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

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

1 Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544; jingbo.verdu@princeton.edu
2 Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, 94720-1770; courtade@berkeley.edu
3 Renaissance Technologies; paul.cuff@gmail.com
* Correspondence: jingbo@princeton.edu

Abstract: Inspired by the forward and the reverse channels from the image-size characterization problem in network information theory, we introduce a functional inequality which 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. 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 $B_i : E \to E_i$ be linear maps. Let $(c_i)_{i=1}^m$ and $D$ be positive real numbers. Then the Brascamp-Lieb inequality

$$\int \prod_{i=1}^m f_i(x_i) \, dx \leq D \prod_{i=1}^m \left( \int f_i(x_i) \, dx_i \right)^{c_i},$$

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. Similarly, for $F$ a positive real number, the reverse Brascamp-Lieb inequality, also known as Barthe’s inequality,

$$\int \sup_{y_i : \sum_{i=1}^m c_i B_i^* y_i = x} \prod_{i=1}^m f_i(y_i) \, dx \geq F \prod_{i=1}^m \left( \int f_i(y_i) \, dy_i \right)^{c_i},$$

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.

1 Not to be confused with the “variance Brascamp-Lieb inequality” (cf. [2][3][4]), which generalizes the Poincaré inequality.
2 [5, Theorem 1] actually contains additional assumptions, which make the best constants $D$ and $F$ positive and finite, but are not really necessary for the conclusion to hold ([5, Remark 1]).
3 A centered Gaussian function is of the form $x \mapsto \exp(-x^\top A x)$, where $A$ is a positive semidefinite matrix and $r \in \mathbb{R}$.
4 $B_i^*$ denotes the adjoint of $B_i$.® 2018 by the author(s). Distributed under a Creative Commons CC BY license.
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 the Barthe’s inequality. Due in part to their utility in establishing impossibility bounds, these functional inequalities have attracted a lot of attention in information theory [10][11][12][13][14][15][16][17], theoretical computer science [18][19][20][21][22], and statistics [23][24][25][26][27][28], to name only a small subset of the literature. Over the years, various proofs of these inequalities have been proposed [1][29][30][31]. 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

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

(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

$$D \cdot F = 1$$

(4)

as long as $D, F < \infty$ [5]. Moreover, both $D$ and $F$ are equal to 1 under Ball’s geometric condition [32]: $E_1, \ldots, E_m$ are dimension 1 and

$$\sum_{i=1}^{m} c_i B_i B_i^* = I$$

(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

$$g(x) \leq \prod_{i=1}^{m} f_i^c (B_i x), \quad \forall x,$$

we have

$$\int_{E} g \leq D \prod_{j=1}^{m} \left( \int_{E_j} f_j \right)^{c_j}.$$

(7)

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

$$\prod_{i=1}^{l} g_i^{b_i} (z_i) \leq f (\sum_{i=1}^{l} b_i B_i^* z_i), \quad \forall z_1, \ldots, z_l,$$

we have

$$\prod_{i=1}^{l} g_i^{b_i} (z_i) \leq f (\sum_{i=1}^{l} b_i B_i^* z_i), \quad \forall z_1, \ldots, z_l,$$

(8)
we have
\[ \prod_{i=1}^{l} \left( \int_{E_i} g_i \right)^{b_i} \leq D \int_E f. \]  

(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 \( X, Y_1, \ldots, Y_m, Z_1, \ldots, Z_l \) be measurable spaces. Consider measurable maps \( \phi_j: X \to Y_j, j = 1, \ldots, m \) and \( \psi_i: X \to Z_i, i = 1, \ldots, l \). 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 \( Z_1, \ldots, Z_l \), and \( \mu_1, \ldots, \mu_m \) be measures on \( Y_1, \ldots, Y_m \), respectively. What is the smallest \( D > 0 \) such that for all nonnegative \( f_1, \ldots, f_m \) on \( Y_1, \ldots, Y_m \) and \( g_1, \ldots, g_l \) on \( Z_1, \ldots, Z_l \) satisfying

\[ \prod_{i=1}^{l} g_i^{b_i}(\psi_i(x)) \leq \prod_{j=1}^{m} f_j^{c_j}(\phi_j(x)), \quad \forall x, \]  

(10)

we have

\[ \prod_{i=1}^{l} \left( \int g_i \, d\nu_i \right)^{b_i} \leq D \prod_{j=1}^{m} \left( \int f_j \, d\mu_j \right)^{c_j}. \]  

(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 Ahlswede, Csiszár and Körner [33][34] 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). Although not the way treated in [33][34], 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, [35] uses functional inequalities that are variants of (11) together with a reverse hypercontractivity machinery to improve the image-size plus blowing-up machinery of [36], 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. [31][37][38][33][39][40][41][42][43]) it is known that certain integral inequalities are equivalent to inequalities involving relative entropies. In particular, Carlen, Loss and Lieb [44] and Carlen and Cordero-Erausquin [31] 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 different in the text). In fact, Ahlswede, Csiszár and Körner [36][34] essentially derived image-size inequalities from similar entropic inequalities. Because of the reverse part, the proof of equivalence of (11) and
corresponding entropic inequality is more involved than the forward case considered in [31] beyond the case of finite $X, \mathcal{Y}_i, \mathcal{Z}_i$, and certain machineries from min-max theory appear 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 forward-reverse Brascamp-Lieb inequality). Assume that

1. If the nonnegative continuous functions $(g_i), (f_j)$ are bounded away from 0 and satisfy
   
   \[ \sum_{i=1}^l b_i Q_{Z_i|X}(g_i) \leq \sum_{j=1}^m c_j Q_{Y_j|X}(f_j) \]  
   (12)

   then

   \[ \prod_{i=1}^l \left( \int g_i \, dv_i \right)^{b_i} \leq \exp(d) \prod_{j=1}^m \left( \int f_j \, d\nu_j \right)^{c_j} \]  
   (13)

2. For any $(P_{Z_i})$ such that $D(P_{Z_i}||v_i) < \infty^5, i = 1, \ldots, l$,
   
   \[ \sum_{i=1}^l b_i D(P_{Z_i}||v_i) + d \geq \inf_{P_X} \sum_{j=1}^m c_j D(P_{Y_j}||\mu_j) \]  
   (14)

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

   for each $i = 1, \ldots, l$.

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 function

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 [45], thereby taking advantage of the dual formulation of Theorem 1.

---

5 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).
Forward-reverse Brascamp-Lieb (13)

Forward part

Strong data processing inequality [33]

Hypercontractivity (108)

Reverse part

Reverse hypercontractivity with one negative parameter (115)

Reverse hypercontractivity with positive parameters (111)

Figure 1. 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].

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 \( B_{ji} : E_i \to E^j \) be a linear map for each \( i \in \{1, \ldots, l\} \) and \( j \in \{1, \ldots, m\} \). Then, for all continuous functions \( f_j : E^j \to [0, +\infty), g_i : E_i \to [0, \infty) \) satisfying

\[
\prod_{i=1}^l g_i^{b_i}(x_i) \leq \prod_{j=1}^m f_j^{c_j} \left( \sum_{i=1}^l m_{ji} x_i \right), \quad \forall x_1, \ldots, x_l, \quad (15)
\]

we have

\[
\prod_{i=1}^l \left( \int g_i \right)^{b_i} \leq D \prod_{j=1}^m \left( \int f_j \right)^{c_j}, \quad (16)
\]

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 1. Next, we also identify a particular case where the best constant in the forward-reverse inequality equals 1:

Theorem 3. Let \( l \) be a positive integer, and let \( M := (m_{ji})_{1 \leq j \leq l, 1 \leq i \leq l} \) be an orthogonal matrix. For any nonnegative continuous functions \( (f_i)_{i=1}^l, (g_i)_{i=1}^l \) on \( \mathbb{R}^l \) such that

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

we have

\[
\prod_{i=1}^l \int g_i(x) dx \leq \prod_{j=1}^l \int f_j(x) dx. \quad (18)
\]

The rest of the paper is organized as follows: Section 2 defines 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, under a certain “non-degenerate” assumption where the linear maps \( B_{ji} \)'s are regularized by an additive noise, which guarantees the existence of extremizers. Then, a limiting
argument in Appendix F lets the noise vanish, which, combined with the equivalence between the functional and entropic formulations, establishes Theorem 2 and Theorem 3.

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. Of course, this 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 [47] based on the KKT condition.

Notation 1. Let $X$ be a topological space.

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

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. [48][49].

Theorem 4 (Riesz-Markov-Kakutani). If $X$ is a locally compact, $\sigma$-compact Polish space, the dual $\mathcal{M}(X)$ of both $C_c(X)$ and $C_0(X)$ is $\mathcal{M}(X)$.

Remark 1. The dual space of $C_b(X)$ can be strictly larger than $\mathcal{M}(X)$, since it also contains those linear functionals that depend on the “limit at infinity” of a function $f \in C_b(X)$ (originally defined for those $f$ that do have a limit at infinity, and then extended to the whole $C_b(X)$ by the Hahn-Banach theorem; see e.g. [48]).

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

$$ f \mapsto \int f \, d\mu $$

where $f$ is a function in $C_0(X)$ or $C_c(X)$. As is well known, Theorem 4 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; this justifies the shorthand notation

$$ \mu(f) := \int f \, d\mu $$

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.

---

6 A Polish space is a complete separable metric space. It enjoys several nice properties that we use heavily in this section, including Prokhorov theorem and Riesz-Kakutani theorem (the latter is related to the fact that every Borel probability measure on a Polish space is inner regular, hence a Radon measure). Short introductions on the Polish space can be found in e.g. [37][46].

7 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).

8 In fact, some authors prefer to construct measure theory by defining a measure as a linear functional on a suitable measure space; see Lax [48] or Bourbaki [50].
Definition 1. Let \( \Lambda : C_b(\mathcal{X}) \to (-\infty, +\infty] \) be a lower semicontinuous, proper convex function. Its Legendre-Fenchel transform \( \Lambda^*: C_b(\mathcal{X})^* \to (-\infty, +\infty] \) is given by

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

(21)

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

\[
\Lambda(f) := \log \nu(\exp(f)) \quad = \log \int \exp(f) \, dv.
\]  

(22)

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

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

(24)

which agrees with the more familiar definition \( D(\mu || \nu) := \mu(\log \frac{d\nu}{d\mu}) \) when \( \nu \) is a probability measure, by the Donsker-Varadhan formula (c.f. [46, 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 \( T : C_b(\mathcal{Y}) \to C_b(\mathcal{X}) \), the dual operator \( T^* : C_b(\mathcal{X})^* \to C_b(\mathcal{Y})^* \) is defined in terms of

\[
T^*\mu_X : C_b(\mathcal{Y}) \to \mathbb{R}; \quad f \mapsto \mu_X(Tf),
\]  

(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. Also, 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 [51],[52]), \( C_b \) is a functor from the category of topological spaces to the category of topological vector spaces, which is contra-variant because for any continuous, \( \phi : \mathcal{X} \to \mathcal{Y} \) (morphism between topological spaces), we have \( C_b(\phi) : C_b(\mathcal{Y}) \to C_b(\mathcal{X}) \), \( u \mapsto \phi \circ u \) where \( \phi \circ u \) 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 \( \mathcal{M}(\phi) : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y}) \), \( \mu \mapsto \mu \circ \phi^{-1} \), where \( \mu \circ \phi^{-1}(B) := \mu(\phi^{-1}(B)) \) for any Borel measurable \( B \subseteq \mathcal{Y} \). “Duality” itself is a contra-variant functor between the category of topological spaces (note the reversal of arrows in Fig. 2). 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 \( \phi : \mathcal{X} \to \mathcal{Y} \) is continuous. Definition 2 can therefore be viewed as the special case where \( \phi \) is the projection map:

Definition 2. Suppose \( \phi : Z_1 \times Z_2 \to Z_1 \), \( (z_1, z_2) \mapsto z_1 \) is the projection to the first coordinate.

- \( C_b(\phi) : C_b(\mathcal{Z}_1) \to C_b(\mathcal{Z}_1 \times Z_2) \) is called a canonical map, whose action is almost trivial: it sends a function of \( z_1 \) to itself, but viewed as a function of \( (z_1, z_2) \).

- \( \mathcal{M}(\phi) : \mathcal{M}(\mathcal{Z}_1 \times Z_2) \to \mathcal{M}(\mathcal{Z}_1) \) is called marginalization, which simply takes a joint distribution to a marginal distribution.

The Fenchel-Rockafellar duality (see [37, Theorem 1.9], or [53] in the case of finite dimensional vector spaces) usually refers to the \( k = 1 \) special case of the following result.
Theorem 5. Assume that $A$ is a topological vector space whose dual is $A^*$. Let $\Theta_j : A \to \mathbb{R} \cup \{+\infty\}$, $j = 0, 1, \ldots, k$, for some positive integer $k$. Suppose there exist some $(u_j)_{j=1}^k$ and $u_0 := -(u_1 + \cdots + u_k)$ such that

$$\Theta_j(u_j) < \infty, \quad j = 0, \ldots, k$$

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

$$\inf_{\ell \in A^*} \left[ \sum_{j=0}^k \Theta_j^*(\ell) \right] = \inf_{u_0, \ldots, u_k \in A} \left[ \Theta_0 \left( -\sum_{j=1}^k u_j \right) + \sum_{j=1}^k \Theta_j(u_j) \right].$$

For completeness, we provide a proof of this result, which is based on the Hahn-Banach theorem (Theorem 6) and is similar to the proof of [37, 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

$$m_0 = \inf_{u_0, \ldots, u_k \in A} \sup_{\ell \in A^*} \left\{ \sum_{j=0}^k \Theta_j(u_j) - \ell(\sum_{j=0}^k u_j) \right\}$$

$$\geq \sup_{\ell \in A^*} \inf_{u_0, \ldots, u_k \in A} \left\{ \sum_{j=0}^k \Theta_j(u_j) - \ell(\sum_{j=0}^k u_j) \right\}$$

$$\Longrightarrow \inf_{\ell \in A^*} \left[ \sum_{j=0}^k \Theta_j^*(\ell) \right].$$

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

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

$$B := \{(0, m) \in A \times \mathbb{R} : m \leq m_0\}.$$ (32)

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

$$C := C_0 + \cdots + C_k$$

is a convex set with a nonempty interior. Moreover, $C \cup B = \emptyset$. By the Hahn-Banach theorem (Theorem 6), there exists $(\ell, s) \in A^* \times \mathbb{R}$ such that

$$sm \leq \ell \left( \sum_{j=0}^k u_j \right) + s \sum_{j=0}^k r_j.$$ (34)

For any $m \leq m_0$ and $(u_j, r_j) \in C_j$, $j = 0, \ldots, k$. From (32) we see (34) can only hold when $s \geq 0$. Moreover, from (26) and the upper semicontinuity of $\Theta_0$ at $u_0$ we see the $\sum_{j=0}^k u_j$ in (34) can take value in a neighbourhood of $0 \in A$, hence $s \neq 0$. Thus, by dividing $s$ on both sides of (34) and setting $\ell \leftarrow -\ell/s$, we see that

$$\inf_{u_0, \ldots, u_k \in A} \left[ -\ell \left( \sum_{j=0}^k u_j \right) + \sum_{j=0}^k \Theta_j(u_j) \right]$$

(35)

$$\Longrightarrow \inf_{\ell \in A^*} \left[ \sum_{j=0}^k \Theta_j^*(\ell) \right].$$ (36)
which establishes $\geq$ in (27).

**Theorem 6** (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
   $$\sup_{u \in B} \ell(u) \leq \inf_{u \in C} \ell(u).$$  \hspace{8cm} (37)

2. If $A$ is locally convex, $B$ is compact, and $C$ is closed, then there exists $\ell \in A^*$ such that
   $$\sup_{u \in B} \ell(u) < \inf_{u \in C} \ell(u).$$  \hspace{8cm} (38)

**Remark 3.** The assumption in Theorem 6 that $C$ has nonempty interior is only necessary in the infinite dimensional case. However, even if $A$ in Theorem 5 is finite dimensional, the assumption in Theorem 5 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 (34).

3. The Entropic-Functional Duality

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

3.1. Compact $X$

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.\footnote{Theorem 1 is not included in the conference paper [47], but was announced in the conference presentation.} Our proof based on the Legendre-Fenchel duality (Theorem 5) was inspired by the proof of the Kantorovich duality in the theory of optimal transportation (see [37, 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 $P_{Y|X} = T_j^*, j = 1, \ldots, m$ and $P_{Z_i|X} = S_i^*$, $i = 1, \ldots, l$.

![Diagrams for Theorem 1](image)

**Figure 2.** Diagrams for Theorem 1.

**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 5, put\(^{10}\)

\[
\Theta_0: u \in C_b(\mathcal{X}) \mapsto \begin{cases} 
0, & u \leq 0; \\
+\infty, & \text{otherwise.}
\end{cases}
\]  

(39)

Then,

\[
\Theta_0^*: \pi \in \mathcal{M}(\mathcal{X}) \mapsto \begin{cases} 
0, & \pi \geq 0; \\
+\infty, & \text{otherwise.}
\end{cases}
\]  

(40)

For each \(j = 1, \ldots, m\), set

\[
\Theta_j(u) := c_j \inf \log \mu_j \left( \exp \left( \frac{1}{c_j} v \right) \right)
\]  

(41)

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 (41) 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

\[
D(P_{Y_j} \| \mu_j) = \sup_v \{ P_{Y_j}(v) - \log \mu_j(\exp(v)) \} = \sup_v \{ P_X(T_j v) - \log \mu_j(\exp(v)) \} = \sup_{u \in C_b(\mathcal{X})} \left\{ P_X(u) - \frac{1}{c_j} \Theta_j(c_j u) \right\} = +\infty
\]  

(42)\((43)\)(44)\(45\)

which contradicts the assumption that \(\sum_{j=1}^m c_j D(P_{Y_j} \| \mu_j) < \infty\) in the theorem;
- From the steps (42)-(44), 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

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

(46)

Notice that

- \(\Theta_{m+1}\) is convex;

\(\)\(^{10}\) In (39), \(u \leq 0\) means that \(u\) is pointwise non-positive.
where \( P_X \) is such that \( S_i^+ P_X = P_{Z_i} \), \( i = 1, \ldots, l \), whose existence is guaranteed by the assumption of the theorem. This also shows that \( \Theta_{m+1} > -\infty \).

- \( \Theta_{m+1} \) is well-defined (that is, the choice of \((w_i)\) in (46) is inconsequential). Indeed if \((w_i)_{i=1}^l\) is such that \( \sum_{i=1}^l S_i w_i = 0 \), then

\[
\sum_{i=1}^l P_{Z_i}(w_i) = \sum_{i=1}^l S_i^+ P_X(w_i) = \sum_{i=1}^l P_X(S_i w_i) = 0,
\]

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

Invoking Theorem 5 (where the \( u_j \) in Theorem 5 can be chosen as the constant function \( u_j \equiv 1 \), \( j = 1, \ldots, m + 1 \):

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

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

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 (54) 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

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

Now invoking (13) with \( f_i := \exp \left( \frac{1}{c_j} v_j \right), j = 1, \ldots, m \) and \( g_i := \exp \left( -\frac{1}{\mu} w_i \right), i = 1, \ldots, l \), we upper bound the left side of (55) by

\[
e - \sum_{i=1}^l b_i \log v_i(g_i) + \sum_{i=1}^l b_i P_{Z_i}(\log g_i) \leq e + \sum_{i=1}^l b_i D(P_{Z_i} \| v_i)
\]

where the last step follows by the Donsker-Varadhan formula. Therefore (14) is established since \( e > 0 \) is arbitrary.
2)⇒1) Since \( v_i \) is finite and \( g_i \) is bounded by assumption, we have \( v_i(g_i) < \infty \), \( i = 1, \ldots, l \). Moreover (13) is trivially true when \( v_i(g_i) = 0 \) for some \( i \), so we will assume below that \( v_i(g_i) \in (0, \infty) \) for each \( i \). Define \( P_{Z_i} \) by

\[
\frac{dP_{Z_i}}{dv_i} = \frac{g_i}{v_i(g_i)}, \quad i = 1, \ldots, l.
\]  

(57)

Then for any \( \epsilon > 0 \),

\[
\sum_{i=1}^{l} b_i \log v_i(g_i) = \sum_{i=1}^{l} b_i [P_{Z_i}(\log g_i) - D(P_{Z_i} || v_i)] < \sum_{j=1}^{m} c_j P_{Y_j}(\log f_j) + \epsilon - \sum_{j=1}^{m} c_j D(P_{Y_j} || \mu_j)
\]

(58)

\[
\leq \epsilon + \sum_{j=1}^{m} c_j \log \mu_j(f_j)
\]

(59)

\[
\leq \epsilon + \sum_{j=1}^{m} c_j \log \mu_j(f_j)
\]

(60)

where

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

\[
\sum_{i=1}^{l} b_i D(P_{Z_i} || v_i) > \sum_{j=1}^{m} c_j D(P_{Y_j} || \mu_j) - \epsilon
\]

(61)

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

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

\( \square \)

Remark 4. Condition iv) in the theorem imposes a rather strong assumption on \((S_i)\): for simplicity, consider the case where \(|X|, |Z| < \infty \). 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 \( P_{Z_i/X} \) being a reverse erasure channel gives a simple example where \( P_{Z_i/X} \) is not a deterministic map.

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

Theorem 7. Theorem 1 continues to hold if Condition iv) therein is weakened to the following:

- For any \( P_X \) such that \( D(S_i^* P_X || v_i) < \infty \), \( i = 1, \ldots, l \), there exists \( \hat{P}_X \) such that \( S_i^* \hat{P}_X = S_i^* P_X \) for each \( i \) and \( \sum_{i=1}^{m} c_i D(T_j^* \hat{P}_X || \mu_j) < \infty \) 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 0 and such that

\[
\sum_{i=1}^{l} b_i S_i \log g_i \leq \sum_{j=1}^{m} c_j T_j \log f_j
\]

(62)

we have

\[
\inf_{(g_i)} \prod_{i=1}^{l} v_i^b(g_i) \leq \exp(d) \prod_{j=1}^{m} \mu_j^c(f_j).
\]

(63)
2. For any \((P_X)\) such that \(D(S_i^T P_X \| v_i) < \infty\), \(i = 1, \ldots, l\),

\[
\sum_{i=1}^l b_i D(S_i^T P_X \| v_i) + d \geq \inf_{P_X: S_i^T P_X = S_i^T \mu} \sum_{j=1}^m c_j D(T_j^* P_X \| \mu_j). \tag{64}
\]

In Appendix A we show that Theorem 7 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 (55), we select \((\hat{g}_i)\) such that

\[
\sum_{i=1}^l b_i S_i \log \hat{g}_i \geq \sum_{i=1}^l b_i S_i \log g_i, \tag{65}
\]

\[
\sum_{i=1}^l b_i \log v_i(\hat{g}_i) \leq \sum_{j=1}^m c_j \log \mu_j(f_j) + \epsilon. \tag{66}
\]

Then, in lieu of (67), we upper-bound the left side of (55) by

\[
2\epsilon - \sum_{i=1}^l b_i \log v_i(\hat{g}_i) + \sum_{i=1}^l b_i P_{Z_i}(\log \hat{g}_i) \leq 2\epsilon + \sum_{i=1}^l b_i D(P_{Z_i} \| v_i) \tag{67}
\]

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

\[
\Lambda_i(u) := \inf_{\hat{g}_i > 0 : b_i \log \hat{g}_i = u} b_i \log v_i(\hat{g}_i). \tag{68}
\]

Then following essentially the same proof as that of \(\Theta_j\) in (41), we see that \(\Lambda_i\) is proper convex and

\[
\Lambda_i^*(\pi) = b_i D(S_i^T \pi \| \mu_j). \tag{69}
\]

Moreover let

\[
\Lambda_{l+1}(u) := \begin{cases} 0 & \text{if } u = -\sum b_i S_i \log g_i; \\ +\infty & \text{otherwise}. \end{cases} \tag{70}
\]

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

\[
\inf_{(\hat{g}_i) : \sum_{i=1}^l b_i S_i \log \hat{g}_i \geq \sum_{i=1}^l b_i S_i \log g_i} \sum_{i=1}^l b_i \log v_i(\hat{g}_i) \leq 2\epsilon - \sum_{i=1}^l b_i \log v_i(\hat{g}_i) + \sum_{i=1}^l b_i P_{Z_i}(\log \hat{g}_i) \leq 2\epsilon + \sum_{i=1}^l b_i D(P_{Z_i} \| v_i) \tag{71}
\]

\[
\inf_{u_1, \ldots, u_{l+1}} \Theta_0 \left( -\sum_{i=1}^{l+1} u_i + \sum_{i=1}^{l+1} \Lambda_i(u_i) \right) \leq \sup_{\pi} \left\{-\sum_{i=0}^{l+1} \Theta_i^*(\pi)\right\} \leq \sup_{\pi \geq 0} \left\{-\sum_{i=0}^{l+1} \Theta_i^*(\pi)\right\} \leq \sup_{\pi \geq 0} \left\{\sum_{i=1}^l b_i S_i^T \pi(\log g_i) - \sum_{i=1}^l b_i D(S_i^T \pi \| v_i)\right\} \leq \sum_{i=1}^l b_i S_i^T P_X(\log g_i) - \sum_{i=1}^l b_i D(S_i^T P_X \| v_i) + \epsilon \tag{72}
\]
\[ \leq \sum_{j=1}^{m} c_j T_j^* \tilde{P}_X(\log f_j) - \sum_{j=1}^{m} c_j D(T_j^* \tilde{P}_X \parallel \mu_j) + 2\epsilon \]  
(76)

\[ \leq 2\epsilon + \sum_{j=1}^{m} c_j \log \mu_j(f_j) \]  
(77)

where

- To see (75) we note that the sup in (74) can be restricted to \( \pi \) which is a probability measure, since otherwise the relative entropy terms in (74) are \( +\infty \) by its definition via the Donsker-Varadhan formula. Then we select \( P_X \) such that (75) holds.
- In (76), we have chosen \( \tilde{P}_X \) such that

\[ S_i^* \tilde{P}_X = S_i^* P_X, \quad 1 \leq i \leq l; \]

\[ \sum_{i=1}^{l} b_i D(S_i^* P_X) > \sum_{j=1}^{m} c_j D(T_j^* P_X \parallel \mu_j) - \epsilon, \]

and then applied the assumption (62). The result follows since \( \epsilon > 0 \) can be arbitrary.

\[ \square \]

Remark 5. The infimum in (14) is in fact achievable: For any \( \{P_{Z_i}\} \), there exists a \( P_X \) that minimizes

\[ \sum_{i=1}^{m} c_j D(P_{Y_i} \parallel \mu_j) \]

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, the set \( \bigcap_{i=1}^{m} (S_i^* P_X) \) is weak*-closed in \( \mathcal{M}(\mathcal{X}) \); 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 [54]. Moreover, by the weak*-lower semicontinuity of \( D(\cdot \parallel \mu_j) \) (easily seen from the variational formula/Donsker-Varadhan formula of the relative entropy, cf. [55]) and the weak*-continuity of \( T_j^* \), \( j = 1, \ldots, m \), we see that \( \sum_{j=1}^{m} c_j D(T_j^* P_X \parallel \mu_j) \) 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) \( \Rightarrow \) 2) part is the non-trivial direction which requires regularity on the spaces. In contrast, the 2) \( \Rightarrow \) 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) \( \Rightarrow \) 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 \( \mathcal{X} = \mathbb{Z}_1 \times \cdots, \times \mathbb{Z}_l \) and each \( S_i \) is the canonical map. Then \( \Theta_{m+1}(u) \) as defined in (46) 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 [37, Chapter 1.3] to prove a particular extension of Theorem 1 to a non-compact case.

---

11 Generally, if \( T : A \to B \) is a continuous map between two topologically vector spaces, then \( T^* : B^* \to A^* \) is a weak* continuous map between the dual spaces. Indeed, if \( y_a \to y \) is a weak*-convergent subsequence in \( B^* \), meaning \( y_a(b) \to y(b) \) for any \( b \in B \), then we must have \( T^* y_a(a) = y_a(Ta) \to y(Ta) = T^* y(a) \) for any \( a \in A \), meaning that \( T^* y_a \) converges to \( T^* y \) in the weak* topology.
Theorem 8. 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;
- \(\mathcal{X} = \prod_{i=1}^{l} Z_i\) and \(S_i : C_b(Z_i) \to C_b(\mathcal{X})\), \(i = 1, \ldots, l\) 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 1)⇒2) part. Let \(\Theta_0\) be defined as before, in (39). Then for any \(\ell \in C_c(\mathcal{X})^\ast\),

\[
\Theta_0^\ast(\ell) = \sup_{u \leq 0} \ell(u)
\]

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

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

\[
\Theta_{m+1}^\ast(\ell) = \begin{cases} 
0 & \text{if } \ell(\sum_i S_i w_i) = \sum_i P_{Z_i}(w_i), \quad \forall w_i \in C_b(Z_i), \ i = 1, \ldots, l;
+\infty & \text{otherwise}.
\end{cases}
\]

For any \(\ell \in C_c^\ast(\mathcal{X})\). Generally, the condition in the first line of (81) does not imply that \(\ell\) is a measure. However, if \(\ell\) is also nonnegative, then using a technical result in [37, Lemma 1.25] we can further simplify:

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

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 \(\Theta_j^\ast(\ell) = c_j D(T_j^\ast \ell || \nu_j)\) still holds with the \(\Theta_j\) defined in (41), \(j = 1, \ldots, m\). Thus the rest of the proof can proceed as before. \(\Box\)

Remark 7. The second assumption is made in order to achieve (82) 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 8), in principle, the rotational invariance argument can be applied either to the functional representation (as in Lieb’s paper [29]) or to the entropic representation (as in Geng-Nair [45]). 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 and Theorem 3.

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 \((v_j)_{j=1}^l\) on \(\mathbb{R}\);

- non-degenerate (Definition 3 below) linear Gaussian random transformation \((P_{Y_j|X})_{j=1}^m\) (where \(X := (X_1, \ldots, X_l)\)) associated with conditional expectation operators \((T_j)_{j=1}^m\).
where the infimum is over Borel measures $P_X$ that has $(P_{X_i})$ as marginals. Note that (83) 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 9.** $\sup_{(P_{X_i})} F_0((P_{X_i}))$, where the supremum is over Borel measures $P_{X_i}$ on $\mathbb{R}$, $i = 1, \ldots, l$, is achieved by some Gaussian $(P_{X_i})_{i=1}^l$, in which case the infimum in (83) is achieved by some Gaussian $P_X$.

Naturally, one would expect that Gaussian optimality can be established when $(\mu_i)^m_{i=1}$ and $(\nu_i)^l_{i=1}$ 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 10.** $\sup_{(P_{X_i})} F_0((P_{X_i}))$ is finite and there exist $\sigma_i^2 \in (0, \infty)$, $i = 1, \ldots, l$ such that it equals

$$
\sup_{(P_{X_i}): \mathbb{E}[X_i^2] \leq \sigma_i^2} F_0((P_{X_i})).
$$

**Proof.** When $\mu_i$ is Lebesgue and $P_{Y_i|X}$ is non-degenerate, $D(P_{Y_i}||\mu_i) = -h(P_{Y_i}) \leq -h(P_{Y_i}|X)$ is bounded above (in terms of the variance of additive noise of $P_{Y_i|X}$). 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 (83) shows that $\sup_{(P_{X_i})} F_0((P_{X_i})) > -\infty$.

To see (84), notice that

$$
D(P_{X_i}||\nu_i) = D(P_{X_i}||\nu_i') + \mathbb{E}[\nu_i'||\nu_i(X)]
$$

$$
= D(P_{X_i}||\nu_i') + D(\nu_i'||\nu_i)
$$

$$
\geq D(\nu_i'||\nu_i)
$$

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 (83) is bounded above by some constant depending only on $(P_{Y_i|X})$. 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 $i$ 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. \[\square\]

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

**Proposition 11.** Under Assumption 1,

1. For any $(P_{X_i})_{i=1}^l$, the infimum in (83) is attained by some Borel $P_X$.  

\[\]
2. If \((P_{Y_i|X_i}^{(m)})_{i=1}^m\) are non-degenerate (Definition 3), then the supremum in (84) is achieved by some Borel

\((P_{X_i})_{i=1}^m\).

The proof of Proposition 11 is given in Section E. After taking care of the existence of the extremizers, we get into the tensorization properties which are the crux of the proof:

**Lemma 12.** 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

\[
\inf_{P_{X^{(1)}}} \sum_{i=1}^m c_i D(P_{Y_i^{(1)}} \parallel \mu_j^{(1)}) = \sum_{i=1}^m c_i D(P_{Y_i^{(1)}} \parallel \mu_j^{(1)})
\]

where for each \(j\),

\[
P_{Y_i^{(1)}} := T_j^* \otimes P_{X_i^{(1)}}
\]

on the left side and

\[
P_{Y_i^{(0)}} := T_j^* \otimes P_{X_i^{(0)}}
\]

on the right side, \(t = 1, 2\).

**Proof.** We only need to prove the nontrivial part. For any \(P_{X_i^{(2)}}\) on the left side, choose \(P_{X_i^{(0)}}\) on the right side by marginalization. Then

\[
D(P_{Y_i^{(1)}} \parallel \mu_j^{(1)}) - \sum_t D(P_{Y_i^{(0)}} \parallel \mu_j^{(0)}) = I(Y_i^{(1)}; Y_i^{(2)}) \geq 0
\]

for each \(j\). □

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

**Proof of Theorem 9.**

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

\[
X^+ := \frac{1}{\sqrt{2}} \left( X^{(1)} + X^{(2)} \right);
\]

\[
X^- := \frac{1}{\sqrt{2}} \left( X^{(1)} - X^{(2)} \right).
\]

Define \((Y_i^+)\) and \((Y_i^-)\) analogously. Then \(Y_i^+ \mid \{X^+ = x^+, X^- = x^-\} \sim Q_{Y_i \mid X=x^+}\) is independent of \(x^+\) and \(Y_i^- \mid \{X^+ = x^+, X^- = x^-\} \sim Q_{Y_i \mid X=x^-}\) is independent of \(x^-\).

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

\[
\sum_{t=1}^2 F_0 \left( \left( P_{X_i^{(t)}} \right)^{1/2} \right) = \inf_{P_{X^{(1,2)}}} \sum_t c_t D(P_{Y_i^{(1,2)}} \parallel \mu_j^{(1,2)}) - \sum_t b_t D(P_{X_i^{(1,2)}} \parallel \nu_i^{(1,2)})
\]

\[
= \inf_{P_{X^+X^-}} \sum_t c_t D(P_{Y_i^{(1,2)}} \parallel \mu_j^{(1,2)}) - \sum_t b_t D(P_{X_i^{(1,2)}} \parallel \nu_i^{(1,2)})
\]

(96)
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

\[
D(P_{X_i^-|X_i^+} \| v_i \| P_{X_i^+}) = D(P_{X_i^-|X_i^+} \| v_i \| P_{X_i+}).
\]

where \((99)\) is because in the above we have constructed the coupling optimally.

\[
D(P_{X_i^-|X_i^+} \| v_i \| P_{X_i^+}) = D(P_{X_i^-} \| v_i \| P_{X_i+}).
\]

where \((103)\) 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 \(I(X_i^+, X_i^-) = D(P_{X_i^-|X_i^+} \| v_i \| P_{X_i^+}) - D(P_{X_i^-} \| v_i) = 0\), 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 12 and the doubling trick one can show that the optimal coupling is also Gaussian.

4.2. A Geometric Forward-Reverse Brascamp-Lieb Inequality

In this section we give a sketch of the proof of Theorem 3 which is simple but certain ‘technicalities’ are not justified. A detailed proof is deferred to Appendix F.

Proof Sketch for Theorem 3. By duality (Theorem 8) it suffices to prove the corresponding entropic inequality. The Gaussian optimality result in Theorem 9 assumed Gaussian reference measures on the output and non-degenerate forward channels in order to simplify the proof of the existence of

\[\text{for a justification that we can select optimal coupling } P_{X^-|X^+} \text{ in a way that } P_{X^-|X^+} \text{ is indeed a regular conditional probability distribution, see [7].}\]
minimizers; however, supposing that Gaussian optimality extends beyond those technical conditions, then we see that it suffices to prove that for any centered Gaussian \((P_{X_i})\),

\[
\sum_{i=1}^l h(P_{X_i}) \leq \sup_{P_{X'}} \sum_{j=1}^l h(P_{Y_j}) \tag{104}
\]

where the supremum is over Gaussian \(P_{X'}\) with the marginals \(P_{X_1}, \ldots, P_{X_l}\), and \(Y_j := \sum_{i=1}^l m_{ij} X_i\). Let \(a_i := \mathbb{E}[X_i^2]\) and choose \(P_{X_i} = \prod_{i=1}^l P_{X_i}\), we see (104) holds if

\[
\sum_{i=1}^l \log a_i \leq \sum_{j=1}^l \log \left( \sum_{i=1}^l m_{ij}^2 a_i \right), \quad \forall a_i > 0, i = 1, \ldots, l, \tag{105}
\]

where \((a_i)\) are the eigenvalues and \(\left( \sum_{i=1}^l m_{ij} a_i \right)_{i=1}^l\) are the diagonal entries of the matrix \(M_{\text{diag}}(a_i)_{1 \leq i \leq l} M^\top\). (106)

Therefore (105) holds. \(\square\)

5. Relation to Hypercontractivity and Its Reverses

As alluded before and illustrated by Figure 1, 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; but 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.

5.1. Hypercontractivity

Fix a joint probability distribution \(Q_{Y_1 Y_2}\) and nonnegative continuous functions \(F_1\) and \(F_2\) on \(Y_1\) and \(Y_2\), respectively, both bounded away from 0. In Theorem 1, take \(l \leftarrow 1, m \leftarrow 2, b_1 \leftarrow 1, d \leftarrow 0, f_1 \leftarrow F_1^{-2}, f_2 \leftarrow F_2^{-2}, \nu_1 \leftarrow Q_{Y_1}, \nu_2 \leftarrow Q_{Y_2}\). Also, put \(Z = X = (Y_1, Y_2)\), and let \(T_1\) and \(T_2\) be the canonical maps (Definition 2). The constraint (12) translates to

\[
g_1(y_1, y_2) \leq f_1(y_1) f_2(y_2), \quad \forall y_1, y_2 \tag{107}
\]
and the optimal choice of $g_1$ is when the equality is achieved. We thus obtain the equivalence between
\begin{equation}
\|F_1\|_1 \cdot \|F_2\|_2 \geq \mathbb{E}[F_1(Y_1)F_2(Y_2)], \quad \forall F_1 \in L^\frac{1}{t_1}(Q_{Y_1}), \; F_2 \in L^\frac{1}{t_2}(Q_{Y_2})
\end{equation}

and
\begin{equation}
\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}).
\end{equation}

This equivalence can also be obtained from Theorem 1. By Hölder’s inequality, (108) is equivalent to saying that the norm of the linear operator sending $F_1 \in L^\frac{1}{t_1}(Q_{Y_1})$ to $\mathbb{E}[F_1(Y_1)Y_2] = J \in L^{\frac{1}{t_2}}(Q_{Y_2})$ does not exceed 1. The interesting case is $\frac{1}{t_2} > 1$, hence the name hypercontractivity. The equivalent formulation of hypercontractivity was shown in [41] using a different proof via the method of types/typicality, which requires that $|Y_1|, |Y_2| < \infty$. 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)

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{reverse_hypercontractivity_diagram.png}
\caption{Diagram for reverse hypercontractivity}
\end{figure}

Let $Q_{Z_1Z_2}$ be a given joint probability distribution, and let $G_1$ and $G_2$ be nonnegative functions on $Z_1$ and $Z_2$, respectively, both bounded away from 0. In Theorem 1, take $l \leftarrow 2, m \leftarrow 1, c_1 \leftarrow 1, d \leftarrow 0, g_1 \leftarrow G_1^{\frac{1}{l}}, g_2 \leftarrow G_2^{\frac{1}{m}}, f_1 \leftarrow Q_{Z_1Z_2}, f_2 \leftarrow Q_{Z_1Z_2}$. Also, put $Y_1 = X = (Z_1, Z_2)$, and let $S_1$ and $S_2$ be the canonical maps (Definition 2). Note that the constraint (12) translates to
\begin{equation}
f_1(z_1, z_2) \geq G_1(z_1)G_2(z_2), \quad \forall z_1, z_2.
\end{equation}

and the equality case yields the optimal choice of $f_1$ for (13). By Theorem 1 we thus obtain the equivalence between
\begin{equation}
\|G_1\|_{\frac{1}{l}} \cdot |G_2|_{\frac{1}{m}} \leq \mathbb{E}[G_1(Z_1)G_2(Z_2)], \quad \forall G_1, G_2
\end{equation}

and
\begin{equation}
\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}).
\end{equation}

Note that in this setup, if $Z_1$ and $Z_2$ 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 [56], where the proof is based on the method of types.

---

13 By a standard dense-subspace argument, we see that it is inconsequential that $F_1$ and $F_2$ in (108) 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 (108).

14 By “positive parameters” we mean the $b_1$ and $b_2$ in (112) are positive.
5.3. Reverse Hypercontractivity (One Negative Parameter)

\[ \mathcal{P}(Z_1) \xrightarrow{S_1^*} \mathcal{P}(Z_1 \times \mathcal{Y}_2) \xrightarrow{T_2} \mathcal{P}(\mathcal{Y}_2) \]

Figure 5. Diagram for reverse hypercontractivity with one negative parameter

In Theorem 1, take \( l \leftarrow 1, m \leftarrow 2, c_1 \leftarrow 1, d \leftarrow 0 \). Let \( \mathcal{Y}_1 = X = (Z_1, Y_2) \), 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 \( G \) be arbitrary nonnegative continuous functions on \( \mathcal{Y}_2 \) and \( Z_1 \), respectively, which are bounded away from 0. Take \( g_1 \leftarrow G \frac{1}{2}, f_2 \leftarrow F^{-\frac{1}{2}} \) in Theorem 1. The constraint (12) translates to

\[ f_1(z_1, y_2) \geq G(z_1)F(y_2), \quad \forall z_1, y_2. \tag{113} \]

Note that (13) translates to

\[ \|G\|_\frac{1}{2} \leq Q_{Y_2Z_1}(f_1)Q_{Y_2}^G(F^{-\frac{1}{2}}) \tag{114} \]

for all \( F, G \), and \( f_1 \) satisfying (113). It suffices to verify (114) for the optimal choice \( f_1 = GF \), so (114) is reduced to

\[ \|F\|_\frac{1}{2} \|G\|_\frac{1}{2} \leq E[GF(Y_2)G(Z_1)], \quad \forall F, G. \tag{115} \]

By Theorem 1, (115) is equivalent to

\[ \forall P_{Z_1}, \exists P_{Z_1Y_2}, \quad D(P_{Z_1Y_2}\|Q_{Z_1Y_2}) \leq b_1 D(P_{Z_1}\|Q_{Z_1}) + (-c_2) D(P_{Y_2}\|Q_{Y_2}). \tag{116} \]

Inequality (115) is called reverse hypercontractivity with a negative parameter in [42], where the entropic version (116) is established for \(|Z_1|, |\mathcal{Y}_2| < \infty \) using the method of types. Multiterminal extensions of (115) and (116) (called reverse Brascamp-Lieb type inequality with negative parameters in [42]) 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 Section 5.1, 5.2 and 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 (108), (111) or (115).

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 Professor Elliott H. Lieb for teaching the Brascamp-Lieb inequality as well as some techniques used in this paper in his graduate class.

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 Jingbo Liu. Parts of the sections on the existence of the minimizer and the Gaussian optimality were written by Thomas Courtade.

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

---

15 By “one negative parameter” we mean the \( b_1 \) is positive and \( -c_2 \) is negative in (116).
Appendix A. Recovering Theorem 1 from Theorem 7 as a Special Case

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

$$C := \left\{ (w_i) : w_i \in C_b(Z_i), \sum_{i=1}^l \inf w_i(z_i) \geq 0 \right\},$$

(A1)

which is a closed convex cone in $C_b(Z_1) \times \cdots \times C_b(Z_l)$. Given $(g_i)$ we show that $\sum_{i=1}^l b_i S_i \log \hat{g}_i \geq \sum_{i=1}^l b_i S_i \log g_i$ implies

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

Indeed, we can verify that the dual cone

$$C^* := \left\{ (\pi_i) : \sum_{i=1}^l \pi_i (w_i) \geq 0, \forall (w_i) \in C \right\}$$

(A3)

$$= \{ \lambda (P_{Z_1}, \ldots, P_{Z_l}) : \lambda \geq 0 \}. \tag{A4}$$

Under the surjectivity assumption, we see

$$\sum_{i=1}^l \pi_i (b_i \log \hat{g}_i - b_i \log g_i) \geq 0, \quad \forall (\pi_i) \in C^*. \tag{A5}$$

Now if (A2) is not true, by the Hahn-Banach theorem (Theorem 6) we find $\pi_i \in M(Z_i), i = 1, \ldots, l$ such that

$$\sum_{i=1}^l \pi_i (b_i \log \hat{g}_i - b_i \log g_i) < \inf_{(w_i) \in C} \sum_{i=1}^l \pi_i (w_i) \tag{A6}$$

so right side of (A6) is not $-\infty$. Since $C$ is a cone containing the origin, the right side of (A6) hence must be nonnegative, and we conclude that $(\pi_i) \in C^*$. But then (A6) contradicts (A5).

Appendix B. Existence of Weakly Convergent Couplings

**Lemma 13.** 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 $P_X$ is a coupling of $(P_{X_i})_{1 \leq i \leq l}$, then, upon extraction of a subsequence, there exist couplings $P_X^{(n)}$ for $(P_{X_i}^{(n)})_{1 \leq i \leq l}$ which converge weakly to $P_X$ as $n \to \infty$.

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

$$\phi_k : x \mapsto \begin{cases} \lfloor 2^k x \rfloor & |x| \leq k, x \not\in 2^{-k} \mathbb{Z}; \\ e & \text{otherwise}, \end{cases} \tag{A7}$$

and let $W_i := (W_i^{(k)})_{i=1}^\infty$. Denote by $W_i^{[k]} := \{-2^k, \ldots, 2^k - 1, e\}$ 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:

$$P_{X_i} \left( \bigcup_{k=1}^\infty 2^{-k} \mathbb{Z} \right) = 0, \quad i = 1, \ldots, l. \tag{A8}$$
Since $P_{X_i}^{(n)} \rightharpoonup P_{X_i}$ 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

$$P_{W_i}^{(n)}(w) \geq \left(1 - \frac{1}{k}\right) P_{W_i}(w), \quad \text{(A9)}$$

for each $i$ and $w \in W_i$. Now define a coupling $(P_{W_i}^{(n)})_i$ compatible with the $(P_{W_i}^{(n)})_i$ induced by $(P_{X_i}^{(n)})_i$ as follows:

$$P_{W_i}^{(n)} := \left(1 - \frac{1}{k}\right) P_{W_i} + k^{l-1} \prod_{i=1}^{l} \left(P_{W_i}^{(n)} - \left(1 - \frac{1}{k}\right) P_{W_i}\right). \quad \text{(A10)}$$

Observe that (A10) is a well-defined probability measure because of (A9), and indeed has marginals $(P_{W_i}^{(n)})_i$. Moreover, by the triangle inequality we have the following bound on the total variation distance

$$|P_{W_i}^{(n)} - P_{W_i}| \leq \frac{2}{k}. \quad \text{(A11)}$$

Next, construct $P_X^{(n)}$:

$$P_X^{(n)} := \sum_{w^i \in W_1 \times \cdots \times W_l} \frac{P_{W_i}^{(n)}(w^i)}{\prod_{i=1}^{l} P_{W_i}^{(n)}(w_i)} \prod_{i=1}^{l} P_{X_i}^{(n)}(\phi_i^{-1}(w_i)). \quad \text{(A12)}$$

Observe that $P_X^{(n)}$ defined in (A12) is compatible with the $P_{W_i}^{(n)}$ defined in (A10), and indeed has marginals $(P_{X_i}^{(n)})_i$. Since $n := n_k$ can be made increasing in $k$, we have constructed the desired sequence $(P_{X_i}^{(n_k)})_{k=1}^{\infty}$ converging weakly to $P_X$. Indeed, for any bounded open dyadic cube $A$, using (A11) and the assumption (A8), we conclude

$$\liminf_{k \to \infty} P_X^{(n_k)}(A) \geq P_X(A). \quad \text{(A13)}$$

Moreover, since bounded open dyadic cubes form a countable basis of the topology in $\mathbb{R}^l$, we see (A13) actually holds for any open set $A$ (by writing $A$ as a countable union of dyadic cubes, using the continuity of measure to pass to a finite disjoint union, and then apply (A13)), as desired. $\square$

**Appendix C. Upper Semicontinuity of the Infimum**

The following is a consequence of Lemma 13.

---

16 We use $P|_A$ to denote the restriction of a probability measure $P$ on measurable set $A$, that is, $P|_A(B) := P(A \cap B)$ for any measurable $B$.

17 That is, a cube whose corners have coordinates being multiples of $2^{-k}$ where $k$ is some integer.
Corollary 14. Consider non-degenerate \((P_{Y|X})\). 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 \(P_{X_i}\) for each \(i\). Then
\[
\limsup_{n \to \infty} \inf_{P_X : S_i P_X = P_{X_i}^{(n)}} \sum_{j=1}^{m} c_j D(T^*_j P_X \| \mu_j) \leq \inf_{P_X : S_i P_X = P_{X_i}^{(n)}} \sum_{j=1}^{m} c_j D(T^*_j P_X \| \mu_j).
\] (A14)

Proof. By passing to a convergent subsequence, we may assume that the limit on the left side of (A14) exists. For any coupling \(P_X^*\) of \((P_{X_i}^*)\), by invoking Lemma 13 and passing to a subsequence, we find a sequence of couplings \(P_X^{(n)}\) of \((P_{X_i}^{(n)})\) that converges weakly to \(P_X^*\). 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. [45, Proposition 18], [57, Theorem 7], or [58, Theorem 1, Theorem 2]). Thus
\[
\lim_{n \to \infty} \sum_{j=1}^{m} c_j D(T^*_j P_X \| \mu_j) = \sum_{j=1}^{m} c_j D(T^*_j P_X \| \mu_j)
\] (A15)
and (A14) follows since \(P_X^*\) was arbitrarily chosen. □

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

Lemma 15. Suppose \((P_{X_n})\) is a sequence of distributions on \(\mathbb{R}^d\) converging weakly to \(P_X^*\), and
\[
E[X_n X_n^\top] \preceq \Sigma
\] (A16)
for all \(n\). Then
\[
\limsup_{n \to \infty} h(X_n) \leq h(X^*).
\] (A17)

Remark 8. The result fails without the condition (A16). Also, related results when the weak convergence is replaced with pointwise convergence of density functions and certain additional constraints was shown in [58, Theorem 1, Theorem 2] (see also the proof of [45, Theorem 5]). Those results are not applicable here since the density functions of \(X_n\) do not converge pointwise. They are applicable for the problems discussed in [45] 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. [55]). This fact and a scaling argument immediately show that, for any \(r > 0\),
\[
\limsup_{n \to \infty} h(X_n \| X_n) \leq r \leq h(X^* \| X^*) \leq r.
\] (A18)

Let \(p_n(r) := \mathbb{P}[\|X_n\| > r]\), then (A16) implies
\[
E[XX^\top \| X_n) > r] \leq \frac{1}{p_n(r)} \Sigma.
\] (A19)
Therefore, since the Gaussian distribution maximizes differential entropy given a second moment upper bound, we have
\[
h(X_n \| X_n) > r \leq \frac{1}{2} \log \frac{(2\pi)^d \Sigma}{p_n(r)}.
\] (A20)
Since \( \lim_{r \to \infty} \sup_n p_n(r) = 0 \) by (A16) and Chebyshev’s inequality, (A20) implies that
\[
\lim_{r \to \infty} \sup_n p_n(r) h(\mathbf{X}_n \| \mathbf{X}_n) > r = 0.
\] (A21)

The desired result follows from (A18), (A21) and the fact that
\[
h(\mathbf{X}_n) = p_n(r) h(\mathbf{X}_n \| \mathbf{X}_n) > r + (1 - p_n(r)) h(\mathbf{X}_n \| \mathbf{X}_n) \leq r + h(p_n(r)).
\] (A22)

\[ \square \]

Appendix E. Proof of Proposition 11

1. For any \( \epsilon > 0 \), by the continuity of measure there exists \( K > 0 \) such that
\[
P_{X_i}([-K, K]) \geq 1 - \frac{\epsilon}{\tau}, \quad i = 1, \ldots, l.
\] (A23)

By the union bound,
\[
P_X([-K, K]^l) \geq 1 - \epsilon
\] (A24)

wherever \( P_X \) is a coupling of \( (P_{X_i}) \). Now let \( P_{X_i}^{(n)}, n = 1, 2, \ldots \) be a such that
\[
\lim_{n \to \infty} \sum_{j=1}^{m} c_j D(P_{X_i}^{(n)} \| \mu_j) = \inf_{P_X} \sum_{j=1}^{m} c_j D(P_X \| \mu_j)
\] (A25)

where \( P_{X_i} : = T_j^i P_X, j = 1, \ldots, m \). The sequence \( (P_{X_i}^{(n)}) \) is tight by (A24), Thus invoking Prokhorov theorem and by passing to a subsequence, we may assume that \( (P_{X_i}^{(n)}) \) converges weakly to some \( P_X^* \). Therefore \( P_{X_i}^{(n)} \) converges to \( P_X^* \) weakly, and by the semicontinuity property in Lemma 15 we have
\[
\sum_{j=1}^{m} c_j D(P_{X_i}^* \| \mu_j) \leq \lim_{n \to \infty} \sum_{j=1}^{m} c_j D(P_{X_i}^{(n)} \| \mu_j)
\] (A26)

establishing that \( P_X^* \) 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 10 and
\[
\lim_{n \to \infty} F_0 \left( \left( P_{X_i}^{(n)} \right)_{i=1}^{j} \right) = \sup_{(P_X) : \Sigma_X \preceq \sigma^2} F_0 \left( \left( P_{X_i} \right)_{i=1}^{j} \right).
\] (A27)

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 \( P_{X_i}^* \). We have the moment bound
\[
\mathbb{E}[X_i^2] = \lim_{K \to \infty} \mathbb{E}[\min\{X_i^2, K\}]
\] (A28)
\[
= \lim_{K \to \infty} \mathbb{E}[\min\{(X_i^{(n)})^2, K\}]
\] (A29)
\[
\leq \sigma_i^2
\] (A30)
where \( X_i \sim P_{X_i}^* \) and \( X_i^{(n)} \sim P_{X_i}^{(n)} \). Then by Lemma 15,
\[
\sum_i b_i D(P_{X_i}^* \| v_i) \leq \lim_{n \to \infty} \sum_i b_i D(P_{X_i}^{(n)} \| v_i) \tag{A31}
\]

Under the covariance regularization and the non-degenerateness assumption, we showed in Proposition 10 that the value of \((84)\) cannot be \(+\infty \) or \(-\infty\). This implies that we can assume (by passing to a subsequence) that \( P_{X_i}^{(n)} \ll \lambda, i = 1, \ldots, l \) since otherwise \( F((P_{X_i}^{(n)})) = -\infty \). Moreover, since \( \left( \sum_i c_i D(P_{Y_i}^{(n)} \| \mu_i) \right)_{n \geq 1} \) is bounded above under the non-degenerateness assumption, the sequence \( \left( \sum_i b_i D(P_{X_i}^{(n)} \| v_i) \right)_{n \geq 1} \) must also be bounded from above, which implies, using (A31), that
\[
\sum_i b_i D(P_{X_i}^* \| v_i) < \infty \tag{A32}
\]

In particular, we have \( P_{X_i}^* \ll \lambda \) for each \( i \). Now Corollary 14 shows that
\[
\inf_{P_X : S_1 \bar{P}_X = P_{X_i}^*} \sum_i c_i D(T_i^T P_X \| \mu_j) \geq \lim_{n \to \infty} \inf_{P_X : S_1 \bar{P}_X = P_{X_i}^{(n)}} \sum_i c_i D(T_i^T P_X \| \mu_j) \tag{A33}
\]

Thus (A31) and (A33) show that \( (P_{X_i}^*) \) is in fact a maximizer.

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

**Appendix F.1. Proof of Theorem 3**

The proof will be based on Theorem 9, 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 \( j = 1, \ldots, l \), defined the linear operator \( T_j^T \) by
\[
(T_j^T \phi)(x_1, \ldots, x_l) := \mathbb{E} \left[ \phi \left( \sum_{i=1}^l m_{ji} x_i + N_i \right) \right] \tag{A34}
\]

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

**Lemma 16.** For any \( \epsilon > 0 \), let \( (T_j^T) \) be defined as in (A34). Then for any Borel \( P_{X_i} \ll \lambda, i = 1, \ldots, l \),
\[
\sum_{i=1}^l D(P_{X_i} \| \gamma_{\frac{1}{2}}) - \frac{1}{2} \log \frac{2\pi \epsilon}{\epsilon} \geq \inf_{P_X : S_1 \bar{P}_X = P_{X_i}} \left\{ - \frac{1}{2} \sum_{j=1}^l h(T_j^T P_X) \right\} \tag{A35}
\]

**Proof.** By Theorem 9, it suffices to prove (A35) when \( P_{X_i} \) is Gaussian, and from (A35) it is easy to see that it suffices to prove the case of centered Gaussian. Let \( P_{X_i} = \mathcal{N}(0, a_i) \), \( i = 1, \ldots, l \). We can upper bound the right side of (A35) by taking \( P_{X_i} = P_{X_i} \times P_{X_i} \) instead of the infimum, so it suffices to prove that
\[
\frac{1}{2} \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i=1}^l \log a_i \geq - \frac{1}{2} \sum_{j=1}^l \log \left( \sum_{i=1}^l m_{ji}^2 a_i + \epsilon \right) \tag{A36}
\]

for any \( a_1, \ldots, a_l \in (0, \infty) \). This is implied by the \( \epsilon = 0 \) case, which we proved in (105). \( \square \)
By the duality of the forward-reverse Brascamp-Lieb inequality (Theorem 8)\(^{18}\), we conclude from Lemma 16 that

**Lemma 17.** For any \( \epsilon > 0 \) and nonnegative continuous \((f_j), (g_i)\) satisfying

\[
\sum_{i=1}^{l} \log g_i(x_i) \leq \sum_{j=1}^{l} (T^\epsilon f_j)(x^j), \quad \forall x^l \in \mathbb{R}^l, \tag{A37}
\]

we have

\[
\left( \frac{2\pi}{\epsilon} \right)^{\frac{l}{2}} \prod_{i=1}^{l} g_i dx \leq \prod_{i=1}^{l} f_j dx. \tag{A38}
\]

Now suppose that Theorem 3 is not true; then there are nonnegative continuous \((f_j)\) and \((g_i)\) satisfying (17) while

\[
\prod_{i=1}^{l} g_i(x) dx > \prod_{i=1}^{l} f_j(x) dx, \tag{A39}
\]

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

\[
g_i(x) = 0, \quad \forall x: |x| \geq R, 1 \leq i \leq l; \tag{A40}
\]
\[
f_j(x) \geq \delta e^{-x^2}, \quad \forall 1 \leq j \leq l, \tag{A41}
\]

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

\[
\sum_{i=1}^{l} m_i x_i \in [-\sqrt{l}R, \sqrt{l}R]. \tag{A42}
\]

Since \( \log f_j \) is uniformly continuous on \([-2\sqrt{l}R, 2\sqrt{l}R]\) for each \( j \) and since we assumed (A41), we have

\[
\lim_{\epsilon \to 0} \inf_{x^l \in [-R, R]^l} \left\{ \sum_{j=1}^{l} (T^\epsilon f_j)(x^j) - \sum_{j=1}^{l} (T^0 f_j)(x^j) \right\} \geq 0. \tag{A43}
\]

But since we assumed (17) and (A40), we must also have

\[
\lim_{\epsilon \to 0} \eta_{\epsilon} \geq 0 \tag{A44}
\]

where

\[
\eta_{\epsilon} := \inf_{x^l \in \mathbb{R}^l} \left\{ \sum_{j=1}^{l} (T^\epsilon f_j)(x^j) - \sum_{i=1}^{l} \log g_i(x_i) \right\}. \tag{A45}
\]

---

\(^{18}\) Although we stated Theorem 8 for finite reference measures \((\mu_i)\), we see from the proof that the relatively easy direction “entropic⇒functional” does not need this assumption. Moreover the assumption in Theorem 1 that \((f_j)\) and \((g_i)\) are bounded away from 0 was made to ensure that \( \log f_j \) and \( \log g_i \) are bounded functions so that the conditional expectation operators as defined in Section 2 can be applied to them. But this assumption can be dispensed with when some specific conditional expectation operators can be applied to noncontinuous functions, as is the case of Lemma 17.
Put
\[
\tilde{g}_1^\varepsilon := \exp(\eta_0)g_1, \quad (A46)
\]
\[
\tilde{g}_i^\varepsilon := g_i, \quad i = 1, \ldots, l. \quad (A47)
\]

Then \((\tilde{g}_1^\varepsilon)\) and \((f_j)\) satisfy the constraint \((A37)\) for any \(\varepsilon > 0\). By applying the monotone convergence theorem and then Lemma 17,
\[
\prod_{i=1}^l \int g_i(x_i) dx_i \leq \lim_{\varepsilon \to 0} \left( \frac{2\pi}{\varepsilon} \right)^{\frac{l}{2}} \prod_{i=1}^l \int \tilde{g}_i^\varepsilon d\gamma_2
\]
\[
\leq \prod_{i=1}^l \int f_j(x) dx
\]
which violates the hypothesis \((A39)\), 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 \(T_j^\varepsilon\) the same as \((A34)\) except that \(N_\varepsilon \sim N(0, \varepsilon I)\), where \(I\) is the identity matrix whose dimension is clear from the context (equal to \(\dim(E_j)\) here), and let \(P_{Y_j|X_1, \ldots, X_l}^\varepsilon\) be the dual operator. For each \(i = 1, \ldots, l\), let \(\nu_i^\varepsilon := \left( \frac{2\pi}{\varepsilon} \right)^{\frac{1}{2}\dim(E_i)} \cdot N(0, \varepsilon^{-1} I)\), whose density converges pointwise to that of \(\nu_0^i\), defined as the Lebesgue measure on \(E_i\). Define
\[
d^\varepsilon := \sup \left\{ \sum_{i=1}^l b_i \log \nu_i^\varepsilon(g_i) - \sum_{j=1}^m c_j \log \int f_j \right\}
\]
\[
(A50)
\]
where the supremum is over nonnegative continuous functions \(f_1, \ldots, f_m\) and \(g_1, \ldots, g_l\) such that the summands in \((A50)\) are finite and
\[
\sum_{i=1}^l b_i \log g_i(x_i) \leq \sum_{j=1}^m c_j (T_j^\varepsilon \log f_j)(x_1, \ldots, x_l), \quad \forall x_1, \ldots, x_l.
\]
\[
(A51)
\]
The same limiting argument \((A39)-(A49)\) extended to the vector case shows that
\[
d^0 \leq \lim_{\varepsilon \to 0} d^\varepsilon.
\]
\[
(A52)
\]
Next, define \(F_0^\varepsilon(\cdot)\) for \((\mu_j)\), \((\nu_i^\varepsilon)\) and \(P_{Y_j|X_1, \ldots, X_l}^\varepsilon\), similarly to \((83)\). The entropic⇒functional argument (see Footnote 18) shows that
\[
d^\varepsilon \leq \sup_{P_{X_1}, \ldots, P_{X_l}} F_0^\varepsilon(P_{X_1}, \ldots, P_{X_l}).
\]
\[
(A53)
\]
But Theorem 9 based on the rotational invariance of the Gaussian measure can be extended to the vector case, so for any \(\varepsilon > 0\),
\[
\sup_{P_{X_1}, \ldots, P_{X_l}} F_0^\varepsilon(P_{X_1}, \ldots, P_{X_l}) = \sup_{P_{X_1}, \ldots, P_{X_l}} F_0^\varepsilon(P_{X_1}, \ldots, P_{X_l}),
\]
\[
(A54)
\]
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 $F_0^c$ is monotonically decreasing in $c$, and in particular

$$\sup_{P_{X_1,\ldots,P_{X_t}} \in c.G.} F_0^c (P_{X_1}, \ldots, P_{X_t}) \leq \sup_{P_{X_1,\ldots,P_{X_t}} \in c.G.} F_0^0 (P_{X_1}, \ldots, P_{X_t}).$$

To finish the proof with the above chain of inequalities, it only remains to show that the right side of (A59) equals to the supremum in (A50) with $(f_j)$ $(g_j)$ taken over center Gaussian functions. This follows by similar steps as the proof of the functional-arrow entropy 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

$$\Lambda(r, M) := \ln \int \exp \left( r + x^\top M x \right) \, dx$$

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

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

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

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

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

The supremum in (A59) equals $+\infty$ if $H$ is not positive-semidefinite. But if $H$ is positive-semidefinite, the supremum equals $-\frac{1}{2} \ln 2\pi e |H|$, which is equal to the relative entropy between $N(0, H)$ and the Lebesgue measure (supremum achieved when $M = -(2H)^{-1}$). 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

29. Lieb, E.H. Gaussian kernels have only Gaussian maximizers. Inventiones Mathematicae 1990, 102, 179–208.


