigg[- \sum_{i=1}^{k} d_i I(X_{i1} + \sqrt{\delta}W_{i1}; X_{i2} + \sqrt{\delta}W_{i2}|U) \\ &\qquad + \sum_{j=1}^{m} c_j I(A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}; A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}|U)\Bigg]. \end{aligned} \tag{34}$$

For simplicity, denote the terms in the square brackets by $T_1(U)$, $T_2(U)$, and $T_3(U)$, respectively, even though they actually depend on $p_{U|X_1,X_2}$. Observe that that $p_{X_1|U}(\cdot|U), p_{X_2|U}(\cdot|U) \in \mathcal{P}(\mathbf{r})$ (see Figure 1). Thus, we conclude that $T_1(U) \leq S_{\epsilon,\delta}(X_1)$ and $T_2(U) \leq S_{\epsilon,\delta}(X_2)$, using the definition in equation (28). Substituting these inequalities, we arrive at

$$s_{\epsilon,\delta}(X_1, X_2|U) \leq S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2) + T_3(U). \tag{35}$$

We now expand $s_{\epsilon,\delta}(X_1, X_2 \mid U)$ in a different way, which comes from applying expansion (B) term by term:

$$s_{\epsilon,\delta}(X_1, X_2 \mid U) \tag{36}$$

$$\begin{aligned} &= \Big[\sum_{i=1}^{k} d_i h(X_{i1} + \sqrt{\delta}W_{i1}|U, X_{i2} + \sqrt{\delta}W_{i2}) - \sum_{j=1}^{m} c_j h(A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}|U, A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2})\Big] \\ &+ \Big[\sum_{i=1}^{k} d_i h(X_{i2} + \sqrt{\delta}W_{i2}|U, X_{i1} + \sqrt{\delta}W_{i1}) - \sum_{j=1}^{m} c_j h(A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}|U, A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1})\Big] \\ &+ \Big[\sum_{i=1}^{k} d_i I(X_{i1} + \sqrt{\delta}W_{i1}; X_{i2} + \sqrt{\delta}W_{i2}|U) - \sum_{j=1}^{m} c_j I(A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}; A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}|U)\Big]. \end{aligned} \tag{37}$$

For ease of notation, call the three terms in the square brackets $R_1(U)$, $R_2(U)$, and $R_3(U) = -T_3(U)$, respectively, even though each term actually depends on $p_{U|X_1,X_2}$. Similar to inequality (35), we would like to upper bound $R_1(U)$ and $R_2(U)$ by $S_{\epsilon,\delta}(X_1)$ and $S_{\epsilon,\delta}(X_2)$ respectively. However, the conditioning in each of the two differential entropy terms in each $R_a(U)$, $a = 1, 2$ is not the same, so we cannot directly conclude such a bound. However, using the chain rule of mutual information and data-processing relations, we may make the conditioning in $R_1(U)$ and $R_2(U)$

uniform by introducing some extra mutual information terms:

$$
\begin{aligned}
&R_1(U) \\
&= \Big[\sum_{i=1}^{k} d_i h(X_{i1} + \sqrt{\delta} W_{i1} | U, X_{i2} + \sqrt{\delta} W_{i2}) - \sum_{j=1}^{m} c_j h(A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1} | U, A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2}) \Big] \\
&= \Big[\sum_{i=1}^{k} d_i h(X_{i1} + \sqrt{\delta} W_{i1} | U, X_2 + \sqrt{\delta} W_2) - \sum_{j=1}^{m} c_j h(A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1} | U, X_2 + \sqrt{\delta} W_2) \Big] \\
&- \sum_{j=1}^{m} c_j I(A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1}; X_2 + \sqrt{\delta} W_2 | U, A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2}) \\
&=: \tilde{R}_1(U) - I_1(U),
\end{aligned}
$$

where we write $\tilde{R}_1(U)$ and $I_1(U)$ for simplicity, even thought the corresponding terms depend on $p_{U|X_1,X_2}$. The above steps are justified as follows. First, it is easy to check that $(X_{i1} + \sqrt{\delta} W_{i1}) \perp\!\!\!\perp \{X_{l2} + \sqrt{\delta} W_{l2}\}_{l \neq i}$ conditioned on $(U, X_{i2} + \sqrt{\delta} W_{i2})$. This means that, for all $1 \leq i \leq k$,

$$
\begin{aligned}
&h(X_{i1} + \sqrt{\delta} W_{i1} | U, X_{i2} + \sqrt{\delta} W_{i2}) \\
&= h(X_{i1} + \sqrt{\delta} W_{i1} | U, X_{12} + \sqrt{\delta} W_{12}, \ldots, X_{i2} + \sqrt{\delta} W_{i2}, \ldots, X_{k2} + \sqrt{\delta} W_{k2}) \\
&= h(X_{i1} + \sqrt{\delta} W_{i1} | U, X_2 + \sqrt{\delta} W_2).
\end{aligned}
$$

Also, we may verify the Markov chain (conditioned on U)

$$
\Big[A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2} \Big] \to \Big[X_2 + \sqrt{\delta} W_2 \Big] \to \Big[A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1} \Big],
$$

which gives the equality

$$
\begin{aligned}
&h(A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1} | U, A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2}) \\
&= h(A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1} | U, X_2 + \sqrt{\delta} W_2) \\
&+ I(A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1}; X_2 + \sqrt{\delta} W_2 | U, A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2}).
\end{aligned}
$$

Similar reasoning for $R_2(U)$ gives

$$
\begin{aligned}
&R_2(U) \\
&= \Big[\sum_{i=1}^{k} d_i h(X_{i2} + \sqrt{\delta} W_{i2} | U, X_{i1} + \sqrt{\delta} W_{i1}) - \sum_{j=1}^{m} c_j h(A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2} | U, A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1}) \Big] \\
&= \Big[\sum_{i=1}^{k} d_i h(X_{i2} + \sqrt{\delta} W_{i2} | U, X_1 + \sqrt{\delta} W_1) - \sum_{j=1}^{m} c_j h(A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2} | U, X_1 + \sqrt{\delta} W_1) \Big] \\
&- \sum_{j=1}^{m} c_j I(A_j(X_2 + \sqrt{\delta} W_2) + \sqrt{\epsilon} Z_{j2}; X_1 + \sqrt{\delta} W_1 | U, A_j(X_1 + \sqrt{\delta} W_1) + \sqrt{\epsilon} Z_{j1}) \\
&=: \tilde{R}_2(U) - I_2(U),
\end{aligned}
$$

where we use the notation $\tilde{R}_2(U)$ and $I_2(U)$ for simplicity, even though the corresponding terms depend on $p_{U|X_1,X_2}$. Substituting the expressions for $R_1(U)$ and $R_2(U)$ in the expansion in equation

(37), we arrive at

$$\begin{aligned} s_{\epsilon,\delta}(X_1, X_2 \mid U) &= \tilde{R}_1(U) + \tilde{R}_2(U) - T_3(U) - I_1(U) - I_2(U) \\ &\overset{(a)}{\leq} S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2) - T_3(U) - I_1(U) - I_2(U) \\ &\overset{(b)}{\leq} S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2) - T_3(U). \end{aligned} \tag{38}$$

Here, in step (a) we used the fact that $p_{X_1|U,X_2+\sqrt{\delta}W_2}(\cdot|U, X_2 + \sqrt{\delta}W_2), p_{X_2|U,X_1+\sqrt{\delta}W_1}(\cdot|U, X_1 + \sqrt{\delta}W_1) \in \mathcal{P}(\mathbf{r})$ and the definition in equation (28). Step (b) follows by noticing that the c_j are non-negative, and so are $I_1(U)$ and $I_2(U)$ since they are nonnegative linear combinations of mutual informations.

We can combine inequalities (35) and (38) to get

$$s_{\epsilon,\delta}(X_1, X_2|U) \leq S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2). \tag{39}$$

Taking the supremum on the left hand side of this inequality over all auxiliary variables U taking values in finite sets $\mathcal{U}$, such that $p_{X_1,X_2|U}(\cdot,\cdot|U) \in \mathcal{P}(2\mathbf{r})$, yields the claimed subadditivity result. □

Proof of Corollary 4.1. When $X_1 \perp\!\!\!\perp X_2$, we have the inequality

$$S_{\epsilon,\delta}(X_1, X_2) \geq S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2). \tag{40}$$

This is because we can always choose $U := (U_1, U_2)$ such that $(U_1, X_1) \perp\!\!\!\perp (U_2, X_2)$ and $p_{X_1|U_1}(\cdot|U_1)$, $p_{X_2|U_2}(\cdot|U_2) \in \mathcal{P}(\mathbf{r})$. The supremum in equation (31) over this restricted class of auxiliaries is simply $S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2)$, which therefore is a lower bound on $S_{\epsilon,\delta}(X_1, X_2)$. Inequality (40) combined with Lemma 4.1 completes the proof of Corollary 4.1. □

Our next lemma serves to some extent as a converse to Corollary 4.1. In particular, we show that if $S_{\epsilon,\delta}(X_1, X_2) = S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2)$, then X_1 and X_2 are independent conditioned on the optimal auxiliary U^*, assuming it exists. The formal statement is as follows:

Lemma 4.2 (Independence relations)**.** *Fix $\epsilon, \delta > 0$. Given $(X_1, X_2) \in \mathcal{P}(2\mathbf{r})$, suppose that $S_{\epsilon,\delta}(X_1, X_2) = S_{\epsilon,\delta}(X_1) + S_{\epsilon,\delta}(X_2)$. Suppose that U^* is such that $p_{X_1,X_2|U^*}(\cdot,\cdot|U^*) \in \mathcal{P}(2\mathbf{r})$ and $s_{\epsilon,\delta}(X_1, X_2 \mid U^*) = S_{\epsilon,\delta}(X_1, X_2)$. Then the following results hold:*

(a) For all $u^* \in \mathcal{U}^*$, we have that $X_1 \perp\!\!\!\perp X_2$ conditioned on $U^* = u^*$,

(b) $s_{\epsilon,\delta}(X_1|U^*) = S_{\epsilon,\delta}(X_1)$ and $s_{\epsilon,\delta}(X_2|U^*) = S_{\epsilon,\delta}(X_2)$.

Proof. Notice that the proof of Lemma 4.1 implies that the optimizing U^*, if it exists, must satisfy $I_1(U^*) = I_2(U^*) = T_3(U^*) = 0$. The first two equalities yield the Markov chains (conditioned on $U^* = u^*$)

$$\begin{aligned} &\Big[A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}\Big] \to \Big[A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}\Big] \to \Big[X_2 + \sqrt{\delta}W_2\Big], \quad \text{and} \\ &\Big[A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}\Big] \to \Big[A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}\Big] \to \Big[X_1 + \sqrt{\delta}W_1\Big]. \end{aligned}$$

However, we have the obvious Markov chains

$$\begin{aligned} &\Big[A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}\Big] \to \Big[X_2 + \sqrt{\delta}W_2\Big] \to \Big[A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}\Big], \quad \text{and} \\ &\Big[A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}\Big] \to \Big[X_1 + \sqrt{\delta}W_1\Big] \to \Big[A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}\Big]. \end{aligned}$$

Using Lemma A.1, we may conclude that, conditioned on U^*, we have

$$\begin{aligned}\Big[A_j(X_1+\sqrt{\delta}W_1)+\sqrt{\epsilon}Z_{j1}\Big] &\perp\!\!\!\perp \Big[X_2+\sqrt{\delta}W_2\Big], \quad \text{and}\\ \Big[A_j(X_2+\sqrt{\delta}W_2)+\sqrt{\epsilon}Z_{j2}\Big] &\perp\!\!\!\perp \Big[X_1+\sqrt{\delta}W_1\Big].\end{aligned}$$

Recall that $T_3(U^*)$ is given by

$$\begin{aligned}T_3(U^*) = \Bigg[&-\sum_{i=1}^{k} d_i I(X_{i1}+\sqrt{\delta}W_{i1}; X_{i2}+\sqrt{\delta}W_{i2}|U^*)\\ &+\sum_{j=1}^{m} c_j I(A_j(X_1+\sqrt{\delta}W_1)+\sqrt{\epsilon}Z_{j1}; A_j(X_2+\sqrt{\delta}W_2)+\sqrt{\epsilon}Z_{j2}|U^*)\Bigg].\end{aligned}$$

Substituting the above independence relations in $T_3(U^*)=0$, we conclude that, conditioned on U^*, we have

$$X_1+\sqrt{\delta}W_1 \perp\!\!\!\perp X_2+\sqrt{\delta}W_2,$$

which by Lemma A.2 implies that, conditioned on U^*, we have

$$X_1 \perp\!\!\!\perp X_2,$$

and concludes the proof of (a).

Having proved (a), rewrite equation (34), with U^* for U, as

$$s_{\epsilon,\delta}(X_1,X_2|U^*) = s_{\epsilon,\delta}(X_1|U^*)+s_{\epsilon,\delta}(X_2|U^*). \tag{41}$$

The above inequality, combined with the assumed equality $s_{\epsilon,\delta}(X_1,X_2|U^*) = S_{\epsilon,\delta}(X_1)+S_{\epsilon,\delta}(X_2)$, immediately yields

$$\begin{aligned}s_{\epsilon,\delta}(X_1|U^*) &= S_{\epsilon,\delta}(X_1) \quad , \text{ and}\\ s_{\epsilon,\delta}(X_2|U^*) &= S_{\epsilon,\delta}(X_2).\end{aligned}$$

□

4.3 Proof of Theorem 3

Having proved the key subadditivity step, the rest of the proof closely follows the steps outlined in [22, Appendix II].

Definition 10. Let $\Sigma := \mathrm{Diag}(\Sigma_1,\Sigma_2,\ldots,\Sigma_k)$ be an $n\times n$ block diagonal matrix such that each Σ_i is an $r_i\times r_i$ positive definite matrix. For $\epsilon,\delta>0$, define

$$v(\Sigma) := \sup_{X\in\mathcal{P}(\mathbf{r}),\mathbb{E}XX^T=\Sigma} s_{\epsilon,\delta}(X), \quad \text{and} \tag{42}$$

$$V(\Sigma) := \sup_{X\in\mathcal{P}(\mathbf{r}),\mathbb{E}XX^T\preceq\Sigma} S_{\epsilon,\delta}(X), \tag{43}$$

where $\preceq$ denotes ordering in the positive semidefinite partial order.

Lemma 4.3. *There exist random variables X^* and U^* satisfying (1) $|\mathcal{U}^*| \le \sum_{i=1}^k \frac{r_i(r_i+1)}{2} + 1$; (2) $X^* \in \mathcal{P}(\mathbf{r})$; and (3) $\mathbb{E}X^*X^{*T} \preceq \Sigma$, such that the following holds:*

$$V(\Sigma) = s_{\epsilon,\delta}(X^* \mid U^*). \tag{44}$$

Proof of Lemma 4.3. Let $(X^{(t)}, t \ge 1)$ be a sequence of random variables such that $\mathbb{E}X^{(t)}(X^{(t)})^T = \widehat{\Sigma}$ and $s_{\epsilon,\delta}(X^{(t)}) \uparrow v(\widehat{\Sigma})$ as $t \to \infty$. This sequence of random variables is tight due to the covariance constraint [22, Proposition 17], and thus we may assume without loss of generality that the $X^{(t)}$ converge weakly to a random variable $X^{\widehat{\Sigma}}$ as $t \to \infty$. Since $X^{(t)} + \sqrt{\delta}W$ satisfies the necessary regularity conditions as in [22, Proposition 18], we also have $h(X_i^{(t)} + \sqrt{\delta}W_i) \to h(X_i^{\widehat{\Sigma}} + \sqrt{\delta}W_i)$ for $i \in [k]$, and $h(A_j(X^{(t)} + \sqrt{\delta}W) + \sqrt{\epsilon}Z_j) \to h(A_j(X^{\widehat{\Sigma}} + \sqrt{\delta}W) + \sqrt{\epsilon}Z_j)$ for $j \in [m]$. Hence we may conclude $s_{\epsilon,\delta}(X^{\widehat{\Sigma}}) = v(\widehat{\Sigma})$.

Recall that $V(\Sigma)$ is defined as

$$\begin{aligned} V(\Sigma) &= \sup_{X \in \mathcal{P}(\mathbf{r}), \mathbb{E}XX^T \preceq \Sigma} S_{\epsilon,\delta}(X) \\ &= \sup_{(U,X), p_{X|U}(\cdot|U) \in \mathcal{P}(\mathbf{r}), \mathbb{E}XX^T \preceq \Sigma} s_{\epsilon,\delta}(X \mid U) \\ &\overset{(a)}{=} \sup_{\alpha_l \ge 0, \widehat{\Sigma}_l : \sum_{l=1}^M \alpha_l = 1, \sum_{l=1}^M \alpha_l \widehat{\Sigma}_l \preceq \Sigma} \sum_{l=1}^M \alpha_l v(\widehat{\Sigma}_l), \end{aligned} \tag{45}$$

where, for the moment, M ranges over positive integers of arbitrary size. The equality in (a) is because, we may restrict $p_{X|U}(\cdot|U)$ to the class of optimizers $X^{\widehat{\Sigma}}$ for $\widehat{\Sigma} \succeq 0$. We now show that we can fix M to be $\sum_{i=1}^k \binom{r_i+1}{2} + 1$ in (45). Let $\mathcal{T}$ denote the connected subset of positive definite matrices Σ of the form $\mathrm{Diag}(\Sigma_1, \ldots, \Sigma_k)$ where Σ_i is an $r_i \times r_i$ positive definite matrix for $i \in [k]$. Consider the connected subset, $\mathcal{V}$, of the M-dimensional Euclidean space obtained using the continuous mapping $\Phi : \mathcal{T} \mapsto \mathbb{R}^M$, defined by $\Phi(\Sigma) = (\{\Sigma_i(j,k)_{1 \le j \le k \le r_i}\}, v(\Sigma))$. Bunt's extension of Carathéodory's Theorem [34] states that any finite convex combination of points in $\mathcal{V}$, can be represented as a convex combination of at most M points in $\mathcal{V}$. Hence for any (U, X^{Σ_U}) we can find a pair $(U', X^{\Sigma_{U'}})$ with U' taking at most M values, such that $E(\Sigma_U) = E(\Sigma_{U'})$ and $E(v(\Sigma_U)) = E(v(\Sigma_{U'}))$. Thus from this point onwards in the proof we define $M := \sum_{i=1}^k \binom{r_i+1}{2} + 1$.

Consider any sequence of convex combinations $\left(\{\alpha_l^{(t)}\}_{l=1}^M, \{\widehat{\Sigma}_l^{(t)}\}_{l=1}^M\right)$ with $\sum_{l=1}^m \alpha_l^{(t)} \widehat{\Sigma}_l^{(t)} \preceq \Sigma$ for all $t \ge 1$, and such that $\sum_{l=1}^m \alpha_l^{(t)} v(\widehat{\Sigma}_l^{(t)})$ converges to $v(\Sigma)$ as $t \to \infty$. Appealing to the compactness of the M-dimensional simplex, we may assume without loss of generality that $\alpha_l^{(t)} \to \alpha_l^*$ for all $i \in [M]$. If any of the α_l^* equals 0, then noticing that $\alpha_l^{(t)} \widehat{\Sigma}_l^{(t)} \preceq \Sigma$ gives us

$$\begin{aligned} v(\widehat{\Sigma}_l^{(t)}) &\overset{(a)}{\le} \sum_{i=1}^k \frac{d_i}{2} \log(2\pi e)^{r_i} |\widehat{\Sigma}_{li}^{(t)} + \delta I_{r_i \times r_i}| - \sum_{j=1}^m \frac{c_j n_j}{2} \log(2\pi e \epsilon) \\ &\le \sum_{i=1}^k \frac{d_i}{2} \log(2\pi e)^{r_i} \left| \frac{\Sigma_{li}}{\alpha_l^{(t)}} + \delta I_{r_i \times r_i} \right| - \sum_{j=1}^m \frac{c_j n_j}{2} \log(2\pi e \epsilon) \\ &= \sum_{i=1}^k \frac{d_i}{2} \log \left| \frac{\Sigma_{li}}{\alpha_l^{(t)}} + \delta I_{r_i \times r_i} \right| + C_0, \end{aligned}$$

where C_0 is some constant that does not depend on t. In (a), we used the fact that each $h(X_i + \sqrt{\delta}W_i)$ is upper-bounded by the entropy of a Gaussian random variable with the same covariance matrix as $X_i + \sqrt{\delta}W_i$, and $h(A_j(X + \sqrt{\delta}W) + \sqrt{\epsilon}Z_j) \ge h(\sqrt{\epsilon}Z_j)$.

It is now clear that the limit $\alpha_l^{(t)} v(\widehat{\Sigma}_l^{(t)})$ as $t \to \infty$ is equal to 0 whenever $\alpha_l^{(t)} \to 0$. Thus, we may assume that $\min_{l\in[M]} \alpha_l^* = \alpha_{\min} > 0$. This implies that $\widehat{\Sigma}_l^{(t)} \preceq \frac{2\Sigma}{\alpha_{\min}}$ for all large enough t. Hence, we can find a convergent subsequence such that $\widehat{\Sigma}_l^{(t)} \to \Sigma_l^*$ for each $l \in [M]$ when $t \to \infty$ along this subsequence. Hence we arrive at

$$V(\Sigma) = \sum_{l=1}^{M} \alpha_l^* v(\Sigma_l^*), \tag{46}$$

or, in other words, we can find a pair of random variables (X^*, U^*) with $|\mathcal{U}^*| \le M$ such that $V(\Sigma) = s_{\epsilon,\delta}(X^*|U^*)$. This completes the proof. □

Lemma 4.4. *Consider random variables (X_1, X_2, U) such that $(X_1, X_2) \in \mathcal{P}(2\mathbf{r})$ for some $\mathbf{r}$-partition of $n > 0$. Define new random variables X_+ and X_- via*

$$X_+ := \frac{X_1 + X_2}{\sqrt{2}}, \quad \text{and} \quad X_- := \frac{X_1 - X_2}{\sqrt{2}}.$$

Then $s_{\epsilon,\delta}(X_1, X_2|U) = s_{\epsilon,\delta}(X_+, X_-|U)$.

Proof. We have the equality

$$\begin{aligned} s_{\epsilon,\delta}(X_1, X_2|U) = &\sum_{i=1}^{k} d_i h(X_{i1} + \sqrt{\delta}W_{i1}, X_{i2} + \sqrt{\delta}W_{i2}|U) \\ &- \sum_{j=1}^{m} c_j h(A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}, A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}|U), \end{aligned} \tag{47}$$

Further, defining $W_{i+} := \frac{W_{i1}+W_{i2}}{\sqrt{2}}$, $W_{i-} := \frac{W_{i1}-W_{i2}}{\sqrt{2}}$, $Z_{j+} := \frac{Z_{j1}+Z_{j2}}{\sqrt{2}}$, and $Z_{j-} := \frac{Z_{j1}-Z_{j2}}{\sqrt{2}}$, we have

$$h(X_{i1} + \sqrt{\delta}W_{i1}, X_{i2} + \sqrt{\delta}W_{i2}|U) = h(X_{i+} + \sqrt{\delta}W_{i+}, X_{i-} + \sqrt{\delta}W_{i-}|U), \tag{48}$$

and

$$\begin{aligned} &h(A_j(X_1 + \sqrt{\delta}W_1) + \sqrt{\epsilon}Z_{j1}, A_j(X_2 + \sqrt{\delta}W_2) + \sqrt{\epsilon}Z_{j2}|U) \\ &= h(A_j(X_+ + \sqrt{\delta}W_+) + \sqrt{\epsilon}Z_{j+}, A_j(X_- + \sqrt{\delta}W_-) + \sqrt{\epsilon}Z_{j-}|U). \end{aligned} \tag{49}$$

(W_1, W_2, Z_1, Z_2) and (W_+, W_-, Z_+, Z_-) are equal in distribution. Multiplying the equations in (48) by d_i and those in (49) by c_j and subtracting the sum of the latter from the sum of the former, we may conclude that $s_{\epsilon,\delta}(X_1, X_2|U) = s_{\epsilon,\delta}(X_+, X_-|U)$. □

Lemma 4.5. *Fix $\epsilon, \delta > 0$. Let the random variables X^* and U^* be as in Lemma 4.3; i.e., satisfying the equality $V(\Sigma) = s_{\epsilon,\delta}(X^*|U^*)$, and with $|\mathcal{U}^*| \le M$. Consider two independent and identically distributed copies of (X^*, U^*), denoted by (X_1, U_1) and (X_2, U_2). Define new random variables X_+ and X_- as follows:*

$$X_+ := \frac{X_1 + X_2}{\sqrt{2}}, \quad \text{and} \quad X_- := \frac{X_1 - X_2}{\sqrt{2}}.$$

Also, define $U := (U_1, U_2)$. Then the following results hold:

(a) X_+ *and* X_- *are conditionally independent given* U*,*

(b) $V(\Sigma) = s_{\epsilon,\delta}(X_+|U)$ *and* $V(\Sigma) = s_{\epsilon,\delta}(X_-|U)$*.*

Proof. We have the following sequence of inequalities:

$$\begin{aligned} 2V(\Sigma) &\overset{(a)}{=} s_{\epsilon,\delta}(X_1|U_1) + s_{\epsilon,\delta}(X_2|U_2) \\ &\overset{(b)}{=} s_{\epsilon,\delta}(X_1, X_2|U_1, U_2) \\ &\overset{(c)}{=} s_{\epsilon,\delta}(X_+, X_-|U_1, U_2) \\ &\overset{(d)}{\leq} S_{\epsilon,\delta}(X_+, X_-) \\ &\overset{(e)}{\leq} S_{\epsilon,\delta}(X_+) + S_{\epsilon,\delta}(X_-) \\ &\overset{(f)}{\leq} V(\Sigma) + V(\Sigma) = 2V(\Sigma). \end{aligned}$$

Here (a) follows from the assumption that $s_{\epsilon,\delta}(X^*|U^*) = V(\Sigma)$. Equality (b) follows from the independence $(X_1, U_1) \perp\!\!\!\perp (X_2, U_2)$. Equality (c) holds because of Lemma 4.4. Inequality (d) follows from the definition of $S_{\epsilon,\delta}(\cdot)$. Inequality (e) follows from the tensorization result in Lemma 4.1. Finally, inequality (f) follows from the definition in equation (43), and the fact that X_+ and X_- have the same covariance as X^*, which is bounded above by Σ in the positive semidefinite partial order.

Since the first and last expressions match, all the inequalities in the above sequence of inequalities must be equalities. In particular, equalities (d) and (e) combined with Lemma 4.2 imply that $X_+ \perp\!\!\!\perp X_-$ conditioned on (U_1, U_2), thus establishing part (a) of the lemma. Lemma 4.2 also gives $s_{\epsilon,\delta}(X_+|U_1, U_2) = S_{\epsilon,\delta}(X_+)$ and $s_{\epsilon,\delta}(X_-|U_1, U_2) = S_{\epsilon,\delta}(X_-)$. Finally, equality in (f) gives $S_{\epsilon,\delta}(X_+) = V(\Sigma)$ and $S_{\epsilon,\delta}(X_-) = V(\Sigma)$. This completes the proof of part (b). □

Lemma 4.6. *There exists* $G^* \sim \mathcal{N}(0, \Sigma^*) \in \mathcal{P}(\mathbf{r})$ *such that* $\Sigma^* \preceq \Sigma$ *and* $V(\Sigma) = s_{\epsilon,\delta}(G^*)$*. Furthermore, the random variable* G^* *is the unique element of the set* $\mathcal{P}(\mathbf{r}) \cap \{X : \mathbb{E}XX^T \preceq \Sigma\}$ *satisfying* $s_{\epsilon,\delta}(X) = V(\Sigma)$*.*

Proof. Consider the setting as in Lemma 4.5. Using Lemma 4.5, we have that $X_+ \perp\!\!\!\perp X_-$ conditioned on $U = (u_1, u_2)$ for any $u_1, u_2 \in \mathcal{U}^*$. However, we also have $X_1 \perp\!\!\!\perp X_2$ conditioned on $U = (u_1, u_2)$. The characterization theorem for Gaussian distributions [33] implies that X_1 and X_2 must be Gaussian with identical covariance matrices, conditioned on $U = (u_1, u_2)$. Recall that (X_1, U_1) is independent of (X_2, U_2), and the covariance matrix of X_i conditioned on $U = (u_1, u_2)$ is simply the covariance matrix of X_i conditioned on $U_i = u_i$ for $i \in \{1, 2\}$. Since u_1 and u_2 may be chosen arbitrarily, we conclude that the covariance matrix of X_1 is some fixed $\Sigma^* \preceq \Sigma$ for all $u_1 \in \mathcal{U}^*$. Let $G^* \sim \mathcal{N}(0, \Sigma^*)$. Thus,

$$\begin{aligned} V(\Sigma) &= \sum_{u_1 \in \mathcal{U}^*} p_{U_1}(u_1) s_{\epsilon,\delta}(X_1|U_1 = u_1) \\ &= \sum_{u_1 \in \mathcal{U}^*} p_{U_1}(u_1) s_{\epsilon,\delta}(G^*) \\ &= s_{\epsilon,\delta}(G^*). \end{aligned}$$

To establish uniqueness, first note that it is enough to only consider Gaussian random variables X satisfying $s_{\epsilon,\delta}(X) = V(\Sigma)$, since our argument above shows that any X that achieves this equality

must be Gaussian. Now suppose that $G_1 \sim \mathcal{N}(0, \Sigma_1)$ and $G_2 \sim \mathcal{N}(0, \Sigma_2)$ are two distinct random variables such that $s_{\epsilon,\delta}(G_1) = s_{\epsilon,\delta}(G_2) = V(\Sigma)$ with $\Sigma_1, \Sigma_2 \preceq \Sigma$. Define (X, U) such that $X = G_1$ when $U = 1$ and $X = G_2$ when $U = 2$. Suppose also that U takes values 1 and 2 with probability $1/2$, each. It is easy to check that X satisfies the covariance constraint, and that $s_{\epsilon,\delta}(X|U) = V(\Sigma)$. As in Lemma 4.5, consider two i.i.d. copies of (X_1, U_1) and (X_2, U_2) of (X, U). Lemma 4.5 states that conditioned on $(U_1 = u_1, U_2 = u_2)$, we have $X_1 + X_2 \perp\!\!\!\perp X_1 - X_2$, for any values of u_1 and u_2. Conditioned on $u_1 = 1$ and $u_2 = 2$, we have $X_1 + X_2 = G_1 + G_2$ and $X_1 - X_2 = G_1 - G_2$. This implies $G_1 + G_2 \perp\!\!\!\perp G_1 - G_2$, which is impossible since $\Sigma_1 \neq \Sigma_2$, and thus there cannot be two distinct Gaussian maximizers. □

Proof of Theorem 3. We now complete the proof of Theorem 3. Recall the definition of M_g:

$$M_g := \sup_{X \in \mathcal{P}_g(\mathbf{r})} \sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X).$$

Clearly, there is nothing to prove if M_g is infinite, so we assume $M_g < \infty$. Let $X \in \mathcal{P}(\mathbf{r})$ be an arbitrary random vector. By choosing a large enough Σ such that $\mathbb{E}XX^T \preceq \Sigma$, we may conclude that

$$s_{\epsilon,\delta}(X) \leq V(\Sigma). \tag{50}$$

Let $G^* \sim \mathcal{N}(0, \Sigma^*) \in \mathcal{P}(\mathbf{r})$, where $\Sigma^* \preceq \Sigma$, be the unique maximizer such that $s_{\epsilon,\delta}(G^*) = V(\Sigma)$, as in Lemma 4.6. Thus, we have the sequence of inequalities

$$\begin{aligned} V(\Sigma) &= \sum_{i=1}^{k} d_i h(G_i^* + \sqrt{\delta} W_i) - \sum_{j=1}^{m} c_j h(A_j(G^* + \sqrt{\delta} W) + \sqrt{\epsilon} Z_j) \\ &\overset{(a)}{\leq} \sum_{i=1}^{k} d_i h(G_i^* + \sqrt{\delta} W_i) - \sum_{j=1}^{m} c_j h(A_j(G^* + \sqrt{\delta} W)) \\ &\overset{(b)}{\leq} M_g. \end{aligned} \tag{51}$$

Here, inequality (a) follows from the entropy inequality

$$h(A_j(G^* + \sqrt{\delta} W) + \sqrt{\epsilon} Z_j) \geq h(A_j(G^* + \sqrt{\delta} W)),$$

for all $j \in [m]$. The inequality in (b) is true because the random variable $\tilde{G}^*$ defined by $\tilde{G}_i^* := G_i^* + \sqrt{\delta} W_i$ for $i \in [k]$ is a Gaussian random variable in $\mathcal{P}_g(\mathbf{r})$. Thus, by the definition of M_g, we must have

$$\begin{aligned} \sum_{i=1}^{k} d_i h(\tilde{G}_i^*) - \sum_{j=1}^{m} c_j h(A_j \tilde{G}^*) &\leq \sup_{X \in \mathcal{P}_g(\mathbf{r})} \sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X) \\ &= M_g. \end{aligned}$$

Combining inequalities (50) and (51), we have

$$s_{\epsilon,\delta}(X) \leq M_g. \tag{52}$$

Recall that $s_{\epsilon,\delta}(X)$ depends on δ, ϵ since

$$s_{\epsilon,\delta}(X) = \sum_{i=1}^{k} d_i h(X_i + \sqrt{\delta} W_i) - \sum_{j=1}^{m} c_j h(A_j(X + \sqrt{\delta} W) + \sqrt{\epsilon} Z_j).$$

If X satisfies certain mild conditions (e.g. bounded second moments) provided in Lemma A.3, we have that

$$\lim_{\epsilon,\delta\to 0} s_{\epsilon,\delta}(X) = \sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X).$$

This means that we may take the limit in inequality (52) as $\epsilon, \delta \to 0$ to conclude

$$\sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X) \leq M_g,$$

and conclude the proof of Theorem 3. □

5 Conditions for $M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d}) < \infty$

Theorem 3 shows that it is enough to find necessary and sufficient conditions for $M_g(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ to be finite, since $M = M_g$. We prove Theorem 4 by finding necessary conditions on the BL-EPI datum for such finiteness in Claim 5.1, and showing that the necessary conditions are also sufficient in Claim 5.2.

Claim 5.1. *If $M_g(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ is finite, then the conditions in equations (14) and (15) must be satisfied.*

Proof. The necessity of the condition in equation (15) is seen as follows. Choose $Z \sim \lambda\mathcal{N}(0, I_{n\times n})$ for some $\lambda > 0$. It is easy to see that $\sum_{i=1}^{k} d_i h(Z_i) - \sum_{j=1}^{m} c_j h(A_j Z)$ scales as $\left(\sum_{i=1}^{k} d_i r_i - \sum_{j=1}^{m} c_j n_j\right) \log(\lambda)$ as a function of λ as $\lambda \to \infty$. Since λ is arbitrary, the above expression is finite only if the condition in equation (15) is satisfied.

To show that the condition in equation (14) is necessary, let V be a subspace of $\mathbb{R}^n$ of $\mathbf{r}$-product form. Consider a Gaussian random variable $Z := (Z_V, Z_{V^\perp})$ such that $Z_V \perp\!\!\!\perp Z_{V^\perp}$, and Z_V is supported on V and $Z_{V^\perp}$ is supported on $V^\perp$. Furthermore, assume $Z_V \sim \mathcal{N}(0, \lambda I_{\dim(V)\times\dim(V)})$ and $Z_{V^\perp} \sim \mathcal{N}(0, I_{\dim(V^\perp\times\dim(V^\perp)})$. Taking the limit as $\lambda \to \infty$ and gathering the coefficients of $\log\lambda$, we see that $M_g(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ scales as

$$\left(\sum_{i=1}^{k} d_i \dim(V_i) - \sum_{j=1}^{k} c_j \dim(A_j V)\right) \log(\lambda),$$

as $\lambda \to \infty$. Thus, M_g is finite only if the condition in equation (14) is satisfied. □

The proof of sufficiency of the conditions in equations (14) and (15) relies on two lemmas which we prove below.

Lemma 5.1. *Let $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ be a BL-EPI datum. Let $U := (U_1, \ldots, U_k)$ be an arbitrary $\mathbf{r}$-product form subspace such that $\dim(U_i) = \tilde{r}_i \leq r_i$ for $i \in [k]$. Let $\tilde{\mathbf{r}} := (\tilde{r}_1, \ldots, \tilde{r}_k)$ and $\tilde{\mathbf{r}}^c := \mathbf{r} - \tilde{\mathbf{r}}$. Define two BL-EPI data as follows:*

(a) $(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d})$ is a BL-EPI datum defined on U. For each $j \in [m]$, define the linear maps $\tilde{A}_j : U \to (A_j U)$ by $\tilde{A}_j x = A_j x$ for $x \in U$.

(b) $(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d})$ *is a BL-EPI datum defined on* $U^\perp$*. For* $j \in [m]$*, the linear maps* $\tilde{\tilde{A}}_j : U^\perp \to (A_jU)^\perp$ *are defined by*

$$\tilde{\tilde{A}}_j x = \Pi_{(A_jU)^\perp} A_j x.$$

We also define the linear maps $\Gamma_j : U^\perp \to (A_jU)$ *as*

$$\Gamma_j x = \Pi_{(A_jU)} A_j x.$$

Here Π_V *denotes the projection on to a subspace* V*. Note that* $A_j x = \tilde{\tilde{\mathbf{A}}} x + \Gamma_j x$*.*

Then the following relation holds:

$$M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d}) \le M(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d}) + M(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d}). \tag{53}$$

Remark 5.1. *Note that it may happen that* $\dim(U_i) = 0$ *for some* $i \in [k]$*. It may also happen that for some* $j \in [m]$*, we have* $\dim((A_jU)^\perp) = 0$*. We do not rule out such cases, and keep our notation the same by instead defining entropy on a 0-dimensional subspace as 0.*

Proof of Lemma 5.1. By definition, the linear transformations in $\tilde{\mathbf{A}}$ and $\tilde{\tilde{\mathbf{A}}}$ are surjective. Also, $\sum_i \tilde{r}_i = \dim(U)$ and $\sum_i \tilde{r}_i^c = \dim(U^\perp)$. This verifies that $(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d})$ and $(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d})$ are indeed valid BL-EPI data on U and $U^\perp$, respectively. Every vector $x \in \mathbb{R}^n$ may be expressed as $x = \Pi_U x + \Pi_{U^\perp} x := \tilde{x} + \tilde{\tilde{x}}$. We use the notation $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_k)$ where $\tilde{x}_i = \Pi_{U_i} x_i$, and similarly for $\tilde{\tilde{x}}_i$. We have the equality

$$\begin{aligned} A_j x &= A_j(\Pi_U x + \Pi_{U^\perp} x) \\ &= A_j(\Pi_U x) + A_j(\Pi_{U^\perp} x) \\ &= \tilde{A}_j \tilde{x} + \Pi_{(A_jU)} A_j \tilde{\tilde{x}} + \Pi_{(A_jU)^\perp} A_j \tilde{\tilde{x}} \\ &= \tilde{A}_j \tilde{x} + \Gamma_j \tilde{\tilde{x}} + \tilde{\tilde{A}}_j \tilde{\tilde{x}}. \end{aligned}$$

For any $X \in \mathcal{P}(\mathbf{r})$,

$$\begin{aligned} \sum_{i=1}^k d_i h(X_i) - \sum_{j=1}^m c_j h(A_j X) &= \sum_{i=1}^k d_i h(X_i) - \sum_{j=1}^m c_j h(\tilde{A}_j \tilde{X} + \Gamma_j \tilde{\tilde{X}} + \tilde{\tilde{A}}_j \tilde{\tilde{X}}) \\ &= \sum_{i=1}^k d_i h(\tilde{X}_i, \tilde{\tilde{X}}_i) - \sum_{j=1}^m c_j h(\tilde{A}_j \tilde{X} + \Gamma_j \tilde{\tilde{X}}, \tilde{\tilde{A}}_j \tilde{\tilde{X}}) \\ &= \sum_{i=1}^k d_i h(\tilde{\tilde{X}}_i) - \sum_{j=1}^m c_j h(\tilde{\tilde{A}}_j \tilde{\tilde{X}}) + \sum_{i=1}^k d_i h(\tilde{X}_i \mid \tilde{\tilde{X}}_i) - \sum_{j=1}^m c_j h(\tilde{A}_j \tilde{X} + \Gamma_j \tilde{\tilde{X}} \mid \tilde{\tilde{A}}_j \tilde{\tilde{X}}) \\ &\le \sum_{i=1}^k d_i h(\tilde{\tilde{X}}_i) - \sum_{j=1}^m c_j h(\tilde{\tilde{A}}_j \tilde{\tilde{X}}) + \sum_{i=1}^k d_i h(\tilde{X}_i \mid \tilde{\tilde{X}}_i) - \sum_{j=1}^m c_j h(\tilde{A}_j \tilde{X} \mid \tilde{\tilde{X}}) \\ &= \sum_{i=1}^k d_i h(\tilde{\tilde{X}}_i) - \sum_{j=1}^m c_j h(\tilde{\tilde{A}}_j \tilde{\tilde{X}}) + \sum_{i=1}^k d_i h(\tilde{X}_i \mid \tilde{\tilde{X}}) - \sum_{j=1}^m c_j h(\tilde{A}_j \tilde{X} \mid \tilde{\tilde{X}}) \\ &\le M(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d}) + M(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d}). \end{aligned}$$

Taking the supremum over all $X \in \mathcal{P}(\mathbf{r})$ completes the proof. ∎

Lemma 5.2. *Suppose that a BL-EPI datum* $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ *satisfies the conditions in equations* (14) *and* (15)*, and suppose that* U *is an* $\mathbf{r}$*-product form critical subspace. Then the BL-EPI data* $(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d})$ *and* $(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d})$ *defined as in Lemma 5.1 also satisfy the conditions in equations* (14) *and* (15)*.*

Proof. Verifying the conditions for $(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d})$ is immediate: the condition in equation (14) restricted to $\tilde{\mathbf{r}}$ product form subspaces of U yields the first condition, and the criticality of U yields the second condition.

For $j \in [m]$, it is not hard to verify that $\dim(\tilde{\tilde{A}}_j U^\perp)$ is $n_j - \dim(\tilde{A}_j U)$. We may now check the second condition for $(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d})$ by observing the equality

$$\sum_{i=1}^{k} d_i(r_i - \dim(U_i)) = \sum_{j=1}^{m} c_j(n_j - \dim(\tilde{A}_j U)),$$

using the criticality of U and the fact that $\sum_{i=1}^{k} d_i r_i = \sum_{j=1}^{m} c_j n_j$. Let V be an arbitrary $\tilde{\mathbf{r}}^c$-product form subspace of $U^\perp$. Consider the new subspace $V_+ = V \oplus U \subset \mathbb{R}^n$, which is the direct sum of the subspace V with the subspace U. Note that V_+ is an $\mathbf{r}$-product form subspace of $\mathbb{R}^n$. Using the condition in equation (14) for V_+, we have

$$\sum_{i=1}^{k} d_i \dim(V_{+i}) \leq \sum_{j=1}^{m} c_j \dim(A_j V_+).$$

Note that $\dim(V_{+i}) = \dim(V_i) + \dim(U_i)$, for all $1 \leq i \leq k$. Moreover, $\dim(A_j V_+) = \dim(A_j U) + \dim(\tilde{\tilde{A}}_j V_i)$. Substituting these equalities in the above inequality, we arrive at

$$\sum_{i=1}^{k} d_i(\dim(V_i) + \dim(U_i)) \leq \sum_{j=1}^{m} c_j(\dim(A_j U) + \dim(\tilde{\tilde{A}}_j V_i)).$$

The criticality of U then implies

$$\sum_{i=1}^{k} d_i \dim(V_i) \leq \sum_{j=1}^{m} c_j \dim(\tilde{\tilde{A}}_j V_i),$$

and this completes the proof. □

We are now in a position to prove the following sufficiency result:

Claim 5.2. *If the conditions in equations* (14) *and* (15) *are satisfied, then* $M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ *is finite.*

Proof. The proof proceeds via a double induction on the dimension n and the number of linear maps m. We first prove the result for $n = 1$ and arbitrary m, and for $m = 1$ and arbitrary n. For $n = 1$, it must be that $\mathbf{r} = \{1\}$ and $\mathbf{d} = \{d_1\}$. The conditions in equations (14) and (15) imply that $d_1 = \sum_{j=1, n_j > 0}^{m} c_j$, because $n_j > 0 \implies n_j = 1$. Thus, $M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ equals

$$\begin{aligned}
\sup_{X \in \mathcal{P}(\mathbf{r})} d_1 h(X) - \sum_{j=1}^{m} c_j h(A_j X) &= \sup_{X \in \mathcal{P}(\mathbf{r})} d_1 h(X) - \sum_{j=1, n_j > 0}^{m} c_j h(A_j X) \\
&= \sup_{X \in \mathcal{P}(\mathbf{r})} - \sum_{j=1, n_j > 0}^{m} c_j \log |A_j| \\
&= - \sum_{j=1, n_j > 0}^{m} c_j \log |A_j| < \infty,
\end{aligned}$$

since $h(A_jX) = h(X) + \log|A_j|$ for all $j \in [m]$ such that $n_j > 0$, and A_j is a nonzero scalar for each such j.

Now fix $m = 1$ and let $n > 0$, $k > 0$, $\mathbf{r}$, and $\mathbf{d}$ be arbitrary. It must be that $\mathbf{c} = \{c_1\}$ and $\mathbf{A} = \{A_1\}$ such that A_1 is a full-rank $n \times n$ matrix. Condition (14) applied to the subspaces $V_i := \phi \times \cdots \times \phi \times \mathbb{R}^{r_i} \times \phi \times \ldots \phi$ for $i \in [k]$ imply that:

$$d_i \leq c_1 \quad \text{for all } i \in [k],$$

where we used the fact that $\dim(A_1V_i) = \dim(V_i) = r_i$. Combined with the condition in equation (15), this forces the equality

$$d_i = c_1 \quad \text{for all } i \in [k].$$

Thus,

$$\begin{aligned}\sup_{X\in\mathcal{P}(\mathbf{r})} \sum_{i=1}^{k} d_i h(X_i) - c_1 h(A_1X) &= \sup_{X\in\mathcal{P}(\mathbf{r})} \sum_{i=1}^{k} c_1 h(X_i) - c_1 h(X) - c_1 \log|\det(A_1)| \\ &= -c_1 \log|\det(A_1)| < \infty.\end{aligned}$$

We have shown that the claim is true for $n = 1$ and all m. Assume that claim is true for all $n < n_0$ and all m. Our goal is to establish the claim for $n = n_0$ and all $m > 0$. To do so, we induct on m. The case of $n = n_0$ and $m = 1$ follows from our calculations above. Now we assume that the claim is true for $n = n_0$ and all $m < m_0$, and show that it also holds for $n = n_0$ and $m = m_0$.

Let $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ be a BL-EPI datum in $\mathbb{R}^{n_0}$ with $m = m_0$. We may assume that $n_j > 0$ for all $j \in [m]$, since otherwise we could have treated the scenario as a BL-EPI datum in $\mathbb{R}^{n_0}$ with $m < m_0$, which is already covered by the inductive hypothesis. For fixed $\mathbf{A}$, $\mathbf{r}$, and $\mathbf{d}$, consider the function defined on $\mathbf{c} \in \mathbb{R}_+^{m_0}$ as

$$M(\mathbf{c}) = \sup_{X\in\mathcal{P}(\mathbf{r})} \sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_jX). \tag{54}$$

Since M is a pointwise supremum of linear functions, M is convex. Let $\mathcal{K}$ be the region of all $\mathbf{c} \in \mathbb{R}_+^{m_0}$ such that $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ satisfy the conditions in equations (14) and (15). Note that $\mathcal{K}$ is a compact, convex set. By Claim 5.1, we have that M takes $+\infty$ values outside $\mathcal{K}$. We wish to show that M takes finite values everywhere on $\mathcal{K}$. Since M is convex and $\mathcal{K}$ is closed, it is enough to show finiteness of M at all points on the boundary of $\mathcal{K}$. Since $n_j > 0$ for all $j \in [m]$, a point $\mathbf{c}$ is a boundary point of $\mathcal{K}$ if and only if at least one of the following two conditions is satisfied: (1) $c_{j_0} = 0$ for some $j_0 \in [m]$; or (2) there exists a proper $\mathbf{r}$-product form subspace of $\mathbb{R}^{n_0}$ that is critical. If a boundary point satisfies (1), then our induction assumption (on m) ensures the finiteness of M evaluated at that BL-EPI datum, since we could have treated the scenario as a BL-EPI datum in $\mathbb{R}^{n_0}$ with $m < m_0$.

Now consider a boundary point that satisfies (2), assuming that $c_j \neq 0$ for all $j \in [m]$. Let $V = (V_1, \ldots, V_k)$ be an $\mathbf{r}$-product form critical subspace of $\mathbb{R}^{n_0}$; i.e., a subspace that satisfies the equality

$$\sum_{i=1}^{k} d_i \dim(V_i) = \sum_{j=1}^{m} c_j \dim(A_jV), \tag{55}$$

with $\dim(V) < n_0$. Lemma 5.1 shows that given any $\mathbf{r}$-product form subspace V, it is possible to define BL-EPI data on V and $V^\perp$ in terms of the original BL-EPI datum $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ that satisfy a

certain subadditivity property. In particular, if the datum on V is denoted by $(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d})$ and that on $V^\perp$ is denoted by $(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d})$, then Lemma 5.1 states that

$$M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d}) \leq M(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d}) + M(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d}).$$

Thus, to show that $M(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ is finite, is enough to show that $M(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d})$ and $M(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d})$ are finite. Lemma 5.2 asserts that since V is a critical $\mathbf{r}$-product form subspace, the BL-EPI data $(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d})$ and $(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d})$ satisfy both the conditions in equations (14) and (15). Since $\dim(V), \dim(V^\perp) < n_0$, we may use the induction assumption (on the dimension) to assert $M(\tilde{\mathbf{A}}, \mathbf{c}, \tilde{\mathbf{r}}, \mathbf{d}) < \infty$ and $M(\tilde{\tilde{\mathbf{A}}}, \mathbf{c}, \tilde{\mathbf{r}}^c, \mathbf{d}) < \infty$, and conclude the proof.

□

6 A special case

We examine a special case here to see what kinds of new inequalities may result from Theorem 3. Let X_1, X_2, and Y be real valued random variables such that $(X_1, X_2) \perp\!\!\!\perp Y$. We would like to lower bound the entropy $h(X_1 + Y, X_2 + Y)$. Note that the regular EPI applied with the independent random vectors (X_1, X_2) and (Y, Y) yields the trivial lower bound

$$e^{h(X_1+Y, X_2+Y)} \geq e^{h(X_1,X_2)} + e^{h(Y,Y)} = e^{h(X_1,X_2)}.$$

Note also that

$$\begin{pmatrix} X_1 + Y \\ X_2 + Y \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ Y \end{pmatrix}.$$

However, it is not possible to use Zamir and Feder's EPI to provide lower bounds on $h(X_1+Y, X_2+Y)$ because of the dependency between X_1 and X_2. We show that Theorem 3 may be used to obtain a family of nontrivial lower bounds that account for this dependency.

Lemma 6.1. *Let $\alpha, \beta, \delta_1, \delta_2 \geq 0$. Consider the inequality*

$$\alpha h(X_1, X_2) + \beta h(Y) \leq h(X_1 + Y, X_2 + Y) + \delta_1 h(X_1) + \delta_2 h(X_2) + C(\alpha, \beta, \delta_1, \delta_2), \tag{56}$$

where $C(\alpha, \beta, \delta_1, \delta_2)$ is some constant that depends only on α, β, and δ_1, δ_2. The above inequality holds for all $(X_1, X_2) \perp\!\!\!\perp Y$ if and only if $\alpha, \beta, \delta_1, \delta_2$ satisfy the following inequalities:

1. $2\alpha + \beta = 2 + \delta_1 + \delta_2$;
2. $\beta \leq 1$;
3. $\alpha \leq 1 + \delta_1$, *and* $\alpha \leq 1 + \delta_2$;
4. $\alpha + \beta \leq 1 + \delta_1 + \delta_2$, *which, combined with condition (1), is equivalent to* $\alpha \geq 1$.

Proof. We shall use Theorem 4 to show this result. The above inequality is easily seen to be of the form in Theorem 3, where $A_1 = [1, 0, 1; 0, 1, 1]$, $A_2 = [1, 0, 0]$, $A_3 = [0, 1, 0]$, $\mathbf{r} = (2, 1)$, $d_1 = \alpha$, and $d_2 = \beta$. An exhaustive search of all possible subspaces V that are in $\mathbf{r}$-product form where $\mathbf{r} = (2, 1)$ is not hard to do. For simplicity, we refer to the axes in $\mathbb{R}^3$ as X_1, X_2, Y. Thus, the subspace X_1 is simply the subspace spanned by $(1, 0, 0)$.

1. Equality (1) follows directly from equation (15) of Theorem 4;
2. Inequality (2) follows from equation (14) of Theorem 4, by choosing $V = \phi \times Y$;
3. Inequality (3) follows from equation (15) of Theorem 4, by choosing $V = X_1 \times \phi$ and $V = X_2 \times \phi$;
4. Inequality (4) is obtained from equation (15) of Theorem 4, by a careful choice of $V = (X_1 + X_2) \times Y$, i.e. the subspace spanned by $(1, 1, 0)$ and $(0, 0, 1)$.

□

Claim 6.1. *For $\alpha, \beta < 1, \delta_1 = \delta_2 = \delta$ satisfying the conditions in Lemma 6.1, the following inequality holds:*

$$h(X_1 + Y, X_2 + Y) \geq (\alpha - \delta)h(X_1, X_2) + \beta h(Y) - \delta I(X_1; X_2) - D,$$

where

$$D = \frac{1}{2} \log \left(\frac{\beta^\beta (1-\beta)^{1-\beta}}{2^\beta} \left(1 + \frac{\beta}{2\delta}\right)^{\alpha+\beta-1} \left(1 - \frac{\beta}{2\delta}\right)^{\alpha-1} \right).$$

Proof. For $\alpha, \beta, \delta_1, \delta_2$, the optimal constant C is given by

$$\begin{aligned} e^{2C} &= \sup_{K_1,K_2,K_3,\rho} \frac{\left(\det \begin{pmatrix} K_1 & \rho\sqrt{K_1K_2} \\ \rho\sqrt{K_1K_2} & K_2 \end{pmatrix}\right)^\alpha \cdot K_3^\beta}{\det \begin{pmatrix} K_1 + K_3 & \rho\sqrt{K_1K_2} + K_3 \\ \rho\sqrt{K_1K_2} + K_3 & K_2 + K_3 \end{pmatrix} K_1^{\delta_1} K_2^{\delta_2}} \\ &= \sup_{K_1,K_2,K_3,\rho} \frac{K_1^{\alpha-\delta_1} K_2^{\alpha-\delta_2} (1-\rho^2)^\alpha \cdot K_3^\beta}{K_1K_2(1-\rho^2) + K_3(K_1 + K_2 - 2\rho\sqrt{K_1K_2})}. \end{aligned}$$

Calculating the above supremum for arbitrary $\alpha, \beta, \delta_1, \delta_2$ is cumbersome so we assume $\delta_1 = \delta_2 = \delta$. The supremum simplifies to

$$e^{2C} = \sup_{K_1,K_2,K_3,\rho} \frac{(K_1K_2)^{\alpha-\delta}(1-\rho^2)^\alpha \cdot K_3^\beta}{K_1K_2(1-\rho^2) + K_3(K_1 + K_2 - 2\rho\sqrt{K_1K_2})}.$$

For a fixed K_1K_2 and fixed K_3, it is clear that the optimal choice of $K_1 = K_2 = \sqrt{K_1K_2}$ maximizes the above expression. Thus, we assume that $K_1 = K_2 = K$ and obtain

$$e^{2C} = \sup_{K,K_3,\rho} \frac{(K)^{2\alpha-2\delta-1}(1-\rho^2)^\alpha \cdot K_3^\beta}{K(1-\rho^2) + 2K_3(1-\rho)}.$$

Let $x := K_3/K$, and noting that $2\alpha - 2\delta - 1 = 1 - \beta$, we obtain

$$\begin{aligned} e^{2C} &= \sup_{x\geq 0,\rho} \frac{x^\beta (1-\rho^2)^\alpha}{(1-\rho^2) + 2x(1-\rho)} \\ &= \sup_{x\geq 0,\rho} \frac{x^\beta (1-\rho)^\alpha (1+\rho)^\alpha}{(1-\rho)(1+\rho) + 2x(1-\rho)} \\ &= \sup_{x\geq 0,\rho} \frac{x^\beta (1-\rho)^{\alpha-1} (1+\rho)^\alpha}{(1+\rho) + 2x}. \end{aligned}$$

For a fixed ρ, the maximum of the above expression is attained when

$$x = \frac{\beta(1+\rho)}{2(1-\beta)}.$$

Substituting this value of x,

$$\begin{aligned} e^{2C} &= \sup_{\rho} \frac{(1+\rho)^{\alpha}(1-\rho)^{\alpha-1}\left(\frac{\beta(1+\rho)}{2(1-\beta)}\right)^{\beta}}{\frac{\beta(1+\rho)}{(1-\beta)}+(1+\rho)} \\ &= \sup_{\rho} \frac{(1+\rho)^{\alpha}(1-\rho)^{\alpha-1}\left(\frac{\beta(1+\rho)}{2(1-\beta)}\right)^{\beta}}{\frac{1+\rho}{1-\beta}} \\ &= \frac{\beta^{\beta}(1-\beta)^{1-\beta}}{2^{\beta}} \sup_{\rho} (1+\rho)^{\alpha+\beta-1}(1-\rho)^{\alpha-1}. \end{aligned}$$

Differentiating with respect to ρ, the supremum is seen to be attained when $\rho = \frac{\beta}{2\alpha+\beta-2} = \frac{\beta}{2\delta}$. Substituting this, we get

$$e^{2C} = \frac{\beta^{\beta}(1-\beta)^{1-\beta}}{2^{\beta}} \left(1+\frac{\beta}{2\delta}\right)^{\alpha+\beta-1} \left(1-\frac{\beta}{2\delta}\right)^{\alpha-1}.$$

This leads to the entropy inequality

$$\begin{aligned} &h(X_1+Y, X_2+Y) \\ &\geq \alpha h(X_1, X_2) + \beta h(Y) - \delta h(X_1) - \delta h(X_2) - \frac{1}{2}\log\left(\frac{\beta^{\beta}(1-\beta)^{1-\beta}}{2^{\beta}} \left(1+\frac{\beta}{2\delta}\right)^{\alpha+\beta-1} \left(1-\frac{\beta}{2\delta}\right)^{\alpha-1}\right) \\ &= (\alpha-\delta) h(X_1, X_2) + \beta h(Y) - \delta I(X_1; X_2) - \frac{1}{2}\log\left(\frac{\beta^{\beta}(1-\beta)^{1-\beta}}{2^{\beta}} \left(1+\frac{\beta}{2\delta}\right)^{\alpha+\beta-1} \left(1-\frac{\beta}{2\delta}\right)^{\alpha-1}\right). \end{aligned}$$

Notice that the mutual information term $I(X_1; X_2)$ accounts for the dependency between X_1 and X_2. □

7 Conclusion

In this paper, we established a new inequality that unifies the BLI and the EPI by developing a variant of the doubling trick proof strategy. There are several interesting research directions that are worth pursuing. We did not address the questions of extremizability and uniqueness of extremizers in this work. One reason for this is that Theorem 3 is established by taking the limit as ϵ and δ go to 0. When ϵ and δ are strictly bounded away from 0, the extremizer of $s_{\epsilon,\delta}(\cdot)$ under a covariance constraint exists and is a unique Gaussian distribution. However, these existence and uniqueness properties need not hold in the limit as $\epsilon, \delta \to 0$. In general, the doubling trick based strategy is a powerful tool for proving inequalities, but may not always succeed in identifying necessary and sufficient conditions for equality. For this reason, alternate proof strategies that rely on heat flow based arguments [17, 13, 16] or optimal transport methods [21, 35] are worth exploring as well. Finally, although our results generalize the BLI and the EPI to vector random variables with more general independence properties, these independence properties are still quite restrictive. It would be interesting to establish similar entropy inequalities under weaker independence conditions.

Acknowledgements

The research of VA was supported by the NSF grants CNS-1527846 and CCF-1618145, the NSF Science & Technology Center grant CCF-0939370 (Science of Information), and the William and Flora Hewlett Foundation supported Center for Long Term Cybersecurity at Berkeley. VJ is supported by the NSF grant CCF-1841190, and is grateful to the Department of Information Engineering at CUHK for hosting him in July 2018, when a part of this work was done. The research of CN was supported by GRF grants 14303714, 14231916, 14206518 and a discretionary fund of the Vice Chancellor of CUHK.

References

[1] C. E. Shannon. A mathematical theory of communication, I and II. *Bell Syst. Tech. J*, 27:379–423, 1948.

[2] N. Blachman. The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, 11(2):267–271, 1965.

[3] M. Costa. A new entropy power inequality. *Information Theory, IEEE Transactions on*, 31(6):751–760, 1985.

[4] R. Zamir and M. Feder. A generalization of the entropy power inequality with applications. *IEEE Transactions on Information Theory*, 39(5):1723–1728, 1993.

[5] T. A. Courtade. A strong entropy power inequality. *IEEE Transactions on Information Theory*, 64(4):2173–2192, 2018.

[6] Y. Polyanskiy and Y. Wu. Strong data-processing inequalities for channels and Bayesian networks. In *Convexity and Concentration*, pages 211–249. Springer, 2017.

[7] Y. Polyanskiy and Y. Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

[8] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.

[9] O. Rioul. Information theoretic proofs of entropy power inequalities. *IEEE Transactions on Information Theory*, 57(1):33–55, 2011.

[10] E. H. Lieb. Proof of an entropy conjecture of Wehrl. In *Inequalities*, pages 359–365. Springer, 2002.

[11] R. Zamir and M. Feder. A generalization of information theoretic inequalities to linear transformations of independent vector. In *Proceedings of the 6-th Joint Swedish-Russian International Workshop on Information Theory*, pages 254–258, 1993.

[12] 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*, 20(2):151–173, 1976.

[13] J. Bennett, A. Carbery, M. Christ, and T. Tao. The Brascamp-Lieb inequalities: Finiteness, structure and extremals. *Geometric and Functional Analysis*, 17(5):1343–1415, 2008.

Lemma A.3. *Let X be an $\mathbb{R}^n$-valued random variable with density $p_X(x)$ and $Z \sim \mathcal{N}(0, I_{n\times n})$ be independent of X. Suppose that $\mathbb{E}[\Psi(X)] < \infty$ for some nonnegative continuous function $\Psi : \mathbb{R}^n \mapsto \mathbb{R}$, satisfying $\int_{\mathbb{R}^n} e^{-\Psi(x)}dx < \infty$ and $\lim_{\delta\to 0} \mathbb{E}[\Psi(X + \sqrt{\delta}Z)] = \mathbb{E}(\Psi(X))$. (Note that, for instance, $\Psi(X) = \|X\|_p, p \geq 1$ satisfies the conditions.) Then the following equality holds:*

$$\lim_{\delta\to 0} h(X + \sqrt{\delta}Z) = h(X). \tag{66}$$

Proof. Our proof relies on the following (lower semi-continuity) result from Posner [**?**, Theorem 1]: If P_m, Q_m are Borel probability distributions on a Polish space with $P_m \overset{w}{\Rightarrow} P$ and $Q_m \overset{w}{\Rightarrow} Q$, then

$$D(P\|Q) \leq \liminf_m D(P_m\|Q_m),$$

where $D(P\|Q)$ denotes the relative entropy of the distribution P with respect to the distribution Q. Picking an arbitrary sequence $\{\delta_m\}_{m\geq 1}$ that converges to 0, let $X_m = X + \sqrt{\delta}_m Z$. Using characteristic functions (or otherwise), it is easy to check that X_m converges to X in distribution. Let P_m denote the distribution of X_m, P denote the distribution of X. Let $Q_m = Q$ be the distribution corresponding to the density function $Ce^{-\Psi(x)}$. Note that

$$D(P_m\|Q) = \mathbb{E}[\Psi(X + \sqrt{\delta_m}Z)] - h(X + \sqrt{\delta_m}Z) - \log C.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}[\Psi(X)] - h(X) - \log C \\
&= D(P\|Q) \overset{(a)}{\leq} \liminf_m D(P_m\|Q) \\
&\quad \leq \liminf_m \Big\{\mathbb{E}[\Psi(X + \sqrt{\delta_m}Z)] - h(X + \sqrt{\delta_m}Z) - \log C\Big\}. \\
&\quad \overset{(b)}{=} \mathbb{E}[\Psi(X)] - \limsup_m h(X + \sqrt{\delta_m}Z) - \log C.
\end{aligned}$$

Here (a) follows from the Posner's result and (b) follows from assumption (2). Hence

$$\limsup_{m\to\infty} h(X + \sqrt{\delta_m}Z) \leq h(X). \tag{67}$$

On the other hand, non-negativity of mutual information $I(Z; X + \sqrt{\delta_m}Z) \geq 0$ yields $h(X + \sqrt{\delta}_m Z) \geq h(X)$. Taking the lim inf on both sides of this equality, we conclude

$$\liminf_{m\to\infty} h(X + \sqrt{\delta}_m Z) \geq h(X). \tag{68}$$

Inequalities (**??**) and (**??**) yield the equality

$$\lim_{m\to\infty} h(X + \sqrt{\delta}_m Z) = h(X), \tag{69}$$

and concludes the proof. □

[30] F. Barthe. Optimal Young's inequality and its converse: a simple proof. *Geometric & Functional Analysis GAFA*, 8(2):234–242, 1998.

[31] S. Beigi and C. Nair. Equivalent characterization of reverse Brascamp-Lieb-type inequalities using information measures. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1038–1042. IEEE, 2016.

[32] C. Nair. *An extremal inequality related to hypercontractivity of Gaussian random variables.* Information Theory and Applications Workshop (2014), 2014. Available at http://chandra.ie.cuhk.edu.hk/pub/papers/manuscripts/ITA14.pdf.

[33] S. G. Ghurye and I. Olkin. A characterization of the multivariate normal distribution. *The Annals of Mathematical Statistics*, 33(2):533–541, 1962.

[34] L.N.H. Bunt. *Bijdrage tot de theorie der convexe puntverzamelingen.* PhD thesis, Univ. Groningne, Amsterdam, 1934.

[35] O. Rioul. Yet another proof of the entropy power inequality. *IEEE Transactions on Information Theory*, 63(6):3595–3599, 2017.

[36] E. Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, July 1975.

A Supporting lemmas for Theorem 3

Lemma A.1. *Let X, Y, and Z be random variables taking values in $\mathbb{R}^{n_X}, \mathbb{R}^{n_Y}$, and $\mathbb{R}^{n_Z}$ respectively, such that the following hold: (a) (X, Y, Z) has a strictly positive density on $\mathbb{R}^{n_X+n_Y+n_Z}$; (b) $X \to Y \to Z$; and (c) $X \to Z \to Y$. Then $X \perp\!\!\!\perp (Y, Z)$.*

Proof. For any $x \in \mathbb{R}^{n_X}$, $y \in \mathbb{R}^{n_Y}$, and $z \in \mathbb{R}^{n_Z}$, we have that

$$p_{X|YZ}(x|y,z) = p_{X|Y}(x|y) = p_{X|Z}(x|z), \tag{57}$$

where we used the assumed strict positivity of the density of (X, Y, Z) to write the above equations. Fix $y_0 \in \mathbb{R}^{n_Y}$. For any $z \in \mathbb{R}^{n_Z}$, we have

$$p_{X|Z}(x|z) = p_{X|Y}(x|y_0).$$

Integrating both sides of the above equality with respect to $p_Z(z)$, we obtain

$$p_X(x) = p_{X|Y}(x|y_0).$$

Since y_0 was chosen arbitrarily, we conclude that $X \perp\!\!\!\perp Y$. A similar argument shows that $X \perp\!\!\!\perp Z$. Using equation (57), we conclude that $X \perp\!\!\!\perp (Y, Z)$. □

Lemma A.2. *Let X_1 and X_2 be $\mathbb{R}^n$-valued random variables and let $(Z_1, Z_2) \perp\!\!\!\perp (X_1, X_2)$ be such that $(Z_1, Z_2) \sim \mathcal{N}(0, I_{2n\times 2n})$. If $(X_1 + Z_1) \perp\!\!\!\perp (X_2 + Z_2)$, then $X_1 \perp\!\!\!\perp X_2$.*

Proof. Using the independence of $(X_1 + Z_1)$ and $(X_2 + Z_2)$, we have that for any $t_1, t_2 \in \mathbb{R}^n$,

$$\begin{aligned}
\phi_{X_1+Z_1,X_2+Z_2}(t_1,t_2) &:= \mathbb{E}e^{i\langle t_1, X_1+Z_1\rangle + i\langle t_2, X_2+Z_2\rangle} && (58)\\
&= \mathbb{E}e^{i\langle t_1, X_1+Z_1\rangle}\mathbb{E}e^{i\langle t_2, X_2+Z_2\rangle} && (59)\\
&= \mathbb{E}e^{i\langle t_1, X_1\rangle}\mathbb{E}e^{i\langle t_2, X_2\rangle}\mathbb{E}e^{i\langle t_1, Z_1\rangle}\mathbb{E}e^{i\langle t_2, Z_2\rangle} && (60)\\
&= \phi_{X_1}(t_1)\phi_{X_2}(t_2)\phi_{Z_1,Z_2}(t_1,t_2). && (61)
\end{aligned}$$

However, using the independence $(X_1, X_2) \perp\!\!\!\perp (Z_1, Z_2)$, we also have

$$\phi_{X_1+Z_1,X_2+Z_2}(t_1,t_2) = \mathbb{E}e^{i\langle t_1, X_1+Z_1\rangle + i\langle t_2, X_2+Z_2\rangle} \tag{62}$$

$$= \mathbb{E}e^{i\langle t_1, X_1\rangle + i\langle t_2, X_2\rangle}\mathbb{E}e^{i\langle t_1, Z_1\rangle + i\langle t_2, Z_2\rangle} \tag{63}$$

$$= \phi_{X_1,X_2}(t_1,t_2)\phi_{Z_1,Z_2}(t_1,t_2). \tag{64}$$

Since $\phi_{Z_1,Z_2}(\cdot,\cdot)$ has no zeros (Z_i's being independent standard Gaussian random variables), we conclude that

$$\phi_{X_1,X_2}(t_1,t_2) = \phi_{X_1}(t_1)\phi_{X_2}(t_2), \tag{65}$$

that is, $X_1 \perp\!\!\!\perp X_2$. □

Lemma A.3. *Let X be an $\mathbb{R}^n$-valued random variable with density $p_X(x)$ and $Z \sim \mathcal{N}(0, I_{n\times n})$ be independent of X. Suppose that $\mathbb{E}[\Psi(X)] < \infty$ for some nonnegative continuous function $\Psi : \mathbb{R}^n \mapsto \mathbb{R}$, satisfying $\int_{\mathbb{R}^n} e^{-\Psi(x)}dx < \infty$ and $\lim_{\delta\to 0}\mathbb{E}[\Psi(X+\sqrt{\delta}Z)] = \mathbb{E}(\Psi(X))$. (Note that, for instance, $\Psi(X) = \|X\|_p, p \geq 1$ satisfies the conditions.) Then the following equality holds:*

$$\lim_{\delta\to 0} h(X+\sqrt{\delta}Z) = h(X). \tag{66}$$

Proof. Our proof relies on the following (lower semi-continuity) result from Posner [36, Theorem 1]: If P_m, Q_m are Borel probability distributions on a Polish space with $P_m \overset{w}{\Rightarrow} P$ and $Q_m \overset{w}{\Rightarrow} Q$, then

$$D(P\|Q) \leq \liminf_m D(P_m\|Q_m),$$

where $D(P\|Q)$ denotes the relative entropy of the distribution P with respect to the distribution Q. Picking an arbitrary sequence $\{\delta_m\}_{m\geq 1}$ that converges to 0, let $X_m = X + \sqrt{\delta}_m Z$. Using characteristic functions (or otherwise), it is easy to check that X_m converges to X in distribution. Let P_m denote the distribution of X_m, P denote the distribution of X. Let $Q_m = Q$ be the distribution corresponding to the density function $Ce^{-\Psi(x)}$. Note that

$$D(P_m\|Q) = \mathbb{E}[\Psi(X+\sqrt{\delta_m}Z)] - h(X+\sqrt{\delta_m}Z) - \log C.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}[\Psi(X)] - h(X) - \log C \\
&= D(P\|Q) \overset{(a)}{\leq} \liminf_m D(P_m\|Q) \\
&\leq \liminf_m \Big\{\mathbb{E}[\Psi(X+\sqrt{\delta_m}Z)] - h(X+\sqrt{\delta_m}Z) - \log C\Big\}. \\
&\overset{(b)}{=} \mathbb{E}[\Psi(X)] - \limsup_m h(X+\sqrt{\delta_m}Z) - \log C.
\end{aligned}$$

Here (a) follows from the Posner's result and (b) follows from assumption (2). Hence

$$\limsup_{m\to\infty} h(X+\sqrt{\delta_m}Z) \leq h(X). \tag{67}$$

On the other hand, non-negativity of mutual information $I(Z; X+\sqrt{\delta_m}Z) \geq 0$ yields $h(X+\sqrt{\delta}_m Z) \geq h(X)$. Taking the lim inf on both sides of this equality, we conclude

$$\liminf_{m\to\infty} h(X+\sqrt{\delta}_m Z) \geq h(X). \tag{68}$$

Inequalities (67) and (68) yield the equality

$$\lim_{m\to\infty} h(X+\sqrt{\delta}_m Z) = h(X), \tag{69}$$

and concludes the proof. □