*entropy*

MDPI

*Article*

# A Forward-Reverse Brascamp-Lieb Inequality: Entropic Duality and Gaussian Optimality

**Jingbo Liu [1], Thomas A. Courtade [2,*], Paul W. Cuff [3] and Sergio Verdú [1]**

[1] Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA; jingbo@princeton.edu (J.L.); verdu@princeton.edu (S.V.)

[2] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720-1770, USA

[3] Renaissance Technologies, LLC 600 Route 25A East Setauket, New York, NY 11733, USA; paul.cuff@gmail.com

* Correspondence: courtade@berkeley.edu

Received: 30 March 2018; Accepted: 25 May 2018; Published: 30 May 2018

**Abstract:** Inspired by the forward and the reverse channels from the image-size characterization problem in network information theory, we introduce a functional inequality that unifies both the Brascamp-Lieb inequality and Barthe's inequality, which is a reverse form of the Brascamp-Lieb inequality. For Polish spaces, we prove its equivalent entropic formulation using the Legendre-Fenchel duality theory. Capitalizing on the entropic formulation, we elaborate on a "doubling trick" used by Lieb and Geng-Nair to prove the Gaussian optimality in this inequality for the case of Gaussian reference measures.

**Keywords:** Brascamp-Lieb inequality; hypercontractivity; functional-entropic duality; Gaussian optimality; network information theory; image size characterization

## 1. Introduction

The Brascamp-Lieb inequality and its reverse [1] concern the optimality of Gaussian functions in a certain type of integral inequality. (Not to be confused with the "variance Brascamp-Lieb inequality" (cf. [2–4]), which generalizes the Poincaré inequality). These inequalities have been generalized in various ways since their discovery, nearly 40 years ago. A modern formulation due to Barthe [5] may be stated as follows:

**Brascamp-Lieb Inequality and Its Reverse** ([5] Theorem 1). *Let $E, E_1, \ldots, E_m$ be Euclidean spaces and $\mathbf{B}_i\colon E \to E_i$ be linear maps. Let $(c_i)_{i=1}^m$ and $D$ be positive real numbers. Then, the Brascamp-Lieb inequality:*

$$\int \prod_{i=1}^m f_i^{c_i}(\mathbf{B}_i \mathbf{x})\, \mathrm{d}\mathbf{x} \le D \prod_{i=1}^m \left( \int f_i(\mathbf{x}_i) \mathrm{d}\mathbf{x}_i \right)^{c_i}, \tag{1}$$

*for all nonnegative measurable functions $f_i$ on $E_i$, $i = 1, \ldots, m$, holds if and only if it holds whenever $f_i$, $i = 1, \ldots, m$ are centered Gaussian functions (a centered Gaussian function is of the form $\mathbf{x} \mapsto \exp(r - \mathbf{x}^\top \mathbf{A}\mathbf{x})$, where $\mathbf{A}$ is a positive semidefinite matrix and $r \in \mathbb{R}$). Similarly, for $F$ a positive real number, the reverse Brascamp-Lieb inequality, also known as Barthe's inequality ($\mathbf{B}_i^*$ denotes the adjoint of $\mathbf{B}_i$),*

$$\int \sup_{(\mathbf{y}_i)\colon \sum_{i=1}^m c_i \mathbf{B}_i^* \mathbf{y}_i = \mathbf{x}} \prod_{i=1}^m f_i^{c_i}(\mathbf{y}_i)\, \mathrm{d}\mathbf{x} \ge F \prod_{i=1}^m \left( \int f_i(\mathbf{y}_i) \mathrm{d}\mathbf{y}_i \right)^{c_i}, \tag{2}$$

*for all nonnegative measurable functions $f_i$ on $E_i$, $i = 1, \ldots, m$, holds if and only if it holds for all centered Gaussian functions.*

For surveys on the history of both the Brascamp-Lieb inequality and Barthe's inequality and their applications, see, e.g., [6,7]. The Brascamp-Lieb inequality can be seen as a generalization of several other inequalities, including Hölder's inequality, the sharp Young inequality, the Loomis-Whitney inequality, the entropy power inequality (cf. [6] or the survey paper [8]), hypercontractivity and the logarithmic Sobolev inequality [9]. Furthermore, the Prékopa-Leindler inequality can be seen as a special case of Barthe's inequality. Due in part to their utility in establishing impossibility bounds, these functional inequalities have attracted much attention in information theory [10–17], theoretical computer science [18–22] and statistics [23–28], to name only a small subset of the literature. Over the years, various proofs of these inequalities have been proposed [1,29–34]. Among these, Lieb's elegant proof [29], which is very close to one of the techniques that will be used in this paper, employs a doubling trick that capitalizes on the rotational invariance property of the Gaussian function: if $f$ is a one-dimensional Gaussian function, then:

$$f(x)f(y) = f\left(\frac{x-y}{\sqrt{2}}\right) f\left(\frac{x+y}{\sqrt{2}}\right). \tag{3}$$

Since (1) and (2) have the same structure modulo the direction of the inequality, a common viewpoint is to consider (1) and (2) as dual inequalities. This viewpoint successfully captures the geometric aspects of (1) and (2). Indeed, it is known that:

$$D \cdot F = 1 \tag{4}$$

as long as $D, F < \infty$ [5]. Moreover, both $D$ and $F$ are equal to one under Ball's geometric condition [35]: $E_1, \ldots, E_m$ are dimension one, and:

$$\sum_{i=1}^{m} c_i \mathbf{B}_i \mathbf{B}_i^* = \mathbf{I} \tag{5}$$

is the identity matrix. While fruitful, this "dual" viewpoint does not fully explain the asymmetry between the forward and the reverse inequalities: there is a sup in (2), but not in (1).

This paper explores a different viewpoint. In particular, we propose a single inequality that unifies (1) and (2). Accordingly, we should reverse both sides of (2) to make the inequality sign consistent with (1). To be concrete, let us first observe that (1) and (2) can be respectively restated in the following more symmetrical forms (with changes of certain symbols):

- For all nonnegative functions $g$ and $f_1, \ldots, f_m$ such that:

$$g(\mathbf{x}) \le \prod_{i=1}^{m} f_j^{c_j}(\mathbf{B}_j \mathbf{x}), \quad \forall \mathbf{x}, \tag{6}$$

  we have:

$$\int_E g \le D \prod_{j=1}^{m} \left( \int_{E_j} f_j \right)^{c_j}. \tag{7}$$

- For all nonnegative measurable functions $g_1, \ldots g_l$ and $f$ such that:

$$\prod_{i=1}^{l} g_i^{b_i}(\mathbf{z}_i) \le f\Big(\sum_{i=1}^{l} b_i \mathbf{B}_i^* \mathbf{z}_i\Big), \quad \forall \mathbf{z}_1, \ldots, \mathbf{z}_l, \tag{8}$$

  we have:

$$\prod_{i=1}^{l} \left( \int_{E_i} g_i \right)^{b_i} \le D \int_E f. \tag{9}$$

Note that in both cases, the optimal choice of one function ($f$ or $g$) can be explicitly computed from the constraints, hence the conventional formulations in (1) and (2). Generalizing further, we can consider the following problem: Let $\mathcal{X}$, $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$, $\mathcal{Z}_1, \ldots, \mathcal{Z}_l$ be measurable spaces. Consider measurable maps $\phi_j \colon \mathcal{X} \to \mathcal{Y}_j$, $j = 1, \ldots, m$ and $\psi \colon \mathcal{X} \to \mathcal{Z}_i$, $i = 1, \ldots, l$. Let $b_1, \ldots, b_l$ and $c_1, \ldots, c_m$ be nonnegative real numbers. Let $\nu_1, \ldots, \nu_l$ be measures on $\mathcal{Z}_1, \ldots, \mathcal{Z}_l$ and $\mu_1, \ldots, \mu_m$ be measures on $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$, respectively. What is the smallest $D > 0$ such that for all nonnegative $f_1, \ldots, f_m$ on $\mathcal{Y}_1, \ldots \mathcal{Y}_m$ and $g_1, \ldots, g_l$ on $\mathcal{Z}_1, \ldots, \mathcal{Z}_l$ satisfying:

$$\prod_{i=1}^{l} g_i^{b_i}(\psi_i(x)) \le \prod_{j=1}^{m} f_j^{c_j}(\phi_j(x)), \quad \forall x, \tag{10}$$

we have:

$$\prod_{i=1}^{l} \left( \int g_i \mathrm{d}\nu_i \right)^{b_i} \le D \prod_{j=1}^{m} \left( \int f_j \mathrm{d}\mu_j \right)^{c_j} ? \tag{11}$$

Except for special case of $l = 1$ (resp. $m = 1$), it is generally not possible to deduce a simple expression from (10) for the optimal choice of $g_i$ (resp. $f_j$) in terms of the rest of the functions. We will refer to (11) as a forward-reverse Brascamp-Lieb inequality.

One of the motivations for considering multiple functions on both sides of (11) comes from multiuser information theory: independently, but almost simultaneously with the discovery of the Brascamp-Lieb inequality in mathematical physics, in the late 1970s, information theorists including Ahslwede, Gács and Körner [36,37] invented the image-size technique for proving strong converses in source and channel networks. An image-size inequality is a characterization of the tradeoff of the measures of certain sets connected by given random transformations (channels); we refer the interested readers to [37] for expositions on the image-size problem. Although not the way treated in [36,37], an image-size inequality can essentially be obtained from a functional inequality similar to (11) by taking the functions to be (roughly speaking) the indicator functions of sets. In the case of (10), the forward channels $\phi_1, \ldots, \phi_m$ and the reverse channels $\psi_1, \ldots, \psi_l$ degenerate into deterministic functions. In this paper, motivated by information theoretic applications similar to those of the image-size problems, we will consider further generalizations of (11) to the case of random transformations. Since the functional inequality is not restricted to indicator functions, it is strictly stronger than the corresponding image-size inequality. As a side remark, [38] uses functional inequalities that are variants of (11) together with a reverse hypercontractivity machinery to improve the image-size plus the blowing-up machinery of [39] and shows that the non-indicator function generalization is crucial for achieving the optimal scaling of the second-order rate expansion.

Of course, to justify the proposal of (11), we must also prove that (11) enjoys certain nice mathematical properties; this is the main goal of the present paper. Specifically, we focus on two aspects of (11): equivalent entropic formulation and Gaussian optimality.

In the mathematical literature, e.g., [32,36,40–46], it is known that certain integral inequalities are equivalent to inequalities involving relative entropies. In particular, Carlen, Loss and Lieb [47] and Carlen and Cordero-Erausquin [32] proved that the Brascamp-Lieb inequality is equivalent to the superadditivity of relative entropy. In this paper, we prove that the forward-reverse Brascamp-Lieb inequality (11) also has an entropic formulation, which turns out to be very close to the rate region of certain multiuser information theory problems (but we will clarify the difference in the text). In fact, Ahlswede, Csiszár and Körner [37,39] essentially derived image-size inequalities from similar entropic inequalities. Because of the reverse part, the proof of the equivalence of (11) and corresponding entropic inequality is more involved than the forward case considered in [32] beyond the case of finite $\mathcal{X}$, $\mathcal{Y}_j$, $\mathcal{Z}_i$, and certain machinery from min-max theory appears necessary. In particular, the proof involves a novel use of the Legendre-Fenchel duality theory. Next, we give a basic version of our

main result on the functional-entropic duality (more general versions will be given later). In order to streamline its presentation, all formal definitions of notation are postponed to Section 2.

**Theorem 1** (Dual formulation of the forward-reverse Brascamp-Lieb inequality)**.** *Assume that:*

*(i)* *$m$ and $l$ are positive integers; $d \in \mathbb{R}$, $\mathcal{X}$ is a compact metric space;*

*(ii)* *$b_i \in (0,\infty)$, $\nu_i$ is a finite Borel measure on a Polish space $\mathcal{Z}_i$, and $Q_{Z_i|X}$ is a random transformation from $\mathcal{X}$ to $\mathcal{Z}_i$, for each $i = 1,\ldots,l$;*

*(iii)* *$c_j \in (0,\infty)$, $\mu_j$ is a finite Borel measure on a Polish space $\mathcal{Y}_j$, and $Q_{Y_j|X}$ is a random transformation from $\mathcal{X}$ to $\mathcal{Y}_i$, for each $j = 1,\ldots,m$;*

*(iv)* *For any $(P_{Z_i})_{i=1}^l$ such that $\sum_{i=1}^l D(P_{Z_i}\|\nu_i) < \infty$, there exists $P_X$ such that $P_X \to Q_{Z_i|X} \to P_{Z_i}$, $i = 1,\ldots,l$ and $\sum_{j=1}^m D(P_{Y_j}\|\mu_j) < \infty$, where $P_X \to Q_{Y_j|X} \to P_{Y_j}$, $j = 1,\ldots,m$.*

*Then, the following two statements are equivalent:*

1. *If the nonnegative continuous functions $(g_i)$, $(f_j)$ are bounded away from zero and satisfy:*

$$\sum_{i=1}^{l} b_i Q_{Z_i|X}(g_i) \le \sum_{j=1}^{m} c_j Q_{Y_j|X}(f_j) \tag{12}$$

*then:*

$$\prod_{i=1}^{l}\left(\int g_i \mathrm{d}\nu_i\right)^{b_i} \le \exp(d)\prod_{j=1}^{m}\left(\int f_j \mathrm{d}\mu_j\right)^{c_j} \tag{13}$$

2. *For any $(P_{Z_i})$ such that $D(P_{Z_i}\|\nu_i) < \infty$ (of course, this assumption is not essential (if we adopt the convention that the infimum in (14) is $+\infty$ when it runs over an empty set)), $i = 1,\ldots,l$,*

$$\sum_{i=1}^{l} b_i D(P_{Z_i}\|\nu_i) + d \ge \inf_{P_X} \sum_{j=1}^{m} c_j D(P_{Y_j}\|\mu_j) \tag{14}$$

*where $P_X \to Q_{Y_j|X} \to P_{Y_j}$, $j = 1,\ldots,m$, and the infimum is over $P_X$ such that $P_X \to Q_{Z_i|X} \to P_{Z_i}$, $i = 1,\ldots,l$.*

Next, in a similar vein as the proverbial result that "Gaussian functions are optimal" for the forward or the reverse Brascamp-Lieb inequality, we show in this paper that Gaussian functions are also optimal for the forward-reverse Brascamp-Lieb inequality, particularized to the case of Gaussian reference measures and linear maps. The proof scheme is based on rotational invariance (3), which can be traced back in the functional setting to Lieb [29]. More specifically, we use a variant for the entropic setting introduced by Geng and Nair [48], thereby taking advantage of the dual formulation of Theorem 1.

**Theorem 2.** *Consider $b_1,\ldots,b_l,c_1,\ldots,c_m,D \in (0,\infty)$. Let $E_1,\ldots,E_l,E^1,\ldots,E^m$ be Euclidean spaces, and let $\mathbf{B}_{ji}\colon E_i \to E^j$ be a linear map for each $i \in \{1,\ldots,l\}$ and $j \in \{1,\ldots,m\}$. Then, for all continuous functions $f_j\colon E^j \to [0,+\infty)$, $g_i\colon E_i \to [0,\infty)$ satisfying:*

$$\prod_{i=1}^{l} g_i^{b_i}(\mathbf{x}_i) \le \prod_{j=1}^{m} f_j^{c_j}\left(\sum_{i=1}^{l}\mathbf{B}_{ji}\mathbf{x}_i\right), \quad \forall \mathbf{x}_1,\ldots,\mathbf{x}_l, \tag{15}$$

*we have:*

$$\prod_{i=1}^{l}\left(\int g_i\right)^{b_i} \le D\prod_{j=1}^{m}\left(\int f_j\right)^{c_j}, \tag{16}$$

*if and only if for all centered Gaussian functions $f_1, \ldots, f_m, g_1, \ldots, g_l$ satisfying (15), we have (16).*

As mentioned, in the literature on the forward or the reverse Brascamp-Lieb inequalities, it is known that a certain geometric condition (5) ensures that the best constant equals one. Now, for the forward-reverse inequality, there is a simple example where the best constant equals one:

**Example 1.** *Let $l$ be a positive integer, and let $\mathbf{M} := (m_{ji})_{1\le j\le l, 1\le i\le l}$ be an orthogonal matrix. For any nonnegative continuous functions $(f_j)_{j=1}^l$ $(g_i)_{i=1}^l$ on $\mathbb{R}$ such that:*

$$\prod_{i=1}^{l} g_i(x_i) \le \prod_{j=1}^{l} f_j\left(\sum_{i=1}^{l} m_{ji}x_i\right), \quad \forall x^l \in \mathbb{R}^l, \tag{17}$$

*we have:*

$$\prod_{i=1}^{l} \int g_i(x)\mathrm{d}x \le \prod_{i=1}^{l} \int f_j(x)\mathrm{d}x. \tag{18}$$

The rest of the paper is organized as follows: Section 2 defines the notation and reviews some basic theory of convex duality. Section 3 proves Theorem 1 and also presents its extensions to the settings of noncompact spaces or general reverse channels. Section 4 proves the Gaussian optimality in the entropic formulation, with the caveat that a certain "non-degenerate" assumption is imposed to ensure the existence of extremizers. At the end of Section 4, we give a proof sketch of Example 1 and also propose a generalization of the example. To completely prove Theorem 2, in Appendix F, we use a limiting argument to drop the non-degenerate assumption and apply the equivalence between the functional and entropic formulations.

## 2. Review of the Legendre-Fenchel Duality Theory

Our proof of the equivalence of the functional and the entropic inequalities uses the Legendre-Fenchel duality theory, a topic from convex analysis. Before getting into that, a recap of some basics on the duality of topological vector spaces seems appropriate. Unless otherwise indicated, we assume Polish spaces and Borel measures. Recall that metric space. It enjoys several nice properties that we use heavily in this section, including the Prokhorov theorem and the Riesz-Kakutani theorem. Of course, the Polish space assumption covers the cases of Euclidean and discrete spaces (endowed with the Hamming metric, which induces the discrete topology, making every function on the discrete set continuous), among others. Readers interested in discrete spaces only may refer to the (much simpler) argument in [49] based on the KKT condition.

**Notation 1.** *Let $\mathcal{X}$ be a topological space.*

- *$C_c(\mathcal{X})$ denotes the space of continuous functions on $\mathcal{X}$ with a compact support;*
- *$C_0(\mathcal{X})$ denotes the space of all continuous functions $f$ on $\mathcal{X}$ that vanish at infinity (i.e., for any $\epsilon > 0$, there exists a compact set $\mathcal{K} \subseteq \mathcal{X}$ such that $|f(x)| < \epsilon$ for $x \in \mathcal{X} \setminus \mathcal{K}$);*
- *$C_b(\mathcal{X})$ denotes the space of bounded continuous functions on $\mathcal{X}$;*
- *$\mathcal{M}(\mathcal{X})$ denotes the space of finite signed Borel measures on $\mathcal{X}$;*
- *$\mathcal{P}(\mathcal{X})$ denotes the space of probability measures on $\mathcal{X}$.*

We consider $C_c$, $C_0$ and $C_b$ as topological vector spaces, with the topology induced from the sup norm. The following theorem, usually attributed to Riesz, Markov and Kakutani, is well known in functional analysis and can be found in, e.g., [50,51].

**Theorem 3** (Riesz-Markov-Kakutani)**.** *If $\mathcal{X}$ is a locally compact, $\sigma$-compact Polish space, the dual (the dual of a topological vector space consists of all continuous linear functionals on that space, which is naturally also topological vector space (with the weak$^*$ topology)) of both $C_c(\mathcal{X})$ and $C_0(\mathcal{X})$ is $\mathcal{M}(\mathcal{X})$.*

**Remark 1.** *The dual space of $C_b(\mathcal{X})$ can be strictly larger than $\mathcal{M}(\mathcal{X})$, since it also contains those linear functionals that depend on the "limit at infinity" of a function $f \in C_b(\mathcal{X})$ (originally defined for those $f$ that do have a limit at infinity and then extended to the whole $C_b(\mathcal{X})$ by the Hahn-Banach theorem; see, e.g., [50]).*

Of course, any $\mu \in \mathcal{M}(\mathcal{X})$ is a continuous linear functional on $C_0(\mathcal{X})$ or $C_c(\mathcal{X})$, given by:

$$f \mapsto \int f \mathrm{d}\mu \tag{19}$$

where $f$ is a function in $C_0(\mathcal{X})$ or $C_c(\mathcal{X})$. As is well known, Theorem 3 states that the converse is also true under mild regularity assumptions on the space. Thus, we can view measures as continuous linear functionals on a certain function space (in fact, some authors prefer to construct measure theory by defining a measure as a linear functional on a suitable measure space; see Lax [50] or Bourbaki [52]); this justifies the shorthand notation:

$$\mu(f) := \int f \mathrm{d}\mu \tag{20}$$

which we employ in the rest of the paper. This viewpoint is the most natural for our setting since in the proof of the equivalent formulation of the forward-reverse Brascamp-Lieb inequality, we shall use the Hahn-Banach theorem to show the existence of certain linear functionals.

**Definition 1.** *Let $\Lambda \colon C_b(\mathcal{X}) \to (-\infty, +\infty]$ be a lower semicontinuous, proper convex function. Its Legendre-Fenchel transform $\Lambda^* \colon C_b(\mathcal{X})^* \to (-\infty, +\infty]$ is given by:*

$$\Lambda^*(\ell) := \sup_{u \in C_b(\mathcal{X})} [\ell(u) - \Lambda(u)]. \tag{21}$$

Let $\nu$ be a nonnegative finite Borel measure on a Polish space $\mathcal{X}$, and define the convex functional on $C_b(\mathcal{X})$:

$$\Lambda(f) := \log \nu(\exp(f)) \tag{22}$$

$$= \log \int \exp(f) \mathrm{d}\nu. \tag{23}$$

Then, note that the relative entropy has the following alternative definition: for any $\mu \in \mathcal{M}(\mathcal{X})$,

$$D(\mu \| \nu) := \sup_{f \in C_b(\mathcal{X})} [\mu(f) - \Lambda(f)] \tag{24}$$

which agrees with the more familiar definition $D(\mu\|\nu) := \mu(\log \frac{\mathrm{d}\mu}{\mathrm{d}\nu})$ when $\nu$ is a probability measure, by the Donsker-Varadhan formula (cf. [53] Lemma 6.2.13). If $\mu$ is not a probability measure, then $D(\mu\|\nu)$ as defined in (24) is $+\infty$.

Given a bounded linear operator $T \colon C_b(\mathcal{Y}) \to C_b(\mathcal{X})$, the dual operator $T^* \colon C_b(\mathcal{X})^* \to C_b(\mathcal{Y})^*$ is defined in terms of:

$$\begin{aligned} T^* \mu_X \colon C_b(\mathcal{Y}) &\to \mathbb{R}; \\ f &\mapsto \mu_X(Tf), \end{aligned} \tag{25}$$

for any $\mu_X \in C_b(\mathcal{X})^*$. Since $\mathcal{P}(\mathcal{X}) \subseteq \mathcal{M}(\mathcal{X}) \subseteq C_b(\mathcal{X})^*$, $T$ is said to be a conditional expectation operator if $T^*P \in \mathcal{P}(\mathcal{Y})$ for any $P \in \mathcal{P}(\mathcal{X})$. The operator $T^*$ is defined as the dual of a conditional expectation operator $T$ and, in a slight abuse of terminology, is said to be a random transformation from $\mathcal{X}$ to $\mathcal{Y}$.

For example, in the notation of Theorem 1, if $g \in C_b(\mathcal{Y})$ and $Q_{Y|X}$ is a random transformation from $\mathcal{X}$ to $\mathcal{Y}$, the quantity $Q_{Y|X}(g)$ is a function on $\mathcal{X}$, defined by taking the conditional expectation. Furthermore, if $P_X \in \mathcal{P}(\mathcal{X})$, we write $P_X \to Q_{Y|X} \to P_Y$ to indicate that $P_Y \in \mathcal{P}(\mathcal{Y})$ is the measure induced on $\mathcal{Y}$ by applying $Q_{Y|X}$ to $P_X$.

**Remark 2.** *From the viewpoint of category theory (see for example [54,55]), $C_b$ is a functor from the category of topological spaces to the category of topological vector spaces, which is contra-variant because for any continuous, $\phi\colon \mathcal{X} \to \mathcal{Y}$ (morphism between topological spaces), we have $C_b(\phi)\colon C_b(\mathcal{Y}) \to C_b(\mathcal{X})$, $u \mapsto u \circ f$ where $u \circ \phi$ denotes the composition of two continuous functions, reversing the arrows in the maps (i.e., the morphisms). On the other hand, $\mathcal{M}$ is a covariant functor and $\mathcal{M}(\phi)\colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$, $\mu \mapsto \mu \circ \phi^{-1}$, where $\mu \circ \phi^{-1}(\mathcal{B}) := \mu(\phi^{-1}(\mathcal{B}))$ for any Borel measurable $\mathcal{B} \subseteq \mathcal{Y}$. "Duality" itself is a contra-variant functor between the category of topological spaces (note the reversal of arrows in Figure 1). Moreover, $C_b(\mathcal{X})^* = \mathcal{M}(\mathcal{X})$ and $C_b(\phi)^* = \mathcal{M}(\phi)$ if $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and $\phi\colon \mathcal{X} \to \mathcal{Y}$ is continuous. Definition 2 can therefore be viewed as the special case where $\phi$ is the projection map:*

$$\begin{array}{ccccc} C_b(\mathcal{Z}_1) \ni g_1 & \xrightarrow{S_1} & & \xleftarrow{T_1} & C_b(\mathcal{Y}_1) \ni f_1 \\ & & C_b(\mathcal{X}) & & \\ C_b(\mathcal{Z}_2) \ni g_2 & \xrightarrow{S_2} & & \xleftarrow{T_2} & C_b(\mathcal{Y}_2) \ni f_2 \end{array}$$

$$\begin{array}{ccccc} \mathcal{P}(\mathcal{Z}_1) \ni P_{Z_1} & \xleftarrow{S_1^*} & & \xrightarrow{T_1^*} & \mathcal{P}(\mathcal{Y}_1) \ni P_{Y_1} \\ & & \mathcal{P}(\mathcal{X}) \ni P_X & & \\ \mathcal{P}(\mathcal{Z}_2) \ni P_{Z_2} & \xleftarrow{S_2^*} & & \xrightarrow{T_2^*} & \mathcal{P}(\mathcal{Y}_2) \ni P_{Y_2} \end{array}$$

**Figure 1.** Diagrams for Theorem 1.

**Definition 2.** *Suppose $\phi\colon \mathcal{Z}_1 \times \mathcal{Z}_2 \to \mathcal{Z}_1$, $(z_1, z_2) \mapsto z_1$ is the projection to the first coordinate.*

- *$C_b(\phi)\colon C_b(\mathcal{Z}_1) \to C_b(\mathcal{Z}_1 \times \mathcal{Z}_2)$ is called a canonical map, whose action is almost trivial: it sends a function of $z_i$ to itself, but viewed as a function of $(z_1, z_2)$.*
- *$\mathcal{M}(\phi)\colon \mathcal{M}(\mathcal{Z}_1 \times \mathcal{Z}_2) \to \mathcal{M}(\mathcal{Z}_1)$ is called marginalization, which simply takes a joint distribution to a marginal distribution.*

The Fenchel-Rockafellar duality (see [40] Theorem 1.9, or [56] in the case of finite dimensional vector spaces) usually refers to the $k = 1$ special case of the following result.

**Theorem 4.** *Assume that $A$ is a topological vector space whose dual is $A^*$. Let $\Theta_j\colon A \to \mathbb{R} \cup \{+\infty\}$, $j = 0, 1, \ldots, k$, for some positive integer $k$. Suppose there exist some $(u_j)_{j=1}^k$ and $u_0 := -(u_1 + \cdots + u_k)$ such that:*

$$\Theta_j(u_j) < \infty, \quad j = 0, \ldots, k \tag{26}$$

*and $\Theta_0$ is upper semicontinuous at $u_0$. Then:*

$$-\inf_{\ell\in A^*}\left[\sum_{j=0}^{k}\Theta_j^*(\ell)\right]=\inf_{u_1,\ldots,u_k\in A}\left[\Theta_0\left(-\sum_{j=1}^{k}u_j\right)+\sum_{j=1}^{k}\Theta_j(u_j)\right]. \tag{27}$$

For completeness, we provide a proof of this result, which is based on the Hahn-Banach theorem (Theorem 5) and is similar to the proof of [40] Theorem 1.9.

**Proof.** Let $m_0$ be the right side of (27). The $\leq$ part of (27) follows trivially from the (weak) min-max inequality since:

$$\begin{aligned} m_0 &= \inf_{u_0,\ldots,u_k\in A}\sup_{\ell\in A^*}\left\{\sum_{j=0}^{k}\Theta_j(u_j)-\ell(\sum_{j=0}^{k}u_j)\right\} \\ &\geq \sup_{\ell\in A^*}\inf_{u_0,\ldots,u_k\in A}\left\{\sum_{j=0}^{k}\Theta_j(u_j)-\ell(\sum_{j=0}^{k}u_j)\right\} \\ &= -\inf_{\ell\in A^*}\left[\sum_{j=0}^{k}\Theta_j^*(\ell)\right]. \end{aligned} \tag{28}$$

It remains to prove the $\geq$ part, and it suffices to assume without loss of generality that $m_0 > -\infty$. Note that (26) also implies that $m_0 < +\infty$. Define convex sets:

$$C_j := \{(u,r)\in A\times\mathbb{R}\colon\ r>\Theta_j(u)\},\quad j=0,\ldots,k; \tag{29}$$

$$B := \{(0,m)\in A\times\mathbb{R}\colon\ m\leq m_0\}. \tag{30}$$

Observe that these are nonempty sets because of (26). Furthermore, $C_0$ has a nonempty interior by the assumption that $\Theta_0$ is upper semicontinuous at $u_0$. Thus, the Minkowski sum:

$$C := C_0+\cdots+C_k \tag{31}$$

is a convex set with a nonempty interior. Moreover, $C\cup B=\emptyset$. By the Hahn-Banach theorem (Theorem 5), there exists $(\ell,s)\in A^*\times\mathbb{R}$ such that:

$$sm\leq \ell\left(\sum_{j=0}^{k}u_j\right)+s\sum_{j=0}^{k}r_j. \tag{32}$$

For any $m\leq m_0$ and $(u_j,r_j)\in C_j$, $j=0,\ldots,k$. From (30), we see (32) can only hold when $s\geq 0$. Moreover, from (26) and the upper semicontinuity of $\Theta_0$ at $u_0$, we see that the $\sum_{j=0}^{k}u_j$ in (32) can take a value in a neighborhood of $0\in A$; hence, $s\neq 0$. Thus, by dividing $s$ on both sides of (32) and setting $\ell\leftarrow -\ell/s$, we see that:

$$\begin{aligned} m_0 &\leq \inf_{u_0,\ldots,u_k\in A}\left[-\ell\left(\sum_{j=0}^{k}u_j\right)+\sum_{j=0}^{k}\Theta_j(u_j)\right] \\ &= -\left[\sum_{j=0}^{k}\Theta_j^*(\ell)\right] \end{aligned} \tag{33}$$

which establishes $\geq$ in (27). □

**Theorem 5** (Hahn-Banach). *Let $C$ and $B$ be convex, nonempty disjoint subsets of a topological vector space $A$.*

1. *If the interior of $C$ is non-empty, then there exists $\ell \in A^*$, $\ell \neq 0$ such that:*

$$\sup_{u \in B} \ell(u) \leq \inf_{u \in C} \ell(u). \tag{34}$$

2. *If $A$ is locally convex, $B$ is compact and $C$ is closed, then there exists $\ell \in A^*$ such that:*

$$\sup_{u \in B} \ell(u) < \inf_{u \in C} \ell(u). \tag{35}$$

**Remark 3.** *The assumption in Theorem 5 that $C$ has a nonempty interior is only necessary in the infinite dimensional case. However, even if $A$ in Theorem 4 is finite dimensional, the assumption in Theorem 4 that $\Theta_0$ is upper semicontinuous at $u_0$ is still necessary, because this assumption was not only used in applying Hahn-Banach, but also in concluding that $s \neq 0$ in (32).*

## 3. The Entropic-Functional Duality

In this section, we prove Theorem 1 and some of its generalizations.

### *3.1. Compact $\mathcal{X}$*

We first state a duality theorem for the case of compact spaces to streamline the proof. Later, we show that the argument can be extended to a particular non-compact case (Theorem 1 is not included in the conference paper [49], but was announced in the conference presentation). Our proof based on the Legendre-Fenchel duality (Theorem 4) was inspired by the proof of the Kantorovich duality in the theory of optimal transportation (see [40] Chapter 1, where the idea was credited to Brenier).

Recall from Section 2 that a random transformation (a mapping between probability measures) is formally the dual of a conditional expectation operator. Suppose $P_{Y_j|X} = T_j^*$, $j = 1, \ldots, m$ and $P_{Z_i|X} = S_i^*$, $i = 1, \ldots, l$.

**Proof of Theorem 1.** We can safely assume $d = 0$ below without loss of generality (since otherwise, we can always substitute $\mu_1 \leftarrow \exp\left(\frac{d}{c_1}\right) \mu_1$).

**1)⇒2)** This is the nontrivial direction, which relies on certain (strong) min-max type results. In Theorem 4, put (in (36), $u \leq 0$ means that $u$ is pointwise non-positive):

$$\Theta_0 \colon u \in C_b(\mathcal{X}) \mapsto \begin{cases} 0 & u \leq 0; \\ +\infty & \text{otherwise.} \end{cases} \tag{36}$$

Then,

$$\Theta_0^* \colon \pi \in \mathcal{M}(\mathcal{X}) \mapsto \begin{cases} 0 & \pi \geq 0; \\ +\infty & \text{otherwise.} \end{cases} \tag{37}$$

For each $j = 1, \ldots, m$, set:

$$\Theta_j(u) := c_j \inf \log \mu_j \left( \exp\left( \frac{1}{c_j} v \right) \right) \tag{38}$$

where the infimum is over $v \in C_b(\mathcal{Y})$ such that $u = T_j v$; if there is no such $v$, then $\Theta_j(u) := +\infty$ as a convention. Observe that:

- $\Theta_j$ is convex: indeed, given arbitrary $u^0$ and $u^1$, suppose that $v^0$ and $v^1$ respectively achieve the infimum in (38) for $u^0$ and $u^1$ (if the infimum is not achievable, the argument still goes through by the approximation and limit argument). Then, for any $\alpha \in [0, 1]$, $v^\alpha :=$

$(1-\alpha)v^0+\alpha v^1$ satisfies $u^\alpha = T_j v^\alpha$ where $u^\alpha := (1-\alpha)u^0+\alpha u^1$. Thus, the convexity of $\Theta_j$ follows from the convexity of the functional in (23);

- $\Theta_j(u) > -\infty$ for any $u \in C_b(\mathcal{X})$. Otherwise, for any $P_X$ and $P_{Y_j} := T_j^* P_X$, we have:

$$\begin{aligned} D(P_{Y_j}\|\mu_j) &= \sup_v \{P_{Y_j}(v) - \log \mu_j(\exp(v))\} && (39)\\ &= \sup_v \{P_X(T_j v) - \log \mu_j(\exp(v))\} && (40)\\ &= \sup_{u\in C_b(\mathcal{X})} \left\{ P_X(u) - \frac{1}{c_j}\Theta_j(c_j u)\right\} && (41)\\ &= +\infty && (42)\end{aligned}$$

  which contradicts the assumption that $\sum_{j=1}^m c_j D(P_{Y_j}\|\mu_j) < \infty$ in the theorem;
- From Steps (39)–(41), we see $\Theta_j^*(\pi) = c_j D(T_j^*\pi\|\mu_j)$ for any $\pi \in \mathcal{M}(\mathcal{X})$, where the definition of $D(\cdot\|\mu_j)$ is extended using the Donsker-Varadhan formula (that is, it is infinite when the argument is not a probability measure).

Finally, for the given $(P_{Z_i})_{i=1}^l$, choose:

$$\Theta_{m+1}\colon u \in C_b(\mathcal{X}) \mapsto \begin{cases} \sum_{i=1}^l P_{Z_i}(w_i) & \text{if } u = \sum_{i=1}^l S_i w_i \text{ for some } w_i \in C_b(\mathcal{Z}_i);\\ +\infty & \text{otherwise.}\end{cases} \quad (43)$$

Notice that:

- $\Theta_{m+1}$ is convex;
- $\Theta_{m+1}$ is well defined (that is, the choice of $(w_i)$ in (43) is inconsequential). Indeed, if $(w_i)_{i=1}^l$ is such that $\sum_{i=1}^l S_i w_i = 0$, then:

$$\begin{aligned}\sum_{i=1}^l P_{Z_i}(w_i) &= \sum_{i=1}^l S_i^* P_X(w_i) && (44)\\ &= \sum_{i=1}^l P_X(S_i w_i)\\ &= 0,\end{aligned}$$

  where $P_X$ is such that $S_i^* P_X = P_{Z_i}$, $i = 1,\dots,l$, whose existence is guaranteed by the assumption of the theorem. This also shows that $\Theta_{m+1} > -\infty$.
-

$$\begin{aligned}\Theta_{m+1}^*(\pi) &:= \sup_u \{\pi(u) - \Theta_{m+1}(u)\} && (45)\\ &= \sup_{w_1,\dots,w_l} \left\{ \pi\left(\sum_{i=1}^l S_i w_i\right) - \sum_{i=1}^l P_{Z_i}(w_i)\right\}\\ &= \sup_{w_1,\dots,w_l} \left\{ \sum_{i=1}^l S_i^*\pi(w_i) - \sum_{i=1}^l P_{Z_i}(w_i)\right\}\\ &= \begin{cases} 0 & \text{if } S_i^*\pi = P_{Z_i}, \quad i = 1,\dots,l;\\ +\infty & \text{otherwise.}\end{cases}\end{aligned}$$

Invoking Theorem 4 (where the $u_j$ in Theorem 4 can be chosen as the constant function $u_j \equiv 1$, $j = 1,\dots,m+1$):

$$\inf_{\pi:\ \pi\geq 0,\, S_i^*\pi = P_{Z_i}} \sum_{j=1}^{m} c_j D(T_j^*\pi \| \mu_j)$$

$$= - \inf_{v^m, w^l:\ \sum_{j=1}^m T_j v_j + \sum_{i=1}^l S_i w_i \geq 0} \left[ \sum_{j=1}^{m} c_j \log \mu_j \left( \exp\left( \frac{1}{c_j} v_j \right) \right) + \sum_{i=1}^{l} P_{Z_i}(w_i) \right] \tag{46}$$

where $v^m$ denotes the collection of the functions $v_1, \ldots, v_m$, and similarly for $w^l$. Note that the left side of (46) is exactly the right side of (14). For any $\epsilon > 0$, choose $v_j \in C_b(\mathcal{Y}_j)$, $j = 1, \ldots, m$ and $w_i \in C_b(\mathcal{Z}_i)$, $i = 1, \ldots, l$ such that $\sum_{j=1}^m T_j v_j + \sum_{i=1}^l S_i w_i \geq 0$ and:

$$\epsilon - \sum_{j=1}^{m} c_j \log \mu_j \left( \exp\left( \frac{1}{c_j} v_j \right) \right) - \sum_{i=1}^{l} P_{Z_i}(w_i) > \inf_{\pi:\ \pi\geq 0,\, S_i^*\pi = P_{Z_i}} \sum_{j=1}^{m} c_j D(T_j^*\pi \| \mu_j) \tag{47}$$

Now, invoking (13) with $f_j := \exp\left(\frac{1}{c_j} v_j\right)$, $j = 1, \ldots, m$ and $g_i := \exp\left(-\frac{1}{b_i} w_i\right)$, $i = 1, \ldots, l$, we upper bound the left side of (47) by:

$$\epsilon - \sum_{i=1}^{l} b_i \log \nu_i(g_i) + \sum_{i=1}^{l} b_i P_{Z_i}(\log g_i) \leq \epsilon + \sum_{i=1}^{l} b_i D(P_{Z_i} \| \nu_i) \tag{48}$$

where the last step follows by the Donsker-Varadhan formula. Therefore, (14) is established since $\epsilon > 0$ is arbitrary.

**2)⇒1)** Since $\nu_i$ is finite and $g_i$ is bounded by assumption, we have $\nu_i(g_i) < \infty$, $i = 1, \ldots, l$. Moreover, (13) is trivially true when $\nu_i(g_i) = 0$ for some $i$, so we will assume below that $\nu_i(g_i) \in (0, \infty)$ for each $i$. Define $P_{Z_i}$ by:

$$\frac{\mathrm{d}P_{Z_i}}{\mathrm{d}\nu_i} = \frac{g_i}{\nu_i(g_i)}, \quad i = 1, \ldots, l. \tag{49}$$

Then, for any $\epsilon > 0$,

$$\sum_{i=1}^{l} b_i \log \nu_i(g_i) = \sum_{i=1}^{l} b_i [P_{Z_i}(\log g_i) - D(P_{Z_i} \| \nu_i)] \tag{50}$$

$$< \sum_{j=1}^{m} c_j P_{Y_j}(\log f_j) + \epsilon - \sum_{j=1}^{m} c_j D(P_{Y_j} \| \mu_j) \tag{51}$$

$$\leq \epsilon + \sum_{j=1}^{m} c_j \log \mu_j(f_j) \tag{52}$$

where:

- (51) uses the Donsker-Varadhan formula, and we have chosen $P_X$, $P_{Y_j} := T_j^* P_X$, $j = 1, \ldots, m$ such that:

$$\sum_{i=1}^{l} b_i D(P_{Z_i} \| \nu_i) > \sum_{j=1}^{m} c_j D(P_{Y_j} \| \mu_j) - \epsilon \tag{53}$$

- (52) also follows from the Donsker-Varadhan formula.

The result follows since $\epsilon > 0$ can be arbitrary.

□

**Remark 4.** *Condition (iv) in the theorem imposes a rather strong assumption on $(S_i)$: for simplicity, consider the case where $|\mathcal{X}|, |Z_i| < \infty$. Then, Condition (iv) assumes that for any $(P_{Z_i})$, there exists $P_X$ such that $P_{Z_i} = S_i^* P_X$. This assumption is certainly satisfied when $(S_i)$ are induced by coordinate projections; the case of $l = 1$ and $P_{Z|X}$ being a reverse erasure channel gives a simple example where $P_{Z|X}$ is not a deterministic map.*

Next, we give a generalization of Theorem 1, which alleviates the restriction on $(S_i)$:

**Theorem 6.** *Theorem 1 continues to hold if Condition (iv) therein is weakened to the following:*

- *For any $P_X$ such that $D(S_i^* P_X \| \nu_i) < \infty$, $i = 1, \ldots, l$, there exists $\tilde{P}_X$ such that $S_i^* \tilde{P}_X = S_i^* P_X$ for each $i$ and $\sum_{j=1}^m c_j D(T_j^* \tilde{P}_X \| \mu_j) < \infty$ for each $j$.*

*and the conclusion of the theorem will be replaced by the equivalence of the following two statements:*

1. *For any nonnegative continuous functions $(g_i)$, $(f_j)$ bounded away from zero and such that:*

$$\sum_{i=1}^{l} b_i S_i \log g_i \le \sum_{j=1}^{m} c_j T_j \log f_j \tag{54}$$

*we have:*

$$\inf_{(\tilde{g}_i):\ \sum_{i=1}^l b_i S_i \log \tilde{g}_i \ge \sum_{i=1}^l b_i S_i \log g_i} \prod_{i=1}^{l} \nu_i^{b_i}(\tilde{g}_i) \le \exp(d) \prod_{j=1}^{m} \mu_j^{c_j}(f_j). \tag{55}$$

2. *For any $(P_X)$ such that $D(S_i^* P_X \| \nu_i) < \infty$, $i = 1, \ldots, l$,*

$$\sum_{i=1}^{l} b_i D(S_i^* P_X \| \nu_i) + d \ge \inf_{\tilde{P}_X:\ S_i^* \tilde{P}_X = S_i^* P_X} \sum_{j=1}^{m} c_j D(T_j^* \tilde{P}_X \| \mu_j). \tag{56}$$

In Appendix A, we show that Theorem 6 indeed recovers Theorem 1 for the more restricted class of random transformations.

**Proof.** Here, we mention the parts of the proof that need to be changed: upon specifying $(f_j)$ and $(g_i)$ right after (47), we select $(\tilde{g}_i)$ such that:

$$\sum_{i=1}^{l} b_i S_i \log \tilde{g}_i \ge \sum_{i=1}^{l} b_i S_i \log g_i \tag{57}$$

$$\sum_{i=1}^{l} b_i \log \nu_i(\tilde{g}_i) \le \sum_{j=1}^{m} c_j \log \mu_j(f_j) + \epsilon. \tag{58}$$

Then, in lieu of (59), we upper-bound the left side of (47) by:

$$2\epsilon - \sum_{i=1}^{l} b_i \log \nu_i(\tilde{g}_i) + \sum_{i=1}^{l} b_i P_{Z_i}(\log \tilde{g}_i) \le 2\epsilon + \sum_{i=1}^{l} b_i D(P_{Z_i} \| \nu_i) \tag{59}$$

which establishes the 1)⇒2) part. For the other direction, for each $i \in \{1, 2, \ldots, l\}$, define:

$$\Lambda_i(u) := \inf_{\tilde{g}_i > 0:\ b_i S_i \log \tilde{g}_i = u} b_i \log \nu_i(\tilde{g}_i). \tag{60}$$

Then, following essentially the same proof as that of $\Theta_j$ in (38), we see that $\Lambda_i$ is proper convex and:

$$\Lambda_i^*(\pi) = b_i D(S_i^* \pi \| \mu_j). \tag{61}$$

Moreover, let:

$$\Lambda_{l+1}(u) := \begin{cases} 0 & \text{if } u = -\sum b_i S_i \log g_i; \\ +\infty & \text{otherwise.} \end{cases} \tag{62}$$

Then, $\Lambda_{l+1}^*(\pi) = -\sum b_i S_i^* \pi(\log g_i)$. Using the Legendre-Fenchel duality, we see that for any $\epsilon > 0$,

$$\begin{aligned}
&\inf_{(\tilde{g}_i):\ \sum_{i=1}^l b_i S_i \log \tilde{g}_i \geq \sum_{i=1}^l b_i S_i \log g_i} \sum_{i=1}^l b_i \log \nu_i(\tilde{g}_i) \\
&= \inf_{u_1,\ldots,u_{l+1}} \left\{ \Theta_0 \left( -\sum_{i=1}^{l+1} u_i \right) + \sum_{i=1}^{l+1} \Lambda_i(u_i) \right\} && (63) \\
&= \sup_{\pi} \left\{ -\sum_{i=0}^{l+1} \Theta_i^*(\pi) \right\} && (64) \\
&= \sup_{\pi \geq 0} \left\{ -\sum_{i=1}^{l+1} \Theta_i^*(\pi) \right\} && (65) \\
&= \sup_{\pi \geq 0} \left\{ \sum_{i=1}^{l} b_i S_i^* \pi(\log g_i) - \sum_{i=1}^{l} b_i D(S_i^* \pi \| \nu_i) \right\} && (66) \\
&\leq \sum_{i=1}^{l} b_i S_i^* P_X(\log g_i) - \sum_{i=1}^{l} b_i D(S_i^* P_X \| \nu_i) + \epsilon && (67) \\
&\leq \sum_{j=1}^{m} c_j T_j^* \tilde{P}_X(\log f_j) - \sum_{j=1}^{m} c_j D(T_j^* \tilde{P}_X \| \mu_j) + 2\epsilon && (68) \\
&\leq 2\epsilon + \sum_{j=1}^{m} c_j \log \mu_j(f_j) && (69)
\end{aligned}$$

where:

- To see (67), we note that the sup in (66) can be restricted to $\pi$, which is a probability measure, since otherwise, the relative entropy terms in (66) are $+\infty$ by its definition via the Donsker-Varadhan formula. Then, we select $P_X$ such that (67) holds.
- In (68), we have chosen $\tilde{P}_X$ such that:

$$S_i^* \tilde{P}_X = S_i^* P_X, \quad 1 \leq i \leq l; \tag{70}$$

$$\sum_{i=1}^{l} b_i D(S_i^* P_X) > \sum_{j=1}^{m} c_j D(T_j^* \tilde{P}_X \| \mu_j) - \epsilon, \tag{71}$$

and then applied the assumption (54). The result follows since $\epsilon > 0$ can be arbitrary.

□

**Remark 5.** *The infimum in (14) is in fact achievable: for any $(P_{Z_i})$, there exists a $P_X$ that minimizes $\sum_{j=1}^m c_j D(P_{Y_j} \| \mu_j)$ subject to the constraints $S_i^* P_X = P_{Z_i}$, $i = 1, \ldots m$, where $P_{Y_j} := T_j^* P_X$, $j = 1, \ldots, m$. Indeed, since the singleton $\{P_{Z_i}\}$ is weak*-closed and $S_i^*$ is weak*-continuous (Generally, if $T\colon A \to B$ is a continuous map between two topologically vector spaces, then $T^*\colon B^* \to A^*$ is a weak* continuous map between the dual spaces. Indeed, if $y_n \to y$ is a weak*-convergent subsequence in $B^*$, meaning $y_n(b) \to y(b)$ for any $b \in B$, then, we must have $T^* y_n(a) = y_n(Ta) \to y(Ta) = T^* y(a)$ for any $a \in A$, meaning that $T^* y_n$ converges to $T^* y$ in the weak* topology.), the set $\bigcap_{i=1}^l (S_i^*)^{-1} P_{Z_i}$ is weak*-closed in $\mathcal{M}(X)$; hence, its intersection with $\mathcal{P}(\mathcal{X})$ is weak*-compact in $\mathcal{P}(\mathcal{X})$, because $\mathcal{P}(\mathcal{X})$ is weak*-compact by (a simple version for the setting of a compact underlying space $\mathcal{X}$ of) the Prokhorov theorem [57]. Moreover, by the weak*-lower*

*semicontinuity of $D(\cdot\|\mu_j)$ (easily seen from the variational formula/Donsker-Varadhan formula of the relative entropy, cf. [58]) and the weak*-continuity of $T_j^*$, $j = 1, \ldots, m$, we see that $\sum_{j=1}^m c_j D(T_j^* P_X \| \mu_j)$ is weak*-lower semicontinuous in $P_X$, and hence, the existence of a minimizing $P_X$ is established.*

**Remark 6.** *Abusing the terminology from min-max theory, Theorem 1 may be interpreted as a "strong duality" result, which establishes the equivalence of two optimization problems. The 1)⇒2) part is the non-trivial direction, which requires regularity on the spaces. In contrast, the 2)⇒1) direction can be thought of as a "weak duality", which establishes only a partial relation, but holds for more general spaces.*

### *3.2. Noncompact $\mathcal{X}$*

Our proof of 1)⇒2) in Theorem 1 makes use of the Hahn-Banach theorem and hence relies crucially on the fact that the measure space is the dual of the function space. Naively, one might want to extend the the proof to the case of locally compact $\mathcal{X}$ by considering $C_0(\mathcal{X})$ instead of $C_b(\mathcal{X})$, so that the dual space is still $\mathcal{M}(\mathcal{X})$. However, this would not work: consider the case when $\mathcal{X} = \mathcal{Z}_1 \times, \ldots, \times \mathcal{Z}_l$ and each $S_i$ is the canonical map. Then, $\Theta_{m+1}(u)$ as defined in (43) is $+\infty$ unless $u \equiv 0$ (because $u \in C_0(\mathcal{X})$ requires that $u$ vanishes at infinity); thus, $\Theta_{m+1}^* \equiv 0$. Luckily, we can still work with $C_b(\mathcal{X})$; in this case, $\ell \in C_b(\mathcal{X})^*$ may not be a measure, but we can decompose it into $\ell = \pi + R$ where $\pi \in \mathcal{M}(\mathcal{X})$ and $R$ is a linear functional "supported at infinity". Below, we use the techniques in [40] (Chapter 1.3) to prove a particular extension of Theorem 1 to a non-compact case.

**Theorem 7.** *Theorem 1 still holds if*

- *The assumption that $\mathcal{X}$ is a compact metric space is relaxed to the assumption that it is a locally compact and $\sigma$-compact Polish space;*
- *$\mathcal{X} = \prod_{i=1}^l \mathcal{Z}_i$ and $S_i \colon C_b(\mathcal{Z}_i) \to C_b(\mathcal{X})$, $i = 1, \ldots, l$ are canonical maps (see Definition 2).*

**Proof.** The proof of the "weak duality" part 2)⇒1) still works in the noncompact case, so we only need to explain what changes need to be made in the proof of the 1)⇒2) part. Let $\Theta_0$ be defined as before, in (36). Then, for any $\ell \in C_b(\mathcal{X})^*$,

$$\Theta_0^*(\ell) = \sup_{u \le 0} \ell(u) \tag{72}$$

which is zero if $\ell$ is nonnegative (in the sense that $\ell(u) \ge 0$ for every $u \ge 0$), and $+\infty$ otherwise. This means that when computing the infimum on the left side of (27), we only need to take into account those nonnegative $\ell$.

Next, let $\Theta_{m+1}$ be also defined as before. Then, directly from the definition, we have:

$$\Theta_{m+1}^*(\ell) = \begin{cases} 0 & \text{if } \ell(\sum_i S_i w_i) = \sum_i P_{Z_i}(w_i), \quad \forall w_i \in C_b(\mathcal{Z}_i),\ i = 1, \ldots l; \\ +\infty & \text{otherwise.} \end{cases} \tag{73}$$

For any $\ell \in C_b^*(\mathcal{X})$. Generally, the condition in the first line of (73) does not imply that $\ell$ is a measure. However, if $\ell$ is also nonnegative, then using a technical result in [40] Lemma 1.25, we can further simplify:

$$\Theta_{m+1}^*(\ell) = \begin{cases} 0 & \text{if } \ell \in \mathcal{M}(\mathcal{X}) \text{ and } S_i^* \ell = P_{Z_i}, \quad i = 1, \ldots, l; \\ +\infty & \text{otherwise.} \end{cases} \tag{74}$$

This further shows that when we compute the left side of (27), the infimum can be taken over $\ell$, which is a coupling of $(P_{Z_i})$. In particular, if $\ell$ is a probability measure, then $\Theta_j^*(\ell) = c_j D(T_j^* \ell \| \mu_j)$ still holds with the $\Theta_j$ defined in (38), $j = 1, \ldots, m$. Thus, the rest of the proof can proceed as before. □

**Remark 7.** *The second assumption is made in order to achieve (74) in the proof.*

## 4. Gaussian Optimality

Recall that the conventional Brascamp-Lieb inequality and its reverse ((1) and (2)) state that centered Gaussian functions exhaust such inequalities, and in particular, verifying those inequalities is reduced to a finite dimensional optimization problem (only the covariance matrices in these Gaussian functions are to be optimized). In this section, we show that similar results hold for the forward-reverse Brascamp-Lieb inequality, as well. Our proof uses the rotational invariance argument mentioned in Section 1. Since the forward-reverse Brascamp-Lieb inequality has dual representations (Theorem 7), in principle, the rotational invariance argument can be applied either to the functional representation (as in Lieb's paper [29]) or the entropic representation (as in Geng-Nair [48]). Here, we adopt the latter approach. We first consider a certain "non-degenerate" case where the existence of an extremizer is guaranteed. Then, Gaussian optimality in the general case follows by a limiting argument (Appendix F), establishing Theorem 2.

### *4.1. Non-Degenerate Forward Channels*

This subsection focuses on the following case:

**Assumption 1.**

- *Fix Lebesgue measures* $(\mu_j)_{j=1}^m$ *and Gaussian measures* $(\nu_i)_{i=1}^l$ *on* $\mathbb{R}$*;*
- *non-degenerate (Definition 3 below) linear Gaussian random transformation* $(P_{Y_j|\mathbf{X}})_{j=1}^m$ *(where* $\mathbf{X} := (X_1, \dots, X_l)$*) associated with conditional expectation operators* $(T_j)_{j=1}^m$*;*
- $(S_i)_{i=1}^l$ *are induced by coordinate projections;*
- *positive* $(c_j)$ *and* $(b_i)$*.*

**Definition 3.** *We say* $(Q_{\mathbf{Y}_1|\mathbf{X}}, \dots, Q_{\mathbf{Y}_m|\mathbf{X}})$ *is non-degenerate if each* $Q_{\mathbf{Y}_j|\mathbf{X}=\mathbf{0}}$ *is an* $n_j$*-dimensional Gaussian distribution with an invertible covariance matrix.*

Given Borel measures $P_{X_i}$ on $\mathbb{R}$, $i = 1, \dots, l$, define:

$$F_0((P_{X_i})) := \inf_{P_{\mathbf{X}}} \sum_{j=1}^m c_j D(P_{Y_j} \| \mu_j) - \sum_{i=1}^l b_i D(P_{X_i} \| \nu_i) \tag{75}$$

where the infimum is over Borel measures $P_{\mathbf{X}}$ that have $(P_{X_i})$ as marginals. Note that (75) is well defined since the first term cannot be $+\infty$ under the non-degenerate assumption, and the second term cannot be $-\infty$. The aim of this subsection is to prove the following:

**Theorem 8.** $\sup_{(P_{X_i})} F_0((P_{X_i}))$*, where the supremum is over Borel measures* $P_{X_i}$ *on* $\mathbb{R}$*, and* $i = 1, \dots, l$*, is achieved by some Gaussian* $(P_{X_i})_{i=1}^l$*, in which case the infimum in (75) is achieved by some Gaussian* $P_{\mathbf{X}}$*.*

Naturally, one would expect that Gaussian optimality can be established when $(\mu_j)_{j=1}^m$ and $(\nu_i)_{i=1}^l$ are either Gaussian or Lebesgue. We made the assumption that the former is Lebesgue and the latter is Gaussian so that certain technical conditions can be justified more easily. More precisely, the following observation shows that we can regularize the distributions by a second moment constraint for free:

**Proposition 1.** $\sup_{(P_{X_i})} F_0((P_{X_i}))$ *is finite and there exist* $\sigma_i^2 \in (0, \infty)$*,* $i = 1, \dots, l$ *such that it equals:*

$$\sup_{(P_{X_i}) \colon \mathbb{E}[X_i^2] \le \sigma_i^2} F_0((P_{X_i})). \tag{76}$$

**Proof.** When $\mu_j$ is Lebesgue and $P_{Y_j|\mathbf{X}}$ is non-degenerate, $D(P_{Y_j}\|\mu_j) = -h(P_{Y_j}) \leq -h(P_{Y_j}|\mathbf{X})$ is bounded above (in terms of the variance of the additive noise of $P_{Y_j|\mathbf{X}}$). Moreover, $D(P_{X_i}\|\nu_i) \geq 0$ when $\nu_i$ is Gaussian, so $\sup_{(P_{X_i})} F_0((P_{X_i})) < \infty$. Further, choosing $(P_{X_i}) = (\nu_i)$ and using the covariance matrix to lower bound the first term in (75) show that $\sup_{(P_{X_i})} F_0((P_{X_i})) > -\infty$.

To see (76), notice that:

$$\begin{aligned} D(P_{X_i}\|\nu_i) &= D(P_{X_i}\|\nu_i') + \mathbb{E}[\imath_{\nu_i'\|\nu_i}(X)] \\ &= D(P_{X_i}\|\nu_i') + D(\nu_i'\|\nu_i) \\ &\geq D(\nu_i'\|\nu_i) \end{aligned} \tag{77}$$

where $\nu_i'$ is a Gaussian distribution with the same first and second moments as $X_i \sim P_{X_i}$. Thus, $D(P_{X_i}\|\nu_i)$ is bounded below by some function of the second moment of $X_i$, which tends to $\infty$ as the second moment of $X_i$ tends to $\infty$. Moreover, as argued in the preceding paragraph, the first term in (75) is bounded above by some constant depending only on $(P_{Y_j|\mathbf{X}})$. Thus, we can choose $\sigma_i^2 > 0$, $i = 1, \ldots, l$ large enough such that if $\mathbb{E}[X_i^2] > \sigma_i^2$ for some of $i$, then $F_0((P_{X_i})) < \sup_{(P_{X_i})} F_0((P_{X_i}))$, irrespective of the choices of $P_{X_1}, \ldots, P_{X_{i-1}}, P_{X_{i+1}}, \ldots, P_{X_l}$. Then, these $\sigma_1, \ldots, \sigma_l$ are as desired in the proposition. □

The non-degenerate assumption ensures that the supremum is achieved:

**Proposition 2.** *Under Assumption 1,*

1. *For any $(P_{X_i})_{i=1}^l$, the infimum in (75) is attained by some Borel $P_{\mathbf{X}}$.*
2. *If $(P_{Y_j|X^l})_{j=1}^m$ are non-degenerate (Definition 3), then the supremum in (76) is achieved by some Borel $(P_{X_i})_{i=1}^l$.*

The proof of Proposition 2 is given in Appendix E. After taking care of the existence of the extremizers, we get into the tensorization properties, which are the crux of the proof:

**Lemma 1.** *Fix $(P_{X_i^{(1)}})$, $(P_{X_i^{(2)}})$, $(\mu_j)$, $(T_j)$, $(c_j) \in [0,\infty)^m$, and let $S_j$ be induced by coordinate projections. Then:*

$$\inf_{P_{\mathbf{X}^{(1,2)}} : S_i^{*\otimes 2} P_{\mathbf{X}^{(1,2)}} = P_{X_i^{(1)}} \times P_{X_i^{(2)}}} \sum_{j=1}^m c_j D(P_{Y_j^{(1,2)}}\|\mu_j^{\otimes 2}) = \sum_{t=1,2} \sum_{j=1}^m c_j \inf_{P_{\mathbf{X}^{(t)}} : S_i^* P_{\mathbf{X}^{(t)}} = P_{X_i^{(t)}}} D(P_{Y_j^{(t)}}\|\mu_j) \tag{78}$$

*where for each $j$,*

$$P_{Y_j^{(1,2)}} := T_j^{*\otimes 2} P_{\mathbf{X}^{(1,2)}} \tag{79}$$

*on the left side and:*

$$P_{Y_j^{(t)}} := T_j^{*\otimes 2} P_{\mathbf{X}^{(t)}} \tag{80}$$

*on the right side, $t = 1, 2$.*

**Proof.** We only need to prove the nontrivial $\geq$ part. For any $P_{\mathbf{X}^{(1,2)}}$ on the left side, choose $P_{\mathbf{X}^{(t)}}$ on the right side by marginalization. Then:

$$\begin{aligned} D(P_{Y_j^{(1,2)}}\|\mu_j^{\otimes 2}) - \sum_t D(P_{Y_j^{(t)}}\|\mu_j) &= I(Y_j^{(1)}; Y_j^{(2)}) \\ &\geq 0 \end{aligned} \tag{81}$$

for each $j$. □

We are now ready to show the main result of this section.

**Proof of Theorem 8.**

1. Assume that $(P_{X_i^{(1)}})$ and $(P_{X_i^{(2)}})$ are maximizers of $F_0$ (possibly equal). Let $P_{X_i^{1,2}} := P_{X_i^{(1)}} \times P_{X_i^{(2)}}$. Define:

$$\mathbf{X}^+ := \frac{1}{\sqrt{2}}\left(\mathbf{X}^{(1)} + \mathbf{X}^{(2)}\right); \tag{82}$$

$$\mathbf{X}^- := \frac{1}{\sqrt{2}}\left(\mathbf{X}^{(1)} - \mathbf{X}^{(2)}\right). \tag{83}$$

Define $(Y_j^+)$ and $(Y_j^-)$ analogously. Then, $Y_j^+|\{\mathbf{X}^+ = \mathbf{x}^+, \mathbf{X}^- = \mathbf{x}^-\} \sim Q_{Y_j|\mathbf{X}=\mathbf{x}^+}$ is independent of $\mathbf{x}^-$, and $Y_j^-|\{\mathbf{X}^+ = \mathbf{x}^+, \mathbf{X}^- = \mathbf{x}^-\} \sim Q_{Y_j|\mathbf{X}=\mathbf{x}^-}$ is independent of $\mathbf{x}^+$.

2. Next, we perform the same algebraic expansion as in the proof of tensorization:

$$\sum_{t=1}^{2} F_0\left(\left(P_{X_i^{(t)}}\right)_{i=1}^{l}\right) = \inf_{P_{\mathbf{X}^{(1,2)}} : S_j^{*\otimes 2} P_{\mathbf{X}^{(1,2)}} = P_{X_j^{(1,2)}}} \sum_j c_j D(P_{Y_j^{(1,2)}} \| \mu_j^{\otimes 2}) - \sum_i b_i D(P_{X_i^{(1,2)}} \| \nu_i^{\otimes 2}) \tag{84}$$

$$= \inf_{P_{\mathbf{X}^+\mathbf{X}^-} : S_j^{*\otimes 2} P_{\mathbf{X}^+\mathbf{X}^-} = P_{X_j^+ X_j^-}} \sum_j c_j D(P_{Y_j^+ Y_j^-} \| \mu_j^{\otimes 2}) - \sum_i b_i D(P_{X_i^+ X_i^-} \| \nu_i^{\otimes 2}) \tag{85}$$

$$\leq \inf_{P_{\mathbf{X}^+\mathbf{X}^-} : S_j^{*\otimes 2} P_{\mathbf{X}^+\mathbf{X}^-} = P_{X_j^+ X_j^-}} \sum_j c_j \left[D(P_{Y_j^+} \| \mu_j) + D(P_{Y_j^-|\mathbf{X}^+} \| \mu_j | P_{\mathbf{X}^+})\right] - \sum_i b_i \left[D(P_{X_i^+} \| \nu_i) + D(P_{X_i^-|X_i^+} \| \nu_i | P_{X_i^+})\right] \tag{86}$$

$$\leq \sum_j c_j \left[D(P^\star_{Y_j^+} \| \mu_j) + D(P^\star_{Y_j^-|\mathbf{X}^+} \| \mu_j | P^\star_{\mathbf{X}^+})\right] - \sum_i b_i \left[D(P^\star_{X_i^+} \| \nu_i) + D(P^\star_{X_i^-|\mathbf{X}^+} \| \nu_i | P^\star_{\mathbf{X}^+})\right] \tag{87}$$

$$= F_0\left(\left(P^\star_{X_i^+}\right)_{i=1}^{l}\right) + \int F_0\left(\left(P^\star_{X_i^-|\mathbf{X}^+}\right)_{i=1}^{l}\right) \mathrm{d}P^\star_{\mathbf{X}^+} \tag{88}$$

$$\leq \sum_{t=1}^{2} F_0\left(\left(P_{X_i^{(t)}}\right)_{i=1}^{l}\right) \tag{89}$$

where:

- (84) uses Lemma 1.
- (86) is because of the Markov chain $Y_j^+ - \mathbf{X}^+ - Y_j^-$ (for any coupling).
- In (87), we selected a particular instance of coupling $P_{\mathbf{X}^+\mathbf{X}^-}$, constructed as follows: first, we select an optimal coupling $P_{\mathbf{X}^+}$ for given marginals $(P_{X_i^+})$. Then, for any $\mathbf{x}^+ = (x_i^+)_{i=1}^l$, let $P_{\mathbf{X}^-|\mathbf{X}^+=x^+}$ be an optimal coupling of $(P_{X_i^-|X_i^+=x_i^+})$ (for a justification that we can select optimal coupling $P_{\mathbf{X}^-|\mathbf{X}^+=\mathbf{x}^+}$ in a way that $P_{\mathbf{X}^-|\mathbf{X}^+}$ is indeed a regular conditional probability distribution, see [7]). With this construction, it is apparent that $X_i^+ - \mathbf{X}^+ - X_i^-$, and hence:

$$D(P_{X_i^-|X_i^+} \| \nu_i | P_{X_i^+}) = D(P_{X_i^-|\mathbf{X}^+} \| \nu_i | P_{\mathbf{X}^+}). \tag{90}$$

- (88) is because in the above, we have constructed the coupling optimally.
- (89) is because $(P^{(t)}_{X_i})$ maximizes $F_0$, $t = 1, 2$.

3. Thus, in the expansions above, equalities are attained throughout. Using the differentiation technique as in the case of forward inequality, for almost all $(b_i)$, $(c_j)$, we have:

$$D(P_{X_i^-|X_i^+}\|\nu_i|P_{X_i^+}) = D(P_{X_i^+}\|\nu_i) \tag{91}$$

$$= D(P_{X_i^-}\|\nu_i), \quad \forall i \tag{92}$$

where (92) is because by symmetry, we can perform the algebraic expansions in a different way to show that $(P_{X_i^-})$ is also a maximizer of $F_0$. Then, $I(X_i^+;X_i^-) = D(P_{X_i^-|X_i^+}\|\nu_i|P_{X_i^+}) - D(P_{X_i^-}\|\nu_i) = 0$, which, combined with $I(X_i^{(1)};X_i^{(2)})$, shows that $X_i^{(1)}$ and $X_i^{(2)}$ are Gaussian with the same covariance. Lastly, using Lemma 1 and the doubling trick, one can show that the optimal coupling is also Gaussian.

□

### *4.2. Analysis of Example 1 Using Gaussian Optimality*

We note that Example 1 is a rather simple setting, where (17) can be proven by integrating the two sides of (18) and applying the change of variables, noting that the absolute value of the Jacobian equals one. Nevertheless, it is illuminating to give an alternative proof using the Gaussian optimality result, as a proof of concept. In this section, we only give a proof sketch where certain "technicalities" are not justified. Details of the justifications are deferred to Appendix F.

**Proof sketch for the claim in Example 1.** By duality (Theorem 7), it suffices to prove the corresponding entropic inequality. The Gaussian optimality result in Theorem 8 assumed Gaussian reference measures on the output and non-degenerate forward channels in order to simplify the proof of the existence of minimizers; however, supposing that Gaussian optimality extends beyond those technical conditions, we see that it suffices to prove that for any centered Gaussian $(P_{X_i})$,

$$\sum_{i=1}^{l} h(P_{X_i}) \le \sup_{P_{X^l}} \sum_{j=1}^{l} h(P_{Y_j}) \tag{93}$$

where the supremum is over Gaussian $P_{X^l}$ with the marginals $P_{X_1},\dots,P_{X_l}$ and $Y_j := \sum_{i=1}^{l} m_{ji}X_i$. Let $a_i := \mathbb{E}[X_i^2]$, and choose $P_{X^l} = \prod_{i=1}^{l} P_{X_i}$; we see that (93) holds if:

$$\sum_{i=1}^{l} \log a_i \le \sum_{j=1}^{l} \log\left(\sum_{i=1}^{l} m_{ji}^2 a_i\right), \quad \forall a_i > 0,\ i = 1,\dots,l, \tag{94}$$

where $(a_i)$ are the eigenvalues and $\left(\sum_{i=1}^{l} m_{ji}a_i\right)_{i=1}^{l}$ are the diagonal entries of the matrix:

$$\mathbf{M}\mathrm{diag}(a_i)_{1\le i\le l}\mathbf{M}^\top. \tag{95}$$

Therefore, (94) holds. □

A generalization of Example 1 is as follows.

**Proposition 3.** *For any orthogonal matrix* $\mathbf{M} := (m_{ji})_{1\le j\le l, 1\le i\le l}$ *with nonzero entries, we claim that there exists a neighborhood* $\mathcal{U}$ *of the uniform probability vector* $(\frac{1}{l},\dots,\frac{1}{l})$*, such that for any* $(b_1,\dots,b_l)$ *and* $(c_1,\dots,c_l)$ *in* $\mathcal{U}$*, the best constant* $D$ *in the FR-BLinequality* (16) *equals* $\exp(H(c^l) - H(b^l))$ *where* $H(\cdot)$ *is the entropy functional.*

The proposition generalizes the claim in Example 1. Indeed, observe that there is no loss of generality in assuming that $(b_1,\dots,b_l)$ and $(c_1,\dots,c_l)$ are probability vectors, since by dimensional

analysis, we see that the best constant is infinite unless $\sum_{i=1}^{l} b_i = \sum_{j=1}^{l} c_j$; and it is also clear that the best constant is invariant when each $b_i$ and $c_j$ is multiplied by the same positive number. Moreover, any orthogonal matrix can be approximated by a sequence of orthogonal $\mathbf{M}$ with nonzero entries, for which the neighborhood $\mathcal{U}$ shrinks, but always contains the uniform probability vector $(\frac{1}{l}, \ldots, \frac{1}{l})$.

**Proof sketch for Proposition 3.** Note that along the same lines as (94), the best constant in the FR-BL inequality equals:

$$D = \sup_{a^l \in \Delta} \frac{\prod_{i=1}^{l} a_i^{b_i}}{\sup_{\mathbf{A} \succeq \mathbf{0}:\ \mathbf{A}_{ii} = a_i} \prod_{j=1}^{l} \left[\mathbf{M}\mathbf{A}\mathbf{M}^\top\right]_{jj}^{c_j}} \tag{96}$$

where without loss of generality, we assumed $a^l \in \Delta$ is in the probability simplex. We first observe that if the positive semidefinite constraint $\mathbf{A} \succeq \mathbf{0}$ in (96) were nonexistent, then the sup in the denominator in (96) would equal $\prod_{j=1}^{l} c_j^{c_j}$, and consequently, (96) would equal $\exp(H(c^l) - H(b^l))$, for any $b^l, c^l \in \Delta$ not necessarily close to the uniform probability vector. Indeed, fixing $\mathbf{A}_{ii} = a_i, i = 1, \ldots, l$, the linear map from the off-diagonal entries to the diagonal entries of $\mathbf{M}\mathbf{A}\mathbf{M}^\top$ is onto the space of $l$-vectors whose entries sum to one; proof of the surjectivity can be reduced to checking the fact that the only diagonal matrix that commutes with $\mathbf{M}$ is a multiple of the identity matrix. Then, the sup in the denominator is achieved when $\left[\mathbf{M}\mathbf{A}\mathbf{M}^\top\right]_{jj} = c_j, j = 1 \ldots l$, which is independent of $a^l$.

Next, we argue that the constraint $\mathbf{A} \succeq \mathbf{0}$ in (96) is not active when $b^l$ and $c^l$ are close to the uniform vector. Denote by $\mathcal{U}(t)$ the set of $l$-vectors whose distance (say in total variation) to the uniform vector $(\frac{1}{l}, \ldots, \frac{1}{l})$ is at most $t$. Observe that:

1. There exists $t > 0$ such that for every $a^l \in \mathcal{U}(t)$,

$$\sup_{\mathbf{A} \succeq \mathbf{0}:\ \mathbf{A}_{ii} = a_i} \prod_{j=1}^{l} \left[\mathbf{M}\mathbf{A}\mathbf{M}^\top\right]_{jj} = 1/l^l \tag{97}$$

   which follows by continuity and the fact that when $a^l$ is uniform, the sup (97) is achieved at the strictly positive definite $\mathbf{A} = l^{-1}\mathbf{I}$.
2. When $b^l = c^l = (\frac{1}{l}, \ldots, \frac{1}{l})$ is the uniform probability vector, (96) equals one, which is uniquely achieved by $a^l = (\frac{1}{l}, \ldots, \frac{1}{l})$. To see the uniqueness, take $\mathbf{A}$ to be diagonal in the denominator and observe that the denominator is strictly bigger than the numerator when the diagonals of $\mathbf{M}\mathbf{A}\mathbf{M}^\top$ are not a permutation of $a^l$. Then, since the extreme value of a continuous functions is achieved on a compact set, we can find $\epsilon > 0$ such that:

$$\frac{\prod_{i=1}^{l} a_i^{1/l}}{\sup_{\mathbf{A} \succeq \mathbf{0}:\ \mathbf{A}_{ii} = a_i} \prod_{j=1}^{l} \left[\mathbf{M}\mathbf{A}\mathbf{M}^\top\right]_{jj}^{1/l}} < 1 - \epsilon \tag{98}$$

   for any $a^l \notin \mathcal{U}(t/2)$.
3. Finally, by continuity, we can choose $s \in (0, t/2)$ small enough such that for any $b^l, c^l \in \mathcal{U}(s)$,

$$\frac{\prod_{i=1}^{l} a_i^{b_i}}{\sup_{\mathbf{A} \succeq \mathbf{0}:\ \mathbf{A}_{ii} = a_i} \prod_{j=1}^{l} \left[\mathbf{M}\mathbf{A}\mathbf{M}^\top\right]_{jj}^{c_j}} < 1 - \epsilon/2, \quad \forall a^l \notin \mathcal{U}(t/2); \tag{99}$$

$$\sup_{\mathbf{A} \succeq \mathbf{0}:\ \mathbf{A}_{ii} = a_i} \prod_{j=1}^{l} \left[\mathbf{M}\mathbf{A}\mathbf{M}^\top\right]_{jj}^{c_j} = \sup_{\mathbf{A}:\ \mathbf{A}_{ii} = a_i} \prod_{j=1}^{l} \left[\mathbf{M}\mathbf{A}\mathbf{M}^\top\right]_{jj}^{c_j}, \quad \forall a^l \in \mathcal{U}(t/2); \tag{100}$$

$$\exp(H(c^l) - H(b^l)) > 1 - \epsilon/2. \tag{101}$$

Taking the neighborhood $\mathcal{U}(s)$ proves the claim. □

## 5. Relation to Hypercontractivity and Its Reverses

As alluded to before and illustrated by Figure 2, the forward-reverse Brascamp-Lieb inequality generalizes several other inequalities from functional analysis and information theory; a more complete discussion on these relationships can be found in [7]. In this section, we focus on hypercontractivity and show how its three cases all follow from Theorem 1. Among these, the case in Section 5.3 can be regarded as an instance of the forward-reverse inequality that cannot be reduced to either the forward or the reverse inequality alone. It is also interesting to note that, from the viewpoint of the forward-reverse Brascamp-Lieb inequality, in each of the three special cases, there ought to be three functions involved in the functional formulation; however, the optimal choice of one function can be computed from the other two. Therefore, the conventional functional formulations of the three cases of hypercontractivity involve only two functions, making it non-obvious to find a unifying inequality.

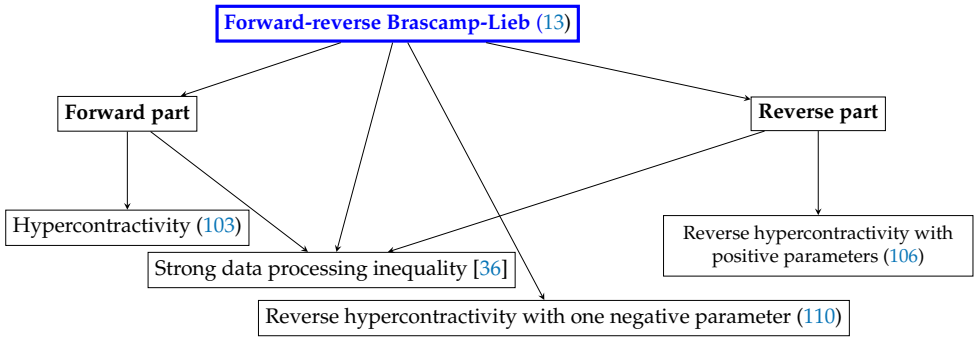


**Figure 2.** The forward-reverse Brascamp-Lieb inequality generalizes several other functional inequalities/information theoretic inequalities. For more discussions on these relations, see the extended version [7].

### 5.1. Hypercontractivity

Fix a joint probability distribution $Q_{Y_1Y_2}$ and nonnegative continuous functions $F_1$ and $F_2$ on $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively, both bounded away from zero. In Theorem 1, take $l \leftarrow 1$, $m \leftarrow 2$, $b_1 \leftarrow 1$, $d \leftarrow 0$, $f_1 \leftarrow F_1^{\frac{1}{c_1}}$, $f_2 \leftarrow F_2^{\frac{1}{c_2}}$, $\nu_1 \leftarrow Q_{Y_1Y_2}$, $\mu_1 \leftarrow Q_{Y_1}$, $\mu_2 \leftarrow Q_{Y_2}$. Furthermore, put $Z_1 = X = (Y_1, Y_2)$, and let $T_1$ and $T_2$ be the canonical maps (Definition 2). The measure spaces and the random transformations are as shown in Figure 3.

$$\mathcal{P}(\mathcal{Z}_1) \xleftarrow{\cong} \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2) \begin{array}{l} \xrightarrow{T_1^*} \mathcal{P}(\mathcal{Y}_1) \\ \xrightarrow{T_2^*} \mathcal{P}(\mathcal{Y}_2) \end{array}$$

**Figure 3.** Diagram for hypercontractivity.

The constraint (12) translates to:

$$g_1(y_1, y_2) \leq F_1(y_1)F_2(y_2), \quad \forall y_1, y_2 \tag{102}$$

and the optimal choice of $g_1$ is when the equality is achieved. We thus obtain the equivalence between:

$$\|F_1\|_{\frac{1}{c_1}} \|F_2\|_{\frac{1}{c_2}} \geq \mathbb{E}[F_1(Y_1)F_2(Y_2)], \quad \forall F_1 \in L^{\frac{1}{c_1}}(Q_{Y_1}),\ F_2 \in L^{\frac{1}{c_2}}(Q_{Y_2}) \tag{103}$$

and:

$$\forall P_{Y_1Y_2}, \quad D(P_{Y_1Y_2}\|Q_{Y_1Y_2}) \geq c_1 D(P_{Y_1}\|Q_{Y_1}) + c_2 D(P_{Y_2}\|Q_{Y_2}). \tag{104}$$

By a standard dense-subspace argument, we see that it is inconsequential that $F_1$ and $F_2$ in (103) are not assumed to be continuous, nor bounded away from zero. It is also easy to see that the nonnegativity of $F_1$ and $F_2$ is inconsequential for (103).

This equivalence can also be obtained from Theorem 1. By Hölder's inequality, (103) is equivalent to saying that the norm of the linear operator sending $F_1 \in L^{\frac{1}{c_1}}(Q_{Y_1})$ to $\mathbb{E}[F_1(Y_1)|Y_2 = \cdot] \in L^{\frac{1}{1-c_2}}(Q_{Y_2})$ does not exceed one. The interesting case is $\frac{1}{1-c_2} > \frac{1}{c_1}$, hence the name hypercontractivity. The equivalent formulation of hypercontractivity was shown in [44] using a different proof via the method of types/typicality, which requires that $|\mathcal{Y}_1|, |\mathcal{Y}_2| < \infty$. In contrast, the proof based on the nonnegativity of relative entropy removes this constraint, allowing one to prove Nelson's Gaussian hypercontractivity from the information-theoretic formulation (see [7]).

### 5.2. Reverse Hypercontractivity (Positive Parameters)

By "positive parameters" we mean the $b_1$ and $b_2$ in (107) are positive.

Let $Q_{Z_1Z_2}$ be a given joint probability distribution, and let $G_1$ and $G_2$ be nonnegative functions on $\mathcal{Z}_1$ and $\mathcal{Z}_2$, respectively, both bounded away from zero. In Theorem 1, take $l \leftarrow 2$, $m \leftarrow 1$, $c_1 \leftarrow 1$, $d \leftarrow 0$, $g_1 \leftarrow G_1^{\frac{1}{b_1}}$, $g_2 \leftarrow G_2^{\frac{1}{b_2}}$, $\mu_1 \leftarrow Q_{Z_1Z_2}$, $\nu_1 \leftarrow Q_{Z_1}$, $\nu_2 \leftarrow Q_{Z_2}$. Furthermore, put $Y_1 = X = (Z_1, Z_2)$, and let $S_1$ and $S_2$ be the canonical maps (Definition 2). The measure spaces and the random transformations are as shown in Figure 4.

$$\begin{array}{ccccc} \mathcal{P}(\mathcal{Z}_1) & \xleftarrow{S_1^*} & & & \\ & & \mathcal{P}(\mathcal{Z}_1 \times \mathcal{Z}_2) & \xrightarrow{\cong} & \mathcal{P}(\mathcal{Y}_1) \\ \mathcal{P}(\mathcal{Z}_2) & \xleftarrow{S_2^*} & & & \end{array}$$

**Figure 4.** Diagram for reverse hypercontractivity.

Note that the constraint (12) translates to:

$$f_1(z_1, z_2) \geq G_1(z_1)G_2(z_2), \quad \forall z_1, z_2. \tag{105}$$

and the equality case yields the optimal choice of $f_1$ for (13). By Theorem 1, we thus obtain the equivalence between:

$$\|G_1\|_{\frac{1}{b_1}} \|G_2\|_{\frac{1}{b_2}} \leq \mathbb{E}[G_1(Z_1)G_2(Z_2)], \quad \forall G_1, G_2 \tag{106}$$

and:

$$\forall P_{Z_1}, P_{Z_2}, \exists P_{Z_1Z_2}, \quad D(P_{Z_1Z_2}\|Q_{Z_1Z_2}) \leq b_1 D(P_{Z_1}\|Q_{Z_1}) + b_2 D(P_{Z_2}\|Q_{Z_2}). \tag{107}$$

Note that in this setup, if $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are finite, then Condition (iv) in Theorem 1 is equivalent to $Q_{Z_1Z_2} \ll Q_{Z_1} \times Q_{Z_2}$. The equivalent formulations of reverse hypercontractivity were observed in [59], where the proof is based on the method of types.

### 5.3. Reverse Hypercontractivity (One Negative Parameter)

By "one negative parameter" we mean the $b_1$ is positive and $-c_2$ is negative in (111).

In Theorem 1, take $l \leftarrow 1$, $m \leftarrow 2$, $c_1 \leftarrow 1$, $d \leftarrow 0$. Let $Y_1 = X = (Z_1, Y_2)$, and let $S_1$ and $T_2$ be the canonical maps (Definition 2). Suppose that $Q_{Z_1Y_2}$ is a given joint probability distribution, and

set $\mu_1 \leftarrow Q_{Z_1Y_2}$, $\nu_1 \leftarrow Q_{Z_1}$, $\mu_2 \leftarrow Q_{Y_2}$ in Theorem 1. Suppose that $F$ and $G$ are arbitrary nonnegative continuous functions on $\mathcal{Y}_2$ and $\mathcal{Z}_1$, respectively, which are bounded away from zero. Take $g_1 \leftarrow G^{\frac{1}{b_1}}$, $f_2 \leftarrow F^{-\frac{1}{c_2}}$. in Theorem 1. The measure spaces and the random transformations are as shown in Figure 5.

$$\begin{array}{ccccc} & & & \xrightarrow{\cong} & \mathcal{P}(\mathcal{Y}_1) \\ \mathcal{P}(\mathcal{Z}_1) & \xleftarrow{S_1^*} & \mathcal{P}(\mathcal{Z}_1 \times \mathcal{Y}_2) & & \\ & & & \xrightarrow[T_2^*]{} & \mathcal{P}(\mathcal{Y}_2) \end{array}$$

**Figure 5.** Diagram for reverse hypercontractivity with one negative parameter.

The constraint (12) translates to:

$$f_1(z_1, y_2) \geq G(z_1)F(y_2), \quad \forall z_1, y_2. \tag{108}$$

Note that (13) translates to:

$$\|G\|_{\frac{1}{b_1}} \leq Q_{Y_2Z_1}(f_1)Q_{Y_2}^{c_2}(F^{-\frac{1}{c_2}}) \tag{109}$$

for all $F$, $G$ and $f_1$ satisfying (108). It suffices to verify (109) for the optimal choice $f_1 = GF$, so (109) is reduced to:

$$\|F\|_{\frac{1}{-c_2}}\|G\|_{\frac{1}{b_1}} \leq \mathbb{E}[F(Y_2)G(Z_1)], \quad \forall F, G. \tag{110}$$

By Theorem 1, (110) is equivalent to:

$$\forall P_{Z_1}, \exists P_{Z_1Y_2}, \quad D(P_{Z_1Y_2}\|Q_{Z_1Y_2}) \leq b_1 D(P_{Z_1}\|Q_{Z_1}) + (-c_2)D(P_{Y_2}\|Q_{Y_2}). \tag{111}$$

Inequality (110) is called reverse hypercontractivity with a negative parameter in [45], where the entropic version (111) is established for $|\mathcal{Z}_1|, |\mathcal{Y}_2| < \infty$ using the method of types. Multiterminal extensions of (110) and (111) (called the reverse Brascamp-Lieb type inequality with negative parameters in [45]) can also be recovered from Theorem 1 in the same fashion, i.e., we move all negative parameters to the other side of the inequality so that all parameters become positive.

In summary, from the viewpoint of Theorem 1, the results in Sections 5.1–5.3 are degenerate special cases, in the sense that in any of the three cases, the optimal choice of one of the functions in (13) can be explicitly expressed in terms of the other functions; hence, this "hidden function" disappears in (103), (106) or (110).

**Author Contributions:** All the authors have contributed to the problem formulation, refinement, structuring or editing of the paper. Most of the sections were written by J.L. Parts of the sections on the existence of the minimizer and the Gaussian optimality were written by T.A.C.

**Acknowledgments:** This work was supported in part by NSF Grants CCF-1528132, CCF-0939370 (Center for Science of Information), CCF-1319299, CCF-1319304, CCF-1350595 and AFOSR FA9550-15-1-0180. Jingbo Liu would like to thank Elliott H. Lieb for teaching the Brascamp-Lieb inequality, as well as some techniques used in this paper in his graduate class.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Recovering Theorem 1 from Theorem 6 as a Special Case

Assume that $P_X \to (P_{Z_i})$ is surjective. Let $1_{\mathcal{Z}_i}$ denote the constant one function on $\mathcal{Z}_i$. Define:

$$\mathcal{C} := \left\{ (w_i) \colon w_i \in C_b(\mathcal{Z}_i),\ \sum_{i=1}^{l} \inf_{z_i} w_i(z_i) \geq 0 \right\}, \tag{A1}$$

which is a closed convex cone in $C_b(\mathcal{Z}_1) \times \cdots \times C_b(\mathcal{Z}_l)$. Given $(g_i)$, we show that $\sum_{i=1}^{l} b_i S_i \log \tilde{g}_i \geq \sum_{i=1}^{l} b_i S_i \log g_i$ implies:

$$(b_i \log \tilde{g}_i - b_i \log g_i)_{i=1}^{l} \in \mathcal{C}. \tag{A2}$$

Indeed, we can verify that the dual cone:

$$\begin{aligned} \mathcal{C}^* &:= \left\{ (\pi_i) \colon \sum_{i=1}^{l} \pi_i(w_i) \geq 0,\ \forall (w_i) \in \mathcal{C} \right\} \\ &= \left\{ \lambda(P_{Z_1}, \ldots, P_{Z_l}) \colon \lambda \geq 0 \right\}. \end{aligned} \tag{A3}$$

Under the surjectivity assumption, we see:

$$\sum_{i=1}^{l} \pi_i (b_i \log \tilde{g}_i - b_i \log g_i) \geq 0, \quad \forall (\pi_i) \in \mathcal{C}^*. \tag{A4}$$

Now, if (A2) is not true, by the Hahn-Banach theorem (Theorem 5), we find $\pi_i \in \mathcal{M}(\mathcal{Z}_i)$, $i = 1, \ldots, l$ such that:

$$\sum_{i=1}^{l} \pi_i (b_i \log \tilde{g}_i - b_i \log g_i) < \inf_{(w_i) \in \mathcal{C}} \sum_{i=1}^{l} \pi_i(w_i) \tag{A5}$$

so the right side of (A5) is not $-\infty$. Since $\mathcal{C}$ is a cone containing the origin, the right side of (A5) hence must be nonnegative, and we conclude that $(\pi_i) \in \mathcal{C}^*$. However, then, (A5) contradicts (A4).

## Appendix B. Existence of Weakly-Convergent Couplings

This section proves an auxiliary result which will be used in Appendix C.

**Lemma A1.** *Suppose that for each $i = 1, \ldots, l$, $P_{X_i}$ is a Borel measure on $\mathbb{R}$ and $P_{X_i}^{(n)}$ converges weakly to some absolutely continuous (with respect to the Lebesgue measure) $P_{X_i}$ as $n \to \infty$. If $P_{\mathbf{X}}$ is a coupling of $(P_{X_i})_{1 \leq i \leq l}$, then, upon extraction of a subsequence, there exist couplings $P_{\mathbf{X}}^{(n)}$ for $(P_{X_i}^{(n)})_{1 \leq i \leq l}$ that converge weakly to $P_{\mathbf{X}}$ as $n \to \infty$.*

**Proof.** For each integer $k \geq 1$, define the random variable $W_i^{[k]} := \phi_k(X_i)$ where $\phi_k \colon \mathbb{R} \to \mathbb{R} \cup \{\mathsf{e}\}$ is the following "dyadic quantization function":

$$\phi_k \colon x \mapsto \begin{cases} \lfloor 2^k x \rfloor & |x| \leq k,\ x \notin 2^{-k}\mathbb{Z}; \\ \mathsf{e} & \text{otherwise,} \end{cases} \tag{A6}$$

and let $\mathbf{W}^{[k]} := (W_i^{[k]})_{i=1}^l$. Denote by $\mathcal{W}^{[k]} := \{-k2^k, \ldots, k2^k - 1, \mathrm{e}\}$ the set from which $W_i^{[k]}$ takes values. Note that since $P_{X_i}$ is assumed to be absolutely continuous, the set of "dyadic points" has measure zero:

$$P_{X_i}\left(\bigcup_{k=1}^{\infty} 2^{-k}\mathbb{Z}\right) = 0, \quad i = 1, \ldots, l. \tag{A7}$$

Since $P_{X_i}^{(n)} \to P_{X_i}$ weakly and the assumption in the preceding paragraph precluded any positive mass on the quantization boundaries under $P_{X_i}$, for each $k \geq 1$, there exists some $n := n_k$ large enough such that:

$$P_{W_i^{[k]}}^{(n)}(w) \geq \left(1 - \frac{1}{k}\right) P_{W_i^{[k]}}(w), \tag{A8}$$

for each $i$ and $w \in \mathcal{W}^{[k]}$. Now, define a coupling $P_{\mathbf{W}^{[k]}}^{(n)}$ compatible with the $\left(P_{W_i^{[k]}}^{(n)}\right)_{i=1}^l$ induced by $\left(P_{X_i}^{(n)}\right)_{i=1}^l$, as follows:

$$P_{\mathbf{W}^{[k]}}^{(n)} := \left(1 - \frac{1}{k}\right) P_{\mathbf{W}^{[k]}} + k^{l-1} \prod_{i=1}^{l} \left(P_{W_i^{[k]}}^{(n)} - \left(1 - \frac{1}{k}\right) P_{W_i^{[k]}}\right). \tag{A9}$$

Observe that (A9) is a well-defined probability measure because of (A8) and indeed has marginals $\left(P_{W_i^{[k]}}^{(n)}\right)_{i=1}^l$. Moreover, by the triangle inequality, we have the following bound on the total variation distance:

$$\left|P_{\mathbf{W}^{[k]}}^{(n)} - P_{\mathbf{W}^{[k]}}\right| \leq \frac{2}{k}. \tag{A10}$$

Next, construct $P_{\mathbf{X}}^{(n)}$ (we use $P|_{\mathcal{A}}$ to denote the restriction of a probability measure $P$ on measurable set $\mathcal{A}$, that is $P|_{\mathcal{A}}(\mathcal{B}) := P(\mathcal{A} \cap \mathcal{B})$ for any measurable $\mathcal{B}$):

$$P_{\mathbf{X}}^{(n)} := \sum_{w^l \in \mathcal{W}^{[k]} \times \cdots \times \mathcal{W}^{[k]}} \frac{P_{\mathbf{W}^{[k]}}^{(n)}\left(w^l\right)}{\prod_{i=1}^l P_{W_i^{[k]}}^{(n)}(w_i)} \prod_{i=1}^{l} P_{X_i}^{(n)}|_{\phi_k^{-1}(w_i)}. \tag{A11}$$

Observe that $P_{\mathbf{X}}^{(n)}$ defined in (A11) is compatible with the $P_{\mathbf{W}^{[k]}}^{(n)}$ defined in (A9) and indeed has marginals $(P_{X_i}^{(n)})_{i=1}^l$. Since $n := n_k$ can be made increasing in $k$, we have constructed the desired sequence $(P_{\mathbf{X}}^{(n_k)})_{k=1}^{\infty}$ converging weakly to $P_{\mathbf{X}}$. Indeed, for any bounded open dyadic cube (that is, a cube whose corners have coordinates being multiples of $2^{-k}$ where $k$ is some integer) $\mathcal{A}$, using (A10) and the assumption (A7), we conclude:

$$\liminf_{k \to \infty} P_{\mathbf{X}}^{(n_k)}(\mathcal{A}) \geq P_{\mathbf{X}}(\mathcal{A}). \tag{A12}$$

Moreover, since bounded open dyadic cubes form a countable basis of the topology in $\mathbb{R}^l$, we see that (A12) actually holds for any open set $\mathcal{A}$. By writing $\mathcal{A}$ as a countable union of dyadic cubes, using the continuity of measure to pass to a finite disjoint union, and then apply (A12), as desired. □

## Appendix C. Upper Semicontinuity of the Infimum

Using Lemma A1 in Appendix B, we prove the following result, which will be used in Appendix E.

**Corollary A1.** *Consider non-degenerate* $(P_{Y_j|\mathbf{X}})$. *For each* $n \geq 1$, $i = 1, \ldots, l$, $P_{X_i}^{(n)}$ *is a Borel measure on* $\mathbb{R}$, *whose second moment is bounded by* $\sigma_i^2 < \infty$. *Assume that* $P_{X_i}^{(n)}$ *converges to some absolutely continuous* $P_{X_i}^\star$ *for each i. Then:*

$$\limsup_{n\to\infty} \inf_{P_{\mathbf{X}}:\, S_i^* P_{\mathbf{X}} = P_{X_i}^{(n)}} \sum_{j=1}^m c_j D(T_j^* P_{\mathbf{X}} \| \mu_j) \leq \inf_{P_{\mathbf{X}}:\, S_i^* P_{\mathbf{X}} = P_{X_i}^\star} \sum_{j=1}^m c_j D(T_j^* P_{\mathbf{X}} \| \mu_j). \tag{A13}$$

**Proof.** By passing to a convergent subsequence, we may assume that the limit on the left side of (A13) exists. For any coupling $P_{\mathbf{X}}^\star$ of $(P_{X_i}^\star)$, by invoking Lemma A1 and passing to a subsequence, we find a sequence of couplings $P_{\mathbf{X}}^{(n)}$ of $(P_{X_i}^{(n)})$ that converges weakly to $P_{\mathbf{X}}^\star$. It is known that under a moment constraint, the differential entropy of the output distribution of a non-degenerate Gaussian channel enjoys weak continuity in the input distribution (see, e.g., [48] Proposition 18, [60] Theorem 7, or [61] Theorem 1 and Theorem 2). Thus:

$$\lim_{n\to\infty} \sum_{j=1}^m c_j D(T_j^* P_{\mathbf{X}}^{(n)} \| \mu_j) = \sum_{j=1}^m c_j D(T_j^* P_{\mathbf{X}} \| \mu_j) \tag{A14}$$

and (A13) follows since $P_{\mathbf{X}}^\star$ was arbitrarily chosen. □

## Appendix D. Weak Semicontinuity of Differential Entropy under a Moment Constraint

This section proves the following result, which will be used in Appendix E.

**Lemma A2.** *Suppose* $(P_{\mathbf{X}_n})$ *is a sequence of distributions on* $\mathbb{R}^d$ *converging weakly to* $P_{\mathbf{X}^\star}$, *and:*

$$\mathbb{E}[\mathbf{X}_n \mathbf{X}_n^\top] \preceq \mathbf{\Sigma} \tag{A15}$$

*for all n. Then*

$$\limsup_{n\to\infty} h(\mathbf{X}_n) \leq h(\mathbf{X}^\star). \tag{A16}$$

**Remark A1.** *The result fails without the condition* (A15). *Furthermore, related results when the weak convergence is replaced with pointwise convergence of density functions and certain additional constraints were shown in [61] (Theorem 1 and Theorem 2) (see also the proof of [48] (Theorem 5)). Those results are not applicable here since the density functions of* $\mathbf{X}_n$ *do not converge pointwise. They are applicable for the problems discussed in [48] because the density functions of the output of the Gaussian random transformation enjoy many nice properties due to the smoothing effect of the "good kernel".*

**Proof.** It is well known that in metric spaces and for probability measures, the relative entropy is weakly lower semicontinuous (cf. [58]). This fact and a scaling argument immediately show that, for any $r > 0$,

$$\limsup_{n\to\infty} h(\mathbf{X}_n | \|\mathbf{X}_n\| \leq r) \leq h(\mathbf{X}^\star | \|\mathbf{X}^\star\| \leq r). \tag{A17}$$

Let $p_n(r) := \mathbb{P}[\|\mathbf{X}_n\| > r]$, then (A15) implies:

$$\mathbb{E}[\mathbf{X}\mathbf{X}^\top | \|\mathbf{X}_n\| > r] \leq \frac{1}{p_n(r)} \mathbf{\Sigma}. \tag{A18}$$

Therefore, since the Gaussian distribution maximizes differential entropy given a second moment upper bound, we have:

$$h(\mathbf{X}_n | \|\mathbf{X}_n\| > r) \leq \frac{1}{2} \log \frac{(2\pi)^d e |\mathbf{\Sigma}|}{p_n(r)}. \tag{A19}$$

Since $\lim_{r\to\infty} \sup_n p_n(r) = 0$ by (A15) and due to Chebyshev's inequality, (A19) implies that:

$$\lim_{r\to\infty} \sup_n p_n(r) h(\mathbf{X}_n | \|\mathbf{X}_n\| > r) = 0. \tag{A20}$$

The desired result follows from (A17), (A20) and the fact that:

$$h(\mathbf{X}_n) = p_n(r) h(\mathbf{X}_n | \|\mathbf{X}_n\| > r) + (1 - p_n(r)) h(\mathbf{X}_n | \|\mathbf{X}_n\| \leq r) + h(p_n(r)). \tag{A21}$$

□

## Appendix E. Proof of Proposition 2

1. For any $\epsilon > 0$, by the continuity of measure, there exists $K > 0$ such that:

$$P_{X_i}([-K, K]) \geq 1 - \frac{\epsilon}{l}, \quad i = 1, \ldots, l. \tag{A22}$$

By the union bound,

$$P_{\mathbf{X}}([-K, K]^l) \geq 1 - \epsilon \tag{A23}$$

wherever $P_{\mathbf{X}}$ is a coupling of $(P_{X_i})$. Now, let $P_{\mathbf{X}}^{(n)}$, $n = 1, 2, \ldots$ be such that:

$$\lim_{n\to\infty} \sum_{j=1}^m c_j D(P_{Y_j}^{(n)} \| \mu_j) = \inf_{P_{\mathbf{X}}} \sum_{j=1}^m c_j D(P_{Y_j} \| \mu_j) \tag{A24}$$

where $P_{Y_j} := T_j^* P_{\mathbf{X}}$, $j = 1, \ldots, m$. The sequence $(P_{\mathbf{X}}^{(n)})$ is tight by (A23). Thus, invoking the Prokhorov theorem and by passing to a subsequence, we may assume that $(P_{\mathbf{X}}^{(n)})$ converges weakly to some $P_{\mathbf{X}}^\star$. Therefore, $P_{Y_j}^{(n)}$ converges to $P_{Y_j}^\star$ weakly, and by the semicontinuity property in Lemma A2, we have:

$$\sum_{j=1}^m c_j D(P_{Y_j}^\star \| \mu_j) \leq \lim_{n\to\infty} \sum_{j=1}^m c_j D(P_{Y_j}^{(n)} \| \mu_j) \tag{A25}$$

establishing that $P_{\mathbf{X}}^\star$ is an infimizer.

2. Suppose $(P_{X_i}^{(n)})_{1\leq i\leq l, n\geq 1}$ is such that $\mathbb{E}[X_i^2] \leq \sigma_i^2$, $X_i \sim P_{X_i}^{(n)}$, where $(\sigma_i)$ is as in Proposition 1 and:

$$\lim_{n\to\infty} F_0\left((P_{X_i}^{(n)})_{i=1}^l\right) = \sup_{(P_{X_i}) \colon \mathbf{\Sigma}_{X_i} \preceq \sigma_i^2} F_0((P_{X_i})_{i=1}^l). \tag{A26}$$

The regularization on the covariance implies that for each $i$, $(P_{X_i}^{(n)})_{n\geq 1}$ is a tight sequence. Thus, upon the extraction of subsequences, we may assume that for each $i$, $(P_{X_i}^{(n)})_{n\geq 1}$ converges to some $P_{X_i}^\star$. We have the moment bound:

$$\mathbb{E}[X_i^2] = \lim_{K\to\infty} \mathbb{E}[\min\{X_i^2, K\}] \tag{A27}$$

$$= \lim_{K\to\infty} \mathbb{E}[\min\{(X_i^{(n)})^2, K\}] \tag{A28}$$

$$\leq \sigma_i^2 \tag{A29}$$

where $X_i \sim P^\star_{X_i}$ and $X_i^{(n)} \sim P^{(n)}_{X_i}$. Then, by Lemma A2,

$$\sum_i b_i D(P^\star_{X_i} \| \nu_i) \leq \lim_{n\to\infty} \sum_i b_i D(P^{(n)}_{X_i} \| \nu_i) \tag{A30}$$

Under the covariance regularization and the nondegenerateness assumption, we showed in Proposition 1 that the value of (76) cannot be $+\infty$ or $-\infty$. This implies that we can assume (by passing to a subsequence) that $P^{(n)}_{X_i} \ll \lambda$, $i = 1, \ldots, l$, since otherwise $F((P_{X_i})) = -\infty$. Moreover, since $\left(\sum_j c_j D(P^{(n)}_{Y_j} \| \mu_j)\right)_{n\geq 1}$ is bounded above under the nondegenerateness assumption, the sequence $\left(\sum_i b_i D(P^{(n)}_{X_i} \| \nu_i)\right)_{n\geq 1}$ must also be bounded from above, which implies, using (A30), that:

$$\sum_i b_i D(P^\star_{X_i} \| \nu_i) < \infty. \tag{A31}$$

In particular, we have $P^\star_{X_i} \ll \lambda$ for each $i$. Now, Corollary A1 shows that:

$$\inf_{P_{\mathbf{X}}:\, S_i^* P_{\mathbf{X}} = P^\star_{X_i}} \sum_j c_j D(T_j^* P_{\mathbf{X}} \| \mu_j) \geq \lim_{n\to\infty} \inf_{P_{\mathbf{X}}:\, S_i^* P_{\mathbf{X}} = P^{(n)}_{X_i}} \sum_j c_j D(T_j^* P_{\mathbf{X}} \| \mu_j) \tag{A32}$$

Thus, (A30) and (A32) show that $(P^\star_{X_i})$ is in fact a maximizer.

## Appendix F. Gaussian Optimality in Degenerate Cases: A Limiting Argument

This section proves Theorem 2. We first give a proof for the choice of parameters in Example 1, merely for the sake of notational simplicity, and then discuss how to extend the argument.

### *Appendix F.1. Proof of the Claim in Example 1*

The proof will be based on Theorem 8, which assumes non-degenerate forward channels and Gaussian measures on the output of the reverse channels. To that end, we will adopt an approximation argument. For each $j = 1, \ldots, l$, define the linear operator $T_j^\epsilon$ by:

$$(T_j^\epsilon \phi)(x_1, \ldots, x_l) := \mathbb{E}\left[\phi\left(\sum_{i=1}^l m_{ji} x_i + N_\epsilon\right)\right] \tag{A33}$$

for any measurable function $\phi$ on $\mathbb{R}$, where $N_\epsilon \sim \mathcal{N}(0, \epsilon)$. Let $\gamma_{\frac{1}{\epsilon}} := \mathcal{N}(0, \epsilon^{-1})$, and note that the density of $\sqrt{\frac{2\pi}{\epsilon}}\gamma_{\frac{1}{\epsilon}}$ converges pointwise to that of the Lebesgue measure.

**Lemma A3.** *For any $\epsilon > 0$, let $(T_j^\epsilon)$ be defined as in (A33). Then, for any Borel $P_{X_i} \ll \lambda$, $i = 1, \ldots, l$,*

$$\sum_{i=1}^l D(P_{X_i} \| \gamma_{\frac{1}{\epsilon}}) - \frac{l}{2} \log \frac{2\pi}{\epsilon} \geq \inf_{P_{X^l}:\, S_i^* P_{X^l} = P_{X_i}} \left\{ -\sum_{j=1}^l h(T_j^{\epsilon *} P_{X^l}) \right\}. \tag{A34}$$

**Proof.** By Theorem 8, it suffices to prove (A34) when $P_{X_i}$ is Gaussian, and from (A34), it is easy to see that it suffices to prove the case of the centered Gaussian. Let $P_{X_i} = \mathcal{N}(0, a_i)$, $i = 1, \ldots, l$. We can

upper bound the right side of (A34) by taking $P_{X^l} = P_{X_1} \times P_{X_l}$ instead of the infimum, so it suffices to prove that:

$$\frac{\epsilon}{2}\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l} \log a_i \geq -\frac{1}{2}\sum_{j=1}^{l} \log\left(\sum_{i=1}^{l} m_{ji}^2 a_i + \epsilon\right) \tag{A35}$$

for any $\epsilon, a_1, \ldots, a_l \in (0, \infty)$. This is implied by the $\epsilon = 0$ case, which we proved in (94). □

By the duality of the forward-reverse Brascamp-Lieb inequality (Theorem 7) , we conclude from Lemma A3 that:

**Lemma A4.** *For any $\epsilon > 0$ and nonnegative continuous $(f_j)$, $(g_i)$ satisfying:*

$$\sum_{i=1}^{l} \log g_i(x_i) \leq \sum_{j=1}^{l} (T_j^{\epsilon} \log f_j)\left(x^l\right), \quad \forall x^l \in \mathbb{R}^l, \tag{A36}$$

*we have:*

$$\left(\frac{2\pi}{\epsilon}\right)^{\frac{l}{2}} \prod_{i=1}^{l} \int g_i \mathrm{d}\gamma_{\frac{1}{\epsilon}} \leq \prod_{i=1}^{l} \int f_j(x)\mathrm{d}x. \tag{A37}$$

Now, suppose that the claim in Example 1 is not true; then there are nonnegative continuous $(f_j)$ and $(g_i)$ satisfying (17) while:

$$\prod_{i=1}^{l} \int g_i(x)\mathrm{d}x > \prod_{i=1}^{l} \int f_j(x)\mathrm{d}x, \tag{A38}$$

By the standard approximation argument, we can assume, without loss of generality, that:

$$g_i(x) = 0, \quad \forall x \colon |x| \geq R, 1 \leq i \leq l; \tag{A39}$$

$$f_j(x) \geq \delta e^{-x^2}, \quad \forall 1 \leq j \leq l, \tag{A40}$$

for some $R$ sufficiently large and $\delta > 0$ sufficiently small. Note that for any $x^l \in [-R, R]^l$,

$$\sum_{i=1}^{l} m_{ji} x_i \in [-\sqrt{l}R, \sqrt{l}R]. \tag{A41}$$

Since $\log f_j$ is uniformly continuous on $[-2\sqrt{l}R, 2\sqrt{l}R]$ for each $j$ and since we assumed (A40), we have:

$$\lim_{\epsilon \to 0} \inf_{x^l \in [-R,R]^l} \left\{ \sum_{j=1}^{l} (T_j^{\epsilon} \log f_j)\left(x^l\right) - \sum_{j=1}^{l} (T_j^{0} \log f_j)\left(x^l\right) \right\} \geq 0. \tag{A42}$$

However, since we assumed (17) and (A39), we must also have:

$$\lim_{\epsilon \to 0} \eta_{\epsilon} \geq 0 \tag{A43}$$

where:

$$\eta_{\epsilon} := \inf_{x^l \in \mathbb{R}^l} \left\{ \sum_{j=1}^{l} (T_j^{\epsilon} \log f_j)\left(x^l\right) - \sum_{i=1}^{l} \log g_i(x_i) \right\}. \tag{A44}$$

Put:

$$\tilde{g}_1^\epsilon := \exp(\eta_\epsilon) g_1, \tag{A45}$$

$$\tilde{g}_i^\epsilon := g_i, \quad i = 1, \ldots, l. \tag{A46}$$

Then, $(\tilde{g}_i^\epsilon)$ and $(f_j)$ satisfy the constraint (A36) for any $\epsilon > 0$. By applying the monotone convergence theorem and then Lemma A4,

$$\prod_{i=1}^{l} \int g_i(x_i) \mathrm{d}x_i \leq \lim_{\epsilon \to 0} \left( \frac{2\pi}{\epsilon} \right)^{\frac{l}{2}} \prod_{i=1}^{l} \int \tilde{g}_i^\epsilon \mathrm{d}\gamma_{\frac{1}{\epsilon}} \tag{A47}$$

$$\leq \prod_{i=1}^{l} \int f_j(x) \mathrm{d}x \tag{A48}$$

which violates the hypothesis (A38), as desired.

*Appendix F.2. Proof of Theorem 2*

The limiting argument can be extended to the vector case to prove Theorem 2. Specifically, for each $j = 1, \ldots, m$, define $T_j^\epsilon$ the same as (A33) except that $\mathbf{N}_\epsilon \sim \mathcal{N}(\mathbf{0}, \epsilon \mathbf{I})$, where $\mathbf{I}$ is the identity matrix whose dimension is clear from the context (equal to $\dim(E^j)$ here), and let $P^\epsilon_{\mathbf{Y}_j | \mathbf{X}_1 \ldots \mathbf{X}_l}$ be the dual operator. For each $i = 1, \ldots, l$, let $\nu_i^\epsilon := \left(\frac{2\pi}{\epsilon}\right)^{\frac{1}{2}\dim(E_i)} \cdot \mathcal{N}(\mathbf{0}, \epsilon^{-1}\mathbf{I})$, whose density convergences pointwise to that of $\nu_i^0$, defined as the Lebesgue measure on $E_i$. Define:

$$d^\epsilon := \sup \left\{ \sum_{i=1}^{l} b_i \log \nu_i^\epsilon(g_i) - \sum_{j=1}^{m} c_j \log \int f_j \right\} \tag{A49}$$

where the supremum is over nonnegative continuous functions $f_1, \ldots, f_m$ and $g_1, \ldots, g_l$ such that the summands in (A49) are finite and:

$$\sum_{i=1}^{l} b_i \log g_i(\mathbf{x}_i) \leq \sum_{j=1}^{m} c_j (T_j^\epsilon \log f_j)(\mathbf{x}_1, \ldots, \mathbf{x}_l), \quad \forall \mathbf{x}_1, \ldots, \mathbf{x}_l. \tag{A50}$$

The same limiting argument (A38)–(A48) extended to the vector case shows that:

$$d^0 \leq \lim_{\epsilon \downarrow 0} d^\epsilon. \tag{A51}$$

Next, define $F_0^\epsilon(\cdot)$ for $(\mu_j)$, $(\nu_i^\epsilon)$ and $P^\epsilon_{\mathbf{Y}_j | \mathbf{X}_1 \ldots \mathbf{X}_l}$, similarly to (75). The entropic⇒functional argument shows that:

$$d^\epsilon \leq \sup_{P_{\mathbf{X}_1}, \ldots, P_{\mathbf{X}_l}} F_0^\epsilon(P_{\mathbf{X}_1}, \ldots, P_{\mathbf{X}_l}). \tag{A52}$$

However, Theorem 8 based on the rotational invariance of the Gaussian measure can be extended to the vector case, so for any $\epsilon > 0$,

$$\sup_{P_{\mathbf{X}_1}, \ldots, P_{\mathbf{X}_l}} F_0^\epsilon(P_{\mathbf{X}_1}, \ldots, P_{\mathbf{X}_l}) = \sup_{P_{\mathbf{X}_1}, \ldots, P_{\mathbf{X}_l} \text{ c.G.}} F_0^\epsilon(P_{\mathbf{X}_1}, \ldots, P_{\mathbf{X}_l}), \tag{A53}$$

where c.G. means that the supremum on the right side is over centered Gaussian measures. The fact that centered distributions exhaust the supremum follows easily from the definition of $F_0$. Moreover, from the definitions, it is easy to see that $F_0^\epsilon$ is monotonically decreasing in $\epsilon$, and in particular:

$$\sup_{P_{\mathbf{X}_1},\ldots,P_{\mathbf{X}_l}\text{ c.G.}} F_0^\epsilon(P_{\mathbf{X}_1},\ldots,P_{\mathbf{X}_l}) \leq \sup_{P_{\mathbf{X}_1},\ldots,P_{\mathbf{X}_l}\text{ c.G.}} F_0^0(P_{\mathbf{X}_1},\ldots,P_{\mathbf{X}_l}). \tag{A54}$$

To finish the proof with the above chain of inequalities, it only remains to show that the right side of (A54) equals the supremum in (A49) with $(f_j)$ $(g_j)$ taken over center Gaussian functions. This follows by similar steps as the proof of the functional⇒entropic part of Theorem 1. We briefly mention how the idea works: suppose $A$ is the linear space defined as the Cartesian product of $\mathbb{R}$ and the set of $n \times n$ symmetric matrices. Let $\Lambda(\cdot)$ be the convex functional on $A$ defined by:

$$\Lambda(r,\mathbf{M}) := \ln \int \exp_e\left(r + \mathbf{x}^\top \mathbf{M}\mathbf{x}\right) \mathrm{d}\,\mathbf{x} \tag{A55}$$

$$= \begin{cases} r + \frac{n}{2}\ln\pi - \frac{1}{2}\ln|-\mathbf{M}| & \mathbf{M} \preceq \mathbf{0}, \\ +\infty & \text{otherwise.} \end{cases} \tag{A56}$$

The dual space of $A$ is itself, and $\Lambda^*$ is given by:

$$\Lambda^*(s,\mathbf{H}) = \sup_{r,\,\mathbf{M}\succeq\mathbf{0}} \{sr + \mathrm{Tr}(\mathbf{H}^\top\mathbf{M}) - \Lambda(r,\mathbf{M})\}. \tag{A57}$$

Then, $\Lambda^*(s,\mathbf{H}) = +\infty$ if $s \neq 1$, and:

$$\Lambda^*(1,\mathbf{H}) = \sup_{\mathbf{M}\preceq\mathbf{0}} \left\{\mathrm{Tr}(\mathbf{H}^\top\mathbf{M}) - \frac{n}{2}\ln\pi + \frac{1}{2}\ln|-\mathbf{M}|\right\}. \tag{A58}$$

The supremum in (A58) equals $+\infty$ if $\mathbf{H}$ is not positive-semidefinite. However, if $\mathbf{H}$ is positive-semidefinite, the supremum equals $-\frac{1}{2}\ln 2\pi e|\mathbf{H}|$, which is equal to the relative entropy between $\mathcal{N}(\mathbf{0},\mathbf{H})$ and the Lebesgue measure (supremum achieved when $\mathbf{M} = -(2\mathbf{H})^{-1}$). Since the proof of Theorem 1, in essence, only uses the duality between convex functionals, the same algebraic steps therein also establish the desired matrix optimization identity.

## References


1. Brascamp, H.J.; Lieb, E.H. Best constants in Young's inequality, its converse, and its generalization to more than three functions. *Adv. Math.* **1976**, *20*, 151–173. [CrossRef]
2. Brascamp, H.J.; Lieb, E.H. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *J. Funct. Anal.* **1976**, *22*, 366–389. [CrossRef]
3. Bobkov, S.G.; Ledoux, M. From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities. *Geom. Funct. Anal.* **2000**, *10*, 1028–1052. [CrossRef]
4. Cordero-Erausquin, D. Transport inequalities for log-concave measures, quantitative forms and applications. *arXiv* **2015**, arXiv:1504.06147.
5. Barthe, F. On a reverse form of the Brascamp-Lieb inequality. *Invent. Math.* **1998**, *134*, 335–361. [CrossRef]
6. Bennett, J.; Carbery, A.; Christ, M.; Tao, T. The Brascamp-Lieb inequalities: finiteness, structure and extremals. *Geom. Funct. Anal.* **2008**, *17*, 1343–1415. [CrossRef]
7. Liu, J.; Courtade, T.A.; Cuff, P.; Verdú, S. Information theoretic perspectives on Brascamp-Lieb inequality and its reverse. *arXiv* **2017**, arXiv:1702.06260.
8. Gardner, R. The Brunn-Minkowski inequality. *Bull. Am. Math. Soc.* **2002**, *39*, 355–405. [CrossRef]
9. Gross, L. Logarithmic Sobolev inequalities. *Am. J. Math.* **1975**, *97*, 1061–1083. [CrossRef]
10. Erkip, E.; Cover, T.M. The efficiency of investment information. *IEEE Trans. Inf. Theory Mar.* **1998**, *44*, 1026–1040. [CrossRef]

11. Courtade, T. Outer bounds for multiterminal source coding via a strong data processing inequality. In Proceedings of the IEEE International Symposium on Information Theory, Istanbul, Turkey, 7–12 July 2013; pp. 559–563.
12. Polyanskiy, Y.; Wu, Y. Dissipation of information in channels with input constraints. *IEEE Trans. Inf. Theory* **2016**, *62*, 35–55. [CrossRef]
13. Polyanskiy, Y.; Wu, Y. A Note on the Strong Data-Processing Inequalities in Bayesian Networks. Available online: http://arxiv.org/pdf/1508.06025v1.pdf (accessed on 25 August 2015).
14. Liu, J.; Cuff, P.; Verdú, S. Key capacity for product sources with application to stationary Gaussian processes. *IEEE Trans. Inf. Theory* **2016**, *62*, 984–1005.
15. Liu, J.; Cuff, P.; Verdú, S. Secret key generation with one communicator and a one-shot converse via hypercontractivity. In Proceedings of the IEEE International Symposium on Information Theory, Hong Kong, China, 14–19 June 2015; pp. 710–714.
16. Xu, A.; Raginsky, M. Converses for distributed estimation via strong data processing inequalities. In Proceedings of the IEEE International Symposium on Information Theory, Hong Kong, China, 14–19 June 2015; pp. 2376–2380.
17. Kamath, S.; Anantharam, V. On non-interactive simulation of joint distributions. *arXiv* **2015**, arXiv:1505.00769.
18. Kahn, J.; Kalai, G.; Linial, N. The influence of variables on Boolean functions. In Proceedings of the 29th Annual Symposium on Foundations of Computer Science, White Plains, NY, USA, 24–26 October 1988; pp. 68–80.
19. Ganor, A.; Kol, G.; Raz, R. Exponential separation of information and communication. In Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS), Philadelphia, PA, USA, 18–21 Otctober 2014; pp. 176–185.
20. Dvir, Z.; Hu, G. Sylvester-Gallai for arrangements of subspaces. *arXiv* **2014**, arXiv:1412.0795.
21. Braverman, M.; Garg, A.; Ma, T.; Nguyen, H.L.; Woodruff, D.P. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. *arXiv* **2015**, arXiv:1506.07216.
22. Garg, A.; Gurvits, L.; Oliveira, R.; Wigderson, A. Algorithmic aspects of Brascamp-Lieb inequalities. *arXiv* **2016**, arXiv:1607.06711.
23. Talagrand, M. On Russo's approximate zero-one law. *Ann. Probab.* **1994**, *22*, 1576–1587. [CrossRef]
24. Friedgut, E.; Kalai, G.; Naor, A. Boolean functions whose Fourier transform is concentrated on the first two levels. *Adv. Appl. Math.* **2002**, *29*, 427–437. [CrossRef]
25. Bourgain, J. On the distribution of the Fourier spectrum of Boolean functions. *Isr. J. Math.* **2002**, *131*, 269–276. [CrossRef]
26. Mossel, E.; O'Donnell, R.; Oleszkiewicz, K. Noise stability of functions with low influences: Invariance and optimality. *Ann. Math.* **2010**, *171*, 295–341. [CrossRef]
27. Garban, C.; Pete, G.; Schramm, O. The Fourier spectrum of critical percolation. *Acta Math.* **2010**, *205*, 19–104. [CrossRef]
28. Duchi, J.C.; Jordan, M.; Wainwright, M.J. Local privacy and statistical minimax rates. In Proceedings of the IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, CA, USA, 26–29 October 2013; pp. 429–438.
29. Lieb, E.H. Gaussian kernels have only Gaussian maximizers. *Invent. Math.* **1990**, *102*, 179–208. [CrossRef]
30. Barthe, F. Optimal Young's inequality and its converse: A simple proof. *Geom. Funct. Anal.* **1998**, *8*, 234–242. [CrossRef]
31. Barthe, F.; Cordero-Erausquin, D. Inverse Brascamp-Lieb inequalities along the Heat equation. In *Geometric Aspects of Functional Analysis*; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 2004; Volume 1850; pp. 65–71.
32. Carlen, E.A.; Cordero-Erausquin, D. Subadditivity of the entropy and its relation to Brascamp-Lieb type inequalities. *Geom. Funct. Anal.* **2009**, *19*, 373–405. [CrossRef]
33. Barthe, F.; Cordero-Erausquin, D.; Ledoux, M.; Maurey, B. Correlation and Brascamp-Lieb inequalities for Markov semigroups. *Int. Math. Res. Notices* **2011**, *2011*, 2177–2216. [CrossRef]
34. Lehec, J. Short probabilistic proof of the Brascamp-Lieb and Barthe theorems. *Can. Math. Bull.* **2014**, *57*, 585–587. [CrossRef]
35. Ball, K. Volumes of sections of cubes and related problems. In *Geometric Aspects of Functional Analysis*; Springer: Berlin/Heidelberg, Germany, 1989; pp. 251–260.
36. Ahlswede, R.; Gács, P. Spreading of sets in product spaces and hypercontraction of the Markov operator. *Ann. Probab.* **1976**, *4*, 925–939. [CrossRef]

37. Csiszár, I.; Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2011.
38. Liu, J.; van Handel, R.; Verdú, S. Beyond the Blowing-Up Lemma: Sharp Converses via Reverse Hypercontractivity. In Proceedings of the IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017; pp. 943–947.
39. Ahlswede, R.; Gács, P.; Körner, J. Bounds on conditional probabilities with applications in multi-user communication. *Probab. Theory Relat. Fields* **1976**, *34*, 157–177. [CrossRef]
40. Villani, C. *Topics in Optimal Transportation*; American Mathematical Soc.: Providence, RI, USA, 2003; Volume 58.
41. Atar, R.; Merhav, N. Information-theoretic applications of the logarithmic probability comparison bound. *IEEE Trans. Inf. Theory* **2015**, *61*, 5366–5386. [CrossRef]
42. Radhakrishnan, J. Entropy and counting. In *Kharagpur Golden Jubilee Volume*; Narosa: New Delhi, India, 2001.
43. Madiman, M.M.; Tetali, P. Information inequalities for joint distributions, with interpretations and applications. *IEEE Trans. Inf. Theory* **2010**, *56*, 2699–2713. [CrossRef]
44. Nair, C. *Equivalent Formulations of Hypercontractivity Using Information Measures*; International Zurich Seminar: Zurich, Switzerland, 2014.
45. Beigi, S.; Nair, C. Equivalent characterization of reverse Brascamp-Lieb type inequalities using information measures. In Proceedings of the IEEE International Symposium on Information Theory, Barcelona, Spain, 10–15 July 2016.
46. Bobkov, S.G.; Götze, F. Exponential integrability and transportation cost related to Logarithmic Sobolev inequalities. *J. Funct. Anal.* **1999**, *163*, 1–28. [CrossRef]
47. Carlen, E.A.; Lieb, E.H.; Loss, M. A sharp analog of Young's inequality on $S^N$ and related entropy inequalities. *J. Geom. Anal.* **2004**, *14*, 487–520. [CrossRef]
48. Geng, Y.; Nair, C. The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages. *IEEE Trans. Inf. Theory* **2014**, *60*, 2087–2104. [CrossRef]
49. Liu, J.; Courtade, T.A.; Cuff, P.; Verdú, S. Brascamp-Lieb inequality and its reverse: An information theoretic view. In Proceedings of the IEEE International Symposium on Information Theory, Barcelona, Spain, 10–15 July 2016; pp. 1048–1052.
50. Lax, P.D. *Functional Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2002.
51. Tao, T. 245B, Notes 12: Continuous Functions on Locally Compact Hausdorff Spaces. Available online: https://terrytao.wordpress.com/2009/03/02/245b-notes-12-continuous-functions-on-locally-compact-hausdorff-spaces/ (accessed on 2 March 2009).
52. Bourbaki, N. *Intégration*; (Chaps. I-IV, Actualités Scientifiques et Industrielles, no. 1175); Hermann: Paris, France, 1952.
53. Dembo, A.; Zeitouni, O. *Large Deviations Techniques and Applications*; Springer: Berlin, Germany, 2009; Volume 38.
54. Lane, S.M. *Categories for the Working Mathematician*; Springer: New York, NY, USA, 1978.
55. Hatcher, A. *Algebraic Topology*; Tsinghua University Press: Beijing, China, 2002.
56. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 2015.
57. Prokhorov, Y.V. Convergence of random processes and limit theorems in probability theory. *Theory Probab. Its Appl.* **1956**, *1*, 157–214. [CrossRef]
58. Verdú, S. *Information Theory*; In preparation, 2018.
59. Kamath, S. Reverse hypercontractivity using information measures. In Proceedings of the 53rd Annual Allerton Conference on Communications, Control and Computing, Champaign, IL, USA, 30 September–2 October 2015; pp. 627–633.
60. Wu, Y.; Verdú, S. Functional properties of minimum mean-square error and mutual information. *IEEE Trans. Inf. Theory* **2012**, *58*, 1289–1301. [CrossRef]
61. Godavarti, M.; Hero, A. Convergence of differential entropies. *IEEE Trans. Inf. Theory* **2004**, *50*, 171–176. [CrossRef]