Title

Gaussian Entropy Inequalities

Author

Aras, Efe

Publication Date

2023

Peer reviewed|Thesis/dissertation

Gaussian Entropy Inequalities

By

Efe Aras

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Thomas Courtade, Chair
Professor Venkat Anantharam
Professor Michael Christ

Fall 2023

Gaussian Entropy Inequalities

Abstract

Gaussian Entropy Inequalities

by

Efe Aras

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Thomas Courtade, Chair

We give a broad overview of Gaussian entropy inequalities, discuss their scope, analyze their structure, and introduce novel ones. Our new inequalities highlight the connection between information theoretic and analytical inequalities. We then derive a Gaussian comparison inequality that unites a bulk of pre-existing Gaussian entropy inequalities. We present the equality conditions for the Anantharam–Jog–Nair inequalities, and thus derive equality conditions for a wide class of inequalities including the entropy power inequality, the Zamir–Feder inequality, and the Brascamp–Lieb inequalities. We conclude with a discussion of the extremizers of Forward-Reverse Brascamp–Lieb inequalities.

Anneme, anneanneme ve dedeme

Contents

Acknowledgments

I am deeply grateful to my advisor, Professor Thomas Courtade, for his unwavering support and constant guidance. Whenever I would feel stuck, I could always go to his office and he would have something exciting to talk about. Whenever I felt like a problem was impossible, if he believed otherwise, he would say so, and his confidence would help me understand papers a bit deeper. His work ethic, standards, and passion for academics is admirable. I have seen him take the hard route over the easy route many times due to his personal and professional standards, such as when he wrote a full textbook for EE226 when he taught the class. I will fondly remember our three-plus hour meetings, and his amazing jokes, such as the time when he congratulated the class for achieving a uniform distribution on their exam scores which maximized the information conveyed.

I would also like to thank my thesis committee members Professor Venkat Anantharam and Professor Michael Christ for their thorough feedback, and for being so accommodating. Without Shirley Salanio, the department would probably not function, and I am grateful that even with all on her plate, she finds the time and energy to reach out to individual students in their times of need. I am also thankful to Professor Satish Rao, who has always welcomed me into the theory group, and also invited me and my mom to his end-of-year party for his students.

I am grateful to my collaborators Kuan-Yun Lee, Ashwin Pananjady, Yeshwanth Cherapanamjeri and Albert Zhang for their camaraderie. The coffee that Ashwin made for me and Kuan-Yun still keeps me up when I think about it; even more so than the Russian paper Kuan-Yun and I translated.

I am thankful for the members of BLISS lab for always being open to random conversations and fun problems. Thanks to Ashwin Pananjady, Vidya Muthukumar, Orhan Ocal, Soham Phade, Vipul Gupta, Avishek Ghosh, Billy Fang, and Banghua Zhu for constant inspiration and always answering my questions. I am happy to report that the new generation (Nived Rajamaran, Landon Butler, Syomantak Chaudhuri, Yigit Efe Erginbas, Justin Kang) have done a great job reviving the social spirit that was on hiatus, with game nights, gym socials and a basketball hoop in the lab.

I am grateful to Professor Kannan Ramchandran for hiring me as his TA, entrusting me with significant teaching responsibilities and being a beloved teacher by students and course staff alike. I am just as grateful to my course staff from then; whether it be during our Tahoe trips, homework parties, or gatherings in SF, we always found ways to bond.

I could not have done my PhD without support from many friends I have made in Berkeley. A special shoutout to friends I spent the most time with: my fellow social butterfly Elizabeth Yang for the random coffee and boba trips, Roy Zhao for our shared academic drive and our valiant effort with our theses, Arun Ganesh for the many puzzle teams we were on, Jingjia Chen for helping me send my qual emails, Yeshwanth Cherapanamjeri for the long conversations on books and life, and Zihao Chen for the many travel stories. I am forever grateful to my housemate Forest Yang, who has been a bedrock of support for most of my

PhD. He always found a way to make me happier and wiser, whether it be via *Jeopardy*, or *Mother of Learning*; he is one of the few people with whom I can talk about anything.

I am further grateful to my friends Siqi Liu, Ke Wang, Suma Anand, Ekin Karasan, and Shreya Ramachandran for letting me stop by their offices to get some sunlight at random times in the day. I am grateful to Emil Albrychiewicz, Vickie Ye, Vitchyr Pong, and Zachary Wu for keeping me active through graduate school. I am grateful to Sinho Chewi for helping edit my qual slides, and Edward Zeng for helping edit my dissertation.

My other friends in Berkeley: Daniel Raban, Gus Callaway, Deniz Korman, and Andrew Chen regularly promised a warm meal and a fun night. I want to thank my college friends Debbie Burdinski, Alex Garvey, Michelle Wang, Tarun Repala, Kelvin Niu, John Dai and Matthew Faw, who have invited me to their homes and let me be a part of their lives.

I am grateful to all my students who have made teaching a joy; watching y'all have an "aha" moment gives an instructor indescribable pride, and I am grateful to all the teachers who have given me my own "aha" moments throughout my life. Lastly, I am grateful to my family. During my PhD, both of my grandparents sadly passed away, but I wish to keep their memory alive: I have always admired my grandfather's ability to talk to anyone, his energy to walk many kilometers even above 70 years old, his innate desire to jump to action and fix things, and his appreciation for academics even though he was unable to get an education beyond fifth grade. I looked up to my grandmother when she could resolve even the deepest of conflicts, lower the tension and bring us together. Finally, I do not have enough words to thank my mom, who has had to raise me as a single mother, working late hours and still ensuring that I got all the care and love a child wants. When I talk to her, I know that she hears me and every day I feel our bond is growing stronger even though we are an ocean apart.

All in all, I want to acknowledge the education and the love I was given, and hope to pay it forward in day-to-day life as best as I can.

The work for this thesis was supported by NSF-CCF 1750430.

Chapter 1

Introduction

"A quoi sont dues les erreurs accidentelles, nous l'ignorons, et c'est justement parce que nous l'ignorons que nous savons qu'elles vont obeir a la loi de Gauss."
- Poincaré [1912]

1.1 Gaussian

Why do we care about Gaussians? One answer might be that Gaussians show up even when we are not looking for them. Indeed, we can trace a direct path from the dawn of probability involving the correspondence of Fermat and Pascal to the discovery of Gaussians. In the 1600s, Fermat and Pascal, thinking about gambling problems, ended up discussing what would today be called binomial distributions. However, binomial probabilities can be hard to compute, so one of the earliest sightings of Gaussian functions came when De Moivre [1733] showed that

$$\binom{n}{\frac{n}{2}+d}\left(\frac{1}{2}\right)^n \approx \frac{2}{\sqrt{2\pi n}} e^{-2d^2/n}.$$

It would only take a few years until Gaussian functions showed up again. In 1801, a suspected new planet vanished behind the Sun. Where would it emerge from? Gauss, to answer this question, assumed that the maximum likelihood estimator of a quantity from noisy i.i.d observations is the empirical mean, and derived that the error must have distribution

$$\phi(x) = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2} \tag{1.1}$$

which is what we now call a Gaussian function (with variance $\frac{1}{2h^2}$)! [1] We refer the reader to Stahl [2006] for the rest of the story and references.

[1] Hopefully the reader will be reassured to hear that he then predicted where the planet would emerge from, and he was right.

Elsewhere in the world, Jean Baptiste Joseph Fourier was trying to understand how heat distributes along a heated rod. He showed that locally, the heat profile along the rod satisfies

$$\frac{\partial}{\partial t}u(t,x) = \frac{\partial^2}{\partial x^2}u(t,x)$$

[Fourier, 1807] which we now know admits a Gaussian Green's function. There are many other stories, ranging from Laplace's central limit theorem Laplace [1810], to how Einstein derived Brownian motion to explain the observations of pollen movements [Einstein et al., 1905], to how Black–Scholes ended up getting Gaussians from an efficient market [Black and Scholes, 1973]. This rich variety of characterizations hint at the universality of Gaussians, so it should come as no surprise that Gaussians arose in ideas related to *entropy*.

1.2 Entropy

Physicists knew about the notion of entropy since the 1800s. The first mention of entropy is credited to Clausius [1865], stemming from the Greek words "en-tropie", meaning intrinsic direction. Maxwell then connected statistics to thermodynamics, giving rise to statistical mechanics [Bernstein, 1963]. Finally, Boltzmann defined entropy as the logarithm of the number of microstates and established it on a mathematical ground. [Boltzmann, 1872]

Entropy got a whole new meaning with regards to information with the pioneering work of Shannon [1948]. He axiomatically defined the idea of entropy and he interpreted it as a measure of information. He defined the entropy of a continuous random variable X with density p as

$$h(X) := -\int p(x)\log p(x)\,dx$$

which is the definition we will be using throughout whenever it is well-defined. The usefulness of this definition was not purely axiomatic; it also showed up as an essential quantity while computing the rate of information that can be sent through a *channel*. Much like the last section, his paper ended up naturally mentioning Gaussians. For starters, Gaussians are the maximum entropy distribution subject to second moment constraints. Also, they are the standard model for white thermal noise in electrical engineering. One important observation he made was the *entropy power inequality*, which we defer to the next section.

1.3 Inequalities

Inequalities form the backbone of mathematics. For instance, the key relations underlying the Arithmetic–Geometric inequality can be traced all the way back to Euclid. [Steele, 2004]. This would be quite a long thesis if we were to consider all inequalities, so here onward, we do our best to specialize to *Gaussian-extremized* inequalities, that is, inequalities that are met with equality when one plugs in a Gaussian (function or random variable) as an

argument. While it sounds like a restricted class, note that this class includes a lot of well-known functional and geometric inequalities, including the Hölder's inequality and the Prékopa–Leindler inequality. As a simple example, consider the claim made in the previous section that Gaussians maximize entropy subject to second moment constraints. An alternate formulation is that for all n-dimensional random vectors X

$$\mathbb{E}|X|^2 \geq N(X) := \frac{1}{2\pi e} e^{\frac{2}{n}h(X)} \tag{1.2}$$

with equality if and only if X is an isotropic Gaussian, where $|\cdot|$ is the standard Euclidean norm on $\mathbb{R}^n$. This inequality gives that the best signal to use in an additive channel with power constraints has to be a Gaussian. Conversely, consider the entropy power inequality (EPI) which states, for independent n-dimensional random variables X_1, X_2

$$N(X_1) + N(X_2) \leq N(X_1 + X_2) \tag{1.3}$$

with equality if and only if X_1 and X_2 are Gaussians with proportional covariances. This inequality shows, among many other implications, that the worst-case noise in an additive channel is Gaussian.

We will save the rest of the inequalities of interest and their implications to Chapter 3, but rest assured there are quite a few of them. We would like to repeat the following idea: There is nothing in (1.2) and (1.3) that *a priori* suggests any extremal behavior by Gaussians. However, (1.3) has the following equivalent form; if we let Z_i be an n-dimensional Gaussian with entropy equal to $h(X_i)$ for $i = 1, 2$, then, for independent Z_1, Z_2, we have

$$h(X_1 + X_2) \geq h(Z_1 + Z_2).$$

Turns out, we can generalize this much further. Indeed, the goal of this thesis is to give a broad overview of major pre-existing Gaussian extremizable inequalities, generalize a portion of them, and analyze their equality conditions.

1.4 Main results of this thesis

For the rest of the thesis, $\mathbf{c} := (c_i)_{i=1}^k$, and $\mathbf{d} := (d_j)_{j=1}^m$ will denote collections of non-negative scalars, and $\mathbf{B} := (B_j)_{j=1}^m$ denotes suitably dimensioned linear maps mapping Euclidean space E_0 to other Euclidean spaces $(E^j)_{j=1}^m$. We will refer to a triplet $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ as a *datum*.

Gaussian comparison inequality

Let $X_1, \ldots, X_k$ be random vectors with finite second moments. Let $\Pi(X_1, \ldots, X_k; \nu)$ denote the set of correlation-constrained couplings[2] of these vectors. Then, we showed in Aras

[2]Full definition can be found in beginning of Chapter 4.

and Courtade [2022] that there exist Gaussians $Z_1, \dots, Z_k$ with $\dim(X_i) = \dim(Z_i)$ and $h(Z_i) = h(X_i)$ such that

$$\max_{X \in \Pi(X_1,\dots,X_k;\nu)} \sum_{j=1}^{m} d_j h(B_j X) \geq \max_{Z \in \Pi(Z_1,\dots,Z_k;\nu)} \sum_{j=1}^{m} d_j h(B_j Z). \tag{1.4}$$

Inequality (1.4) has a number of consequences. It implies the entropy power inequality, the Brascamp–Lieb inequalities, and the Barthe inequalities. An interesting corollary of (1.4) is a novel entropy power inequality for dependent random variables. In particular, by a judicious choice of parameters, we get, for $X_1, X_2 \in \mathbb{R}^d$, $\zeta \in [0, +\infty]$

$$N(X_1) + N(X_2) + 2\sqrt{(1 - e^{-2\zeta/d}) N(X_1) N(X_2)} \leq \max_{\substack{\Pi(X_1,X_2): \\ I(X_1;X_2) \leq \zeta}} N(X_1 + X_2), \tag{1.5}$$

where the max is over all the couplings of X_1 and X_2 satisfying the given mutual information constraint. We quickly mention that at $\zeta = 0$, this inequality reduces to the EPI, and at $\zeta = +\infty$, it implies the Brunn–Minkowski inequality by specializing to uniform random variables supported on compact sets. Quantitatively linking the Brunn–Minkowski and the EPI using only Shannon entropies had proved elusive, and had been somewhat of a looming question, which our comparison inequality answers.

By taking ν be identically zero, we recover the Anantharam–Jog–Nair inequality [Anantharam et al., 2019], which is the main inequality of interest for the second part of the thesis.

Anantharam–Jog–Nair inequality: Extremizability and structure of extremizers

Anantharam, Jog and Nair characterized the best (i.e., smallest) constant C such that the entropy inequality

$$\sum_{i=1}^{k} c_i h(X_i) \leq \sum_{j=1}^{m} d_j h(B_j X) + C \tag{1.6}$$

holds for any choice of independent $\mathbb{R}^{n_i}$-valued random variables X_i with finite entropies and second moments, $1 \leq i \leq k$, with $X := (X_1, \dots, X_k)$.

Anantharam, Jog and Nair had left open the question of extremizability. That is, when do there exist random vectors $(X_i)_{i=1}^k$ such that (1.6) is met with equality, and what form do any such extremizers take?

We answer both questions in this thesis, which will be based on Aras et al. [2022]. The precise characterization of extremizers is somewhat complicated, but the general idea is easily understood in the context of a toy example. For $\lambda \in (0, 1)$, the following holds: If (X, Y) is independent of Z, and Y and Z are of the same dimension, then

$$\lambda h(X, Y) + (1 - \lambda) h(Z) \leq \lambda h(X) + h(\lambda^{1/2} Y + (1 - \lambda)^{1/2} Z). \tag{1.7}$$

This inequality is obtained by a concatenation of subadditivity of entropy and the EPI. Restricting attention to cases where all entropies are finite, we can use known equality cases for both to assert that (X, Y) and Z are extremizers in (1.7) if and only if (i) X and Y are independent; and (ii) Y and Z are Gaussian with identical covariances.

Roughly speaking, all extremizers of the AJN inequality (1.6) resemble the above example. That is, extremizers are characterized by a rigid factorization into independent components, where some components can have any distribution, and the remaining are necessarily Gaussian with covariances that are typically linked in some way.

Just as a quick example, our results immediately imply the equality conditions for the Zamir–Feder inequality. Its extremizers are characterized by all present non-recoverable components being Gaussian [Rioul and Zamir, 2019, Theorem 1], which has a geometric interpretation that our characterization immediately captures.

Forward-Reverse Brascamp–Lieb inequalities: Structure of extremizers

Similar to the development above, Courtade and Liu [2021] (see also Liu et al. [2016]) characterized the smallest constant D such that the followng entropy inequality holds:

$$\sum_{i=1}^{k} c_i h(X_i) \leq \max_{X \in \Pi(X_1,\dots,X_k)} \sum_{j=1}^{m} d_j h(B_j X) + D \tag{1.8}$$

where we again work with X_i with finite entropies and second moments, and $\Pi(X_1, \dots, X_k)$ denote the set of couplings of $(X_i)_{i=1}^k$. We give an outline of how to characterize the extremizers of this inequality for a vast class of $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ that satisfy a structural assumption. The class we consider includes the Brascamp–Lieb inequalities, the Barthe inequalities, and the Barthe–Wolff inverse Brascamp–Lieb inequalities. We note that a special case of the Barthe inequalities, the Prékopa–Leindler inequality, has log-concave extremizers, so we would expect the extremizers of (1.8) to potentially have components that give rise to log-concave extremizers.

It turns out for a special set of data, the extremizers admit a factorization into independent components, where some components are arbitrary, some have to be log-concave, and the remainder are Gaussian with a particular covariance structure. We further observe another subclass where the extremizers are only *some* of the log-concave distributions. We recover the results of Boroczky et al. [2022] and Valdimarsson [2008], and further characterize equality conditions for the Gaussian-extremizable instances Barthe–Wolff inequality from Barthe and Wolff [2018], which were previously not known.

1.5 Outline of the thesis

We will first outline the notation that will prevail throughout the thesis in Chapter 2. We will then do a brief literature review in Chapter 3 while showing a duality between functional and entropic inequalities. In Chapter 4, we will discuss the comparison principle outlined before, and Chapter 5 will be dedicated to equality conditions of AJN inequality in detail. Finally, we reserve Chapter 6 to summarize very recent results on FRBL extremizers.

A summary of key contributions is presented below:

1. We provide a compendium for various entropic inequalities, and their relation to existing functional inequalities. As an important example, we note that the well-known Cover–Zhang inequality readily admits a generalization to include non-identically distributed random variables which we dub Cover–Zhang type inequalities. Our generalization relies on the duality between Prékopa–Leindler and Cover–Zhang type inequalities.

2. We present a Gaussian comparison inequality that unifies a broad landscape of existing entropy inequalities. As a specific example, we use our comparison inequality to interpolate between the EPI and Brunn–Minkowski just by using entropies. We extend a Gaussian saddle point property of a well-known information game to an asymmetric information game.

3. We analyze the structure of the Anantharam–Jog–Nair inequality. We outline the full connection between extremizability, geometricity, and Gaussian extremizability; and we give the full structure of the extremizers. In particular, we show that all the extremizers have a rigid independence structure, with certain factors being isotropic Gaussians. We note that this implies recent important structural results such as equality conditions for the Zamir–Feder inequality [Rioul and Zamir, 2019].

4. We give a brief overview of how to establish the extremizers for the Forward-Reverse Brascamp–Lieb inequality. We argue that the extremizers again have an induced independence structure, and in each component, they either have arbitrary distributions, or they have to be log-concave. We further note that our structural results unite the results on equality conditions of Barthe's inequality [Boroczky et al., 2022], and the Brascamp–Lieb inequality [Valdimarsson, 2008], and further characterize equality conditions for the Gaussian-extremizable instances of Barthe–Wolff inequality [Barthe and Wolff, 2018].

Parts of the dissertation are based on Aras and Courtade [2022], Aras et al. [2022].

Chapter 2

Preliminaries

In this chapter, we will set up the main notation that we will be using throughout the thesis. In particular, we will be establishing the underlying Euclidean structure, and defining relative entropy and differential entropy.

2.1 Basic setup

Spaces

We will be working on Euclidean spaces, and to be precise, we will want to work with finite dimensional real vector spaces E equipped with an inner product, which we will denote by the usual notation $x, y \in E \mapsto x^T y$. We will denote the Euclidean norm by $|\cdot|$. The (external) **direct sum** of finite dimensional vector spaces $(E_i)_{i=1}^k$, denoted $E_0 := \oplus_{i=1}^k E_i$, is the Cartesian product of E_is. A subspace V of E_0 is **product-form** if $V = \oplus_{i=1}^k V_i$ where $V_i \subset E_i$

Remark 1. *Throughout the thesis, the reader can replace E_i with $\mathbb{R}^{n_i}$ where $n_i = \dim(E_i)$. Similarly, the reader can thus view E_0 as $\oplus_{i=1}^k E_i \cong \mathbb{R}^{\sum_{i=1}^k n_i}$ [1]. Referring to the spaces more abstractly helps to clearly identify them in discussion.*

Note that we have a few canonical maps that come from this construction. The identity map, denoted id_E, is the identity map $E \to E$. A vector $x \in E_0$ will frequently be written in its coordinate representation $x = (x_1, \ldots, x_k)$, where $x_i = \pi_{E_i}(x)$, $1 \leq i \leq k$, is the natural coordinate projection of E_0 onto E_is. We will also rely on (orthogonal) projections and restricted projections. If V is a subspace of a Euclidean space E, then we say the linear map $P_V : E \to E$ is an orthogonal projection if $P_V^2 = P_V = P_V^T$. Finally, we have the restricted projection map $\pi_V : E \to V$ if $P_V = \pi_V^T \pi_V$. Note that the restricted projection map notation agrees with the coordinate projections. We will need the set of quadratic forms on E, denoted $\mathbf{S}(E)$, which consists of symmetric bilinear forms defined on E. We will also

[1] We will slightly alter this definition in Chapter 6.

need to consider the set of positive definite quadratic forms on E, denoted $\mathbf{S}^+(E)$, which are quadratic forms that are strictly positive; and the set of positive semi-definite ones, denoted $\mathbf{S}_0^+(E)$ which is when the quadratic form is allowed to be zero for non-zero inputs.

If $(A_i : E_i \to E_i)_{i=1}^k$, are linear maps, then the direct sum of operators $A = \oplus_{i=1}^k A_i$ is a linear map from E_0 to itself and, without confusion, can be denoted as the block-diagonal operator

$$A = \mathrm{diag}(A_1, \ldots, A_k).$$

So, as an example of the above, we have $\mathrm{id}_{E_0} = \oplus_{i=1}^k \mathrm{id}_{E_i} \equiv \mathrm{diag}(\mathrm{id}_{E_1}, \ldots, \mathrm{id}_{E_k})$. Again, this is all compatible with the representation of linear operators as matrices. Lastly, if $K_i \in \mathbf{S}_0^+(E_i)$, $1 \leq i \leq k$, then we let $\Pi(K_1, \ldots, K_k)$ denote the subset of $\mathbf{S}_0^+(E_0)$ consisting of those matrices K such that

$$\pi_{E_i} K \pi_{E_i}^T = K_i, \quad \forall 1 \leq i \leq k.$$

Gaussians

It is useful to define what is meant by a Gaussian, as we will be relying on them throughout. For a given space E, $\mu \in E$, $\Sigma \in \mathbf{S}_0^+(E)$, a Gaussian distribution defined on E denoted by $\gamma(\mu, \Sigma)$, or $N(\mu, \Sigma)$, is a distribution such that if $Z \sim \gamma(\mu, \Sigma)$, then $\mathbb{E}[e^{it^T Z}] = e^{\frac{-t^T \Sigma t + i t^T \mu}{2}}$. We reserve γ to refer to a Gaussian. Furthermore, we will use the shorthand $\gamma_E \equiv \gamma(0, \mathrm{id}_E)$, and this distribution is referred to as the **standard Gaussian distribution** on E. We will also use the shorthand $\gamma_i \equiv \gamma_{E_i}$. Note that a standard Gaussian admits a density $\propto e^{-\frac{|x|^2}{2}}$. We will occasionally refer to functions that are proportional to Gaussian densities as Gaussian functions.

Measures and couplings

We will equip E_is (and E_0) with their Borel σ-algebras. We will denote the set of all signed measures on E_0 as $M(E_0)$, and the set of all positive measures as $M^+(E_0)$. Let $(X_i)_{i=1}^k$ be random vectors defined on $(E_i)_{i=1}^k$. We say that X, a random variable on E_0, is a **coupling of** $X_1, \ldots, X_k$ if the law of $\pi_i(X)$ is the same as the law of X_i for all $i \in [k]$. We denote the set of couplings by $\Pi(X_1, \ldots, X_k)$. Note that this overloaded notation is consistent with our notation for matrices. Indeed, if $X_i \sim N(0, K_i)$, $1 \leq i \leq k$, then $X \sim N(0, K)$ is a coupling in $\Pi(X_1, \ldots, X_k)$ if and only if $K \in \Pi(K_1, \ldots, K_k)$. In any integral where the measure is unspecified (i.e. $\int_V f$), the integration is done with respect to Lebesgue measure on the space V specified by the context.

2.2 Shannon information quantities

Kullback–Leiber divergence and differential entropy

Let μ, ν be probability measures on E. We define the **Kullback–Leiber (KL) divergence** (also referred to as **relative entropy**) to be

$$D(\mu\|\nu) := \begin{cases} \int_E \frac{d\mu}{d\nu} \log\left(\frac{d\mu}{d\nu}\right) d\nu & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise.} \end{cases}$$

The celebrated result by Donsker and Varadhan [1983] allows writing KL divergence as a supremum:

$$D(\mu\|\nu) := \sup_{f \in C_b(E)} \left\{ \int_E f d\mu - \log \int_E e^f d\nu \right\} \tag{2.1}$$

where we can take $C_b(E)$ to be continuous, bounded, real-valued functions on E. We note that the quantity inside the sup is weakly continuous in μ for a fixed ν, and suprema of weakly continuous functions being lower semi-continuous immediately implies lower semi-continuity of KL divergence. If X has law μ, and Y has law ν, we will regularly use the shorthand $D(X\|Y) \equiv D(\mu\|\nu)$.

A particularly important property of KL divergence, that results from its joint convexity with respect to (μ, ν), is the **data processing inequality**, which says that for any measurable map between two measurable spaces $T : E \to E'$,

$$D(T\#\mu\|T\#\nu) \leq D(\mu\|\nu).$$

Let X be a random variable with law μ having density f with respect to the Lebesgue measure on E. We will define the **differential entropy** as follows:

$$h(\mu) \equiv h(X) = -\int_E f \log f$$

provided the integral is well-defined in the Lebesgue sense. This will always be the case in our settings. Much like KL divergence, differential entropy also admits a variational formula:

$$h(\mu) = \inf_f \left\{ \log \int_E e^f - \int_E f d\mu \right\} \tag{2.2}$$

where the infimum over f is taken for measurable $f : E \to \mathbb{R}$ bounded from above. We let $\mathcal{P}(E)$ denote probability measures with finite entropy and second moment, and we let $\mathcal{G}(E)$ denote the subset of $\mathcal{P}(E)$ that consists of Gaussian measures. When there is no cause for ambiguity, we adopt the hybrid notation where a random vector X and its law μ are denoted interchangeably. So, for example, writing $X \in \mathcal{P}(E)$ means that X is a random vector taking values in E, having finite entropy and finite second moments. We use this

notation to emphasize that we are restricting our domain of interest; we do not make any claims of having results for more general distributions.

Related to differential entropy, we will define the **entropy power** of a random variable in E, which is

$$N(X) := \frac{1}{2\pi e} e^{\frac{2}{\dim(E)} h(X)}.$$

One can view $N(X)$ as sort of a volume, which is highlighted by

$$\mathbb{E}|X|^2 \geq N(X)$$

with equality iff $X \sim N(0, \sigma^2 \operatorname{id}_E)$ for some $\mu \in E$, $\sigma^2 \in \mathbb{R}_+$. While at it, we note another special aspect of Gaussians that we will be using multiple times. Relative entropy of μ with respect to γ_E is given by the differential entropy of μ and a second-moment term:

$$D(\mu \| \gamma_E) = -h(\mu) - \frac{1}{2}(\mathbb{E}_{X\sim\mu}|X|^2 - d \log 2\pi) \tag{2.3}$$

Proof. Let $f := \frac{d\mu}{d\lambda}$, where λ is the Lebesgue measure on E. We note that $\frac{d\mu}{d\gamma_E} = \frac{\frac{d\mu}{d\lambda}}{\frac{d\gamma_E}{d\lambda}}$, so we can write

$$D(\mu \| \gamma_E) = \int_E f \log \frac{f}{\frac{1}{\sqrt{2\pi}} e^{-\frac{|x|^2}{2}}} = \underbrace{\int_E f \log f}_{-h(X)} - \frac{1}{2}(\mathbb{E}_\mu |X|^2 + \frac{d}{2} \log(2\pi)).$$

□

Remark 2. *Note that if μ and γ_E have the same mean and second moments, (2.3) can be simplified to $h(Z) - h(X)$, where Z is the standard Gaussian on E.*

We will finally define two other quantities that the reader will run into while reading this thesis. For a random pair $(X, Y) \in \mathcal{P}(E_1 \oplus E_2)$, we can define their **mutual information** as

$$I(X;Y) = h(X) + h(Y) - h(X,Y).$$

Note that $I(X;Y)$ admits a simple expression using KL divergence as well. In particular, if X and Y have laws μ and ν respectively and their joint law is ζ, we have

$$I(X;Y) = D(\zeta \| \mu \times \nu)$$

where we use $\mu \times \nu$ to denote the product measure. Note that this form immediately implies that $I(X;Y) = 0$ iff X and Y are independent.

Finally, we remind the reader that information theorists sometimes refer to L^p norms of functions via so-called **Rényi entropies**. If $f \in L^p(\mathbb{R})$ is the density of a random variable X for some $p > \mathbb{R}^+ \setminus \{1\}$, then the Rényi entropy of X is given by

$$h_p(X) := \frac{p}{1-p} \log \|f\|_p.$$

Under mild conditions, h_p converges to Shannon entropy as $p \to 1$. Note that Rényi entropies do not satisfy many nice properties of Shannon entropies, such as subadditivity, and in most information theoretic applications, Shannon entropies, and not Rényi entropies appear naturally. See the survey by Van Erven and Harremos [2014] for more properties of Rényi entropies and divergences.

We end this section with a few basic examples of Shannon entropies.

Example 3 (Entropy of a Gaussian distribution)**.** *Let $\gamma \sim N(\mu, \Sigma)$ for some $\mu \in E$, $\Sigma \in \mathbf{S}^+(E)$*

$$h(\gamma) = \frac{1}{2} \log \det (2\pi e \Sigma) .$$

Example 4 (Entropy of a uniform distribution)**.** *Let X be uniformly distributed over a measurable set S in E with nonzero finite measure $|S|$,*

$$h(X) = -\int_E \frac{1}{|S|} \log \frac{1}{|S|} = \log |S|.$$

We remark that uniform distributions maximize entropy subject to a support constraint.

Chapter 3

Examples of Gaussian Inequalities

In this chapter, we will lay out some of the existing inequalities, with a particular emphasis on chains of implications. We will demonstrate how entropic inequalities naturally give rise to statistical, geometric and functional inequalities. We will also showcase the power of the variational principle for entropy by going between functional inequalities and entropic ones.

3.1 Entropic examples

Entropy power inequality

Let us set the stage with the inequality that has been around since the inception of information theory, the entropy power inequality (EPI):

Theorem 5 (EPI)**.** *Let X and Y be independent random vectors in $\mathcal{P}(\mathbb{R}^n)$. Then,*

$$N(X) + N(Y) \leq N(X + Y). \tag{3.1}$$

This inequality was first observed by Shannon [1948], and rigorously proved by Stam [1959], Blachman [1965]. As such, it is sometimes referred to as the Shannon–Stam inequality. Later on, Lieb observed an equivalent form in the context of statistical mechanics:

Theorem 6 (Lieb [1990])**.** *Let X and Y be independent random vectors in $\mathcal{P}(\mathbb{R}^n)$. Then, for all $\lambda \in [0, 1]$,*

$$\lambda h(X) + (1 - \lambda)h(Y) \leq h(\sqrt{\lambda}X + \sqrt{(1 - \lambda)}Y).$$

The entropy power inequality is used throughout information theory to characterize capacities of various channels, such as broadcast channels [Bergmans, 1974, Weingarten et al., 2006, Mohseni and Cioffi, 2006], wiretap channels [Leung-Yan-Cheong and Hellman, 1978, Tekin and Yener, 2006], and interference channels [Costa, 1985]. The reader interested in

applications of the entropy power inequality in information theory are referred to the introduction of Rioul [2010] and references therein. The EPI also has found use and extensions in convex geometry. [Madiman et al., 2017]

Throughout the years, information theorists have developed more refined versions of the EPI. Costa [1985] fixed one of the variables to be a Gaussian to strengthen the EPI (see also Villani [2000]), which was further strengthened by Courtade [2017] to give a concise proof of the rate region for the two-encoder quadratic Gaussian source coding problem. Artstein et al. [2004] developed yet another strong EPI that they used to show monotonicity of entropy along the sequence of standardized sums appearing in the classical central limit theorem. Another generalization was given by Zamir and Feder [1993] which can be stated as follows: Let $X = (X_1, \ldots, X_k)$ be a random vector in $\mathbb{R}^k$ with independent coordinates $(X_i)_{i=1}^k$. If $Z = (Z_1, \ldots, Z_k)$ is a Gaussian vector with independent coordinates $(Z_i)_{i=1}^k$ and entropies satisfying $h(Z_i) = h(X_i)$, $1 \leq i \leq k$, then for any linear map $B : \mathbb{R}^k \to \mathbb{R}^n$, we have

$$h(BX) \geq h(BZ). \tag{3.2}$$

A generalization of the Zamir–Feder inequality is given by Anantharam et al. [2019].

The Anantharam–Jog–Nair inequality

We remind the reader of the definition of a datum given in Section 1.4. For a datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$, Anantharam, Jog and Nair (AJN) characterized the best (i.e., smallest) constant $C_{AJN}(\mathbf{c}, \mathbf{d}, \mathbf{B})$ such that the entropy inequality

$$\sum_{i=1}^{k} c_i h(X_i) \leq \sum_{j=1}^{m} d_j h(B_j X) + C_{AJN}(\mathbf{c}, \mathbf{d}, \mathbf{B}) \tag{3.3}$$

holds for any choice of independent random vectors $X_i \in \mathcal{P}(E_i)$, $1 \leq i \leq k$, with $X := (X_1, \ldots, X_k)$. This inequality unifies the Zamir–Feder inequality (3.2) (and consequently Shannon–Stam inequality (3.1)), and the entropic formulation of the (Euclidean) Brascamp–Lieb inequalities [1] under a common framework. In particular, Anantharam, Jog and Nair showed that the best constant can be computed by considering only Gaussian X_i's, and gave necessary and sufficient conditions for finiteness. As we will be working with their inequality extensively, we present their main result thoroughly below:

Theorem 7 (Anantharam et al. [2019])**.** *Fix a datum* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$*. For any independent random vectors* $X_i \in \mathcal{P}(E_i)$*,* $1 \leq i \leq k$ *and* $X = (X_1, \ldots, X_k)$*,*

$$\sum_{i=1}^{k} c_i h(X_i) - \sum_{j=1}^{m} d_j h(B_j X) \leq C_g(\mathbf{c}, \mathbf{d}, \mathbf{B}), \tag{3.4}$$

[1] that we will further present in Theorem 16

where $C_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is defined as the supremum of the LHS over independent Gaussian vectors $(X_i)_{i=1}^k$. Moreover, the constant $C_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is finite if and only if the following two conditions are satisfied.

(i) **Scaling condition:** *It holds that*

$$\sum_{i=1}^{k} c_i \dim(E_i) = \sum_{j=1}^{m} d_j \dim(E^j). \tag{3.5}$$

(ii) **Dimension condition:** *For all product-form subspaces $T \subset E_0$,*

$$\sum_{i=1}^{k} c_i \dim(\pi_{E_i} T) \leq \sum_{j=1}^{m} d_j \dim(B_j T). \tag{3.6}$$

Cover–Zhang

If we relax the independence assumption between the random variables, then the EPI as stated cannot hold (nor should one expect it to hold). However, considering the couplings between random variables gives rise to a similar inequality. We will first give the original form of the Cover and Zhang [1994] result.

Theorem 8 (Cover–Zhang)**.** *Let X, X' be random variables in $\mathcal{P}(\mathbb{R})$ with a common density f. Then, the inequality*

$$h(2X) \leq \max_{\Pi(X,X')} h(X + X')$$

holds with equality iff f is log-concave.

Remark 9. *The notation $\max_{\Pi(X,X')} h(X + X')$ indicates the entropy is maximized over all couplings of X, X'. This notation is consistent with that commonly employed in the literature of optimal transport [Villani, 2003].*

We can actually generalize this result to include non-identically distributed random variables.

Theorem 10 (Cover–Zhang type inequalities)**.** *Let X and Y be random vectors in $\mathcal{P}(\mathbb{R}^n)$, and $\lambda \in (0, 1)$*

(i) The following inequality holds:

$$\lambda h(X) + (1 - \lambda) h(Y) \leq \max_{\Pi(X,Y)} h(\lambda X + (1 - \lambda) Y)$$

(ii) Equality holds iff $X - \mathbb{E}X$ and $Y - EY$ are identically distributed with a log-concave distribution.

We defer an informal proof to the appendix. We will however note that this inequality is a dual form of the Prékopa–Leindler (PL) inequality that we recall here for convenience: given $\lambda \in [0,1]$ and non-negative $f, g, h \in L^1(\mathbb{R}^n)$, satisfying

$$h(\lambda x + (1-\lambda)y) \geq f^\lambda(x) g^{1-\lambda}(y) \qquad \forall x, y \in \mathbb{R}^n,$$

we have

$$\left(\int_{\mathbb{R}^n} f\right)^\lambda \left(\int_{\mathbb{R}^n} g\right)^{1-\lambda} \leq \int_{\mathbb{R}^n} h.$$

Equality conditions for the Prékopa–Leindler inequality have been shown in Dubuc [1977], which also requires log-concavity. An immediate corollary of Theorem 10 is the Brunn–Minkowski inequality [Brunn, 1887, Minkowski, 1910].

Theorem 11 (Brunn–Minkowski)**.** *Let $\lambda \in (0,1)$, and K and L be compact subsets of $\mathbb{R}^n$. Then, we have*

$$\lambda V(K)^{1/n} + (1-\lambda) V(L)^{1/n} \leq V(\lambda K + (1-\lambda)L)^{1/n}.$$

Moreover, in the case that K, L have non-empty interior, equality holds iff K, L are convex, with $K = L + x_0$ for some $x_0 \in \mathbb{R}^n$.

Proof of Theorem 11. Take X and Y to be uniform on compact sets $K, L \subset \mathbb{R}^n$ that have non-empty interior and note that for any coupling of X and Y, $\lambda X + (1-\lambda)Y$ has to be supported on $\lambda K + (1-\lambda)L$. Note that K, L having non-empty interiors ensures $h(X), h(Y) > -\infty$, and compactness ensures the boundedness of second moments of X and Y, and $h(X), h(Y) < \infty$. Thus, we see that X and Y are in $\mathcal{P}(\mathbb{R}^n)$. Now, invoking Theorem 10, we get

$$V(K)^\lambda V(L)^{1-\lambda} \leq \max_{\Pi(X,Y)} h(\lambda X + (1-\lambda)Y) \leq V(\lambda K + (1-\lambda)L).$$

This is equivalent to the form given above by a simple rescaling of sets [Madiman et al., 2017]. The equality condition follows immediately from Theorem 10. □

The implications of Brunn–Minkowski inequality are numerous [Gardner, 2002, Barthe, 2006]. It is a foundational inequality in convex geometry.

Remark 12. *It has long been observed that there is a striking similarity between the Brunn–Minkowski inequality and the EPI (see, e.g., Costa and Cover [1984] and citing works). It is well-known that each can be obtained from convolution inequalities involving Rényi entropies (e.g., the sharp Young inequality [Brascamp and Lieb, 1976, Lieb, 1990], or rearrangement inequalities [Wang and Madiman, 2014]), when the orders of the involved Rényi entropies are taken to the limit* 0 *or* 1*, respectively.*

3.2 Functional examples

Brascamp–Lieb Inequalities

Hölder's inequality, the sharp Young inequalities and the Loomis–Whitney inequality are all special cases of a wide class of functional inequalities commonly known as Brascamp–Lieb inequalities.

Definition 13 (Brascamp–Lieb inequalities)**.** *Define* $C_{BL}(\mathbf{d}, \mathbf{B})$ *as the smallest constant such that the following inequality is true for all non-negative* $f_j \in L^1(E^j), 1 \leq j \leq m$*:*

$$\int_{E_0} \prod_{j=1}^{m} f_j^{d_j}(B_j x)\, dx \leq e^{C_{BL}(\mathbf{d},\mathbf{B})} \prod_{j=1}^{m} \left(\int_{E^j} f_j \right)^{d_j}. \tag{3.7}$$

Inequalities of the form (3.7) *are referred to as Brascamp–Lieb inequalities*

Brascamp and Lieb [1976] proved that $C_{BL}(\mathbf{d}, \mathbf{B})$ can be computed by considering Gaussian functions for $(f_i)_{i=1}^k$ sharing the same domain E^j. They then used it to show sharp Young's inequality (that Beckner [1975] had shown a bit prior), and Nelson's hypercontractivity [Nelson, 1973]. Then, Lieb [1990] generalized to arbitrary $(B_j, E^j)_{j=1}^m$.

Beyond the applications listed, the Brascamp–Lieb inequality has been used in Ball [1989] to compute and give bounds on volumes of various convex sets. The Brascamp–Lieb inequality has also seen a surge of interest from theoretical computer scientists [Garg et al., 2017], ranging from robust subspace recovery [Hardt and Moitra, 2013], to analyzing determinants of submatrices [Nikolov and Singh, 2016].

Of course, with such a high amount of interest, it is tempting to ask various structural questions, such as decomposability, extremizability, and finiteness. Bennett et al. [2008] addresses them to great detail. In particular, they develop a precise notion of criticality that essentially allows decomposing any BL inequality into inequalities on smaller spaces. They further establish the idea of geometricity; in particular, when the linear maps behave like projections and abide by a frame condition, one can get an orthogonal decomposition of the underlying space into critical subspaces. Finally, a major result is that for Brascamp–Lieb inequalities, geometricity, Gaussian-extremizability and extremizability are all equivalent modulo an equivalence that amounts to a linear change of variables. We refer the reader to Carbery [2007] for a gentle summary of the main results in Bennett et al. [2008]. One question they left unanswered was the structure of all extremizers of Brascamp–Lieb inequalities, which was answered by Valdimarsson [2008].

We would be remiss to not mention the diversity of proof techniques used in the literature. The original paper of Brascamp and Lieb [1976] used rearrangement inequalities, and Lieb [1990] used a doubling argument show Gaussian-extremizability. Barthe [1998] used an optimal transport argument, and Bennett et al. [2008] used a heat flow argument. In discussing equality conditions, Valdimarsson [2008] analyzed the stationary behavior of an appropriate heat flow where the analysis further relied on properties of Fourier transforms

of tempered distributions. Later on, Lehec [2013] used Föllmer drifts (Appendix A) to give a concise proof in the geometric setting. All of these techniques are fundamentally reliant on various characterizations of Gaussians.

Barthe's Reverse Brascamp–Lieb Inequalities

Given that the Brascamp–Lieb inequality has seen so much attention, it should come as no surprise that their reverse versions, namely Barthe's Reverse Brascamp–Lieb Inequalities are just as influential. We start out with their definition:

Definition 14. *Let $f_i \in L^1(E_i), 1 \leq i \leq k$, be non-negative functions, $B_i : E_i \to E^1, 1 \leq i \leq k$, be linear maps, and $h : E^1 \to \mathbb{R}$ be a measurable function that satisfies*

$$\prod_{i=1}^{k} f_i^{c_i}(x_i) \leq h(\sum_{i=1}^{k} c_i B_i x_i) \qquad \forall x_i \in E_i, i = 1, \ldots, k.$$

Define $C_{RBL}(\mathbf{d}, \mathbf{B})$ as the smallest constant such that the following inequality is true for all such $(f_i)_{i=1}^k, h$:

$$\prod_{i=1}^{k} \left(\int_{E_i} f_i \right)^{c_i} \leq e^{C_{RBL}(\mathbf{d},\mathbf{B})} \int_E h. \tag{3.8}$$

Inequalities of the form (3.8) are referred to as Barthe's (Reverse Brascamp–Lieb) inequalities.

Remark 15. *One can pick $h(y) := \sup_{y=\sum c_i B_i x_i} \prod_{i=1}^k f_i^{c_i}(x_i)$ whenever h is measurable. The theorem was originally stated using outer integrals [Barthe, 1998].*

Barthe [1998] showed that the smallest constant $C_{RBL}(\mathbf{d}, \mathbf{B})$ in the inequality above can be computed by considering Gaussian functions. While doing so, he gave an optimal transport proof that proved both Brascamp–Lieb and the reverse form (3.8). Note that it is easy to see that (3.8) implies the Prékopa–Leindler inequality.

Furthermore, Barthe's inequality has been used in convex geometry by Ball [1989]. Similar to Brascamp–Lieb, the structure and the form of extremizers has been studied by Boroczky et al. [2022].

3.3 Entropic duality and Forward-Reverse Brascamp–Lieb inequalities

We start this section by noting that Lieb [1978] outlined a connection between the EPI and the sharp Young's inequality, through the realization of Shannon entropy as the limit of certain L^p norms. This idea was placed into information theory context thanks to Dembo

et al. [1991]. Later on, Carlen et al. [2004] went essentially the other way, deducing a sharp Young inequality from certain sharp entropy inequalities. This connection between functional and entropic inequalities was demystified by Carlen and Cordero-Erausquin [2009] when they highlighted the Legendre duality principle between the Brascamp–Lieb inequalities and their entropic dual inequalities:

Theorem 16 (Entropic Brascamp–Lieb Inequalities)**.** *For all* $X \in \mathcal{P}(E_0)$

$$h(X) \le \sum_{j=1}^{m} d_j h(B_j X) + C_{BL}(\mathbf{d}, \mathbf{B}),$$

with $C_{BL}(\mathbf{d}, \mathbf{B})$ *defined as in* (3.7)

Afterwards, Lehec [2013] has provided simultaneous proofs of functional and entropic inequalities through Föllmer drifts, and this duality between functional and entropic inequalities was further extended in Liu et al. [2016] to give the proof of a general inequality that implies the other inequalities in this chapter.

Forward-Reverse Brascamp–Lieb Inequalities

Liu et al. [2016] established the Forward-Reverse Brascamp–Lieb inequalities. Later on Courtade and Liu [2021] established the necessary conditions for finiteness, and gave the full structural analysis. We give the main result of Courtade and Liu [2021] below:

The entropic forward-reverse Brascamp–Lieb inequality

Define

$$D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) := \sup_{Z_i \in \mathcal{G}(E_i), 1 \le i \le k} \left(\sum_{i=1}^{k} c_i h(Z_i) - \max_{Z \in \Pi(Z_1, \ldots, Z_k)} \sum_{j=1}^{m} d_j h(B_j Z) \right),$$

and

$$D(\mathbf{c}, \mathbf{d}, \mathbf{B}) := \sup_{X_i \in \mathcal{P}(E_i), 1 \le i \le k} \left(\sum_{i=1}^{k} c_i h(X_i) - \max_{X \in \Pi(X_1, \ldots, X_k)} \sum_{j=1}^{m} d_j h(B_j X) \right).$$

Theorem 17. *Fix a datum* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$*. For random vectors* $X_i \in \mathcal{P}(E_i)$*,* $1 \le i \le k$*, we have*

$$\sum_{i=1}^{k} c_i h(X_i) \le \max_{X \in \Pi(X_1, \ldots, X_k)} \sum_{j=1}^{m} d_j h(B_j X) + D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}). \tag{3.9}$$

Moreover, the constant $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is finite if and only if the following two conditions hold.*

(i) **Scaling condition:** *It holds that*

$$\sum_{i=1}^{k} c_i \dim(E_i) = \sum_{j=1}^{m} d_j \dim(E^j). \tag{3.10}$$

(ii) **Dimension condition:** *For all subspaces $T_i \subset E_i$, $1 \le i \le k$,*

$$\sum_{i=1}^{k} c_i \dim(T_i) \le \sum_{j=1}^{m} d_j \dim(B_j T), \qquad \text{where } T = \oplus_{i=1}^{k} T_i. \tag{3.11}$$

Remark 18. *Note that if $X_i \in \mathcal{G}(E_i)$, then the maximum coupling in (3.9) is a joint Gaussian, as joint Gaussians maximize entropy subject to second moment constraints. (In particular, for any $X \in \Pi(X_1, \dots, X_k)$, we can take a Gaussian coupling $X_g \in \Pi(X_1, \dots, X_k)$ with $\mathrm{cov}(X_g) = \mathrm{cov}(X)$, which ensures that $\mathrm{cov}(B_j X_g) = \mathrm{cov}(B_j X)$ for all $j = 1, \dots, m$.)*

By a suitable choice of datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$, and with appropriate regularization, this implies all the entropic inequalities we have seen in this chapter so far. Of note, we recall from Liu et al. [2018], Courtade and Liu [2021] that Theorem 17 has the following equivalent (dual) functional form.

Theorem 19. *Fix a datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$. If measurable functions $f_i : E_i \to \mathbb{R}^+$, $1 \le i \le k$ and $g_j : E^j \to \mathbb{R}^+$, $1 \le j \le m$ satisfy*

$$\prod_{i=1}^{k} f_i^{c_i}(\pi_{E_i}(x)) \le \prod_{j=1}^{m} g_j^{d_j}(B_j x) \qquad \forall x \in E_0, \tag{3.12}$$

then

$$\prod_{i=1}^{k} \left(\int_{E_i} f_i(x_i)\, dx_i \right)^{c_i} \le e^{D_g(\mathbf{c},\mathbf{d},\mathbf{B})} \prod_{j=1}^{m} \left(\int_{E^j} g_j(y_j)\, dy_j \right)^{d_j}. \tag{3.13}$$

Moreover, the constant $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is best possible.

By changing $(\mathbf{c}, \mathbf{d}, \mathbf{B})$, this implies many geometric inequalities such as the Brascamp–Lieb inequalities [Brascamp et al., 1974, Brascamp and Lieb, 1976, Lieb, 1990], and the Barthe inequalities discussed before. Even beyond them, it includes the sharp reverse Young inequality [Brascamp and Lieb, 1976], the Chen–Dafnis–Paouris inequalities [Chen et al., 2015], and a form of the Barthe–Wolff inequalities [Barthe and Wolff, 2018]. Readers are referred to Courtade and Liu [2021] for a more detailed account of these implications and further references. An important result we want to discuss is the general entropic duality principle that holds for Forward-Reverse Brascamp Inequalities. Namely, Theorems 17 and

19 are formally equivalent via Fenchel–Rockafellar duality [Courtade and Liu, 2021]. Theorem 16 is one such example, and as stated before, Theorem 10 (i) is the entropic dual of the Prékopa-Leindler inequality. Here, we want to note that Theorem 19 follows from Theorem 17 by weak duality. The other direction is more difficult, relying on strong duality for convex functions. Hence, the entropic verison (Theorem 17) can be regarded as a formally stronger result.

Remark 20. *It has been brought to our attention by Courtade [2023] that the Gaussian saturation property of (3.4) follows from the Gaussian saturation property of the Brascamp–Lieb inequalities. To give a highlight of the argument, let $X_i \in \mathcal{P}(E_i)$ for $i = 1 \leq i \leq k$ and consider the following instances of the Brascamp–Lieb inequalities for any $X' \in \Pi(X_1, \dots, X_k)$ indexed by $\lambda \geq 0$:*

$$(1+\lambda)h(X') \leq \sum_{j=1}^{m} d_j h(B_j X') + \lambda \sum_{i=1}^{k} h(X_i) + C_\lambda. \tag{3.14}$$

Note that $C_\lambda \geq C_{AJN}(\mathbf{c}, \mathbf{d}, \mathbf{B})$ for any $\lambda \geq 0$ where C is the upper bound given in (3.3), that is

$$C_\lambda \geq C_{AJN}(\mathbf{c}, \mathbf{d}, \mathbf{B}) := \sup\{\sum_{i=1}^{k} h(X_i) - \sum_{j=1}^{m} d_j h(B_j X)\},$$

where the supremum is taken over all $X_i \in \mathcal{P}(E_i)$ for $i = 1 \leq i \leq k$, and X is their independent coupling. One can further show $\lim_{\lambda\to\infty} C_\lambda = C_{AJN}(\mathbf{c}, \mathbf{d}, \mathbf{B})$ using the Gaussian saturation property for finite λ [Lieb, 1990]. In particular, this shows the Gaussian saturation property of AJN data ($\mathbf{c}$,$\mathbf{d}$,$\mathbf{B}$), when the $\mathbf{c}$ is the all ones vector. One can get to the general AJN data by normalizing to ensure $c_i \leq 1$, and adding $(1 - c_i)h(X_i)$ for all $1 \leq i \leq k$. Furthermore, the finiteness conditions on the Brascamp–Lieb inequalities from Bennett et al. [2008] go through the same argument, recovering (3.5) and (3.6). Note that the characterization of extremizers do not pass through this limiting argument.

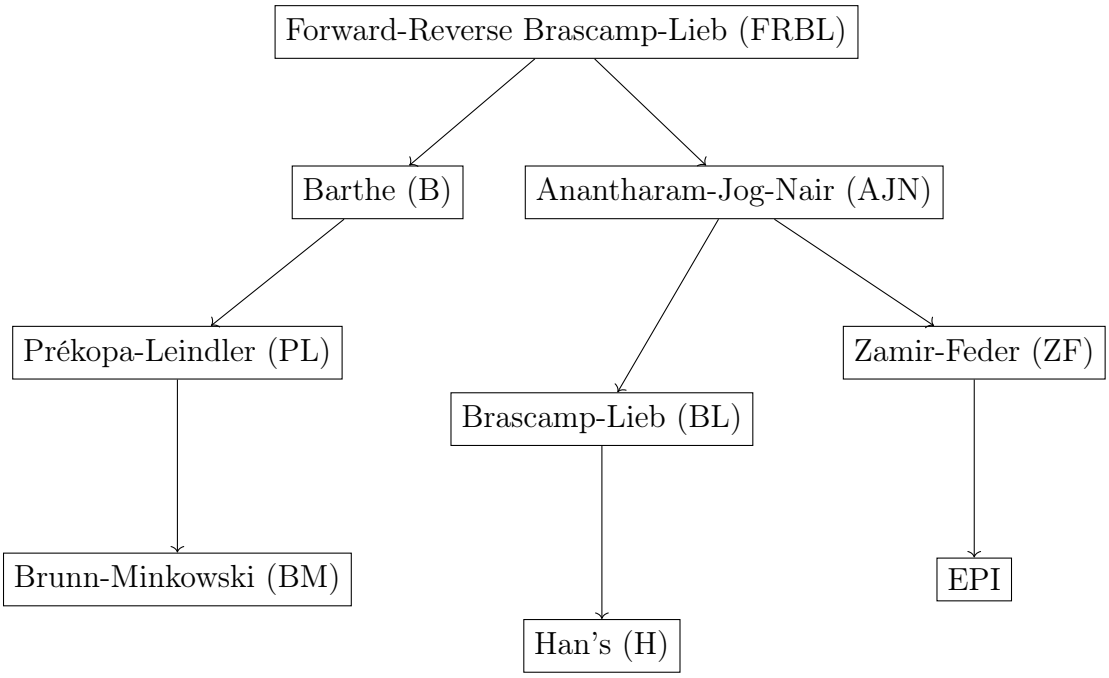


Figure 3.1: A depiction of landscape of a few major entropic inequalities. A solid arrow from A to B means that A implies B.

Chapter 4

Gaussian Comparisons

In this chapter, we establish a general class of entropy inequalities that take the concise form of Gaussian comparisons. The main result unifies many classical and recent results discussed in Chapter 3, including the Shannon–Stam inequality, the Brunn–Minkowski inequality, the Zamir–Feder inequality, the Brascamp–Lieb and Barthe inequalities, the Anantharam–Jog–Nair inequality, and others.

For $X \in \Pi(X_1, \dots, X_k)$ and $S \subset \{1, \dots, k\}$, we define the S-correlation[1]

$$I_S(X) := \sum_{i \in S} h(X_i) - h(\pi_S(X)),$$

where we let π_S denote the canonical projection of E_0 onto $\oplus_{i \in S} E_i$. To avoid ambiguity, we adopt the convention that $I_\emptyset(X) = 0$. Observe that that I_S is the relative entropy between the law of $\pi_S(X)$ and the product of its marginals, so is always nonnegative. Moreover, $I_S(X) = 0$ iff $(X_i)_{i \in S}$ are independent.

For a given constraint function $\nu : 2^{\{1,\dots,k\}} \to [0, +\infty]$, and $X_i \in \mathcal{P}(E_i)$, $1 \le i \le k$, we can now define the set of **correlation-constrained couplings**

$$\Pi(X_1, \dots, X_k; \nu) := \left\{ X \in \Pi(X_1, \dots, X_k) : I_S(X) \le \nu(S) \text{ for each } S \in 2^{\{1,\dots,k\}} \right\}.$$

With notation established, our main result is the following.

Theorem 21. *Fix* $(\mathbf{d}, \mathbf{B})$ *and* $\nu : 2^{\{1,\dots,k\}} \to [0, +\infty]$. *For any* $X_i \in \mathcal{P}(E_i)$, $1 \le i \le k$, *there exist* $Z_i \in \mathcal{G}(E_i)$ *with* $h(Z_i) = h(X_i)$, $1 \le i \le k$ *satisfying*

$$\max_{X \in \Pi(X_1,\dots,X_k;\nu)} \sum_{j=1}^m d_j h(B_j X) \ge \max_{Z \in \Pi(Z_1,\dots,Z_k;\nu)} \sum_{j=1}^m d_j h(B_j Z). \tag{4.1}$$

[1] The S-correlation I_S seems to have no generally agreed-upon name, and has been called different things in the literature. Our choice of terminology reflects that of Watanabe [1960], who used the term *total correlation* to describe I_S when $S = \{1, \dots, k\}$. However, it might also be called S-information, to reflect the "multi-information" terminology preferred by some (see, e.g., Csiszár and Körner [2011]).

Remark 22. *The special case where* $\dim(E_i) = 1$ *for all* $1 \le i \le k$ *appeared in the earlier work by Aras and Courtade [2021].*

Remark 23. *Note that certain* ν *can have redundant constraints due to submodularity of entropy.*

Observe that when $m = 1$, $\nu \equiv 0$ and $\dim(E_i) = 1$ for all $1 \le i \le k$, we recover the Zamir–Feder inequality (3.2). Indeed, taking $\nu \equiv 0$ renders the set of couplings equal to the singleton consisting of the independent coupling, and the one-dimensional nature of the E_i's means that the variances of the Z_i's are fully determined by the entropy constraints. Hence, it is clear that Theorem 21 generalizes the Zamir–Feder inequality (3.2).

As a second and slightly more substantial example, we explain the connection between the EPI and the Brunn–Minkowski inequality alluded to in the introduction. Denote the entropy power of $X \in \mathcal{P}(\mathbb{R}^n)$ by

$$N(X) := \frac{1}{2\pi e} e^{2h(X)/n}.$$

For a coupling $X = (X_1, X_2)$, note that the mutual information $I(X_1; X_2)$ is equal to $I_S(X)$ with $S = \{1, 2\}$.

Corollary 24. *For any* $X_1, X_2 \in \mathcal{P}(\mathbb{R}^n)$ *and* $\zeta \in [0, +\infty]$*, it holds that*

$$N(X_1) + N(X_2) + 2\sqrt{(1 - e^{-2\zeta/n}) N(X_1) N(X_2)} \le \max_{\substack{\Pi(X_1, X_2): \\ I(X_1; X_2) \le \zeta}} N(X_1 + X_2), \tag{4.2}$$

where the maximum is over couplings of X_1, X_2 *such that* $I(X_1; X_2) \le \zeta$*. Equality holds for Gaussian* X_1, X_2 *with proportional covariances.*

Proof. We apply Theorem 21 with $E_1 = E_2 = \mathbb{R}^n$ and $\nu(\{1, 2\}) = \zeta$ to give existence of Gaussian Z_1, Z_2 satisfying $N(Z_i) = N(X_i)$ and

$$\max_{\substack{(X_1, X_2) \in \Pi(X_1, X_2): \\ I(X_1; X_2) \le \zeta}} N(X_1 + X_2) \ge \max_{\substack{(Z_1, Z_2) \in \Pi(Z_1, Z_2): \\ I(Z_1; Z_2) \le \zeta}} N(Z_1 + Z_2).$$

Now, suppose $Z_i \sim N(0, \Sigma_i)$, $i \in \{1, 2\}$ and consider the coupling

$$Z_1 = \rho \Sigma_1^{1/2} \Sigma_2^{-1/2} Z_2 + (1 - \rho^2)^{1/2} W,$$

where $W \sim N(0, \Sigma_1)$ is independent of Z_2, and $\rho := (1 - e^{-2\zeta/n})^{1/2}$. This ensures $I(Z_1; Z_2) = \zeta$, and

$$\begin{aligned} N(Z_1 + Z_2) &= \det(\Sigma_1 + \Sigma_2 + \rho \Sigma_1^{1/2} \Sigma_2^{1/2} + \rho \Sigma_2^{1/2} \Sigma_1^{1/2})^{1/n} \\ &\ge \left(\det(\Sigma_1)^{1/n} + \det(\Sigma_2)^{1/n} + 2\rho \det(\Sigma_1^{1/2})^{1/n} \det(\Sigma_2^{1/2})^{1/n} \right) \\ &= N(X_1) + N(X_2) + 2\sqrt{(1 - e^{-2\zeta/n}) N(X_1) N(X_2)}, \end{aligned}$$

where the inequality follows by Minkowski's determinant inequality. It is easy to see that we have equality in the expressions above if X_1, X_2 are Gaussian with proportional covariances. □

Remark 25. *Theorem 24 may be considered as an extension of the EPI that holds for certain dependent random variables; it appeared in the preliminary work [Aras and Courtade, 2021] by the authors. We remark that Takano et al. [1995] and Johnson [2004] have established that the EPI holds for dependent random variables which have positively correlated scores. However, given the different hypotheses, those results are not directly comparable to Theorem 24.*

Now, we observe that the EPI and the Brunn–Minkowski inequality naturally emerge from (4.2) by considering the endpoints of independence ($\zeta = 0$) and maximal dependence ($\zeta = +\infty$). Of course, (4.2) also gives a sharp inequality for the whole spectrum of cases in between.

Example 26 (EPI)**.** *Taking $\zeta = 0$ enforces the independent coupling in Theorem 24, and recovers the EPI (3.1), which we reproduce here for visual convenience. For independent $X_1, X_2 \in \mathcal{P}(\mathbb{R}^n)$,*

$$e^{2h(X_1)/n} + e^{2h(X_2)/n} \leq e^{2h(X_1+X_2)/n}. \tag{4.3}$$

Hence, Theorem 24 may be regarded as an extension of the EPI for certain dependent random variables with a sharp correction term.

Example 27 (Brunn–Minkowski inequality)**.** *Taking $\zeta = +\infty$ in Theorem 24 allows for unconstrained optimization over couplings, giving*

$$e^{h(X_1)/n} + e^{h(X_2)/n} \leq \sup_{\Pi(X_1,X_2)} e^{h(X_1+X_2)/n},$$

where we emphasize the change in exponent from 2 to 1, relative to (4.3). This may be regarded as an entropic improvement of the Brunn–Minkowski inequality. Indeed, if X_1, X_2 are uniform on compact subsets $K, L \subset \mathbb{R}^n$ that have non-empty interiors respectively, we obtain the familiar Brunn–Minkowski inequality

$$\mathrm{Vol}_n(K)^{1/n} + \mathrm{Vol}_n(L)^{1/n} \leq \sup_{\Pi(X_1,X_2)} e^{h(X_1+X_2)/n} \leq \mathrm{Vol}_n(K+L)^{1/n},$$

where $K + L$ denotes the Minkowski sum of K and L, and $\mathrm{Vol}_n(\cdot)$ denotes the n-dimensional Lebesgue volume. The last inequality follows since $X_1 + X_2$ is supported on the Minkowski sum $K + L$, and hence the entropy is upper bounded by that of the uniform distribution on that set.

Remark 28. *The above form of Brunn–Minkowski is equivalent to the one given in Theorem 11. However, note that the equality conditions are slightly different, with this version only requiring K and L be homothetic rather than identical up to shifts.*

In the above, the Brunn–Minkowski inequality and the EPI are obtained as logical endpoints of a family of inequalities which involve only Shannon entropies instead of Rényi entropies of varying orders. In contrast to derivations involving Rényi entropies where summands are always independent (corresponding to the convolution of densities), the idea here is to allow dependence between the random summands.

We remark that equality cases for (4.1) in the special case where $\nu \equiv 0$ follow from the main results in Chapter 5, and the general case follows from the results in Chapter 6.

4.1 Proof of the main result

This section is dedicated to the proof of Theorem 21. There are several preparations to make before starting the proof; this is done in the first subsection. The second subsection brings everything together to prove an unconstrained version of Theorem 21 where $\nu \equiv +\infty$. The third and final subsection proves Theorem 21 on the basis of its unconstrained variation.

Preliminaries

Here we quote the preparatory results that we shall need, and the definitions required to state them. The various results are organized by subsection, and proofs are only given where necessary.

Some additional notation

A datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is said to be **extremizable** if $D(\mathbf{c}, \mathbf{d}, \mathbf{B}) < \infty$ and there exist $X_i \in \mathcal{P}(E_i)$, $1 \leq i \leq k$ which attain equality in (3.9). Likewise, a datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is said to be **Gaussian-extremizable** if there exist Gaussian $X_i \in \mathcal{G}(E_i)$, $1 \leq i \leq k$ which attain equality in (3.9). Necessary and sufficient conditions for Gaussian-extremizability of a datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ can be found in Courtade and Liu [2021]. Clearly Gaussian-extremizability implies extremizability on account of Theorem 17. We shall need the converse, which was not proved in Courtade and Liu [2021].

Theorem 29. *If a datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is extremizable, then it is Gaussian-extremizable.*

The proof follows a doubling argument similar to what appears in Liu et al. [2018, Proof of Theorem 8]. We will need the following Lemma.

Lemma 30. *For each $1 \leq i \leq k$, let $Z_i \sim N(0, K_i)$ and let $(X_{n,i})_{n \geq 1}$ be a sequence of zero-mean random vectors satisfying*

$$\lim_{n \to \infty} W_2(X_{n,i}, Z_i) = 0,$$

where $W_2 : \mathcal{P}(E_i) \times \mathcal{P}(E_i) \to \mathbb{R}$ is the Wasserstein distance of order 2. For any $K \in \Pi(K_1, \ldots, K_k)$, there exists a sequence of couplings $X_n \in \Pi(X_{n,1}, \ldots, X_{n,k})$, $n \geq 1$ such that $\|\mathrm{Cov}(X_n) - K\|_{\mathrm{HS}} \to 0$.

Proof. Let $Z \sim N(0, K)$, and observe that $Z \in \Pi(Z_1, \dots, Z_k)$. Let $T_{n,i}$ be the optimal transport map sending $N(0, K_i)$ to $\mathrm{law}(X_{n,i})$ (see, e.g., Villani [2003]). Then $X_n = (T_{n,1}(Z_1), \dots, T_{n,k}(Z_k)) \in \Pi(X_{n,1}, \dots, X_{n,k})$ satisfies

$$\begin{aligned} T_{n,i}(Z_i)T_{n,i'}(Z_{i'})^T - Z_i Z_{i'}^T &= Z_i(T_{n,i'}(Z_{i'}) - Z_{i'})^T + (T_{n,i}(Z_i) - Z_i)Z_{i'}^T \\ &\quad + (T_{n,i}(Z_i) - Z_i)(T_{n,i'}(Z_{i'}) - Z_{i'})^T. \end{aligned}$$

Taking expectations of both sides and applying Cauchy–Schwarz, we conclude

$$\|\mathrm{Cov}(X_n) - K\|_{\mathrm{HS}} \to 0$$

since $\mathbb{E}|T_{n,i}(Z_i) - Z_i|^2 = W_2(X_{n,i}, Z_i)^2 \to 0$ for each $1 \le i \le k$. □

Proof of Theorem 29. The approach will be to show that extremizers are closed under convolutions, and apply the entropic central limit theorem. Toward this end, let $X_i \sim \mu_i \in \mathcal{P}(E_i)$ be independent of $Y_i \sim \nu_i \in \mathcal{P}(E_i)$, $1 \le i \le k$, both assumed to be extremal in (3.9). Define

$$Z_i^+ := X_i + Y_i, \quad Z_i^- := X_i - Y_i, \quad 1 \le i \le k,$$

and let

$$Z^+ \in \arg \max_{Z \in \Pi(Z_1^+, \dots, Z_k^+)} \sum_{j=1}^m d_j h(B_j Z).$$

Let $Z_i^-|z_i^+$ denote the random variable Z_i^- conditioned on $\{Z_i^+ = z_i^+\}$, which has law in $\mathcal{P}(E_i)$ for $\mathrm{law}(Z_i^+)$-a.e. $z_i^+ \in E_i$ by disintegration. Note that the a.s. finiteness of entropy follows from the chain rule $h(Z_i^+, Z_i^-) = h(Z_i^+) + h(Z_i^- \mid Z_i^+)$ and independence of X_i, Y_i which ensures that $h(Z_i^+, Z_i^-)$ is finite, and finiteness of second moments a.s. follows by iterated expectation. Next, for $z^+ = (z_1^+, \dots, z_k^+) \in E_0$, let

$$Z^-|z^+ \in \arg \max_{Z \in \Pi(Z_1^-|z_1^+, \dots, Z_k^-|z_k^+)} \sum_{j=1}^m d_j h(B_j Z).$$

We can assume these couplings are such that $z^+ \mapsto \mathrm{law}(Z^-|z^+)$ is Borel measurable (i.e., $\mathrm{law}(Z^-|z^+)$ is a regular conditional probability). This can be justified by measurable selection theorems, as in Villani et al. [2008, Cor. 5.22] and Liu et al. [2018, p. 42]. With this assumption, definitions imply

$$\begin{aligned} \sum_{i=1}^k c_i h(Z_i^+) &\le \sum_{j=1}^m d_j h(B_j Z^+) + D(\mathbf{c}, \mathbf{d}, \mathbf{B}) \\ \sum_{i=1}^k c_i h(Z_i^-|z_i^+) &\le \sum_{j=1}^m d_j h(B_j Z^-|z^+) + D(\mathbf{c}, \mathbf{d}, \mathbf{B}), \end{aligned}$$

where the latter holds for law(Z^+)-a.e. z^+. Integrating the second inequality against the distribution of Z^+ gives the inequality for conditional entropies:

$$\begin{aligned}\sum_{i=1}^{k} c_i h(Z_i^-|Z_i^+) &\le \sum_{j=1}^{m} d_j h(B_j Z^-|Z^+) + D(\mathbf{c},\mathbf{d},\mathbf{B}) \\ &\le \sum_{j=1}^{m} d_j h(B_j Z^-|B_j Z^+) + D(\mathbf{c},\mathbf{d},\mathbf{B}),\end{aligned}$$

where the second inequality follows since conditioning reduces entropy. Now, define

$$X = \frac{1}{2}\left(Z^+ + (Z^-|Z^+)\right), \quad Y = \frac{1}{2}\left(Z^+ - (Z^-|Z^+)\right).$$

Observe that $X \in \Pi(X_1,\dots,X_k)$ and $Y \in \Pi(Y_1,\dots,Y_k)$. So, using the above inequalities and definitions, we have

$$\begin{aligned}2D(\mathbf{c},\mathbf{d},\mathbf{B}) &\le \sum_{i=1}^{k} c_i h(X_i,Y_i) - \sum_{j=1}^{m} d_j h(B_j X) - \sum_{j=1}^{m} d_j h(B_j Y) \\ &\le \sum_{i=1}^{k} c_i h(X_i,Y_i) - \sum_{j=1}^{m} d_j h(B_j X, B_j Y) \\ &= \sum_{i=1}^{k} c_i h(Z_i^+) + \sum_{i=1}^{k} c_i h(Z_i^-|Z_i^+) \\ &\quad - \sum_{j=1}^{m} d_j h(B_j Z^+) - \sum_{j=1}^{m} d_j h(B_j Z^-|B_j Z^+) \\ &\le 2D(\mathbf{c},\mathbf{d},\mathbf{B})\end{aligned}$$

Thus, we conclude

$$\sum_{i=1}^{k} c_i h(Z_i^+) = \sum_{j=1}^{m} d_j h(B_j Z^+) + D(\mathbf{c},\mathbf{d},\mathbf{B}),$$

showing that $Z_i^+ \sim \mu_i * \nu_i \in \mathcal{P}(E_i)$, $1 \le i \le k$ are extremal in (3.9) as desired. The scaling condition (3.5) is necessary for $D(\mathbf{c},\mathbf{d},\mathbf{B}) < \infty$, so it follows by induction and scale invariance that, for every $n \ge 1$, marginally specified $(Z_{n,i})_{i=1}^k$ are extremal in (3.9), where

$$Z_{n,i} := \frac{1}{\sqrt{n}} \sum_{\ell=1}^{n} (X_{\ell,i} - \mathbb{E}[X_i]),$$

and $(X_{\ell,i})_{\ell \ge 1}$ are i.i.d. copies of X_i.

Define $K_i = \mathrm{Cov}(X_i)$ (which is positive definite since $h(X_i)$ is finite), and fix any $K \in \Pi(K_1, \ldots, K_k)$. For any $\epsilon > 0$, Lemma 30 together with the central limit theorem for W_2 implies there exists $N \geq 1$ and a coupling $Z_N \in \Pi(Z_{N,1}, \ldots, Z_{N,k})$ such that $\|\mathrm{Cov}(Z_N) - K\|_{\mathrm{HS}} < \epsilon$. Letting $Z_N^{(n)}$ denote the standardized sum of n i.i.d. copies of Z_N, we have $Z_N^{(n)} \in \Pi(Z_{nN,1}, \ldots, Z_{nN,k})$ for each $n \geq 1$. Thus, by the entropic central limit theorem [Barron, 1986, Carlen and Soffer, 1991], we have

$$\limsup_{n\to\infty} \max_{Z_n \in \Pi(Z_{n,1},\ldots,Z_{n,k})} \sum_{j=1}^m d_j h(B_j Z_n) \geq \lim_{n\to\infty} \sum_{j=1}^m d_j h(B_j Z_N^{(n)}) = \sum_{j=1}^m d_j h(B_j Z_N^*)$$

where $Z_N^* \sim N(0, \mathrm{Cov}(Z_N))$. Our arbitrary choice of K and ϵ together with continuity of determinants implies

$$\limsup_{n\to\infty} \max_{Z_n \in \Pi(Z_{n,1},\ldots,Z_{n,k})} \sum_{j=1}^m d_j h(B_j Z_n) \geq \max_{K \in \Pi(K_1,\ldots,K_k)} \sum_{j=1}^m \frac{d_j}{2} \log\left((2\pi e)^{\dim(E^j)} \det(B_j K B_j^T)\right).$$

Invoking the entropic central limit theorem, and using the fact that $(Z_{n,i})_{i=1}^k$ are extremal in (3.9) for each $n \geq 1$, we conclude

$$\begin{aligned}
&\sum_{i=1}^k \frac{c_i}{2} \log\left((2\pi e)^{\dim(E_i)} \det(K_i)\right) \\
&= \lim_{n\to\infty} \sum_{i=1}^k c_i h(Z_{n,i}) \\
&= \lim_{n\to\infty} \max_{Z_n \in \Pi(Z_{n,1},\ldots,Z_{n,k})} \sum_{j=1}^m d_j h(B_j Z_n) + D(\mathbf{c}, \mathbf{d}, \mathbf{B}) \\
&\geq \max_{K \in \Pi(K_1,\ldots,K_k)} \sum_{j=1}^m \frac{d_j}{2} \log\left((2\pi e)^{\dim(E^j)} \det(B_j K B_j^T)\right) + D(\mathbf{c}, \mathbf{d}, \mathbf{B}).
\end{aligned}$$

Thus, by definitions, we have equality throughout, and $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is Gaussian-extremizable. □

Properties of the max-entropy term

Let us briefly make a few technical observations related to the max-entropy quantity that appears in (3.9). First, we quote a technical lemma that will be needed several times. A proof can be found in Liu et al. [2018, Lemma A2].

Lemma 31. *Let $(\mu_n)_{n\geq 1} \subset \mathcal{P}(E)$ converge in distribution to μ. If $\sup_{n\geq 1} \int_E |x|^2 d\mu_n < \infty$, then*

$$\limsup_{n\to\infty} h(\mu_n) \leq h(\mu).$$

Now, we point out that the max-entropy term is well-defined as a maximum.

Proposition 32. *Fix* $(\mathbf{d}, \mathbf{B})$ *and* $X_i \in \mathcal{P}(E_i)$, $1 \le i \le k$. *The function*

$$X \in \Pi(X_1, \dots, X_k) \longmapsto \sum_{j=1}^m d_j h(B_j X)$$

achieves a maximum at some $X^* \in \Pi(X_1, \dots, X_k)$. *Moreover, if each* X_i *is Gaussian, then* X^* *is Gaussian.*

Proof. We have $\sup_{X \in \Pi(X_1,\dots,X_k)} \mathbb{E}|B_j X|^2 < \infty$ for each $1 \le j \le m$ since each X_i has bounded second moments. The second moment constraint also implies $\Pi(X_1, \dots, X_k)$ is tight, and it is easily checked to be closed in the weak topology. Thus, Prokhorov's theorem ensures $\Pi(X_1, \dots, X_k)$ is sequentially compact. So, if $(X^{(n)})_{n \ge 1} \subset \Pi(X_1, \dots, X_k)$ is such that

$$\lim_{n \to \infty} \sum_{j=1}^m d_j h(B_j X^{(n)}) = \sup_{X \in \Pi(X_1,\dots,X_k)} \sum_{j=1}^m d_j h(B_j X),$$

we can assume $X^{(n)} \to X^* \in \Pi(X_1, \dots, X_k)$ weakly, by passing to a subsequence if necessary. This implies $B_j X^{(n)} \to B_j X^*$ weakly for each $1 \le j \le m$. The first claim follows by an application of Lemma 31.

The second claim now follows from the first, together with the fact that Gaussians maximize entropy under a covariance constraint. □

Next, if $X_i \sim N(0, K_i)$ for $K_i \in \mathbf{S}^+(E_i)$, $1 \le i \le k$, then the entropy maximization in (3.9) is equivalent to the following optimization problem

$$(K_i)_{i=1}^k \mapsto \max_{K \in \Pi(K_1,\dots,K_k)} \sum_{j=1}^m d_j \log\det(B_j K B_j^T). \tag{4.4}$$

This maximization enjoys a certain strong duality property, which is a consequence of the Fenchel–Rockafellar theorem. The following can be found in Courtade and Liu [2021, Theorem 2.8].

Theorem 33. *Fix* $(\mathbf{d}, \mathbf{B})$. *For any* $K_i \in \mathbf{S}^+(E_i)$, $1 \le i \le k$, *it holds that*

$$\begin{aligned} &\max_{K \in \Pi(K_1,\dots,K_k)} \sum_{j=1}^m d_j \log\det\left(B_j K B_j^T\right) + \sum_{j=1}^m d_j \dim(E^j) \\ &= \inf_{(U_i, V_j)_{1 \le i \le k, 1 \le j \le m}} \left(\sum_{i=1}^k \langle U_i, K_i \rangle_{\mathrm{HS}} - \sum_{j=1}^m d_j \log\det V_j \right), \end{aligned} \tag{4.5}$$

where the infimum is over $U_i \in \mathbf{S}^+(E_i), 1 \le i \le k$ *and* $V_j \in \mathbf{S}^+(E^j), 1 \le j \le m$ *satisfying*

$$\sum_{j=1}^m d_j B_j^T V_j B_j \le \mathrm{diag}(U_1, \dots, U_k). \tag{4.6}$$

Corollary 34. *The function in (4.4) is continuous on* $\prod_{i=1}^k \mathbf{S}^+(E_i)$.

Proof. By (4.5), we see that the mapping in (4.4) is a pointwise infimum of functions that are affine in $(K_i)_{i=1}^k$, so it follows that it is upper semi-continuous on $\prod_{i=1}^k \mathbf{S}^+(E_i)$. On the other hand, each $K \in \Pi(K_1, \ldots, K_k)$ can be factored as $K = K_d^{1/2} \Sigma K_d^{1/2}$, for $K_d^{1/2} := \mathrm{diag}(K_1^{1/2}, \ldots, K_k^{1/2})$ and $\Sigma \in \Pi(\mathrm{id}_{E_1}, \ldots, \mathrm{id}_{E_k})$. Since the map $K_i \mapsto K_i^{1/2}$ is continuous on $\mathbf{S}^+(E_i)$, and determinants are also continuous, it follows that (4.4) is a pointwise supremum of continuous functions. As such, it is lower semi-continuous, completing the proof. □

4.2 Convexity properties of $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$

For $(\mathbf{d}, \mathbf{B})$ fixed, define the function $F : \mathbb{R}^k \times \prod_{i=1}^k \mathbf{S}^+(E_i) \to \mathbb{R} \cup \{-\infty\}$ via

$$F\left(\mathbf{c}, (K_i)_{i=1}^k\right) := \max_{K \in \Pi(K_1, \ldots, K_k)} \sum_{j=1}^m d_j \log\det(B_j K B_j^T) - \sum_{i=1}^k c_i \log\det(K_i).$$

The motivation for the above definition is that we have

$$-2D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = \inf_{(K_i)_{i=1}^k \in \prod_{i=1}^k \mathbf{S}^+(E_i)} F\left(\mathbf{c}, (K_i)_{i=1}^k\right) \tag{4.7}$$

by definition of $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ and the fact that the scaling condition (3.5) is a necessary condition for finiteness of $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$. The optimization problem above is not convex in the K_i's, however it is *geodesically-convex*. This property was mentioned to my advisor by Jingbo Liu in a discussion of the geodesically convex formulation of the Brascamp–Lieb constant. We do not know what argument he had in mind, but we'd like to credit the basic observation to him.

Let us first explain what is meant by geodesic convexity. Given a metric space (M, ρ) and points $x, y \in M$, a geodesic is a path $\gamma : [0,1] \to M$ with $\gamma(0) = x$, $\gamma(1) = y$ and

$$d_M\left(\gamma(t_1), \gamma(t_2)\right) = |t_1 - t_2| \rho(x, y), \quad \forall t_1, t_2 \in [0,1].$$

A function $f : M \to \mathbb{R}$ is geodesically-convex if, for any geodesic γ,

$$f(\gamma(t)) \le t f(\gamma(0)) + (1-t) f(\gamma(1)), \quad \forall t \in [0,1].$$

The space (M, ρ) is a unique geodesic metric space if every two points $x, y \in M$ are joined by a unique geodesic.

This is relevant to us as follows. For a Euclidean space E, the space $(\mathbf{S}^+(E), \delta_2)$ is a unique geodesic metric space, where for $A, B \in \mathbf{S}^+(E)$,

$$t \in [0,1] \mapsto A \#_t B := A^{1/2} (A^{-1/2} B A^{-1/2})^t A^{1/2}$$

is the unique geodesic joining A and B with respect to the metric

$$\delta_2(A,B) := \left(\sum_{i=1}^{\dim(E)} \log(\lambda_i(A^{-1}B))^2 \right)^{1/2}.$$

The matrix $A\#B := A\#_{1/2}B$ is referred to as the geometric mean of $A, B \in \mathbf{S}^+(E)$.

The topology on $\mathbf{S}^+(E)$ generated by δ_2 is the usual one, in the sense that $\delta_2(A_n, A) \to 0$ if and only if $\|A_n - A\|_{\mathrm{HS}} \to 0$. Hence, there are no subtleties with regards to the notions of continuity, etc. In particular, if $f : \mathbf{S}^+(E) \to \mathbb{R}$ is continuous and geodesically midpoint-convex, i.e.,

$$f(A\#B) \leq \frac{1}{2}f(A) + \frac{1}{2}f(B), \quad A, B \in \mathbf{S}^+(E),$$

then it is geodesically convex.

Theorem 35. *Fix* $(\mathbf{d}, \mathbf{B})$.

(i) The function $\mathbf{c} \mapsto D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is convex and lower semi-continuous.*

(ii) For fixed $\mathbf{c}$*, the function* $(K_i)_{i=1}^k \mapsto F\left(\mathbf{c}, (K_i)_{i=1}^k\right)$ *is geodesically-convex and continuous on* $\prod_{i=1}^k \mathbf{S}^+(E_i)$.

Remark 36. *As a subspace of* $\mathbf{S}^+(E_0)$, $\prod_{i=1}^k \mathbf{S}^+(E_i)$ *inherits its metric from* $\mathbf{S}^+(E_0)$.

Remark 37. *It may be the case that* $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = +\infty$ *for each* $\mathbf{c}$*, e.g., if some* B_j *fails to be surjective.*

Before the proof, we recall a few basic facts about the geometric mean $A\#B$. A linear transformation $\Phi : \mathbf{S}(E) \to \mathbf{S}(E')$ is said to be *positive* if it sends $\mathbf{S}^+(E)$ into $\mathbf{S}^+(E')$.

Proposition 38. *Let* E, E' *be Euclidean spaces. For* $A_1, A_2, B_1, B_2 \in \mathbf{S}^+(E)$*, the following hold.*

(i) (Monotone Property) If $A_1 \geq B_1$ *and* $A_2 \geq B_2$*, then* $(A_1\#A_2) \geq (B_2\#B_2)$.

(ii) (Cauchy–Schwarz) We have

$$\langle A_1, B_1\rangle_{\mathrm{HS}} + \langle A_2, B_2\rangle_{\mathrm{HS}} \geq 2\langle (A_1\#A_2), (B_1\#B_2)\rangle_{\mathrm{HS}}.$$

(iii) (Ando's inequality) If $\Phi : \mathbf{S}(E) \to \mathbf{S}(E')$ *is a positive linear map, then*

$$\Phi(A_1\#A_2) \leq \Phi(A_1)\#\Phi(A_2).$$

(iv) (Geodesic linearity of $\log\det$*) It holds that*

$$\log\det(A_1\#A_2) = \frac{1}{2}\log\det(A_1) + \frac{1}{2}\log\det(A_2).$$

Proof. The monotonicity property can be found, e.g., in Lawson and Lim [2001, p. 802]. By a change of variables using Lawson and Lim [2001, Lem. 3.1] and Ando [1979, Cor. 2.1(ii)], it suffices to prove (ii) under the assumption that $B_1 = \mathrm{id}_E$. In particular, Cauchy–Schwarz gives

$$\begin{aligned}|\langle (A_1 \# A_2), (\mathrm{id}_E \# B_2)\rangle_{\mathrm{HS}}|^2 &= |\langle (A_2^{-1/2} A_1 A_2^{-1/2})^{1/2} A_2^{1/2}, A_2^{1/2} B_2^{1/2}\rangle_{\mathrm{HS}}|^2 \\ &\leq \|(A_2^{-1/2} A_1 A_2^{-1/2})^{1/2} A_2^{1/2}\|_{\mathrm{HS}} \|A_2^{1/2} B_2^{1/2}\|_{\mathrm{HS}} \\ &= \langle A_1, \mathrm{id}_E\rangle_{\mathrm{HS}} \langle A_2, B_2\rangle_{\mathrm{HS}}.\end{aligned}$$

Thus, the claim follows by taking square roots of both sides and invoking the AM-GM inequality $\sqrt{ab} \leq (a+b)/2$ for $a, b \geq 0$. Ando's inequality can be found in Ando [1979, Thm. 3(i)]. Claim (iv) is trivial. □

Theorem 35 now follows as an easy consequence of the above properties and Theorem 33.

Proof of Theorem 35. Claim (i) follows immediately from (4.7), since $-D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is a pointwise infimum of functions that are affine in $\mathbf{c}$.

To prove (ii), we note that geodesic-linearity of log det implies it suffices to show geodesic midpoint-convexity of the continuous (by Corollary 34) function

$$(K_i)_{i=1}^k \mapsto \max_{K \in \Pi(K_1, \ldots, K_k)} \sum_{j=1}^m d_j \log\det(B_j K B_j^T). \tag{4.8}$$

Invoking Theorem 33, this is the same as establishing geodesic-convexity of

$$(K_i)_{i=1}^k \mapsto \inf_{(U_i, V_j)_{1 \leq i \leq k, 1 \leq j \leq m}} \left(\sum_{i=1}^k \langle U_i, K_i\rangle_{\mathrm{HS}} - \sum_{j=1}^m d_j \log\det V_j \right), \tag{4.9}$$

where the infimum is over $U_i \in \mathbf{S}^+(E_i), 1 \leq i \leq k$ and $V_j \in \mathbf{S}^+(E^j), 1 \leq j \leq m$ satisfying

$$\mathrm{diag}(U_1, \ldots, U_k) \geq \sum_{j=1}^m d_j B_j^T V_j B_j. \tag{4.10}$$

For $\ell \in \{1, 2\}$, let $U_i^{(\ell)} \in \mathbf{S}^+(E_i), 1 \leq i \leq k$ and $V_j^{(\ell)} \in \mathbf{S}^+(E^j), 1 \leq j \leq m$ satisfy (4.10) with strict inequality. As such, there exists $\epsilon > 0$ sufficiently small such that

$$\mathrm{diag}(U_1^{(\ell)}, \ldots, U_k^{(\ell)}) \geq \sum_{j=1}^m d_j B_j^T V_j^{(\ell)} B_j + \epsilon \sum_{j=1}^m \mathrm{Tr}(V_j^{(\ell)}) \, \mathrm{id}_{E_0}, \quad \ell \in \{1, 2\}.$$

Define the positive linear map $\Phi : \mathbf{S}^+(E^0) \to \mathbf{S}^+(E_0)$ via

$$\Phi(V) := \sum_{j=1}^m d_j B_j^T \pi_{E^j} V \pi_{E_j}^T B_j + \epsilon \, \mathrm{Tr}(V) \, \mathrm{id}_{E_0}, \quad V \in \mathbf{S}^+(E^0).$$

By the monotone property and Ando's inequality in Proposition 38,

$$\begin{aligned}\operatorname{diag}(U_1^{(1)}\#U_1^{(2)},\dots,U_k^{(1)}\#U_k^{(2)}) &\geq \Phi\left(\operatorname{diag}(V_1^{(1)},\dots,V_m^{(1)})\right)\#\Phi\left(\operatorname{diag}(V_1^{(2)},\dots,V_m^{(2)})\right)\\ &\geq \Phi\left(\operatorname{diag}(V_1^{(1)}\#V_1^{(2)},\dots,V_m^{(1)}\#V_m^{(2)})\right)\\ &\geq \sum_{j=1}^m d_j B_j^T(V_j^{(1)}\#V_j^{(2)})B_j.\end{aligned}$$

In particular, $(U_i^{(1)}\#U_i^{(2)})\in\mathbf{S}^+(E_i), 1\leq i\leq k$ and $(V_j^{(1)}\#V_j^{(2)})\in\mathbf{S}^+(E^j)$, $1\leq j\leq m$ satisfy (4.10). Therefore, let $(K_i^{(\ell)})_{i=1}^k\in\prod_{i=1}^k\mathbf{S}^+(E_i)$ and use Proposition 38 to write

$$\begin{aligned}&\frac{1}{2}\sum_{\ell\in\{1,2\}}\left(\sum_{i=1}^k\langle U_i^{(\ell)},K_i^{(\ell)}\rangle_{\mathrm{HS}}-\sum_{j=1}^m d_j\log\det V_j^{(\ell)}\right)\\ &\geq\sum_{i=1}^k\langle(U_i^{(1)}\#U_i^{(2)}),(K_i^{(1)}\#K_i^{(2)})\rangle_{\mathrm{HS}}-\sum_{j=1}^m d_j\log\det(V_j^{(1)}\#V_j^{(2)})\\ &\geq\inf_{(U_i,V_j)_{1\leq i\leq k,1\leq j\leq m}}\left(\sum_{i=1}^k\langle U_i,(K_i^{(1)}\#K_i^{(2)})\rangle_{\mathrm{HS}}-\sum_{j=1}^m d_j\log\det V_j\right).\end{aligned}$$

By continuity of the objective in (4.9) with respect to the U_i's, the value of the infimum in (4.9) remains unchanged if we take infimum over U_i's and V_j's satisfying (4.10) with strict inequality. Hence, by the arbitrary choice of $U_i^{(\ell)}\in\mathbf{S}^+(E_i), 1\leq i\leq k$ and $V_j^{(\ell)}\in\mathbf{S}^+(E^j), 1\leq j\leq m$ subject to (4.10) with strict inequality, geodesic midpoint-convexity of (4.9) is proved. □

Sion's theorem for geodesic metric spaces

We will need the following version of Sion's minimax theorem, found in Zhang et al. [2022].

Theorem 39 (Sion's theorem in geodesic metric spaces)**.** *Let (M,d_M) and (N,d_N) be geodesic metric spaces. Suppose $\mathcal{X}\subset M$ is a compact and geodesically convex set, $\mathcal{Y}\subset N$ is a geodesically convex set. If following conditions hold for $f:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$:*

1. *$f(\cdot,y)$ is geodesically-convex and l.s.c. for each $y\in\mathcal{Y}$;*
2. *$f(x,\cdot)$ is geodesically-concave and u.s.c. for each $x\in\mathcal{X}$,*

then

$$\min_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}f(x,y)=\sup_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}f(x,y).$$

Unconstrained comparisons

With all the pieces in place, we can take a big step toward proving Theorem 21 by first establishing the result in the unconstrained case. Namely, the goal of this section is to prove the following.

Theorem 40. *Fix* $(\mathbf{d}, \mathbf{B})$. *For any* $X_i \in \mathcal{P}(E_i)$, $1 \le i \le k$, *there exist* $Z_i \in \mathcal{G}(E_i)$ *with* $h(Z_i) = h(X_i)$ *for* $1 \le i \le k$ *such that*

$$\max_{X \in \Pi(X_1,\dots,X_k)} \sum_{j=1}^{m} d_j h(B_j X) \ge \max_{Z \in \Pi(Z_1,\dots,Z_k)} \sum_{j=1}^{m} d_j h(B_j Z). \tag{4.11}$$

Remark 41. *It is a part of the theorem that each maximum is attained.*

Before we start the proof, let's first describe the high-level idea. To do this, recall that Lieb's form of the EPI from Theorem 6 is as follows: For independent random vectors $X_1, X_2 \in \mathcal{P}(\mathbb{R})$ and any $\lambda \in (0,1)$,

$$h(\sqrt{\lambda} X_1 + \sqrt{1-\lambda} X_2) \ge \lambda h(X_1) + (1-\lambda) h(X_2). \tag{4.12}$$

Motivated by the similarity between the entropy power inequality and the Brunn–Minkowski inequality, Costa and Cover [1984] reformulated (4.12) as the following concise Gaussian comparison[2].

Proposition 42 (Comparison form of Shannon–Stam inequality)**.** *For independent random variables* $X_1, X_2 \in \mathcal{P}(\mathbb{R})$, *we have*

$$h(X_1 + X_2) \ge h(Z_1 + Z_2), \tag{4.13}$$

where Z_1, Z_2 *are independent Gaussian random variables with variances chosen so that* $h(Z_i) = h(X_i)$.

To understand how this comes about, observe that a change of variables in (4.12) yields the equivalent formulation

$$c h(X_1) + (1-c) h(X_2) + \frac{1}{2} h_2(c) \le h(X_1 + X_2), \quad \text{for all } c \in [0,1],$$

where $h_2(c) := -c \log(c) - (1-c)\log(1-c)$ is the binary entropy function. Since the RHS does not depend on c, we may maximize the LHS over $c \in [0,1]$, yielding (4.13). Now, we draw the reader's attention to the formal similarity to (3.9). Namely, we can apply the same logic to bound

$$\sup_{\mathbf{c} \ge 0} \left\{ \sum_{i=1}^{k} c_i h(X_i) - D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) \right\} \le \max_{X \in \Pi(X_1,\dots,X_k)} \sum_{j=1}^{m} d_j h(B_j X). \tag{4.14}$$

[2]The comparison also holds in the multidimensional setting, distinguishing it from the Zamir–Feder inequality.

The difficulty encountered is that, unlike $c \mapsto h_2(c)$, the function $\mathbf{c} \mapsto D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is not explicit, complicating the optimization problem to be solved. Nevertheless, the task can be accomplished with all the ingredients we have at hand.

Proof of Theorem 40. We start by noting each maximum is attained due to Proposition 32. Now, without loss of generality, we can assume $\mathbf{d}$ is scaled so that

$$\sum_{j=1}^{m} d_j \dim(E^j) = 1. \tag{4.15}$$

Also, since there are no qualifications on the linear maps in $\mathbf{B}$, a simple rescaling argument reveals that we can assume without loss of generality that $h(X_i) = \frac{\dim(E_i)}{2} \log(2\pi e)$; this will allow us to consider $Z_i \sim N(0, K_i)$ with $\det(K_i) = 1$ for each $1 \le i \le k$. Thus, by Theorem 17, we have

$$\max_{X \in \Pi(X_1,\dots,X_k)} \sum_{j=1}^{m} d_j h(B_j X) \ge \sum_{i=1}^{k} c_i h(X_i) - D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) \tag{4.16}$$

$$= \frac{1}{2} \log(2\pi e) \sum_{i=1}^{k} c_i \dim(E_i) - D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) \tag{4.17}$$

for any $\mathbf{c}$. Define the simplex

$$A := \left\{ \mathbf{c} \ge 0 : \sum_{i=1}^{k} c_i \dim(E_i) = \sum_{j=1}^{m} d_j \dim(E^j) = 1 \right\},$$

which is compact and convex. By Theorem 17, we have $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) < \infty$ only if $\mathbf{c} \in A$, so our task in maximizing the RHS of (4.17) is to compute

$$\max_{\mathbf{c} \in A} -D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = -\min_{\mathbf{c} \in A} D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}),$$

where the use of max and min is justified, since $\mathbf{c} \mapsto D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is l.s.c. by Theorem 35 and A is compact. For $\mathbf{c} \in A$ and $(K_1, \dots, K_k) \in \prod_{i=1}^{k} \mathbf{S}^+(E_i)$, define

$$F\left(\mathbf{c}, (K_i)_{i=1}^{k}\right) := \max_{K \in \Pi(K_1,\dots,K_k)} \sum_{j=1}^{m} d_j \log\det(B_j K B_j^T) - \sum_{i=1}^{k} c_i \log\det(K_i),$$

which is the same as that in (4.7). Theorem 35 ensures that F satisfies the hypotheses of Theorem 39. Thus, by an application of the latter and definition of $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$, we have

$$\begin{aligned}
&\max_{\mathbf{c}\in A} -2D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) \\
&= \max_{\mathbf{c}\in A} \inf_{(K_i)_{i=1}^k \in \prod_{i=1}^k \mathbf{S}^+(E_i)} F\left(\mathbf{c}, (K_i)_{i=1}^k\right) \\
&= \inf_{(K_i)_{i=1}^k \in \prod_{i=1}^k \mathbf{S}^+(E_i)} \max_{\mathbf{c}\in A} F\left(\mathbf{c}, (K_i)_{i=1}^k\right) \\
&= \inf_{(K_i)_{i=1}^k \in \prod_{i=1}^k \mathbf{S}^+(E_i)} \max_{K\in\Pi(K_1,\dots,K_k)} \sum_{j=1}^m d_j \log\det(B_j K B_j^T) - \min_{1\le i\le k} \frac{\log\det(K_i)}{\dim(E_i)} \\
&= \inf_{\substack{(K_i)_{i=1}^k \in \prod_{i=1}^k \mathbf{S}^+(E_i): \\ \min_{1\le i\le k}\det(K_i)=1}} \max_{K\in\Pi(K_1,\dots,K_k)} \sum_{j=1}^m d_j \log\det(B_j K B_j^T),
\end{aligned}$$

where the last line made use of the observation that the function

$$(K_i)_{i=1}^k \mapsto \max_{K\in\Pi(K_1,\dots,K_k)} \sum_{j=1}^m d_j \log\det(B_j K B_j^T) - \min_{1\le i\le k} \frac{\log\det(K_i)}{\dim(E_i)}$$

is invariant to rescaling $(K_i)_{i=1}^k \mapsto (\alpha K_i)_{i=1}^k$ for $\alpha > 0$ by (4.15).

Now, invoking Theorem 33, we have

$$\begin{aligned}
&\inf_{\substack{(K_i)_{i=1}^k \in \prod_{i=1}^k \mathbf{S}^+(E_i): \\ \min_{1\le i\le k}\det(K_i)=1}} \max_{K\in\Pi(K_1,\dots,K_k)} \sum_{j=1}^m d_j \log\det(B_j K B_j^T) \\
&= \inf_{\substack{(K_i)_{i=1}^k \in \prod_{i=1}^k \mathbf{S}^+(E_i): \\ \min_{1\le i\le k}\det(K_i)=1}} \inf_{(U_i)_{i=1}^k, (V_j)_{j=1}^m} \left(\sum_{i=1}^k \langle U_i, K_i\rangle_{\mathrm{HS}} - \sum_{j=1}^m d_j \log\det V_j \right),
\end{aligned}$$

where the second infimum is over all $U_i \in \mathbf{S}^+(E_i), 1 \le i \le k$ and $V_j \in \mathbf{S}^+(E^j), 1 \le j \le m$ satisfying

$$\sum_{j=1}^m d_j B_j^T V_j B_j \le \mathrm{diag}(U_1, \dots, U_k).$$

Written in this way, it evidently suffices to consider $\det(K_i) = 1$ for all $1 \le i \le k$ in the last line, so we conclude

$$\max_{\mathbf{c}\in A} -2D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = \inf_{\substack{(K_i)_{i=1}^k \in \prod_{i=1}^k \mathbf{S}^+(E_i): \\ \det(K_i)=1, 1\le i\le k}} \max_{K\in\Pi(K_1,\dots,K_k)} \sum_{j=1}^m d_j \log\det(B_j K B_j^T). \tag{4.18}$$

Now, let $\mathbf{c}^* \in \arg\min_{\mathbf{c}\in A} D_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$. By (4.17) and (4.15), we have

$$\max_{X\in\Pi(X_1,\ldots,X_k)} \sum_{j=1}^{m} d_j h(B_j X) \geq \frac{1}{2}\log(2\pi e) - D_g(\mathbf{c}^*, \mathbf{d}, \mathbf{B}). \tag{4.19}$$

If the LHS of (4.19) is equal to $-\infty$, then it is easy to see that one of the B_j's must fail to be surjective. Indeed, suppose each B_j is surjective and factor $B_j = R_j Q_j$, where Q_j has orthonormal rows and R_j is full rank. Letting $Q_j^\perp$ denote the matrix with orthonormal rows and rowspace equal to the orthogonal complement of the rowspace of Q_j, for the independent coupling X we have

$$\sum_{i=1}^{k} h(X_i) = h(X) = h(Q_j X, Q_j^\perp X) \leq h(Q_j X) + h(Q_j^\perp X).$$

Since $h(Q_j^\perp X)$ is bounded from above due to finiteness of second moments and the LHS is finite by assumption, $h(Q_j X)$ is finite, and so is $h(B_j X)$. Therefore, (4.11) holds trivially if the LHS of (4.19) is equal to $-\infty$, so we assume henceforth that the LHS of (4.19) is finite. If $(\mathbf{c}^*, \mathbf{d}, \mathbf{B})$ is extremizable, then by Theorem 29 and (4.18), there exist Gaussians $Z_i^* \sim N(0, K_i)$ with $\det(K_i) = 1$ such that

$$\begin{aligned}\max_{X\in\Pi(X_1,\ldots,X_k)} \sum_{j=1}^{m} d_j h(B_j X) &\geq \frac{1}{2}\log(2\pi e) - D_g(\mathbf{c}^*, \mathbf{d}, \mathbf{B}) \\ &= \max_{Z\in\Pi(Z_1^*,\ldots,Z_k^*)} \sum_{j=1}^{m} d_j h(B_j Z),\end{aligned}$$

where we used the identity $\frac{1}{2}\log(2\pi e) = \sum_{i=1}^k c_i^* h(X_i) = \sum_{i=1}^k c_i^* h(Z_i^*)$. On the other hand, if $(\mathbf{c}^*, \mathbf{d}, \mathbf{B})$ is not extremizable, then we have strict inequality in (4.19), and it follows by (4.18) that there are Gaussians $Z_i \sim N(0, K_i)$ with $\det(K_i) = 1$ such that (4.11) holds (with strict inequality, in fact). □

Proof of Theorem 21

With Theorem 40 at our disposal, it is a straightforward matter to self-strengthen it to produce Theorem 21.

First, observe that lower semicontinuity of relative entropy implies $X \in \Pi(X_1, \ldots, X_k) \mapsto I_S(X)$ is weakly lower semicontinuous, and therefore $\Pi(X_1, \ldots, X_k; \nu)$ is a compact subset of $\Pi(X_1, \ldots, X_k)$ when equipped with the weak topology. Hence, repeating the argument in the Proposition 32, we find that each maximum is achieved the statement of the Theorem.

Now, by the method of Lagrange multipliers,

$$\begin{aligned}
&\max_{X\in\Pi(X_1,\dots,X_k;\nu)} \sum_{j=1}^m d_j h(B_j X) \\
&= \max_{X\in\Pi(X_1,\dots,X_k)} \inf_{\lambda\geq 0} \left(\sum_{j=1}^m d_j h(B_j X) - \sum_{S:\nu(S)<\infty} \lambda(S)(I_S(X) - \nu(S)) \right) \\
&= \inf_{\lambda\geq 0} \max_{X\in\Pi(X_1,\dots,X_k)} \underbrace{\left(\sum_{j=1}^m d_j h(B_j X) - \sum_{S:\nu(S)<\infty} \lambda(S)(I_S(X) - \nu(S)) \right)}_{=:G(\lambda,X)},
\end{aligned}$$

where the infimum is over functions $\lambda : 2^{\{1,\dots,k\}} \to [0,+\infty)$. The exchange of max and inf follows by an application of the classical Sion minimax theorem. Indeed, for any fixed $X \in \Pi(X_1,\dots,X_k)$, the function $\lambda \mapsto G(\lambda, X)$ is linear in λ. On the other hand, $\Pi(X_1,\dots,X_k)$ is a convex subset of $\mathcal{P}(E_0)$ that is compact with respect to the weak topology. For fixed $\lambda \geq 0$, the functional $X \mapsto G(\lambda, X)$ is concave upper semicontinuous on $\Pi(X_1,\dots,X_k)$ by concavity of entropy and Lemma 31.

Using the definition of I_S, for any $\lambda \geq 0$, Theorem 40 applies to give existence of Gaussian $(Z_i)_{i=1}^k$ satisfying

$$\begin{aligned}
&\max_{X\in\Pi(X_1,\dots,X_k)} \left(\sum_{j=1}^m d_j h(B_j X) - \sum_{S:\nu(S)<\infty} \lambda(S)(I_S(X) - \nu(S)) \right) \\
&\geq \max_{Z\in\Pi(Z_1,\dots,Z_k)} \left(\sum_{j=1}^m d_j h(B_j Z) - \sum_{S:\nu(S)<\infty} \lambda(S)(I_S(Z) - \nu(S)) \right) \\
&\geq \max_{Z\in\Pi(Z_1,\dots,Z_k;\nu)} \sum_{j=1}^m d_j h(B_j Z).
\end{aligned}$$

The last inequality follows since we are taking the maximum over a smaller set and because $\lambda \geq 0$. This proves the theorem.

4.3 Applications

Constrained multi-marginal inequalities

In this section, we introduce a constrained version of the multi-marginal inequality considered in (3.9) and demonstrate how the results transfer almost immediately with the help of Theorem 21.

Fix a datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$. For a constraint function $\nu : 2^{\{1,\dots,k\}} \to [0, +\infty]$, let $D(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$ denote the smallest constant D such that the inequality

$$\sum_{i=1}^{k} c_i h(X_i) \leq \max_{X \in \Pi(X_1,\dots,X_k;\nu)} \sum_{j=1}^{m} d_j h(B_j X) + D \tag{4.20}$$

holds for all choices of $X_i \in \mathcal{P}(E_i)$, $1 \leq i \leq k$. Call $(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$ **extremizable** if there are $X_i \in \mathcal{P}(E_i)$, $1 \leq i \leq k$ which achieve equality in (4.20) with $D = D(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$. Similarly, let $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$ denote the smallest constant D such that (4.20) holds for all Gaussian $X_i \in \mathcal{G}(E_i)$, $1 \leq i \leq k$, and call $(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$ **Gaussian-extremizable** if there are $X_i \in \mathcal{G}(E_i)$, $1 \leq i \leq k$ which achieve equality in (4.20) with $D = D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$.

The following generalizes Theorem 17 and 29 to the correlation-constrained setting.

Theorem 43. *For any datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ and constraint function ν,*

(i) $D(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu) = D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$*;*

(ii) $(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$ *is extremizable if and only if it is Gaussian-extremizable; and*

(iii) $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu)$ *is finite if and only if the scaling condition (3.5) and the dimension condition (3.6) hold.*

Proof. For any $X_i \in \mathcal{P}(E_i)$ and any $\mathbf{c}$, an application of Theorem 21 ensures existence of $Z_i \in \mathcal{G}(E_i)$ with $h(Z_i) = h(X_i)$ satisfying

$$\begin{aligned}
&\sum_{i=1}^{k} c_i h(X_i) - \max_{X \in \Pi(X_1,\dots,X_k;\nu)} \sum_{j=1}^{m} d_j h(B_j X) \\
&\leq \sum_{i=1}^{k} c_i h(Z_i) - \max_{Z \in \Pi(Z_1,\dots,Z_k;\nu)} \sum_{j=1}^{m} d_j h(B_j Z) \leq D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu),
\end{aligned}$$

where the final inequality follows by definition of D_g. This establishes both (i) and (iii). As for finiteness, observe that definitions imply

$$D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) \equiv D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}; +\infty) \leq D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}; \nu) \leq D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}; 0) \tag{4.21}$$

for any ν. Now, for any $K \in \Pi(K_1, \dots, K_k)$ with $K_i \in \mathbf{S}^+(E_i)$, $1 \leq i \leq k$, observe that

$$K \leq k \operatorname{diag}(K_1, \dots, K_k).$$

Indeed, for $Z \sim N(0, K)$ and $u = (u_1, \dots, u_k) \in E_0$, Jensen's inequality yields

$$u^T K u = \mathbb{E}|u^T Z|^2 \leq k \sum_{i=1}^{k} \mathbb{E}|u_i^T Z_i|^2 = k u^T \operatorname{diag}(K_1, \dots, K_k) u.$$

This implies, for Gaussian $(Z_i)_{i=1}^k$, that

$$\max_{Z\in\Pi(Z_1,\ldots,Z_k)} \sum_{j=1}^m d_j h(B_j Z) \le \sum_{j=1}^m d_j h(B_j Z^{\mathrm{ind}}) + \log(k) \sum_{j=1}^m d_j \dim(E^j),$$

where Z^{ind} denotes the independent coupling of the Z_i's. Thus,

$$D_g(\mathbf{c},\mathbf{d},\mathbf{B};0) \le D_g(\mathbf{c},\mathbf{d},\mathbf{B}) + \log(k) \sum_{j=1}^m d_j \dim(E^j),$$

so that finiteness of $D_g(\mathbf{c},\mathbf{d},\mathbf{B};\nu)$ is equivalent to finiteness of $D_g(\mathbf{c},\mathbf{d},\mathbf{B})$ by (4.21). Invoking Theorem 17 completes the proof. □

When $\nu \equiv 0$, then the only allowable coupling in (4.20) is the independent one. Thus, we recover the main results of Anantharam et al. [2019, Theorems 3 & 4], which simultaneously capture the entropic Brascamp–Lieb inequalities and the EPI.

When $\nu \equiv +\infty$, then we immediately recover Theorems 17 and 29. We remark that, while Theorem 17 admits an equivalent functional form as in Theorem 19, there is no obvious functional equivalent when ν induces nontrivial correlation constraints. In particular, the comparison (4.1) seems to be most naturally expressed in the language of entropies (even in the unconstrained case).

Gaussian saddle point

The EPI has been successfully applied many times to prove coding theorems, particularly in the field of network information theory. However, it also provides the essential ingredient in establishing that a certain mutual information game admits a saddle point (see Pinsker [1956], Ihara [1978], and also [Cover, 1999, Problem 9.21]). Namely, for numbers $P, N \ge 0$, we have

$$\sup_{P_X:\mathbb{E}|X|^2\le P} \inf_{P_Z:\mathbb{E}|Z|^2\le N} I(X;X+Z) = \inf_{P_Z:\mathbb{E}|Z|^2\le N} \sup_{P_X:\mathbb{E}|X|^2\le P} I(X;X+Z),$$

where the sup (resp. inf) is over $X \sim P_X \in \mathcal{P}(\mathbb{R}^n)$ such that $\mathbb{E}|X|^2 \le P$ (resp. $Z \sim P_Z \in \mathcal{P}(\mathbb{R}^n)$ such that $\mathbb{E}|Z|^2 \le N$), and the mutual information is computed under the assumption that $X \sim P_X$ and $Z \sim P_X$ are independent. It turns out that the game admits a Gaussian saddle point, which together with Shannon's capacity theorem, implies that worst-case additive noise subject to a power-constraint is Gaussian.

In this section, we extend this saddle point property to a game with payoff given by

$$G_\zeta(P_X,P_Z) := \sup_{\substack{(X,Z)\in\Pi(P_X,P_Z):\\ I(X;Z)\le\zeta}} I(X;X+Z),$$

for a parameter $\zeta \geq 0$, where the supremum is over couplings (X, Z) with given marginals $X \sim P_X$ and $Z \sim P_Z$. Of course, by taking $\zeta = 0$, we will recover the classical saddle-point result above. This may be interpreted as a game where the signal and noise players fix their strategies P_X and P_Z, but the signal player has the benefit during game-play of adapting their transmission using side information obtained about the noise player's action.

Theorem 44. *For $0 < P, N < \infty$ and $\zeta \geq 0$,*

$$\sup_{P_X:\mathbb{E}|X|^2\leq P} \inf_{P_Z:\mathbb{E}|Z|^2\leq N} G_\zeta(P_X, P_Z) = \inf_{P_Z:\mathbb{E}|Z|^2\leq N} \sup_{P_X:\mathbb{E}|X|^2\leq P} G_\zeta(P_X, P_Z).$$

Moreover, $P_X = N\left(0, \frac{P}{n}\,\mathrm{id}_{\mathbb{R}^n}\right)$ and $P_Z = N\left(0, \frac{N}{n}\,\mathrm{id}_{\mathbb{R}^n}\right)$ is a saddle point.

Proof of Theorem 44. In a slight abuse of notation, we will write $\Pi(X_1, X_2; \zeta)$ to denote couplings of X_1, X_2 satisfying $I(X_1; X_2) \leq \zeta$.

Let X and Z be a random variables with finite variance, and let X^*, Z^* be centered isotropic Gaussians with $\mathbb{E}|X^*|^2 = \mathbb{E}|X|^2$ and $\mathbb{E}|Z^*|^2 = \mathbb{E}|Z|^2$. Now, observe that Theorem 24 implies

$$\begin{aligned}
\max_{\Pi(X^*,Z;\zeta)} (h(X^* + Z) - h(Z)) &\geq \frac{n}{2} \log\left(1 + \frac{N(X^*)}{N(Z)} + 2\sqrt{(1 - e^{-\frac{2\zeta}{n}})\frac{N(X^*)}{N(Z)}}\right) \\
&\geq \frac{n}{2} \log\left(1 + \frac{N(X^*)}{N(Z^*)} + 2\sqrt{(1 - e^{-\frac{2\zeta}{n}})\frac{N(X^*)}{N(Z^*)}}\right) \\
&= \max_{\Pi(X^*,Z^*;\zeta)} (h(X^* + Z^*) - h(Z^*)),
\end{aligned}$$

where the second inequality follows since $h(Z) \leq h(Z^*)$, and the last equality follows by the equality conditions in Theorem 24. In particular, this gives

$$\begin{aligned}
\sup_{\Pi(X^*,Z;\zeta)} I(X^*; X^* + Z) &= \sup_{\Pi(X^*,Z;\zeta)} (h(X^* + Z) - h(Z) + I(X^*; Z)) \\
&= \sup_{\Pi(X^*,Z;\zeta)} (h(X^* + Z) - h(Z)) + \zeta \qquad (4.22) \\
&\geq \sup_{\Pi(X^*,Z^*;\zeta)} (h(X^* + Z^*) - h(Z^*)) + \zeta \qquad (4.23) \\
&= \sup_{\Pi(X^*,Z^*;\zeta)} I(X^*; X^* + Z^*),
\end{aligned}$$

where (4.22) can be justified using the supremum[3], and (4.23) follows from the previous computation. For any pair (X, Z^*), couple (X^*, Z^*) to have the same covariance. By the

[3]This sounds obvious, but we don't know of a simple argument to justify the assertion. A proof is given in Proposition 45.

max-entropy property of Gaussians, $I(X^*;Z^*) \le I(X;Z^*)$ and $h(X+Z^*) \le h(X^*+Z^*)$. As a result, we have

$$\sup_{\Pi(X,Z^*;\zeta)} I(X;X+Z^*) \le \sup_{\Pi(X^*,Z^*;\zeta)} I(X^*;X^*+Z^*) \le \sup_{\Pi(X^*,Z;\zeta)} I(X^*;X^*+Z).$$

This implies

$$\inf_{P_Z:\mathbb{E}|Z|^2\le N} \sup_{P_X:\mathbb{E}|X|^2\le P} G_\zeta(P_X,P_Z) \le \sup_{P_X:\mathbb{E}|X|^2\le P} \inf_{P_Z:\mathbb{E}|Z|^2\le N} G_\zeta(P_X,P_Z),$$

and the reverse direction follows by the max-min inequality. The fact that the asserted distributions coincide with the saddle point subject to the constraints follows by direct computation. □

We now tie up loose ends by justifying (4.22), which is an easy consequence of the proposition below.

Proposition 45. *Let $X \sim N(0,\mathrm{id}_{\mathbb{R}^n})$ and $Z \in \mathcal{P}(\mathbb{R}^n)$ be jointly distributed with $I(X;Z) \le \zeta < +\infty$. For any $\epsilon > 0$, there is a coupling $(X',Z') \in \Pi(X,Z)$ with $h(X'+Z') \ge h(X+Z)-\epsilon$ and $I(X';Z') = \zeta$.*

Proof. We'll work in dimension $n = 1$ for simplicity of exposition. It suffices to establish existence of $(X',Z') \in \Pi(X,Z)$ with $h(X'+Z') \ge h(X+Z) - \epsilon$ and $\zeta \le I(X';Z') < +\infty$. Indeed, if there is such (X',Z'), then we can let π_0 denote the joint distribution of (X,Z) and π_1 denote the joint distribution of (X',Z'). For $\theta \in [0,1]$ define the mixture

$$\pi_\theta = (1-\theta)\pi_0 + \theta\pi_1.$$

Evidently, $\pi_\theta \in \Pi(X,Z)$ for all $\theta \in [0,1]$. For $(X^{(\theta)},Z^{(\theta)}) \sim \pi_\theta$, concavity of entropy gives

$$h(X^{(\theta)}+Z^{(\theta)}) \ge (1-\theta)h(X+Z) + \theta h(X'+Z') \ge h(X+Z) - \epsilon.$$

Now, convexity of relative entropy ensures that $\theta \mapsto I(X^{(\theta)};Z^{(\theta)})$ is continuous on $(0,1)$. Weak lower semicontinuity of mutual information together with finiteness of $I(X';Z')$ establishes continuity at the endpoints, so that the above mapping is continuous on $[0,1]$. As a result, the intermediate value theorem ensures there is some $\theta \in [0,1]$ such that $h(X^{(\theta)}+Z^{(\theta)}) \ge h(X+Z) - \epsilon$ and $I(X^{(\theta)};Z^{(\theta)}) = \zeta$.

Toward establishing the above ansatz, fix $\epsilon > 0$, and consider the interval $I := (-\epsilon,\epsilon]$. Define $p(\epsilon) := \Pr\{X \in I\}$, and note that $p(\epsilon) = \Theta(\epsilon)$ since X is assumed Gaussian. For fixed parameters $n \ge 1$ and ϵ, we'll rearrange the joint distribution of (X,Z) on the event $\{X \in I\}$. To this end, consider two partitions

$$-\epsilon = t_0 < t_1 < \cdots < t_n = \epsilon$$

and

$$-\infty = s_0 < s_1 < \cdots < s_n = +\infty$$

such that

$$\Pr\{X \in (t_{i-1}, t_i] | X \in I\} = \Pr\{Z \in (s_{i-1}, s_i] | X \in I\} = \frac{1}{n}, \quad 1 \leq i \leq n.$$

This is always possible since X and Z are (marginally) continuous random variables. We now define a random variable Z_n, jointly distributed with X, by rearranging the distribution of (X, Z) as follows. On the event $\{X \notin I\}$, we let $Z_n = Z$. Conditioned on the event $\{X \in I\}$, we let the joint density of (X, Z_n) be supported on the union of rectangles $R := \cup_{i=1}^n (t_{i-1}, t_i] \times (s_{i-1}, s_i]$, given explicitly by

$$f_{X,Z_n|X\in I}(x, z) = n f_{X|X\in I}(x) f_{Z|X\in I}(z) 1_{\{(x,z)\in R\}}.$$

This is well-defined since the conditional densities $f_{X|X\in I}, f_{Z|X\in I}$ exist by marginal continuity of X and Z, and the fact that $\Pr\{X \in I\} > 0$.

Observe that this rearrangement preserves marginals, so $(X, Z_n) \in \Pi(X, Z)$. Further, note that $I(X; Z_n | X \in I) = n$ by construction, therefore

$$\begin{aligned} I(X; Z_n) &= p(\epsilon) I(X; Z_n | X \in I) + (1 - p(\epsilon)) I(X; Z_n | X \notin I) + I(Z; 1_{\{X\in I\}}) \\ &\leq p(\epsilon) n + I(X; Z) + O(\epsilon \log \epsilon). \end{aligned}$$

By nonnegativity of mutual information, the first identity above also implies

$$I(X; Z_n) \geq p(\epsilon) I(X; Z_n | X \in I) = p(\epsilon) n.$$

Since $I(X; Z)$ is finite by assumption, the combination of the above estimates imply

$$I(X; Z_n) = \Theta(\epsilon n), \tag{4.24}$$

where the asymptotics are understood in the sense that $\epsilon > 0$ is fixed and n allowed to increase.

For $x \in I$, let $k(x)$ denote the integer $k \in \{1, \ldots, n\}$ such that $x \in (t_{k-1}, t_k]$. Observe that, conditioned on $\{X \in I\}$, the index $k(X)$ is almost surely equal to a function of $X + Z_n$. This follows since for any $c \in \mathbb{R}$, the line $\{(x, z) : x + z = c\} \subset \mathbb{R}^2$ intersects a unique rectangle of the form

$$(t_{i-1}, t_i] \times (s_{i-1}, s_i].$$

Conditioned on $\{X \in I\}$, the distribution of (X, Z_n) is supported on such rectangles by construction, so the claim follows.

With the above observation together with the fact that X and Z_n are conditionally independent given $\{k(X), X \in I\}$ by construction, we have

$$\begin{aligned} h(X + Z_n | X \in I) &= h(X + Z_n | k(X), X \in I) + I(X + Z_n; k(X) | X \in I) \\ &\geq h(X | k(X), X \in I) + I(X; k(X) | X \in I) \\ &= h(X | X \in I). \end{aligned}$$

In particular,

$$\begin{aligned} h(X+Z_n) &\geq (1-p(\epsilon))h(X+Z_n|X\notin I) + p(\epsilon)h(X+Z_n|X\in I) \\ &\geq (1-p(\epsilon))h(X+Z_n|X\notin I) + p(\epsilon)h(X|X\in I) \\ &= (1-p(\epsilon))h(X+Z|X\notin I) + O(\epsilon\log\epsilon), \end{aligned}$$

where the first line follows since conditioning reduces entropy, and the last line follows since X is nearly uniform on I for ϵ small.

Now, upper bounding entropy in terms of second moments, we have

$$\begin{aligned} h(X+Z|X\in I) &\leq \frac{1}{2}\log\left(2\pi e\mathbb{E}[(X+Z)^2|X\in I]\right) \\ &\leq \frac{1}{2}\log\left(4\pi e(\mathbb{E}[X^2|X\in I] + \mathbb{E}[Z^2|X\in I])\right) \\ &\leq \frac{1}{2}\log\left(\frac{4\pi e}{p(\epsilon)}(\mathbb{E}[X^2]+\mathbb{E}[Z^2])\right). \end{aligned}$$

So, by finiteness of second moments,

$$p(\epsilon)h(X+Z|X\in I) \leq O(\epsilon\log\epsilon).$$

Since $I(X+Z;1_{\{X\in I\}}) \leq H(1_{\{X\in I\}}) = O(\epsilon\log\epsilon)$, we put everything together to find

$$\begin{aligned} h(X+Z) &= h(X+Z|1_{\{X\in I\}}) + I(X+Z;1_{\{X\in I\}}) \\ &\leq (1-p(\epsilon))h(X+Z|X\notin I) + O(\epsilon\log\epsilon) \\ &\leq h(X+Z_n) + O(\epsilon\log\epsilon). \end{aligned}$$

Combining with (4.24) establishes the ansatz, and completes the proof. □

Chapter 5

Anantharam-Jog-Nair Inequality

Anantharam, Jog and Nair (AJN) left open the question of extremizability in their inequality (3.4). That is, when do there exist random vectors $(X_i)_{i=1}^k$ such that (3.4) is met with equality, and what form do any such extremizers take? The goal of this chapter is to answer both questions completely. The first question is addressed in Section 5.1, and the second in Section 5.2, specifically in Theorem 61.

5.1 Extremizability and geometricity

A few preliminary definitions

We start this section by recording a few associated definitions for convenience.

Definition 46. *A subspace $T \subset E_0$ is said to be* **product-form** *if it can be written as $T = \oplus_{i=1}^k T_i$, where $T_i \subset E_i$ for $1 \leq i \leq k$.*

Definition 47. *A subspace $T \subset E_0$ is said to be* **critical** *for $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ if it is product-form, and*

$$\sum_{i=1}^{k} c_i \dim(\pi_{E_i} T) = \sum_{j=1}^{m} d_j \dim(B_j T).$$

Definition 48. *Two data $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ and $(\mathbf{c}', \mathbf{d}', \mathbf{B}')$ are said to be* **equivalent** *if $\mathbf{c} = \mathbf{c}'$, $\mathbf{d} = \mathbf{d}'$, and there exist invertible linear transformations $A_j : E^j \to E^j$ and $C_i : E_i \to E_i$ such that*

$$B'_j = A_j^{-1} B_j C^{-1} \quad \text{for each } 1 \leq j \leq m, \tag{5.1}$$

where $C := \mathrm{diag}(C_1, \ldots, C_k)$.

We remark that, in the special case of $k = 1$, the definitions of critical subspaces and equivalent data coincide with those found in Bennett et al. [2008]. For general k, all three definitions coincide with those in Courtade and Liu [2021].

We first address the question of when (3.4) is extremizable. To make things precise, we say that a datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is **extremizable** if $C_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is finite and there exist independent $X_i \in \mathcal{P}(E_i)$, $1 \leq i \leq k$ such that (3.4) is met with equality. We say that $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is **Gaussian-extremizable** if $C_g(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is finite and there exist independent Gaussian $(X_i)_{i=1}^k$ meeting (3.4) with equality.

In analogy to definitions made in the context of Brascamp–Lieb inequalities, we define the class of AJN-geometric data below. Their significance to (3.4) is the same as that of geometric data to inequalities of Brascamp–Lieb-type. In particular, we will see that all (Gaussian-)extremizable instances of (3.4) are equivalent to AJN-geometric data.

Definition 49 (AJN-Geometric datum)**.** *A datum* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is said to be* **AJN-geometric** *if*

(i) $B_j B_j^T = \mathrm{id}_{E^j}$ *for each* $1 \leq j \leq m$*; and*

(ii) we have the operator identity

$$\sum_{j=1}^{m} d_j \pi_{E_i} B_j^T B_j \pi_{E_i}^T = c_i \, \mathrm{id}_{E_i}, \quad \text{for each } 1 \leq i \leq k. \tag{5.2}$$

Remark 50. *Conditions (i)-(ii) together imply the scaling condition* (3.5)*. This can be seen by taking traces in* (5.2)*, summing from* $i = 1, \ldots, k$*, and using the cyclic and linearity properties of trace together with (ii).*

AJN-geometric data have the convenient property that $C_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = 0$, and they are extremizable by standard Gaussians. We summarize as a formal proposition.

Proposition 51. *If* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is AJN-geometric, then* $C_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = 0$ *and* $X \sim N(0, \mathrm{id}_{E_0})$ *achieves equality in* (3.4)*.*

Proof. We'll use the properties of the Föllmer drift summarized in the Appendix. Begin by fixing centered $\mu_i \in \mathcal{P}(E_i)$, $1 \leq i \leq k$, and let $(W_t)_{t \geq 0}$ be a Brownian motion on E_0 with $\mathrm{Cov}(W_1) = \mathrm{id}_{E_0}$. By Theorem 94, there is a drift $(u_t)_{t=0}^1$ such that $\mathbb{E}[u_t] = 0$ and $(\pi_{E_i}(u_t))_{i=1}^k$ are independent for all $0 \leq t \leq 1$,

$$(W_1 + \int_0^1 u_s ds) \sim \mu_1 \otimes \cdots \otimes \mu_k, \tag{5.3}$$

and $D(\mu_i\|\gamma_{E_i}) = \frac{1}{2}\int_0^1 \mathbb{E}|\pi_{E_i}(u_s)|^2 ds$ for each $1 \le i \le k$. Therefore,

$$\begin{aligned}\sum_{i=1}^k c_i D(\mu_i\|\gamma_{E_i}) &= \frac{1}{2}\mathbb{E}\int_0^1 \sum_{i=1}^k c_i |\pi_{E_i}(u_s)|^2 ds \\ &= \frac{1}{2}\mathbb{E}\int_0^1 \sum_{j=1}^m d_j |B_j u_s|^2 ds \qquad (5.4)\\ &\ge \sum_{j=1}^m d_j D(B_j\sharp(\mu_1\otimes\cdots\otimes\mu_k)\|\gamma_{E^j}), \qquad (5.5)\end{aligned}$$

where (5.4) follows from (5.2) and the properties of u_t, and (5.5) follows from (5.3) and Proposition 92 because $B_j W_1 \sim \gamma_{E^j}$, due to $B_j B_j^T = \mathrm{id}_{E^j}$ by assumption. Now, expanding the relative entropies in terms of Shannon entropies and second moments, the second-moment terms cancel due to independence and (5.2), giving

$$\sum_{i=1}^k c_i h(X_i) \le \sum_{j=1}^m d_j h(B_j X) \qquad (5.6)$$

for any $X_i \sim \mu_i \in \mathcal{P}(E_i)$ and $X \sim \otimes_{i=1}^k \mu_i$, where the centering assumption can be removed due to translation invariance of Shannon entropy. The fact that $X \sim \gamma_{E_0}$ is an extremizer follows immediately from the scaling condition (3.5) (see Remark 50) and the observation that $B_j X \sim \gamma_{E^j}$ (since $B_j B_j^T = \mathrm{id}_{E^j}$). □

Remark 52. *In the case where the datum is such that (3.4) coincides with the Shannon–Stam inequality, the above proof reduces to that of Lehec [2013]. The new idea is identifying and incorporating the "correct" definition of AJN-geometricity. When $k = 1$, the AJN inequality (3.4) coincides with the entropic form of the Brascamp–Lieb inequalities, and the definition of AJN-geometricity reduces to the the definition of geometricity for Brascamp–Lieb data found in Bennett et al. [2008].*

AJN-geometric data have a relatively straightforward geometric interpretation. In particular, first note that each E_i has a natural isometric embedding into E_0 via the inclusion $\pi_{E_i}^T : E_i \to E_0$. If $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is AJN-geometric then $B_j B_j^T = \mathrm{id}_{E^j}$, which means that each E^j can be isometrically embedded into E_0 by the map $B_j^T : E^j \to E_0$. In this way, we can consider $(E_i)_{i=1}^k$ and $(E^j)_{j=1}^m$ to be subspaces of E_0, and $\Pi_{E_i} := \pi_{E_i}^T \pi_{E_i}$ and $\Pi_{E^j} := B_j^T B_j$ define the orthogonal projections of E_0 onto E_i and E^j, respectively. Thus, the geometric instances of the AJN inequality (3.4) can be restated in a way that dispenses with the specific linear maps $\mathbf{B}$ as follows.

Corollary 53. *Let $E^1, \dots, E^m$ be subspaces of $E_0 = \oplus_{i=1}^k E_i$. If $\mathbf{c}$ and $\mathbf{d}$ satisfy*

$$\sum_{j=1}^m d_j \Pi_{E_i}\Pi_{E^j}\Pi_{E_i} = c_i \Pi_{E_i}, \quad \text{for each } 1 \le i \le k, \qquad (5.7)$$

then for any independent $X_i \in \mathcal{P}(E_i)$, $1 \le i \le k$, *and* $X = (X_1, \dots, X_m)$,

$$\sum_{i=1}^{k} c_i h(\Pi_{E_i} X) \le \sum_{j=1}^{m} d_j h(\Pi_{E^j} X). \tag{5.8}$$

Equality is achieved for $X \sim N(0, \mathrm{id}_{E_0})$.

Remark 54. *Entropies in* (5.8) *are computed with respect to Lebesgue measure on the subspace being projected upon. In particular, we have* $h(\Pi_{E_i} X) = h(X_i)$, *but have chosen to write* (5.8) *in a way to emphasize the symmetry of the inequality.*

Above definitions allow us to fully characterize the (Gaussian-)extremizable instances of Theorem 7. It is the main result of this section, and specializes to the extremizability results in Bennett et al. [2008] for the Brascamp–Lieb functional inequalities when $k = 1$.

Theorem 55. *The following are equivalent:*

(i) $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is extremizable.*

(ii) $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is Gaussian-extremizable.*

(iii) There are $K_i \in \mathbf{S}^+(E_i)$, $1 \le i \le k$, *satisfying*

$$\sum_{j=1}^{m} d_j \pi_{E_i} B_j^T (B_j K B_j^T)^{-1} B_j \pi_{E_i}^T = c_i K_i^{-1}, \quad 1 \le i \le k, \tag{5.9}$$

where $K := \mathrm{diag}(K_1, \dots, K_k)$.

(iv) $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is equivalent to an AJN-geometric datum.*

Remark 56. *For* $(K_i)_{i=1}^k$ *satisfying* (5.9), *the Gaussians* $X_i \sim N(0, K_i)$, $1 \le i \le k$ *are extremal in* (3.4). *In fact, the proof of Theorem 55 will show that if* $X_i \in \mathcal{P}(E_i)$, $1 \le i \le k$ *are extremal in* (3.4), *then the covariances* $K_i = \mathrm{Cov}(X_i)$ *necessarily satisfy* (5.9).

As a preliminary observation, we note that the extremizers in (3.4) are closed under convolutions. This fact can be extracted from the doubling argument in Anantharam et al. [2019]; we state and prove it here for completeness.

Proposition 57. *Fix a datum* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *that is extremizable for the AJN inequality* (3.4). *Let* $X = (X_1, \dots, X_k)$ *and* $Y = (Y_1, \dots, Y_k)$ *each satisfy* (3.4) *with equality. If* X, Y *are independent, then* $X + Y = (X_1 + Y_1, \dots, X_k + Y_k)$ *also satisfies* (3.4) *with equality.*

Proof. Define $Z^+ = (Z_1^+, \dots, Z_k^+)$ and $Z^- = (Z_1^-, \dots, Z_k^-)$, where

$$Z_i^+ := \frac{1}{\sqrt{2}}(X_i + Y_i), \quad Z_i^- := \frac{1}{\sqrt{2}}(X_i - Y_i), \quad 1 \le i \le k.$$

Observe that

$$\sum_{i=1}^{k} c_i(h(X_i)+h(Y_i)) = \sum_{i=1}^{k} c_i h(X_i,Y_i) \tag{5.10}$$

$$= \sum_{i=1}^{k} c_i \left(h(Z_i^+)+h(Z_i^-|Z_i^+)\right) \tag{5.11}$$

$$\leq \sum_{j=1}^{m} d_j \left(h(B_jZ^+)+h(B_jZ^-|Z^+)\right) + 2C_g(\mathbf{c},\mathbf{d},\mathbf{B}) \tag{5.12}$$

$$\leq \sum_{j=1}^{m} d_j \left(h(B_jZ^+)+h(B_jZ^-|B_jZ^+)\right) + 2C_g(\mathbf{c},\mathbf{d},\mathbf{B}) \tag{5.13}$$

$$= \sum_{j=1}^{m} d_j \left(h(B_jX,B_jY)\right) + 2C_g(\mathbf{c},\mathbf{d},\mathbf{B}) \tag{5.14}$$

$$= \sum_{j=1}^{m} d_j (h(B_jX)+h(B_jY)) + 2C_g(\mathbf{c},\mathbf{d},\mathbf{B}). \tag{5.15}$$

In the above, (5.10) is due to independence; (5.11) follows due to orthogonality of the transformation $(X_i,Y_i) \to (Z_i^+,Z_i^-)$ and the chain rule; (5.12) is two applications of (3.4); (5.13) follows because conditioning reduces entropy; (5.14) is due to the chain rule and orthogonality of the transformation $(B_jZ^+,B_jZ^-) \to (B_jX,B_jY)$; (5.15) is again due to independence.

Since X and Y are extremal by assumption, we have equality throughout. This implies Z^+ is also extremal, and hence we conclude $X+Y$ is extremal by the scaling condition (3.5). □

Proof of Theorem 55. $(i) \Rightarrow (ii)$:

Let X be an extremizer in (3.4), and put $Z_n := n^{-1/2}\sum_{\ell=1}^{n} X^{(i)}$, where $X^{(1)},X^{(2)},\ldots$ are i.i.d. copies of X, which we assume to be zero-mean without loss of generality. By an application of Proposition 57 and the scaling condition (3.5) (which holds by finiteness of $C_g(\mathbf{c},\mathbf{d},\mathbf{B})$), we have that Z_n is an extremizer in (3.4) for all $n \geq 1$. By an application of the entropic central limit theorem [Barron, 1986, Carlen and Soffer, 1991], it follows that $Z \sim N(0,\mathrm{Cov}(X))$ is also an extremizer in (3.4).

$(ii) \Rightarrow (i)$: This follows immediately from Theorem 7.

$(ii) \Rightarrow (iii)$: If $(\mathbf{c},\mathbf{d},\mathbf{B})$ is Gaussian-extremizable, then there exist $K_i^* \in \mathbf{S}^+(E_i)$, $1 \leq i \leq k$ which maximize

$$(K_i)_{i=1}^k \mapsto \sum_{i=1}^{k} c_i \log\det(K_i) - \sum_{j=1}^{m} d_j \log\det(B_jKB_j^T),$$

where $K := \mathrm{diag}(K_1, \ldots, K_k)$ (note this implies $B_j K^* B_j^T$ is invertible for each $1 \le j \le m$). This means, for any index i and any $A_i \in \mathbf{S}(E_i)$, we can consider the perturbation $K_i = K_i^* + \epsilon A_i$ for ϵ sufficiently small, and the function value cannot increase. By first-order Taylor expansion, this implies

$$\begin{aligned} c_i \langle A_i, (K_i^*)^{-1} \rangle &= \sum_{j=1}^m d_j \langle B_j \pi_{E_i}^T A_i \pi_{E_i} B_j^T, (B_j K^* B_j^T)^{-1} \rangle \\ &= \Big\langle A_i, \sum_{j=1}^m d_j \pi_{E_i} B_j^T (B_j K^* B_j^T)^{-1} B_j \pi_{E_i}^T \Big\rangle, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the Hilbert–Schmidt (trace) inner product. By arbitrariness of A_i, we conclude (5.9).

$(iii) \Rightarrow (iv)$: Let K be as in (5.9). The equivalent datum $(\mathbf{c}, \mathbf{d}, \mathbf{B}')$ defined by

$$B_j' = (B_j K B_j^T)^{-1/2} B_j K^{1/2}, \quad 1 \le j \le m$$

is AJN-geometric. Indeed, $B_j' B_j'^T = \mathrm{id}_{E^j}$ and (5.9) gives

$$\sum_{j=1}^m d_j \pi_{E_i} B_j'^T B_j' \pi_{E_i}^T = \sum_{j=1}^m d_j K_i^{1/2} \pi_{E_i} B_j (B_j K B_j^T)^{-1} B_j \pi_{E_i}^T K_i^{1/2} = c_i \, \mathrm{id}_{E_i} .$$

$(iv) \Rightarrow (ii)$: Let $(\mathbf{c}, \mathbf{d}, \mathbf{B}')$ be the geometric datum equivalent to $(\mathbf{c}, \mathbf{d}, \mathbf{B})$. In the notation of (5.1), for any $X_i \in \mathcal{P}(E_i)$, $1 \le i \le k$ and $X = (X_1, \ldots, X_k)$, we have by a change of variables

$$\begin{aligned} &\sum_{i=1}^k c_i h(X_i) - \sum_{j=1}^m d_j h(B_j X) \\ &= \sum_{i=1}^k c_i h(C_i X_i) - \sum_{i=1}^k c_i \log\det(C_i) - \sum_{j=1}^m d_j h(B_j' C X) - \sum_{j=1}^m d_j \log\det(A_j) \\ &= \sum_{i=1}^k c_i h(Y_i) - \sum_{j=1}^m d_j h(B_j' Y) - \sum_{i=1}^k c_i \log\det(C_i) - \sum_{j=1}^m d_j \log\det(A_j), \end{aligned}$$

where we have defined $Y_i := C_i X_i$, and $Y = (Y_1, \ldots, Y_k)$. Since each C_i is invertible, it is clear that X is a (Gaussan-)extremizer for $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ if and only if Y is a (Gaussan-)extremizer for $(\mathbf{c}, \mathbf{d}, \mathbf{B}')$. The latter is Gaussian-extremizable by the assumption of geometricity and Proposition 51, so the claim follows. □

Remark 58. *We remark that Theorem 7 can be derived as a limiting case of the forward-reverse Brascamp–Lieb inequalities [Liu et al., 2018]; details can be found in [Courtade and*

Liu, 2021, Section 4]. There is a counterpart notion of geometricity for the forward-reverse Brascamp–Lieb inequalities, for which a result parallel to Theorem 55 holds. However, the notion of "geometricity" in the context of Courtade and Liu [2021] does not easily pass through the aforementioned limit, so it seems the simplest proof of Theorem 55 is a more direct one, as given here.

We end this section with a final structural lemma, namely, we note that critical subdata of geometric data consist of geometric data. To this end, we can define the notion of a subdata. Fix a datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$, and a critical subspace $\oplus_{i=1}^{k} T_i = T \subset E_0$. Define the linear maps:

$$B_{j,T} : x \in T \mapsto B_j x \in B_j T$$

Now, define $\mathbf{B}_T := (B_{j,T})_{j=1}^m$. With notation set, we can define $(\mathbf{c}, \mathbf{d}, \mathbf{B}_T)$ as a datum on T.

Theorem 59. *Let $(\boldsymbol{c}, \boldsymbol{d}, \boldsymbol{B})$ be a geometric AJN instance on E_0, and let T be a critical subspace. Then, $(\boldsymbol{c}, \boldsymbol{d}, \boldsymbol{B}_T)$ is a geometric AJN instance on T.*

Proof of Theorem 59. Note that for the restricted maps, (5.2) still holds with equality, as $\pi_{T_i}^T T_i \subset T$ since T is product form. So,

$$\sum_{j=1}^{m} d_j \pi_{T_i} B_{j,T}^T B_{j,T} \pi_{T_i}^T = c_i \, \mathrm{id}_{T_i}, \qquad \text{for each } 1 \le i \le k. \tag{5.16}$$

Now, we only need to check that

$$B_{j,T} B_{j,T}^T = \mathrm{id}_{B_j T}$$

holds for each $1 \le j \le m$. We note that $B_{j,T} B_{j,T}^T = B_j \Pi_T B_j^T$ on $B_j T$, so we have $B_{j,T} B_{j,T}^T = B_j \Pi_T B_j^T \le B_j B_j^T = \mathrm{id}_{B_j T}$ on $B_j T$. In particular, we have $\mathrm{Tr}(B_{j,T} B_{j,T}^T) \le \dim(B_j T)$, with equality iff $B_{j,T} B_{j,T}^T = \mathrm{id}_{B_j T}$. Now, we take traces of (5.16), and sum up over i and invoke criticality to note:

$$\sum_{j=1}^{m} d_j \mathrm{Tr}(B_{j,T} B_{j,T}^T) = \sum_{i=1}^{k} c_i \dim(T_i) = \sum_{j=1}^{m} d_j \dim(B_j T)$$

from which the result follows. □

5.2 Characterization of extremizers

The goal of this section is to give a complete characterization of the extremizers in (3.4). In view of Theorem 55, it suffices to consider geometric instances of the AJN inequality; indeed, the extremizers of any other extremizable instance of the AJN inequality will be linear transformations of the extremizers for an equivalent AJN-geometric datum.

Toward this end, let $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ be AJN-geometric, and regard $(E_i)_{i=1}^k$ and $(E^j)_{j=1}^m$ as subspaces of E_0, as in the discussion preceding Corollary 53. We now extend definitions found in Valdimarsson [2008] to the present setting. For any subspace $W \subset E_0$, we define $W^o := W$, and $W^\perp$ to be the orthogonal complement of W in E_0. Now, a nonzero subspace $K \subset E_0$ is said to be **independent** if it can be written as

$$K = E_i \cap \bigcap_{j=1}^m (E_j)^{b_j},$$

for some $i \in \{1, \ldots, k\}$ and $b = (b_1, \ldots, b_m) \in \{o, \perp\}^m$. Each independent subspace is contained in some E_i, and distinct independent subspaces are orthogonal by construction. So, if $K_1^i, \ldots, K_{n_i}^i$ is an enumeration of independent subspaces of E_i, then we can uniquely decompose

$$E_i = K_0^i \oplus K_1^i \oplus \cdots \oplus K_{n_i}^i, \tag{5.17}$$

where K_0^i is defined to be the orthogonal complement of $\oplus_{\ell=1}^{n_i} K_\ell^i$ in E_i. Now, we can uniquely define the **dependent** subspace K_{dep} as the product-form subspace

$$K_{dep} := \oplus_{i=1}^k K_0^i. \tag{5.18}$$

Proposition 60. *If K_{dep} is nonzero, there is an orthogonal decomposition*

$$K_{dep} = \oplus_{\ell=1}^n K_{dep}^\ell, \tag{5.19}$$

where each K_{dep}^ℓ is critical for the datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$.

A decomposition of the form (5.19) is said to be a **critical decomposition**; we remark that critical decompositions are not necessarily unique. Together with Theorem 55, the following completely characterizes the extremizers in the AJN inequality (3.4). In the statement, we let $\Pi_V : E_0 \to E_0$ denote the orthogonal projection onto the indicated subspace V.

Theorem 61. *Let $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ be AJN-geometric, and decompose each E_i as in (5.17). Independent $X_i \sim \mathcal{P}(E_i)$, $1 \le i \le k$ and $X = (X_1, \ldots, X_k)$ satisfy (3.4) with equality iff*

(i) $\Pi_{K_0^i}(X), \ldots, \Pi_{K_{n_i}^i}(X)$ are independent for each $1 \le i \le k$; and

(ii) there is a critical decomposition $K_{dep} = \oplus_{\ell=1}^n K_{dep}^\ell$ such that $\Pi_{K_{dep}^1}(X)$, …, $\Pi_{K_{dep}^n}(X)$ are independent isotropic Gaussians on their respective subspaces.

In words, (i) says that each random vector X_i splits into independent factors on the orthogonal decomposition of E_i given by (5.17). Condition (ii) tells us that the factor of X supported on K_{dep} is Gaussian with $\mathrm{Cov}(\Pi_{K_{dep}}(X)) = \sum_{\ell=1}^n \sigma_\ell^2 \Pi_{K_{dep}^\ell}$, for some critical decomposition (5.19) and choice of variances $(\sigma_\ell^2)_{\ell=1}^n$. In effect, this links the covariances of the Gaussian factors of the X_i's.

Remark 62. *In the case of $k = 1$, the above characterization of extremizers is compatible with that articulated by Valdimarsson [2008] for the functional Brascamp–Lieb inequalities. As noted in Remark 58, the AJN inequality is formally implied by the Euclidean forward-reverse Brascamp–Lieb inequalities. A characterization of extremizers for the latter remains unknown at the moment, but will necessarily involve a new ingredient of log-concavity (since, e.g., the Prékopa–Leindler inequality is realized as a special case (see Theorem 10), and the extremizers are log-concave [Dubuc, 1977]).*

Before giving the proof, let us consider a few quick examples to demonstrate the result.

Example 63. *Consider the EPI on $E_1 = E_2 = \mathbb{R}^n$ with $\lambda \in (0, 1)$, stated as*

$$\lambda h(X_1) + (1 - \lambda)h(X_2) \leq h(\lambda^{1/2} X_1 + (1 - \lambda)^{1/2} X_2),$$

for independent X_1, X_2 with finite entropies and second moments. There are no independent subspaces, and every maximal critical decomposition of $K_{dep} = E_0 = \mathbb{R}^n \oplus \mathbb{R}^n$ can be written as

$$\mathbb{R}^n \oplus \mathbb{R}^n = \bigoplus_{\ell=1}^{n} (\operatorname{span}\{e_\ell\} \oplus \operatorname{span}\{e_\ell\}),$$

with $(e_\ell)_{\ell=1}^n$ an orthonormal basis of $\mathbb{R}^n$. Thus, (ii) is equivalent to the assertion that X_1 and X_2 must be Gaussian, with identical covariances.

Example 64. *The Zamir–Feder inequality [Zamir and Feder, 1993] can be stated as follows (see, e.g., [Rioul, 2010]). If a matrix $B \in \mathbb{R}^{k \times n}$ satisfying $BB^T = \mathrm{id}_{\mathbb{R}^n}$ has columns $(b_i)_{i=1}^k \subset \mathbb{R}^n$, then any random vector $X = (X_1, \ldots, X_k) \in \mathcal{P}(\mathbb{R}^k)$ with independent coordinates satisfies*

$$h(BX) \geq \sum_{i=1}^{k} |b_i|^2 h(X_i). \tag{5.20}$$

Observe that this is a geometric instance of the AJN inequality, with $B_1 = B$, $d_1 = 1$, and $c_i = |b_i|^2$. Letting $(e_i)_{i=1}^k$ denote the natural basis for $\mathbb{R}^k$, it follows by definitions that any independent subspace must be equal to $\operatorname{span}\{e_i\}$ for some $1 \leq i \leq k$, and $\operatorname{span}\{e_i\}$ is an independent subspace iff $e_i \in \ker(B) \cup \ker(B)^\perp$. Hence, any $X \in \mathcal{P}(\mathbb{R}^k)$ with independent coordinates meeting (5.20) with equality has the following form:

1. *If $e_i \in \ker(B) \cup \ker(B)^\perp$, then X_i can have any distribution in $\mathcal{P}(\mathbb{R})$.*
2. *Otherwise, X_i is Gaussian.*

Observe that $e_i \in \ker(B) \Leftrightarrow b_i = 0$; in this case, coordinate X_i is not present in (5.20). If $e_i \in \ker(B)^\perp$, then X_i is recoverable from BX in the sense that there exists $u \in \mathbb{R}^n$ such that $u^T BX = X_i$. Hence, we might say that the extremizers in (5.20) are characterized by all present non-recoverable components being Gaussian. This is precisely the statement given by Rioul and Zamir in their recent work [Rioul and Zamir, 2019, Theorem 1], which gave the first characterization of extremizers in the Zamir–Feder inequality.

To give an application that yields a new result, consider the following inequality proposed in Anantharam et al. [2019]:

$$c_1 h(Z_1, Z_2) + c_2 h(Y) \leq h(Z_1 + Y, Z_2 + Y) + d_2 h(Z_1) + d_3 h(Z_2) + C_g, \tag{5.21}$$

where the Z_1, Z_2, Y are random variables with (Z_1, Z_2) independent of Y, and all coefficients are assumed to be strictly positive. An immediate consequence of Theorem 7 is that the sharp constant C_g can be computed by considering only Gaussians, and conditions on the coefficients $\mathbf{c}, \mathbf{d}$ ensuring finiteness of C_g can be deduced from (3.5) and (3.6). Using Theorem 61, we can further conclude that if $\mathbf{c}$ and $\mathbf{d}$ are such that (5.21) is extremizable, then it admits only Gaussian extremizers.

To see that this is the case, let $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ denote the datum corresponding to (5.21). In matrix notation with respect to the natural choice of basis, we have

$$B_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}.$$

Assuming $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is extremizable, let C and $(A_j)_{j=1}^3$ be the matrices in (5.1) that transform $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ to an AJN-geometric datum $(\mathbf{c}, \mathbf{d}, \mathbf{B}')$. By rescaling, we can assume without loss of generality that $C = \operatorname{diag}(C_1, 1)$, where C_1 is an invertible 2×2 matrix. In order to show (5.21) admits only Gaussian extremizers, we need to show that $(\mathbf{c}, \mathbf{d}, \mathbf{B}')$ admits no independent subspaces. To do this, we will show the stronger claim that

$$\bigcap_{j=1}^{3} V_j = \{0\}$$

for any choice of V_j equal to E^j or $E^{j\perp}$, where we identify $E^j = \operatorname{col}(C^{-T} B_j^T A_j^{-T}) = \operatorname{col}(C^{-T} B_j^T)$, with $\operatorname{col}(\cdot)$ denoting the columnspace of its argument. Explicitly, we have

$$E^1 = \operatorname{col}\left(\begin{bmatrix} C_1^{-T} \\ 1 \quad 1 \end{bmatrix} \right), \quad E^2 = \operatorname{col}\left(\begin{bmatrix} C_1^{-T} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} \right), \quad E^3 = \operatorname{col}\left(\begin{bmatrix} C_1^{-T} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ 0 \end{bmatrix} \right).$$

Direct computation shows

$$E^{1\perp} = \operatorname{col}\left(\begin{bmatrix} C_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ -1 \end{bmatrix} \right), \quad E^{2\perp} = \operatorname{col}\left(\begin{bmatrix} 0 & C_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ 0 & \\ 1 & 0 \end{bmatrix} \right), \quad E^{3\perp} = \operatorname{col}\left(\begin{bmatrix} 0 & C_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ 0 & \\ 1 & 0 \end{bmatrix} \right).$$

The problem now reduces to casework. By inspection, we have $E^{1\perp} \cap E^2 = E^{1\perp} \cap E^3 = \{0\}$. Next, since C_1 is invertible, we have $E^2 \cap E^3 = \{0\}$, and it similarly follows that $E^1 \cap E^2 = E^1 \cap E^3 = E^{1\perp} \cap E^{2\perp} = \{0\}$. It only remains to show that $E^1 \cap E^{2\perp} \cap E^{3\perp} = \{0\}$. To this end, invertibility of C_1 allows us to write

$$E^{2\perp} \cap E^{3\perp} = \operatorname{col}\left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right).$$

However, the only vector in E^1 that is zero in the first two components is the all-zero vector (again, by invertibility of C_1), so it follows that $E^1 \cap E^{2\perp} \cap E^{3\perp} = \{0\}$, and we conclude that the datum $(\mathbf{c}, \mathbf{d}, \mathbf{B}')$ admits no independent subspaces.

Although the above shows (5.21) can only admit Gaussian extremizers, it does not tell us whether any exist, or their structure if they do. This is, however, the content of Theorem 55. Namely, the covariances of Gaussian extremizers are characterized completely by solutions K to (5.9) for the datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$; see Remark 56. This emphasizes the complementary nature of Theorems 61 and 55.

Proof of Theorem 61

The remainder of this section is dedicated to the proof of Theorem 61. We establish the assertion of sufficiency first, and necessity second. The assumption that the datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ is AJN-geometric prevails throughout. Accordingly we will regard E^j as a subspace of E_0, with $\Pi_{E^j} = B_j^T B_j$ denoting the orthogonal projection onto E^j.

Lemma 65. *Let the notation of (5.17) and (5.18) prevail. For each $1 \leq j \leq m$, we have the orthogonal decomposition*

$$E^j = (\Pi_{E^j} K_{dep}) \oplus \left(\bigoplus_{i=1}^{k} \bigoplus_{\substack{1 \leq \ell \leq n_i: \\ K_\ell^i \subset E^j}} K_\ell^i \right). \tag{5.22}$$

Moreover, for any critical decomposition $K_{dep} = \oplus_{\ell=1}^n K_{dep}^\ell$, we have the orthogonal decomposition

$$\Pi_{E^j} K_{dep} = \oplus_{\ell=1}^n \Pi_{E^j} K_{dep}^\ell. \tag{5.23}$$

Proof of Proposition 60 and Lemma 65. We first note that $\Pi_{E^j} K_{dep}$ is orthogonal to $\Pi_{E^j} K$, for any independent subspace K. Indeed, by definition of an independent subspace, we either have $\Pi_{E^j} K = \{0\}$ or $\Pi_{E^j} K = K$. The former is trivially orthogonal to $\Pi_{E^j} K_{dep}$, and the latter is orthogonal to $\Pi_{E^j} K_{dep}$ since K_{dep} is orthogonal to K by definition and Π_{E^j} is self-adjoint. Indeed,

$$(\Pi_{E^j} x)^T y = (\Pi_{E^j} x)^T y = x^T (\Pi_{E^j} y) = x^T y = 0, \quad \forall x \in K_{dep}, y \in K.$$

This establishes (5.22).

Now, using the decomposition (5.17) and the scaling condition (3.5) (which holds by AJN-geometricity), we have

$$\begin{aligned} \sum_{i=1}^k c_i \sum_{\ell=0}^{n_i} \dim(K_\ell^i) = \sum_{i=1}^k c_i \dim(E_i) &= \sum_{j=1}^m d_j \dim(E^j) \\ &= \sum_{j=1}^m d_j \dim(\Pi_{E^j} K_{dep}) + \sum_{j: K_\ell^i \subset E^j}^m d_j \dim(K_\ell^i). \end{aligned}$$

To summarize,

$$\sum_{i=1}^{k} c_i \sum_{\ell=0}^{n_i} \dim(K_\ell^i) = \sum_{j=1}^{m} d_j \dim(\Pi_{E^j} K_{dep}) + \sum_{j:K_\ell^i \subset E^j}^{m} d_j \dim(K_\ell^i). \tag{5.24}$$

Since each independent subspace is of product form, the dimension condition (3.6) implies, for each $1 \le i \le k$ and $1 \le \ell \le n_i$,

$$c_i \dim(K_\ell^i) \le \sum_{j:K_\ell^i \subset E^j}^{m} d_j \dim(K_\ell^i). \tag{5.25}$$

Likewise, since $K_{dep} = \oplus_{i=1}^{k} K_0^i$ is of product form, (3.6) also implies

$$\sum_{i=1}^{k} c_i \dim(K_0^i) \le \sum_{j=1}^{m} d_j \dim(\Pi_{E^j} K_{dep}). \tag{5.26}$$

Comparing against (5.24), we necessarily have equality in (5.25) and (5.26), which proves that K_{dep} is critical. Thus, there exists at least one critical decomposition of K_{dep} (the trivial one), and Proposition 60 follows.

It remains to show (5.23). By induction, it suffices to show if $K \subset E_0$ is a critical subspace, and $K = K_1 \oplus K_2$ is a critical decomposition, then $\Pi_{E^j} K_1$ and $\Pi_{E^j} K_2$ are orthogonal complements in $\Pi_{E^j} K$. The proof is similar to that of Bennett et al. [2008, Lemma 7.12]. Letting $\Pi_{K_1} : E_0 \to E_0$ denote the orthogonal projection onto K_1, we have that $\Pi_{E^j} \Pi_{K_1}$ is a contraction in E_0, so $\mathrm{Tr}(\Pi_{E^j} \Pi_{K_1}) \le \dim(\Pi_{E^j} K_1)$. Since K_1 is critical, it is product-form by definition and therefore $\Pi_{K_1} = \sum_{i=1}^{k} \Pi_{E_i} \Pi_{K_1} \Pi_{E_i}$. From (5.2), this implies

$$\sum_{i=1}^{k} c_i \dim(\Pi_{E_i} K_1) = \sum_{i=1}^{k} c_i \,\mathrm{Tr}(\Pi_{E_i} \Pi_{K_1}) = \sum_{j=1}^{m} d_j \,\mathrm{Tr}(\Pi_{E^j} \Pi_{K_1}) \le \sum_{j=1}^{m} d_j \dim(\Pi_{E^j} K_1).$$

Since K_1 is critical, we have equality throughout, implying $\mathrm{Tr}(\Pi_{E^j} \Pi_{K_1}) = \dim(\Pi_{E^j} K_1)$ for each j. From this, we can conclude that $\Pi_{K_1} \Pi_{E^j}$ is an isometry from $\Pi_{E^j} K_1$ into K_1, and similarly $\Pi_{K_2} \Pi_{E^j}$ is an isometry from $\Pi_{E^j} K_2$ into K_2. Since K_1 and K_2 are orthogonal complements in K, it follows that $\Pi_{E^j} K_1$ and $\Pi_{E^j} K_2$ are orthogonal complements in $\Pi_{E^j} K$. □

Sufficiency of conditions (i)-(ii) in Theorem 61. Let $X_i \sim \mathcal{P}(E_i)$, $1 \le i \le k$ be independent and satisfy (i)-(ii), and let $X = (X_1, \ldots, X_k)$. By the orthogonal decomposition (5.22) and the independence assumptions imposed by (i), we can decompose

$$h(B_j X) = h(B_j \Pi_{K_{dep}}(X)) + \sum_{i=1}^{k} \sum_{\substack{1 \le \ell \le n_i: \\ K_\ell^i \subset E^j}} h(\Pi_{K_\ell^i}(X_i)), \tag{5.27}$$

where all entropies are computed with respect to the subspace being projected upon. In the proof of Lemma 65, we found (5.25) was met with equality. So, whenever E_i contains an independent subspace (i.e., $n_i \geq 1$), we have

$$c_i = \sum_{j:K_\ell^i \subset E^j}^{m} d_j.$$

Now, using the decomposition (5.17) and the independence assumptions imposed by (i), an application of the above identity followed by (5.27) reveals

$$\begin{aligned}
\sum_{i=1}^{k} c_i h(X_i) &= \sum_{i=1}^{k} \sum_{\ell=0}^{n_i} c_i h(\Pi_{K_\ell^i}(X_i)) \\
&= \sum_{i=1}^{k} c_i h(\Pi_{K_0^i}(X_i)) + \sum_{i=1}^{k} \sum_{\ell=1}^{n_i} \sum_{j:K_\ell^i \subset E^j}^{m} d_j h(\Pi_{K_\ell^i}(X_i)) \\
&= \sum_{i=1}^{k} c_i h(\Pi_{K_0^i}(X_i)) + \sum_{j=1}^{m} d_j \sum_{i=1}^{k} \sum_{\substack{1 \leq \ell \leq n_i: \\ K_\ell^i \subset E^j}} h(\Pi_{K_\ell^i}(X_i)) \\
&= \sum_{i=1}^{k} c_i h(\Pi_{K_0^i}(X_i)) + \sum_{j=1}^{m} d_j \left(h(B_j X) - h(B_j \Pi_{K_{dep}}(X)) \right).
\end{aligned}$$

In summary,

$$\sum_{i=1}^{k} c_i h(X_i) - \sum_{j=1}^{m} d_j h(B_j X) = \sum_{i=1}^{k} c_i h(\Pi_{K_0^i}(X_i)) - \sum_{j=1}^{m} d_j h(B_j \Pi_{K_{dep}}(X)), \tag{5.28}$$

where any entropies over the trivial subspace $\{0\}$ are to be neglected.

It remains to show the right hand side of (5.28) is zero. While we can show it by plugging in isotropic Gaussians, it is more straightforward to see that the datum $(\mathbf{c}, \mathbf{d}, \mathbf{B}_{K_{dep}})$ is a geometric datum by criticality of the dependent space as in (5.26) and Theorem 59, and as such, the right hand side of (5.28) is upper bounded by zero, and is exactly zero when the input is an isotropic Gaussian, as in Corollary 53.

Putting everything together shows

$$\sum_{i=1}^{k} c_i h(X_i) = \sum_{j=1}^{m} d_j h(B_j X),$$

so that (i) and (ii) are sufficient conditions for the X_i's to be extremal, since $C_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = 0$ by Proposition 51. □

We now turn our attention to the necessity part of Theorem 61. We will need the following lemma.

Lemma 66. *Let* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *be AJN-geometric, and* $A_i \in \mathbf{S}_0^+(E_i)$, $1 \le i \le k$. *For any* $A \in \Pi(A_1, \dots, A_k)$, *we have*

$$\sum_{i=1}^{k} c_i \operatorname{Tr}\left((A_i - \mathrm{id}_{E_i})^2\right) \ge \sum_{j=1}^{m} d_j \operatorname{Tr}\left(((B_j A^2 B_j^T)^{1/2} - \mathrm{id}_{E^j})^2\right), \tag{5.29}$$

with equality if and only if $(\mathrm{id}_{E_0} - \Pi_{E^j}) A \Pi_{E^j} = 0$ *for each* $1 \le j \le m$.

Proof. Using the block-diagonal structure of A and the definition of AJN-geometricity, we have

$$\begin{aligned}
\sum_{i=1}^{k} c_i \operatorname{Tr}\left((A_i - \mathrm{id}_{E_i})^2\right) &= \sum_{j=1}^{m} d_j \operatorname{Tr}(B_j (A - \mathrm{id}_{E_0})^2 B_j) \\
&= \sum_{j=1}^{m} d_j \operatorname{Tr}(B_j A^2 B_j^T - 2 B_j A B_j^T + \mathrm{id}_{E^j}) \\
&\ge \sum_{j=1}^{m} d_j \operatorname{Tr}(B_j A^2 B_j^T - 2 (B_j A^2 B_j^T)^{1/2} + \mathrm{id}_{E^j}) \\
&= \sum_{j=1}^{m} d_j \operatorname{Tr}\left(((B_j A^2 B_j^T)^{1/2} - \mathrm{id}_{E^j})^2\right),
\end{aligned}$$

where the inequality follows because square root is operator monotone. More precisely, AJN-geometricity implies

$$(B_j A B_j^T)^2 = B_j A B_j^T B_j A B_j^T \le B_j A^2 B_j^T,$$

so that operator monotonicity of square root gives $B_j A B_j^T \le (B_j A^2 B_j^T)^{1/2}$. Equality in (5.29) is therefore equivalent to equality above, which can be rewritten as

$$B_j A (\mathrm{id}_{E_0} - B_j^T B_j) A B_j^T = 0 \quad \Leftrightarrow \quad (\mathrm{id}_{E_0} - \Pi_{E^j}) A \Pi_{E^j} = 0.$$

□

Necessity of conditions (i)-(ii) in Theorem 61. Let $\mu_i \in \mathcal{P}(E_i)$, $1 \le i \le k$ satisfy

$$\sum_{i=1}^{k} c_i D(\mu_i \| \gamma_{E_i}) = \sum_{j=1}^{m} d_j D(B_j \sharp (\mu_1 \otimes \cdots \otimes \mu_k) \| \gamma_{E^j}) \tag{5.30}$$

under the prevailing assumption of AJN-geometricity; this is the same as equality in (5.6). Without loss of generality, we can assume each μ_i is centered. Moreover, since the extremizers

of the AJN inequality are closed under convolutions (Proposition 57) and standard Gaussians are extremal in the geometric AJN inequality (Proposition 51), we can assume without loss of generality that each μ_i is of the form

$$\mu_i = \widetilde{\mu}_i * \gamma_{E_i} \tag{5.31}$$

for some extremal $\widetilde{\mu}_i \in \mathcal{P}(E_i)$, $1 \leq i \leq k$. Indeed, $X \sim \otimes_{i=1}^k \mu_i$ satisfies (i)-(ii) if and only if $X + Z$ satisfies (i)-(ii) for $Z \sim \gamma_{E_0}$, independent of X.

Necessity of condition (i): In the proof of Proposition 51, the sole inequality is (5.5). Hence, properties of the drift u_t warrant a closer inspection; we follow the approach developed in Eldan and Mikulincer [2020]. Toward this end, let f denote the density of $\mu_1 \otimes \cdots \otimes \mu_k$ with respect to γ_{E_0}, and define the function

$$u_t(x) := \nabla \log P_{1-t} f(x), \qquad x \in E_0,\ 0 \leq t \leq 1,$$

where $(P_t)_{t \geq 0}$ denotes the heat semigroup. Note that this is the Föllmer drift in Theorem 94. Define the matrix-valued function

$$\Gamma_t(x) := (1-t)\nabla u_t(x) + \mathrm{id}_{E_0}, \qquad x \in E_0,\ 0 \leq t \leq 1, \tag{5.32}$$

which, for each $0 \leq t \leq 1$, takes the block-diagonal form $\Gamma_t = \mathrm{diag}(\Gamma_t^1, \ldots, \Gamma_t^k)$ with $\Gamma_t^i \in \mathbf{S}^+(E_i)$ due to the product form of the density f and Lemma 95 applied to (5.31). Now, recall Theorem 96, which implies that

$$D(\mu_i \| \gamma_{E_i}) = \frac{1}{2} \int_0^1 \frac{\mathbb{E}\,\mathrm{Tr}\left((\Gamma_t^i - \mathrm{id}_{E_i})^2\right)}{1-t} dt. \tag{5.33}$$

Next, positive-definiteness of Γ_t and the assumption that $B_j B_j^T = \mathrm{id}_{E^j}$ together justify the definition of a new process $(\widetilde{W}_t^j)_{0 \leq t \leq 1}$ via

$$d\widetilde{W}_t^j = (B_j \Gamma_t^2 B_j^T)^{-1/2} B_j \Gamma_t dW_t, \quad 1 \leq j \leq m.$$

By Lévy's characterization, this process is a Brownian motion, since it has quadratic covariation

$$[\widetilde{W}^j]_t = \int_0^t (B_j \Gamma_s^2 B_j^T)^{-1/2} B_j \Gamma_s^2 B_j^T (B_j \Gamma_s^2 B_j^T)^{-1/2} ds = t\,\mathrm{id}_{E^j}\,.$$

Putting things together, observe that definitions give

$$\int_0^1 (B_j \Gamma_t^2 B_j^T)^{1/2} d\widetilde{W}_t^j = B_j \int_0^1 \Gamma_t dW_t \sim B_j \sharp (\mu_1 \otimes \cdots \otimes \mu_k).$$

Thus, by (5.33) and an application of Lemmas 66 and 93, we have

$$\begin{aligned}\sum_{i=1}^{k} c_i D(\mu_i\|\gamma_{E_i}) &= \frac{1}{2}\int_0^1 \frac{\sum_{i=1}^k c_i \mathbb{E}\,\mathrm{Tr}\,((\Gamma_t^i - \mathrm{id}_{E_i})^2)}{1-t}dt \\ &\geq \frac{1}{2}\int_0^1 \frac{\sum_{j=1}^m d_j \mathbb{E}\,\mathrm{Tr}\,\left(((B_j\Gamma_t^2 B_j^T)^{1/2} - \mathrm{id}_{E^j})^2\right)}{1-t}dt \\ &\geq \sum_{j=1}^m d_j D(B_j\sharp(\mu_1\otimes\cdots\otimes\mu_k)\|\gamma_{E^j})\end{aligned}$$

We have equality throughout due to (5.30). Since X_t has full support for each $0 < t \leq 1$ and $(t,x) \mapsto \Gamma_t(x)$ is smooth by the regularizing properties of the heat semigroup, Lemma 66 and the above equality implies that

$$(\mathrm{id}_{E_0} - \Pi_{E^j})\Gamma_t(x)\Pi_{E^j} = 0, \quad x \in E_0,\ 0 < t < 1,\ 1 \leq j \leq m. \tag{5.34}$$

By definition, this implies that, for each $t \in (0,1)$, we have

$$(\mathrm{id}_{E_0} - \Pi_{E^j})\ \nabla^2 \log P_{1-t} f(x)\Pi_{E^j} = 0, \quad x \in E_0,\ 1 \leq j \leq m.$$

Since f is assumed regular by virtue of (5.31), the above also holds for $t = 1$ by continuity of the derivatives of the heat semigroup. Since $f = \prod_{i=1}^k f_i$ by definition, where each f_i is a density on E_i with respect to γ_{E_i}, the above imposes a block-diagonal structure on the Hessian of $\log f_i$, which can be summarized as

$$D^2(\log f_i)(x,y) = 0,$$

whenever x, y are vectors from distinct spaces in the decomposition (5.17). This implies, for each $1 \leq i \leq k$, that the density f_i has product form

$$f_i(x) = \prod_{\ell=0}^{n_i} f_{i,\ell}(\Pi_{K_\ell^i}(x)), \quad x \in E_i, \tag{5.35}$$

establishing necessity of (i).

Remark 67. *The above proof can be viewed as a modification of Eldan and Mikulincer [2020]'s argument for bounding the deficit in the Shannon–Stam inequality, suitable for setting of the AJN inequality. The emergence of the factorization (5.35) is new, and results from AJN-geometricity via the matrix inequality in Lemma 66. Although Valdimarsson's arguments in the context of the functional Brascamp–Lieb inequalities are slightly different, the same basic factorization emerges in Valdimarsson [2008, Lemma 13]. Hence, the above might be regarded as a combination of ideas from both Eldan and Mikulincer [2020] and Valdimarsson [2008]. In the next step, the Fourier analytic argument is effectively the same as that found in Valdimarsson [2008, Lemma 14], with the drift u_t playing the role of what Valdimarsson calls $\nabla \log F$.*

Necessity of condition (ii): Having established necessity of (i), the initial calculations in the proof of sufficiency hold, leading to the conclusion (5.28). The reduced datum $(\mathbf{c}, \mathbf{d}, \mathbf{B}_{K_{dep}})$ obtained by restricting the maps in $\mathbf{B}$ to domain K_{dep} remains AJN-geometric, so without loss of generality, we can assume for simplicity that there are no independent subspaces henceforth; i.e., $K_{dep} \equiv E_0$. As in the previous step, we let f denote the density of $X \sim \mu_1 \otimes \cdots \otimes \mu_k$ with respect to γ_{E_0}.

Letting definitions from the previous step prevail, Lemma 95 implies that u_t has linear growth in x for each $0 < t < 1$. Hence, we are justified in taking the Fourier transform, which we denote by $\hat{u}_t$. By (5.35), u_t is additively separable in the variables $\Pi_{E_j} x$ and $(\mathrm{id}_{E_0} - \Pi_{E_j})x$, and therefore $\hat{u}_t$ is supported on $H^j \cup (H^j)^\perp$ for each $1 \leq j \leq m$ (where H^j denotes the complex Hilbert space $E^j + \mathbf{i}E^j$). Similarly, since u_t is additively separable in the variables $\pi_{E_1}(x), \ldots, \pi_{E_k}(x)$, it follows that $\hat{u}_t$ is supported on $\cup_{i=1}^k H_i$ (where, $H_i := E_i + \mathbf{i}E_i$). Taking intersections, we find $\hat{u}_t$ is supported on the set

$$(H_1 \cup \cdots \cup H_k) \cap \bigcap_{j=1}^{m} (H^j \cup (H^j)^\perp) = \{0\},$$

where the equality follows by the assumption that there are no independent subspaces. A tempered distribution with Fourier transform supported at the origin is a polynomial [Rudin, 1991, p. 194], so the linear growth estimate in Lemma 95 implies that $x \mapsto u_t(x)$ is affine for each $0 < t < 1$. As a consequence of its defnition, Γ_t is therefore deterministic for each $0 < t < 1$, in the sense that $\Gamma_t(x)$ does not depend on x. We conclude from the representation $\int_0^1 \Gamma_t dW_t \overset{\text{law}}{=} X$ that X is Gaussian with covariance

$$\Sigma := \mathrm{Cov}(X) = \int_0^1 (\Gamma_t)^2 dt.$$

Note that Σ has diagonal form

$$\Sigma = \Pi(\Sigma_1, \ldots, \Sigma_k), \quad \Sigma_i \in \mathbf{S}_0^+(E_i), 1 \leq i \leq k \tag{5.36}$$

due to independence of the coordinates of X.

From (5.34), we have $\Pi_{E^j}\Sigma = \Pi_{E^j}\Sigma\Pi_{E^j}$ for each $1 \leq j \leq m$. This implies that if $v \in E_0$ is an eigenvector of Σ with eigenvalue λ, then $\Pi_{E^j} v$ is an eigenvector of $\Pi_{E^j}\Sigma\Pi_{E^j}$ with eigenvalue λ. In particular, if we consider the spectral decomposition $\Sigma = \sigma_1^2 \Pi_{K_{dep}^1} + \cdots \sigma_n^2 \Pi_{K_{dep}^n}$ with $\sigma_1^2, \ldots, \sigma_n^2$ distinct, then we have the orthogonal decomposition

$$B_j E_0 = \oplus_{\ell=1}^n B_j K_{dep}^\ell, \quad 1 \leq j \leq m, \tag{5.37}$$

where we note each K_{dep}^ℓ is product-form due to (5.36). To see that $E_0 = \oplus_{\ell=1}^n K_{dep}^\ell$ is a

critical decomposition, observe that

$$\sum_{i=1}^{k} c_i h(X_i) = \sum_{j=1}^{m} d_j h(B_j X) \tag{5.38}$$

$$= \sum_{\ell=1}^{n} \frac{1}{2} \log(2\pi e \sigma_\ell^2) \sum_{j=1}^{m} d_j \dim(B_j K_{dep}^\ell) \tag{5.39}$$

$$\geq \sum_{\ell=1}^{n} \frac{1}{2} \log(2\pi e \sigma_\ell^2) \sum_{i=1}^{k} c_i \dim(\pi_{E_i} K_{dep}^\ell) \tag{5.40}$$

$$= \sum_{i=1}^{k} c_i \sum_{\ell=1}^{n} \frac{\dim(\pi_{E_i} K_{dep}^\ell)}{2} \log(2\pi e \sigma_\ell^2) = \sum_{i=1}^{k} c_i h(X_i), \tag{5.41}$$

where (5.38) is the extremality assumption; (5.39) is due to (5.37) and the spectral decomposition of Σ; (5.40) is the dimension condition (3.6); and (5.41) follows due to the orthogonal decomposition $E_i = \oplus_{\ell=1}^{n} \pi_{E_i} K_{dep}^\ell$ for each $1 \leq i \leq k$, because each K_{dep}^ℓ is of product-form. Since we have equality throughout, this implies $K_{dep} \equiv E_0 = \oplus_{\ell=1}^{n} K_{dep}^\ell$ is a critical decomposition, as desired. Since $K_{dep}^1, \ldots, K_{dep}^n$ are eigenspaces of Σ, (ii) holds. □

Chapter 6

Extremizers of the Forward-Reverse Brascamp–Lieb inequalities

We recall from the introduction that a unifying generalization of various important inequalities such as sharp reverse Young Inequality, Brascamp–Lieb inequalities, and the Barthe inequalities is given by Forward-Reverse Brascamp–Lieb (FRBL) inequalities (1.8). As these inequalities have rather disparate equality conditions ranging from requiring Gaussianity, to requiring log-concavity, to requiring independence along various components, it is a priori not clear what form the extremizers of FRBL should take. As such, we discuss the most recent results we have on the extremizers of FRBL.

6.1 Preliminary definitions

We will use the same definitions of criticality, and equivalency of data as those of AJN, so we refer the reader to Chapter 5 for those. We recall the definition of an FRBL-geometric data from Courtade and Liu [2021]. Much like in AJN, extremizability of FRBL data means their equivalence to a geometric data.

Definition 68 (Geometric datum)**.** *A datum* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is said to be* **geometric** *if it satisfies the scaling condition* (3.10)*, and there exists a* $K \in \Pi(\mathrm{id}_{E_1}, \dots, \mathrm{id}_{E_k})$ *such that*

(i)

$$B_j K B_j^T = \mathrm{id}_{E^j} \quad \textit{for each } 1 \le j \le m \tag{6.1}$$

(ii) we have the operator identity

$$\sum_{j=1}^{m} d_j B_j^T B_j \le \mathrm{diag}(c_1 \,\mathrm{id}_{E_1}, \dots, c_k \,\mathrm{id}_{E_k}). \tag{6.2}$$

Furthermore, such a K *is called a* ***witness*** *to the geometricity of* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$*.*

Example 69. *Geometric Brascamp-Lieb data correspond to geometric* $(1, \mathbf{d}, \mathbf{B})$*, and Geometric Barthe data correspond to geometric* $(\mathbf{c}, 1, \mathbf{B})$*.*

Geometric data have the convenient property that $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = 0$, and they are extremizable by standard Gaussians. We summarize that as a formal proposition, which can be found in Courtade and Liu [2021]

Proposition 70. *If* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *is geometric, then* $D_g(\mathbf{c}, \mathbf{d}, \mathbf{B}) = 0$ *and* $X_i \sim N(0, \mathrm{id}_{E_i})$ *achieves equality in* (3.9)*, where an optimal coupling is given by* $X \sim N(0, K)$*, with* K *a witness to geometricity.*

The proof follows almost exactly as in the AJN case, so we omit the proof. Also, much like the AJN case, all extremizable instances are equivalent to geometric instances, so from now, we will restrict our attention to the geometric instances. [Courtade and Liu, 2021]. For notational convenience, we define $\Lambda_c := \mathrm{diag}(c_1 \,\mathrm{id}_{E_1}, \ldots, c_k \,\mathrm{id}_{E_k})$. We let

$$V := \ker(\Lambda_c - \sum_{j=1}^{m} d_j B_j^T B_j).$$

In particular, this is the subspace in which (6.2) is tight.

Now, we make the following assumption that holds in many important cases:

Assumption 71. *There exists a witness* K *such that*

$$\mathrm{range}(K) = V \tag{A}$$

In plain English, this assumes that there is an extremizer that is supported on all of V. Note that it does not assert that *every* extremizer will satisfy this property; just the existence of one that does. For each datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$, there is a naturally associated Hilbert space, which we will denote by H. H consists of vectors in E_0 equipped with the inner product

$$\langle x, y \rangle_H = x^T \Lambda_c y, \quad x, y \in E_0. \tag{6.3}$$

For $V \subset E_0$, we define the projection P_V to be the orthogonal projection of H onto V, where orthogonality is with respect to the inner product (6.3). For each datum $(\mathbf{c}, \mathbf{d}, \mathbf{B})$, we can now define the **reduced datum** $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ through the collection of maps $\mathbf{B}_V = (B_{V,j} : E_0 \to E^j)_{j=1}^m$ where

$$B_{V,j} := B_j P_V \quad 1 \le j \le m.$$

Datum is reduced if it is its own reduction; i.e. $\mathbf{B} = \mathbf{B}_V$. We can now extend the definitions of dependent and independent subspaces in Chapter 5 to FRBL. We start by defining the spaces

$$H_i := \mathrm{range}(P_V \pi_{E_i}^T), \quad \text{and} \quad H^j := \mathrm{range}(P_V (\Lambda_c)^{-1} B_j^T).$$

We further define $W^o := W$, and $W^{\perp_H}$ to be the orthogonal complement of W in H according to inner product (6.3). Next, given $a = (a_1, \dots, a_k) \in \{o, \perp_H\}^k$ and $b = (b_1, \dots, b_m) \in \{o, \perp_H\}^m$, we can define

$$H_{ab} := P_V \bigcap_{i=1}^{k} (H^j)^{a_i} \cap \bigcap_{j=1}^{m} (H^j)^{b_j}$$

We define the **independent subspace** T_{ab} be the smallest product-form subspace that contains H_{ab}. It is an expected (but not obvious) fact that distinct independent subspaces are orthogonal in H. Thus, we can define the **independent decomposition** of E_0:

$$E_0 = T_{dep} \oplus (\oplus_{a,b} T_{ab}), \tag{6.4}$$

where we define **the dependent space**, T_{dep}, to be the orthogonal complement of the direct sum of all independent subspaces. Finally, let $T_{i,ab} = \pi_{E_i} T_{ab}$, and $T_{i,dep} = \pi_{E_i} T_{dep}$ be the projections of these spaces to individual E_is. This induces the independent decomposition of E_i as

$$E_i = T_{i,dep} \oplus (\oplus_{a,b} T_{i,ab}), \tag{6.5}$$

which is an orthogonal decomposition since each T_a is product-form. Finally, we say that the law of X defined on a Hilbert space S with decomposition $\oplus_{l=1}^N S_l$ **splits** along that decomposition if $(\pi_{S_l}(X))_{l=1}^N$ are independent.

Remark 72. *Our definitions agree with the decompositions in Valdimarsson [2008] for Brascamp–Lieb data and Boroczky et al. [2022] for the Barthe data.*

6.2 Main result and examples

Now, we are ready to write the main result that is being advertised with this chapter:

Theorem 73. *Let $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ be a geometric datum. If $(X_i)_{i=1}^k$ are centered and extremal for $(\mathbf{c}, \mathbf{d}, \mathbf{B})$, then, under Assumption (A), we must have:*

(i) The law of X_i splits along the independent decomposition of E_i for each $1 \le i \le k$.

(ii) There exists an optimal coupling $X^* \in \Pi(X_1, \dots, X_k)$ such that the law of X^* splits along the independent decomposition of E_0.

(iii) If for any $a \in \{o, \perp_H\}^k$ such that $\#\{i \mid a_i = o\} > 1$, then $\pi_{T_{ab}}(X^*) = 0$ almost surely, or it is log-concave.

(iv) Each of the components $(\pi_{T_{i,dep}}(X_i))_{i=1}^k$ is Gaussian.

Furthermore, if the datum is reduced, then $(i) - (iv)$ are sufficient for the extremizers.

Examples

Example 74. *We described the extremizers for Brascamp–Lieb in Chapter 5. We note that as an instance of FRBL, Brascamp-Lieb data can be represented as $K = \mathrm{id}_{E_0}$. Furthermore, since K is full rank, the assumption (A) holds. Therefore, we can indeed characterize all the extremizers. Therefore, any extremal X splits along the independent decomposition. This is compatible with Valdimarsson [2008]'s characterization of the extremizers in the functional form of the Brascamp–Lieb inequalities.*

Example 75. *Note that Barthe inequalities (3.8) are a special case of FRBL as well. We note that their data is reduced, and it follows that we can characterize their extremizers exactly. In particular, we know from Boroczky et al. [2022] that in the case of Barthe, E^1 splits into independent components that are either free, required to be log-concave, or required to be Gaussian, which then pull back to individual E_is through isometries $B_{E_i,1} : E_i \to E^1$. As the Barthe data are always reduced, we get the full necessity and sufficiency of the extremizers.*

Example 76. *An important case of the Barthe-Wolff inequalities [Barthe and Wolff, 2018] looks like the following:*

$$\prod_{i=1}^{k}\left(\int_{E_i} f_i(x_i)\,dx_i\right)^{c_i} \le \prod_{j=1}^{m}\left(\int g_j\right)^{d_j}\int_{E_0}\prod_{i=1}^{k} f_i^{c_i}(x_i)\prod_{j=1}^{m} g_j^{-d_j}(B_j x)\,dx. \tag{6.6}$$

We note that the above corresponds to the FRBL datum $(\mathbf{c}, (1, \mathbf{d}), \{\mathrm{id}_{E_0}\} \cup \mathbf{B})$ (see Courtade and Liu [2021]). Since one of the linear maps is the identity map, we immediately get that the equivalent geometric datum has to have a full rank witness. This full-rank property implies that there are no nontrivial independent subspaces T_{ab} with $\#\{i \mid a_i = o\} > 1$. As such, extremizers of (6.6) split along the independent decomposition, with the dependent space consisting of a Gaussian component, and the independent spaces split to be freely chosen. Note that these are the only Gaussian-extremizable instances of Barthe–Wolff inequalities.

Example 77. *Consider the following entropic inequality for $X_1, X_2 \in \mathcal{P}(\mathbb{R}^n)$:*

$$(\lambda + t)h(X_1) + (1 - \lambda + t)h(X_2) \le \max_{\Pi(X_1,X_2)} h(\lambda X_1 + (1-\lambda)X_2) + t h(X_1, X_2)$$

which is an entropic instance of the reverse Young inequality. From our result, we know that the only extremizer of the inequality above for any $t > 0$ is a Gaussian X_t whose covariance tends to $\mathrm{id}_{\mathbb{R}^{2n}}$ as $t \to \infty$. As such, for any $\zeta > 0$, there exists a finite $t > 0$ such that the extremizer of

$$\lambda h(X_1) + (1 - \lambda)h(X_2) \le \max_{\Pi(X_1,X_2)} h(\lambda X_1 + (1-\lambda)X_2) + t I(X_1; X_2)$$

has $I(X_1; X_2) \le \zeta$, and that extremizer is a Gaussian. This is precisely the the case of the dependent EPI (4.2) for when $\zeta \in (0, \infty)$. For the end conditions $\zeta = 0$ and $\zeta = \infty$, we

remember that (4.2) reduces to the EPI and the Cover–Zhang type inequality discussed in the introduction (the entropic form of the Prékopa–Leindler inequality), whose extremizers are Gaussian, and log-concave, respectively.

6.3 Ideas of the proof

Structural Theory

We will give a sketch of the proof for Theorem 73. First off, note that we have to show that the characterization we have given is actually well-defined, so we first need to show that $T_{ab} \perp T_{(ab)'}$ in H for distinct choices of (a, b) and $(a, b)'$. Once we do that, a crucial idea that helps move forward with the proof is the observation that for any critical subspace, the subdata $(\mathbf{c}, \mathbf{d}, \mathbf{B}_T)$ and $(\mathbf{c}, \mathbf{d}, \mathbf{B}_{E_0/T})$ are critical, which allows constructing *a good witness*. While we will not define what a good witness is, we will illustrate it with an example:

Example 78. *Consider the simple example*

$$h(X_1) + h(X_2) \leq \sup_{X \in \Pi(X_1, X_2)} h(\pi_{E_1}(X)) + h(\pi_{E_2}(X))$$

For this example,

$$K_\rho = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \qquad |\rho| \leq 1$$

are all valid witnesses, but given the inequality essentially splits along E_1 and E_2, a good witness that exhibits that behaviour is K_0.

This idea of a good witness helps us split our instances and data into sub-instances, so we can specialize into cases where our domain is a single T_{ab}, or T_{dep}. Furthermore, we can show that our decomposition of E_0 implies a decomposition of $(E^j)_{j=1}^m$s, namely,

$$B_{V,j} T_{ab} \perp B_{V,j} T_{(a,b)'}$$

for all $1 \leq j \leq m$, and $(a, b) \in \{o, \perp_H\}^{k \times m}$.

Note that in the reduced case, when we restrict our domain to a single T_{ab}, we will be working in a setting where range(B_j^T) agrees with range$(B_{j'}^T)$, which follows from the definition of independent subspaces. Turns out, in these settings, we can characterize the unique witness to geometricity.

Theorem 79. *Let $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ be a geometric datum, with $\dim(E_i) = \dim(E^j) = d$ and $R(B_j^T) = R(B_1^T)$ for all $i, j \in [k] \times [m]$. Further, let $\sum_{i=1}^k c_i = 1$. The matrix*

$$K = (\Lambda_c)^{-1} B_1^T B_1 (\Lambda_c)^{-1}$$

is the unique witness to geometricity.

Remark 80. *The assumption* $\sum_{i=1}^k c_i = 1$ *is without any loss of generality.*

Note that in this restrictive condition $\mathrm{range}(B_j^T) = \mathrm{range}(B_1^T)$ for all $j = 1, \dots, m$, for some suitably chosen linear maps $(A_j)_{j=1}^m$, we can work with $B_j' := B_j A_j$s so that $B_1' X^* = B_j' X^*$ for all $1 \le j \le m$, i.e. we can already observe that the distributions of $(B_j' X^*)_{j=1}^m$ have to be equal almost surely.

Stochastic Argument

We will essentially reproduce the stochastic argument in Chapter 5 under Assumption (A). Let us set the relevant notation. Let f_i denote the density of $\mu_i := \mathcal{L}(X_i)$ with respect to γ_{E_i}, and define the function

$$u_t^i(x) := \nabla \log P_{1-t} f_i(x), \qquad x \in E_0,\ 0 \le t \le 1,$$

where $(P_t)_{t \ge 0}$ denotes the heat semigroup. Note that this is again the Föllmer drift in Theorem 94, driving the SDE

$$dX_t^i = du_t^i(X_t^i)dt + dW_t^i$$

where we note that $(W_t^i)_{t=0}^1$ is a Brownian motion on E_i stemming from Brownian motion $(W_t)_{t=0}^1$ on E_0. As mentioned in Chapter 5, we can also define the matrix-valued function

$$\Gamma_t^i(x) := (1-t)\nabla u_t^i(x) + \mathrm{id}_{E_i}, \qquad x \in E_0,\ 0 \le t \le 1, \tag{6.7}$$

Note that $\Gamma_t^i \in \mathbf{S}^+(E_i)$ because we can work with regularized μ_i as done in (5.31) and then reapply Lemma 95. We define

$$\Gamma_t := \mathrm{diag}(\Gamma_t^1, \dots, \Gamma_t^k).$$

Note that if we define $X_t^* = \int_0^t \Gamma_s(X_s) K^{1/2} dW_s$, then we observe that

$$\pi_i \int_0^1 \Gamma_t(X_t) K^{1/2} dW_t \sim \mu_i, \quad 1 \le i \le k$$

so that X_1^* is indeed a coupling of $(X_i)_{i=1}^k$. We denote $\Gamma_{V,t}(x) := \Gamma_t \circ P_V$. Finally, we use the shorthand $L_d := (d_1 \,\mathrm{id}_{E^1}, \dots, d_m \,\mathrm{id}_{E^m})$.

Our main result of this section is the following:

Theorem 81. *Let the setting of Theorem 73 hold. Then, for a good witness* K*, we have*

(i) There exists an extremal coupling X^* *supported on* V*.*

(ii) $B_j \Gamma_{V,t} K B_j^T = B_j K \Gamma_{V,t} B_j^T$*.*

(iii) $(\mathrm{id}_{E_0} - K B_j B_j^T) K \Gamma_{V,t} B_j^T = 0$*.*

(iv) $(\Lambda_c - B^T L_d B)\Gamma_{V,t} K = 0$.

Proof of Theorem 81. For the proof, we use the shorthand $\Gamma_t := \Gamma_t(X_t)$. Similar to the AJN case, we construct a drift on $(E_j)_{j=1}^m$ by considering $(B_j\Gamma_t K \Gamma_t B_j^T)_{j=1}^m$. We can define [1]

$$d\widetilde{W}_t^j = (B_j\Gamma_t K\Gamma_t B_j^T)^{-1/2} B_j \Gamma_t K^{1/2} dW_t, \quad 1 \le j \le m.$$

By Lévy's characterization, this process is a Brownian motion, since it has quadratic covariation

$$[\widetilde{W}^j]_t = \int_0^t (B_j\Gamma_s^2 B_j^T)^{-1/2} B_j \Gamma_s^2 B_j^T (B_j\Gamma_s^2 B_j^T)^{-1/2} ds = t \ \mathrm{id}_{E^j}.$$

Now, remembering (5.33) reproduced below

$$D(\mu_i \| \gamma_{E_i}) = \frac{1}{2}\int_0^1 \frac{\mathbb{E}\,\mathrm{Tr}\left((\Gamma_t^i - \mathrm{id}_{E_i})^2\right)}{1-t} dt. \tag{6.8}$$

we have

$$\begin{aligned}
\sum_{i=1}^k c_i D(\mu_i\|\gamma_{E_i}) &= \sum_{i=1}^k c_i \frac{1}{2}\int_0^1 \frac{\mathbb{E}\,\mathrm{Tr}\left((\Gamma_t^i - \mathrm{id}_{E_i})^2\right)}{1-t} dt \\
&= \frac{1}{2}\int_0^1 \frac{1}{1-t}\mathbb{E}\mathrm{Tr}(\Lambda_c(\Gamma_t - \mathrm{id}_{E_0})K(\Gamma_t - \mathrm{id}_{E_0}))dt \\
&\ge \frac{1}{2}\int_0^1 \frac{1}{1-t}\mathbb{E}\sum_{j=1}^m d_j \mathrm{Tr}(B_j^T B_j(\Gamma_t - \mathrm{id}_{E_0})K(\Gamma_t - \mathrm{id}_{E_0})) \\
&= \sum_{j=1}^m d_j \frac{1}{2}\int_0^1 \frac{1}{1-t}\mathbb{E}\mathrm{Tr}(B_j\Gamma_t K\Gamma_t B_j^T - 2B_j K\Gamma B_j^T + \mathrm{id}_{E_j}) \\
&\ge \sum_{j=1}^m d_j \frac{1}{2}\int_0^1 \frac{1}{1-t}\mathbb{E}\mathrm{Tr}(((B_j\Gamma_t K\Gamma_t B_j^T)^{1/2} - \mathrm{id}_{E_j})^2) \\
&\ge \sum_{j=1}^m d_j D(B_j\mu\|\gamma_{E^j})
\end{aligned}$$

where the equality conditions in the above proof match the $(ii) - (iv)$ in Theorem 81 for Γ_t instead of $\Gamma_{V,t}$ on the support of domain of Γ_t, which is the support of $\mathcal{L}(X_t)$. Now, under assumption (A), the good witness will ensure that this coupling is supported over all of V, and we can replace Γ_t with $\Gamma_{V,t}$. We end the proof by mentioning that the equality conditions imply the tightness of the frame condition along the coupling, which implies the equality condition for the entropy inequality is the same as the relative entropy inequality. □

[1] Here, we assume that $(B_j\Gamma_t K\Gamma_t B_j^T)$ is invertible, which will indeed be the case when K is a good witness.

Now that we can work with $\Gamma_{V,t}$, we observe that the invariant spaces of $\Gamma_{V,t}$ are precisely the independent and the dependent spaces, which ensures that $\Gamma_{V,t}$ splits along those subspaces, giving the necessary conditions for the extremizers. Now, for the dependent data, we run a similar Fourier transform argument like we did in Chapter 5, where we use the fact that the only tempered distribution with Fourier transform supported at the origin is a polynomial, so the linear growth estimate we have implies $\Gamma_{V,t}$ is constant, implying Gaussianity. For the independent subspaces, by weak duality, the best we can say is the following:

Theorem 82. *Let* $(\mathbf{c}, \mathbf{d}, \mathbf{B})$ *be geometric with a single independent subspace and no dependent subspace. Then, any extremizer* $(X_i)_{i=1}^k$ *share a common density* e^{-f} *where* f *satisfies*

$$\sum_{i=1}^{k} c_i f(x_i) \geq \sum_{j=1}^{m} d_j f(A_j^{-1/2} B_j x), \quad x \in E_0 \tag{6.9}$$

for some isometries $(A_j)_{j=1}^m$ *depending on the datum.*

We can further show that this condition implies log-concavity when $k > 1$, and if the data is reduced, then every log-concave distribution satisfies (6.9). However, for non-reduced data, we can get rather unexpected behavior. For instance, consider the following entropy inequality:

$$\frac{1}{3}h(X_1) + \frac{2}{3}h(X_2) \leq \sup_{X \in \Pi(X_1, X_2)} \frac{1}{3}h((\underbrace{\begin{bmatrix} -\frac{2}{30} & \frac{32}{30} \end{bmatrix}}_{B_1} X) + \frac{2}{3}h(\underbrace{\begin{bmatrix} \frac{16}{30} & \frac{14}{30} \end{bmatrix}}_{B_2} X)$$

We note that the associated data is geometric with the unique witness $K = \mathbf{1}\mathbf{1}^T$. As such, the coupling would have to have $X_1 = X_2$ a.s., implying they share a common density e^{-f} with f satisfying

$$\frac{1}{3}f(x_1) + \frac{2}{3}f(x_2) \geq \frac{1}{3}f(B_1 x) + \frac{2}{3}f(B_2 x) \tag{6.10}$$

Equivalently, we could have deduced (6.10) from Theorem 82 where we would note that $A_1 = A_2 = 1$ for this particular datum. We observe that while $f(x) = |x|$ satisfies (6.10), $f(x) = x^4$ does not. Thus, extremizers of FRBL inequalities arguably involve a finer class than those of Barthe or Brascamp–Lieb inequalities.

Chapter 7

Concluding remarks and outlook

While we have covered a variety of new results within the dissertation, there remains many interesting questions to be answered.

7.1 Outlook

We mark several avenues for possible future work.

1. An even bigger generalization than Theorem 21

We make no assertion that Theorem 21 is a grand unification of all entropy inequalities on Euclidean space. Indeed, there are several important examples of inequalities that are not obviously subsumed. Results in Liu and Viswanath [2007], Geng and Nair [2014], Courtade [2017] provide representative examples. We concede that there may be some clever application of Theorem 21 that can recover some of these results, but we do not know of one at the time of this writing. Thus, at the moment, it seems that Theorem 21 may be another piece in a larger puzzle still wanting to be put together.

2. Removing the Euclidean structure on various inequalities

Throughout the thesis, we have assumed a Euclidean structure. However, it has been known for some time that Brascamp-Lieb inequality holds on the sphere [Carlen et al., 2004]. More recently, there has been work that has pushed the Brascamp-Lieb inequalies onto a broader class of Riemannian manifolds [Barthe et al., 2011]. Since Brascamp-Lieb is but a special case of a larger class of inequalities, there is hope that there can be an FRBL-like inequality defined on other, more interesting spaces.

3. Stability of the entropic inequalities

Once we establish the equality conditions for entropic inequalities, it is a natural question to ask for stability results. This was done for Prékopa-Leindler in Böröczky and De [2021] and for EPI in Courtade et al. [2018], Eldan and Mikulincer [2020]. One can reasonably use duality or the Föllmer drifts to extend the results to more general entropic inequalities, such as AJN and FRBL.

4. Quantum Entropy Inequalities

An interesting area where Brascamp-Lieb inequalities have been extended to is quantum physics, as (von Neumann)-entropy has a physical meaning regarding the wave-function [Von Neumann, 1932]. We note that analogues of EPI [König and Smith, 2014], and Brascamp-Lieb inequalities [Berta et al., 2019] exist in the quantum setting already, so there is potential for quantum information theorists to investigate further generalizations of entropic inequalities.

5. Extensions to convex geometry

As discussed in Chapter 3, Brunn-Minkowski inequality is just a special case of Cover-Zhang result. Another important geometric inequality, Santaló inequality [Santaló, 1949], was recently discovered to be a dual of an entropic inequality in Fathi [2018]. It is reasonable to assume that clever uses of Theorem 21 may give rise to other useful inequalities that may find use in convex geometry.

Concluding remarks

Why do we care about Gaussians? As Poincaré remarked about Gaussians, "*Tout le monde y croit cedependant, me disait un jour M. Lippmann, car les experimentaeurs, s'imaginent que c'est un théoreme de mathématiques, et les mathématiciens que c'est un fait expérimental.*" [Poincaré, 1896].

Appendix A

A.1 Markov Semigroups

While we will not extensively rely on particular properties of semigroups, there are many semigroups that are inherent in our analyses. The following is taken from Bakry et al. [2014] specialized to Euclidean spaces.

For the rest of the section $(P_t)_{t\geq 0}$ will be a collection of maps where $P_t : (E \to \mathbb{R}) \to (E \to \mathbb{R})$ for each $t \geq 0$. We say that P_t has a **stationary measure** μ if

$$\int_E P_t f d\mu = \int_E f d\mu$$

for all bounded, positive $f : E \to \mathbb{R}$, and for all $t \geq 0$.

Remark 83. *μ need not be a probability measure.*

Definition 84. *Let $(P_t)_{t\geq 0}$ have a stationary measure μ. We say $(P_t)_{t\geq 0}$ is a **Markov semigroup** if it satisfies the following:*

(i) *For every bounded measurable $f : E \to \mathbb{R}$, $P_t f$ is also bounded measurable.*

(ii) *$P_t f \geq 0$ if $f \geq 0$.*

(iii) *$P_t(1) = 1$ where $1(x) = 1$ for all $x \in E$.*

(iv) *$P_{t+s} = P_t \circ P_s$ for all $t, s \geq 0$.*

(v) *$P_0 f = f$*

(vi) *$P_t(f) \to f$ in $L^2(E, \mu)$ as $t \to 0$.*

Note that semigroups can be characterized by their (infinitesimal) **generators** $L := \frac{\partial}{\partial t} P_t \mid_{t=0}$. The properties above ensure that generators are well-defined.

We say $(P_t)_{t\geq 0}$ is the **heat semigroup** if its generator is $L = \Delta$. As such, the evolution of a density under the heat semigroup solves the heat equation. This allows the heat semigroup to inherit various smoothness properties from the heat equation from PDEs. Here is an example from Polyanskiy and Wu [2016, Proposition 2]).

Lemma 85. *Let $(P_t)_{t\geq 0}$ denote the heat semigroup, and let ρ denote the density of a random variable X with finite variance. Then,*

$$|\nabla \log P_s\rho(x)| \leq c_s(|x|+1), \quad s>0$$

for some finite constant c_s depending only on s and the second moment of X.

Alternatively, $P_t\rho$ is the density of $X+tZ$ where ρ is the density of X, $Z\sim\gamma$, and X and Z are independent. (Heat semigroup is quite special in that it is self-adjoint.)

Similarly, another semigroup of interest is the **Ornstein–Uhlenbeck (O–U) semigroup**, which has the generator $L := \Delta - x\cdot\nabla$. Alternatively, it has the representation:

$$P_t f(x) := \mathbb{E}_{Z\sim\gamma}[f(e^{-t}x + \sqrt{1-e^{-2t}}Z)]$$

Note that one can immediately deduce from the represenation above that a stationary measure of the O-U semigroup is γ. Indeed, by solving the associated PDE derived from L, one can show that it is the only stationary distribution. We end this section with an important result that concerns both analysts and information theorists ([Bakry et al., 2014, Proposition 5.2.2]):

Theorem 86 (de Bruijn's identity for O–U)**.** *Let $(P_t)_{t\geq 0}$ be the O–U semigroup. Let $v_t := (P_t)^*v$, i.e. $dv_t = P_t f d\mu$ where $f = \frac{dv}{d\mu}$. Then,*

$$\frac{\partial}{\partial t}D(v_t\|\gamma) = -I(v_t\|\gamma) := -\int \frac{|\nabla f|^2}{f}\,d\gamma.$$

Let us put it to use. One of the most concise proofs of the Log–Sobolev inequality (LSI) for Gaussians, which has found extensive use in statistics literature, especially in the realm of convergence of diffusions used for sampling [Vempala and Wibisono, 2019], is a few lines with the EPI and de Bruijn. Note that the Gaussian LSI also has been used to construct a Bayesian Cramer–Rao bound, that can be used to further prove various minimax bounds [Lee and Courtade, 2020].

Theorem 87 (Log–Sobolev Inequality)**.** *For any measure μ defined on $\mathbb{R}^n$ such that $D(\mu\|\gamma) < \infty$, we have*

$$D(\mu\|\gamma) \leq \frac{1}{2}I(\mu\|\gamma).$$

Proof of Theorem 87. Take $X\sim\mu$, $Y\sim\gamma$, with X and Y independent. For convenience, denote μ_t the distribution of $e^{-t}X + \sqrt{1-e^{-2t}}Y$, or equivalently as $(P_t)^*\mu$, where $(P_t)_{t\geq 0}$ is the O-U semigroup. We consider the instance of entropy power inequality with $\lambda = e^{-2t}$. Converting the entropies into relative entropies as in equation (2.3), and noting that $\mathrm{var}(\mu) = \mathrm{var}(\mu_t)$, we have

$$-e^{-2t}D(\mu\|\gamma) \leq -D(\mu_t\|\gamma)$$

where we note that there is equality when $t = 0$. As such, assuming differentiability,

$$\frac{d}{dt} - e^{-2t} D(\mu \| \gamma) \mid_{t=0} \leq \frac{d}{dt} - D(\mu_t \| \gamma) \mid_{t=0} .$$

Now, using Theorem (86), the result follows. □

Log–Sobolev inequalities form a strong foundation for concentration inequalities, as normally subgaussianity does not tensorize, but the above inequality very much does! [Van Handel, 2014]

A.2 Itô calculus

Itô calculus allows doing stochastic partial differential equations. To do so, we first need to set the spaces we want to work in.

Let $\mathbb{W} := C^0([0,1], E_0) = \{\omega \mid \omega : [0,\infty) \to E_0, \omega(0) = 0\}$ denote the set of continuous functions from $[0,\infty)$ to E_0, and equip it with the topology of uniform convergence. Furthermore, let $\mathcal{B}$ be the associated Borel σ-algebra. Finally, let $\mathcal{G} := (\mathcal{G}_t)_{t\in[0,1]}$ be the filtration generated by the coordinate maps $w \to w_t$. We will assume our Wiener space $(\mathbb{W}, \mathcal{B})$ will always come equipped with this filtration.

For any measure on this space, we will denote a Wiener process by $W = (W_t)_{t\in[0,1]}$. Similarly, we let $\mathbb{P}$ be the Wiener measure on $(\mathbb{W}, \mathcal{B})$. Finally, we denote the coordinate process $X = (X_t)_{t\in[0,1]}$ where $X_t(\omega) = \omega_t$. We note that one can set $W = X$, i.e. construct a Wiener process by the coordinate maps on $\mathbb{W}$, if the underlying measure is the Wiener measure $\mathbb{P}$.

We call $(u_t)_{t\geq 0}$ **a drift** if it is a process adapted to $\mathcal{G}$ and is in $L_2([0,1])$ $\mathbb{P}-$almost surely, that is, $\int_0^1 |u_t|^2 dt < \infty$ $\mathbb{P}$-almost surely.

Remark 88. *Lehec [2013] refers to $U_t := \int_0^t u_s ds$ as the drift, however, Eldan and Mikulincer [2020] refers to u_t as the drift. We will consciously stick with the latter notation, as u_t will be our drift coefficient in an upcoming SDE.*

We will regularly be writing $\int_0^1 u_t dW_t$ in this thesis where $(W_t)_{t\in[0,1]}$ is a Brownian motion, and u_t is some drift. For the reader uninitiated in Itô calculus, those integrals should be read in a Riemann–Stieltjes sense, that is:

$$\int_0^1 u_t dW_t := \lim_{n\to\infty} \sum_{i=0}^{n} u_t(W_{i/n})(W_{i+1/n} - W_{i/n}).$$

Note that for regular Riemann integrals, it does not matter which endpoint one chooses, but here it does. In particular, the choice of the left-hand point ensures that integrals of form $\int u_t dB_t$ are all martingales, and we shall use that quite extensively. Of course, for this definition to be valid, we need some sort of a control on how "big" the differences

$(W_{i+1/n} - W_{i/n})$ can be, and having that control helps form the backbone of Itô calculus. In particular, we have the following from Øksendal [2003, Lemma 3.1.7]:

Theorem 89 (Itô Isometry)**.**

$$\mathbb{E}([\int_0^1 u_t dW_t)^2] = \mathbb{E}[\int_0^1 \|u_t\|^2 ds]$$

Föllmer Drifts

Let's now setup Girsanov's theorem for our setting along with Novikov's condition, taken from Øksendal [2003]:

Theorem 90. *Let $(W_t)_{t\in[0,1]}$ be a standard Brownian motion defined on a filtered probability space with underlying measure $\mathbb{P}$ (i.e. $(W_t)_{t\in[0,1]}$ is a standard Brownian motion under $(\mathbb{W}, \mathcal{B}, \mathcal{G}, \mathbb{P})$), and let $u = (u_t)_{t\in[0,1]}$ be a drift adapted to the same filtration satisying Novikov's condition ($\mathbb{E}[e^{\int_0^1 |u_s|^2 ds}] < \infty$). Construct a random measure $\mathbb{Q}$ such that*

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = e^{\int_0^1 u_s dW_s + \int_0^1 |u_s|^2 ds}.$$

Under $\mathbb{Q}$, $(W_t + \int_0^t u_s ds)_{t\in[0,1]}$ is a standard Brownian motion.

Remark 91. *In the above, we have explicitly noted the Novikov's condition, but for us, by localization and lower semicontinuity of entropy, we will just need the $\mathbb{P}$-almost sure finiteness written before.*

The following proposition is an important reason for us looking into drifts:

Proposition 92. *Let $(W_t)_{t\in[0,1]}$ be a standard Brownian motion defined on a filtered probability space with underlying measure $\mathbb{P}$, and let $u = (u_t)_{t\in[0,1]}$ be a drift adapted to the same filtration. Then, if $W_1 + \int_0^1 u_s ds \sim \mu$, we have*

$$D(\mu\|\gamma) \leq \int_0^1 \mathbb{E}_{\mathbb{P}}|u_s|^2 ds$$

Proof. Note that under Girsanov's theorem, we have,

$$D(\mathbb{P}\|\mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[\log \frac{d\mathbb{Q}}{d\mathbb{P}}] = \int_0^1 \mathbb{E}_{\mathbb{P}}[|u_s|^2] ds$$

where we note that the martingale term disappears.

Now, let $T : C([0,1]) \to \mathbb{R}$ with $T(f) = f(1)$. Note that $\mu = T\#\mathbb{P}$, and $\gamma = T\#\mathbb{Q}$. The result follows from the data processing inequality for KL divergence.

□

Sometimes, we will want finer control in constructing our drift. In particular, we want to be able to construct matrix-valued processes that still allow us to construct a bound on KL divergence. We will need the following adaptation of the previous proposition taken from Eldan and Mikulincer [2020]; we sketch the proof for completeness.

Lemma 93. *Fix a Euclidean space E. Consider a filtered probability space carrying an E-valued Brownian motion $(W_t)_{\geq 0}$, and let $(F_t)_{\geq 0}$ be an adapted process taking values in $\mathbf{S}_0^+(E)$. If $\int_0^1 F_t dW_t \sim \mu$, then*

$$D(\mu\|\gamma_E) \leq \frac{1}{2}\int_0^1 \frac{\mathbb{E}\,\mathrm{Tr}\left((F_t - \mathrm{id}_E)^2\right)}{1-t}dt.$$

Proof. Define the drift

$$u_t = \int_0^t \frac{F_s - \mathrm{id}_E}{1-s} dW_s.$$

We claim that $W_1 + \int_0^1 u_t dt \sim \mu$. To see this, write

$$\begin{aligned}\int_0^1 F_t dW_t = \int_0^1 \mathrm{id}_E\, dW_t + \int_0^1 (F_t - \mathrm{id}_E) dW_t &= W_1 + \int_0^1 \int_t^1 \frac{F_t - \mathrm{id}_E}{1-t} ds dt \\ &= W_1 + \int_0^1 u_s ds,\end{aligned}$$

where we used the stochastic Fubini theorem. Now, by Proposition 92 and the data processing inequality, Itô isometry, and Fubini's theorem, we have

$$\begin{aligned}D(\mu\|\gamma_E) \leq \frac{1}{2}\int_0^1 \mathbb{E}|u_t|^2 dt &= \frac{1}{2}\int_0^1 \int_0^t \frac{\mathbb{E}\,\mathrm{Tr}\left((F_s - \mathrm{id}_E)^2\right)}{(1-s)^2} ds dt \\ &= \frac{1}{2}\int_0^1 \frac{\mathbb{E}\,\mathrm{Tr}\left((F_s - \mathrm{id}_E)^2\right)}{1-s} ds.\end{aligned} \tag{A.1}$$

□

Circling back to the entropy bound above, a natural question to ask is when is Proposition 92 tight? Following the proof, we see that the only inequality invoked is data processing. In particular, if the Radon-Nikodym derivative between the (random) measure $\mathbb{Q}$ and $\mathbb{P}$ only depends on the last coordinate of ω, we would have equality in the data processing inequality. Föllmer's result is that for measures μ such that $D(\mu\|\gamma) < \infty$, there is always a drift that meets Proposition 92 with equality.

Theorem 94. *Let μ be a measure such that $D(\mu\|\gamma) < \infty$, with $\mu = f d\gamma$. Consider the Wiener space $(\mathbb{W}, \mathcal{B})$ equipped with the measure $\mathbb{P}_\mu$, where $\mathbb{P}_\mu$ that satisfies*

$$\frac{d\mathbb{P}_\mu}{d\mathbb{P}} := f(w_1)$$

where $\mathbb{P}$ is the Wiener measure on $(\mathbb{W}, \mathcal{B})$.

Then, the following holds:

(i) $u_t := \nabla \log P_{1-t} f(X_t)$ *is a drift.*

(ii) $(X_t - \int_0^t u_s ds)_{t\in[0,1]}$ *is a Brownian motion under* $\mathbb{P}_\mu$.

(iii) For the drift $u_t := \nabla \log P_{1-t} f(W_t)$*, we have:*

$$D(\mu\|\gamma) = \int_0^1 \mathbb{E}_{\mathbb{P}}|u_s|^2 ds$$

As we have now established the form of the Föllmer drift, it makes sense to control how big it is. It turns out, it is quite well-behaved.

Lemma 95. *Let* $(P_t)_{t\geq 0}$ *be the heat semigroup, and let* $X \sim \mu \in \mathcal{P}(E)$ *have density* $d\mu = f d\gamma_E$*. For each* $0 < t < 1$*, there is a constant* C *depending only on* t *and the second moments of* X *such that*

$$|\nabla \log P_{1-t} f(x)| \leq C(|x| + 1), \quad x \in E.$$

If, moreover, μ *is of the form* $\mu = \nu * \gamma_E$*, then*

$$\nabla^2 \log(P_{1-t} f(x)) \in \mathbf{S}_0^+(E), \quad x \in E, 0 < t < 1.$$

Proof. Let ρ denote the density of X with respect to Lebesgue measure on E. By direct calculation, we can reparametrize $P_{1-t}f$ in terms of ρ as

$$P_{1-t} f(x) = \left(\frac{2\pi}{t}\right)^{\dim(E)/2} e^{\frac{|x|^2}{2t}} P_{\frac{1-t}{t}} \rho(x/t)$$

Hence,

$$\nabla \log P_{1-t} f(x) = \frac{1}{t} x + \frac{1}{t} \nabla (\log P_{\frac{1-t}{t}} \rho)(x/t). \tag{A.2}$$

Now, Lemma 85 states

$$|\nabla \log P_s \rho(x)| \leq c_s(|x| + 1), \quad s > 0$$

for some finite constant c_s depending only on s and the second moments of ρ. Hence, the first claim follows.

For the second claim, we have $\rho = P_1 \rho_0$ for some density ρ_0. Hence, by the semigroup property combined with (A.2), we have

$$\nabla^2 \log P_{1-t} f(x) = \frac{1}{t} \operatorname{id}_E + \frac{1}{t^2} \nabla^2 (\log P_{\frac{1}{t}} \rho_0)(x/t).$$

By a simple convexity calculation [Eldan and Lee, 2014, Lemma 1.3], it holds that $\nabla^2(\log P_s g) \geq -\frac{1}{s} \operatorname{id}_E$ for any density g and $s > 0$, so we find

$$\nabla^2 \log P_{1-t} f(x) \geq \left(\frac{1}{t} - \frac{1}{t}\right) \operatorname{id}_E = 0.$$

□

We will conclude this chapter with the construction of a matrix valued process that meets the upper bound in Lemma 93 with equality.

Theorem 96. *Fix a Euclidean space E. Consider a filtered probability space carrying an E-valued Brownian motion $(W_t)_{\geq 0}$, and let $(u_t)_{\geq 0}$ denote the Föllmer process. Consider the matrix valued function*

$$\Gamma_t := (1-t)\nabla u_t + \mathrm{id}_E\,.$$

Then, the following holds:

(i)

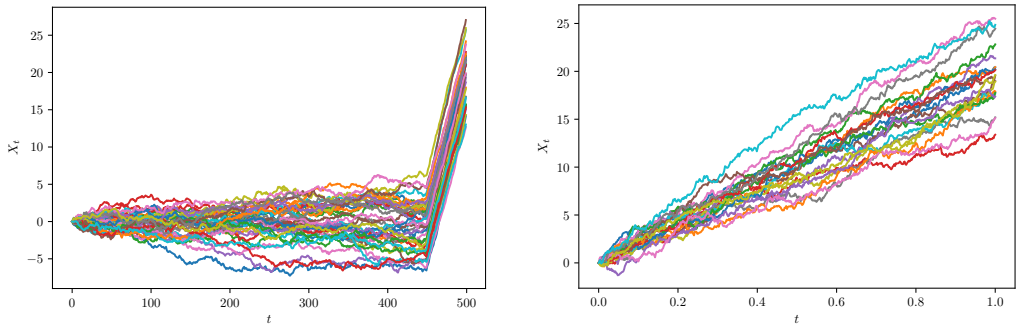
$$\int_0^1 \Gamma_t dW_t \sim W_1 + \int_0^1 u_t dt$$

(ii)

$$D(\mu\|\gamma_E) = \frac{1}{2}\int_0^1 \frac{\mathbb{E}\,\mathrm{Tr}\left((\Gamma_t - \mathrm{id}_E)^2\right)}{1-t}dt.$$

If, moreover, μ is of the form $\mu = \nu * \gamma_E$, then

$$\Gamma_t \in \mathbf{S}_0^+(E), \quad x \in E, 0 < t < 1.$$



(a) Drift $u_1(t) = 20 \cdot 1_{t\in[0.9,1]}$ (b) Drift $u_2(t) = 2$

Figure A.1: Example sample paths from two diffusions spanning from two drift coefficients. Note that they give rise to the same marginal distribution at $t = 1$, namely, $\gamma(2, 1)$. Here, we note that just because two drifts give rise to the same distribution at time 1 does not mean that their time-averaged expected square norms are equal. Föllmer's result essentially states that there is a minimum-norm drift whose expected squared norm will indeed be its relative entropy from a Gaussian.

A.3 Auxillary Proofs

Informal Proof of Theorem 10

While one can piece a proof from the existing literature, and it does follow from Liu et al. [2018] and Dubuc [1977], we will provide an original (albeit only informal) information-theoretic one, relying on the heavy machinery of optimal transport and duality principles. It is most similar to Barthe [1998], but we do our best to stick with entropic forms, as opposed to functional versions. We will assume that all possible densities for X and Y are restricted to densities that are sufficiently regular [Caffarelli, 1992], though normally one can just regularize later [Villani, 2003, Remark 6.6]. In particular, we will assume the existence of the optimal transport maps, and that the densities satisfy the so-called Mange-Ampere equations.

Proposition 97. *Let $\phi : E \to \mathbb{R}$ be a convex map, and let $h(X)$, $h(\nabla\phi(X)) < \infty$. Then, we have:*

$$h(\nabla\phi(X)) = h(X) + \mathbb{E}[\log|\nabla^2\phi(X)|$$

where ∇ is the gradient, and ∇^2 is the Hessian.

Informal Proof of Proposition 97. We note that if the density of X is f, and f is nice, then, the density of $\nabla\phi(X)$ is $f(\nabla\phi(X))|\nabla^2\phi(X)|$. The result then follows from definitions. □

Informal Proof of Theorem 10. (i) We will first show the inequality by constructing a coupling for which the result holds. In particular, consider coupling laws of X and Y such that X and Y be the pushforwards of a standard Gaussian Z on $\mathbb{R}^n$, that is, $X = T_1(Z)$, $Y = T_2(Z)$. We then have:

$$\lambda h(X) + (1-\lambda)h(Y) = h(Z) + \lambda\mathbb{E}[\log|\nabla T_1(Z)|] + (1-\lambda)\mathbb{E}[\log|\nabla T_2(Z)|]$$

Similarly, we have:

$$h(X+Y) = h(Z) + \mathbb{E}[\log|(\lambda\nabla T_1 + (1-\lambda)\nabla T_2)(Z)|] \leq \sup_{(X',Y')\in\Pi((X,Y)} h(\lambda X' + (1-\lambda)Y')$$

By strict concavity of logdet, we have that

$$\lambda h(X) + (1-\lambda)h(Y) \leq h(\lambda X + (1-\lambda)Y) \leq \sup_{(X',Y')\in\Pi((X,Y)} h(\lambda X' + (1-\lambda)Y')$$

as desired.

(ii) Note that since Z has full-support, we need

$$\log|(\lambda\nabla T_1 + (1-\lambda)T_2)(x)| = \lambda|\nabla T_1(Z)| + (1-\lambda)|\nabla T_2(Z)|.$$

Since logdet is strongly concave, this implies $\nabla T_1 = \nabla T_2$, i.e. X and Y have to have the same distribution ν with density h for equality. For notational convenience, let $B := [\lambda, 1-\lambda]$. We will further assume that a strong duality principle holds in the argument below.

$$\begin{aligned}
&\sup_{(X',Y')\in\Pi((X,Y)} h(\lambda X' + (1-\lambda)Y') \\
&= \sup_{(X',Y')\in\Pi((X,Y)} \inf_f \int_E f d(B\#\mu) - \log\int_E e^f \\
&= \sup_{\mu\in M_+(\mathbb{R}^{2n})} \inf_{f,g} \int_E f d\mu - \log\int_E e^f + \lambda(\int g_1 d\nu - \int g_1\mu_1) + (1-\lambda)(\int g_2 d\nu - \int g_2\mu_2) \\
&= \inf_{\lambda g_1(x_1)+(1-\lambda)g_2(x_2)\leq f(Bx)} -\log\int_E e^f + \int(\lambda g_1 + (1-\lambda)g_2)d\nu \\
&\geq \inf_{\lambda g_1(x_1)+(1-\lambda)g_2(x_2)\leq f(Bx)} -\log\int_E e^{g_1} - \log\int_E e^{g_2} + \int(\lambda g_1 + (1-\lambda)g_2)d\nu \\
&\geq h(X)
\end{aligned}$$

where we invoke the Prékopa-Leindler inequality for the first inequality, and variational form for entropy second. Note that the equality in the swapping of sup and inf is strong duality. Now, to enforce equality for the variational principle, we will want to pick $g_1 = g_2 = \log h$ by Jensen's so now we want:

$$\inf_{\lambda \log h(x_1)+(1-\lambda)\log h(x_2)\le f(Bx)} -\log \int_E e^f = 0$$

By monotonicity of the integrand, we want

$$f(y) = \sup_{x_1,x_2|y=\lambda x_1+(1-\lambda)x_2} \left[\lambda \log h(x_1) + (1-\lambda)\log h(x_2)\right].$$

In particular, we can always take $x_1 = x_2 = y$ to observe that $f(y) \ge \log h(y)$. However, to ensure that $\int_E e^f = 0$, this inequality better be strict almost surely, i.e. $f = \log h$. Now, we are done, as

$$\log h(y) = \sup_{x_1,x_2|y=\lambda x_1+(1-\lambda)x_2} \left[\lambda \log h(x_1) + (1-\lambda)\log h(x_2)\right]$$

is the definition of concavity.

□

Informal proof of FRBL duality

In the above example of finding equality conditions for the Cover–Zhang-type inequalities, we invoked a powerful duality principle. It turns out, we can go much farther. The following is a proof that mimics the formal proof for Kantorovich Duality in Villani [2003]. While only informal, it captures the essence of the proof in a few lines. The reader can look at Liu et al. [2018], Liu [2018] for a rigorous exposition involving Fenchel (strong) duality.

Theorem 17 $\implies$ Theorem 19

Let $(f_i)_{i=1}^k$, $(g_j)_{j=1}^m$ be given. We can construct $(X_i)_{i=1}^k$ such that

$$\sum_{i=1}^k c_i h(X_i) = \sum_{i=1}^k c_i \left(\log \int f_i + \mathbb{E}[\log f_i(X_i)]\right)$$

Assume the f_is give rise to a distribution $X_i \in \mathcal{P}(E_i)$.

Now, we have

$$\begin{aligned}
&\sum_{i=1}^{k} c_i \left(\log \int f_i - \mathbb{E}[\log f_i(X_i)] \right) \\
&= \sum_{i=1}^{k} c_i h(X_i) \\
&\leq \sum_{j=1}^{m} d_j h(B_j X) + D(\mathbf{c}, \mathbf{d}, \mathbf{B}) \\
&\leq \sum_{j=1}^{m} d_j \left(\log \int g_j - \mathbb{E}[\log g_j(B_j X^*)] \right) + D(\mathbf{c}, \mathbf{d}, \mathbf{B}) + \varepsilon
\end{aligned}$$

where we pick $X^* \in \Pi(X_1, \ldots, X_k)$ to be a coupling that is at most ε away from $\max_{X \in \Pi(X_1,\ldots,X_k)} \sum_{j=1}^{m} d_j h(B_j X)$. Of course, if given $(f_i)_{i=1}^k$, $(g_j)_{j=1}^m$ satisfy (3.12), then, for any coupling X^*, we have $\mathbb{E}[\log g_j(B_j X^*)] \geq \mathbb{E}[\log f_i(X_i)]$. As $\varepsilon > 0$ is arbitrary, the result follows.

Theorem 19 $\implies$ Theorem 17

We have, formally,

$$\begin{aligned}
&\max_{X \in \Pi(\mu_1,\ldots,\mu_k)} \sum_{j=1}^{m} d_j h(B_j X) \\
&= \sup_{X \in M_+} \inf_{v_i \in C(E_i)} \sum_{j=1}^{m} d_j h(B_j X) + \sum_{i=1}^{k} c_i \int v_i \mu_i - \sum_{i=1}^{k} c_i \mathbb{E}[v_i(X_i)] \\
&= \sup_{X \in M_+} \inf_{v_i \in C(E_i)} \inf_{u_j \in C(E^j)} \sum_{j=1}^{m} d_j (\log \int e^{-u_j} + \mathbb{E} u_j(B_j X)) + \sum_{i=1}^{k} c_i \int v_i \mu_i - \sum_{i=1}^{k} c_i \mathbb{E}[v_i(X_i)] \\
&= \inf_{v_i, u_j} \sum d_j \log \int e^{-u_j} + \sum_{i=1}^{k} \int v_i \mu_i + \sup_X (\mathbb{E} u_j(B_j X)) - \mathbb{E}[\sum_{i=1}^{k} c_i v_i(X_i)]) \\
&= \inf_{v_i, u_j, \sum v_i \circ \pi_i \geq \sum d_j u_j \circ B_j} \sum_{j=1}^{m} d_j \log \int e^{-u_j} + \sum_{i=1}^{k} c_i \int v_i \mu_i \\
&\geq \inf_{v_i} \sum_{i=1}^{k} c_i \log \int e^{-v_i} + \sum_{i=1}^{k} c_i \int v_i \mu_i \\
&\geq \sum_{i=1}^{k} c_i h(X_i)
\end{aligned}$$

For the proof above, note that the first equality follows from representing a joint probability measure as a solution to a restricted optimization problem, and the rest follow from (2.2), and the asserted strong duality.

Bibliography

[1] Venkat Anantharam, Varun Jog, and Chandra Nair. Unifying the Brascamp–Lieb inequality and the entropy power inequality. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1847–1851. IEEE, 2019.

[2] Tsuyoshi Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear algebra and its applications*, 26:203–241, 1979.

[3] Efe Aras and Thomas A Courtade. Sharp maximum-entropy comparisons. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1504–1509. IEEE, 2021.

[4] Efe Aras and Thomas A Courtade. Entropy inequalities and gaussian comparisons. *arXiv preprint arXiv:2206.14182*, 2022.

[5] Efe Aras, Thomas A Courtade, and Albert Zhang. Equality cases in the Anantharam–Jog–Nair inequality. *arXiv preprint arXiv:2206.11809*, 2022.

[6] Shiri Artstein, Keith Ball, Franck Barthe, and Assaf Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004.

[7] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

[8] Keith Ball. Volumes of sections of cubes and related problems. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1987–88*, pages 251–260. Springer, 1989.

[9] Andrew R Barron. Entropy and the central limit theorem. *The Annals of probability*, pages 336–342, 1986.

[10] Franck Barthe. On a reverse form of the Brascamp–Lieb inequality. *Inventiones mathematicae*, 134(2):335–361, 1998.

[11] Franck Barthe. The Brunn–Minkowski theorem and related geometric and functional inequalities. In *International Congress of Mathematicians*, volume 2, pages 1529–1546. Citeseer, 2006.

[12] Franck Barthe and Pawel Wolff. Positive Gaussian kernels also have Gaussian minimizers. *arXiv preprint arXiv:1805.02455*, 2018.

[13] Franck Barthe, Dario Cordero-Erausquin, Michel Ledoux, and Bernard Maurey. Correlation and Brascamp–Lieb inequalities for Markov semigroups. *International Mathematics Research Notices*, 2011(10):2177–2216, 2011.

[14] William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, 102(1): 159–182, 1975.

[15] Jonathan Bennett, Anthony Carbery, Michael Christ, and Terence Tao. The Brascamp–Lieb inequalities: finiteness, structure and extremals. *Geometric and Functional Analysis*, 17(5):1343–1415, 2008.

[16] Patrick Bergmans. A simple converse for broadcast channels with additive white Gaussian noise (corresp.). *IEEE Transactions on Information Theory*, 20(2):279–280, 1974.

[17] Henry T Bernstein. J. Clerk Maxwell on the history of the kinetic theory of gases, 1871. *Isis*, 54(2):206–216, 1963.

[18] Mario Berta, David Sutter, and Michael Walter. Quantum Brascamp–Lieb dualities. *arXiv preprint arXiv:1909.02383*, 2019.

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

[20] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

[21] Ludwig Boltzmann. Weitere studien über das wärmegleichgewicht unter gasmolekülen. sitzungberichte der kaiserlichen akademie der wissenschaften. mathematisch-naturwissenschaftliche classe 66, 275–370. *History of modern physical sciences*, 1: 262–349, 1872.

[22] Károly J Böröczky and Apratim De. Stability of the prékopa-leindler inequality for log-concave functions. *Advances in Mathematics*, 386:107810, 2021.

[23] Karoly J Boroczky, Pavlos Kalantzopoulos, and Dongmeng Xi. The case of equality in geometric instances of Barthe's reverse Brascamp–Lieb inequalit. *arXiv preprint arXiv:2203.01428*, 2022.

[24] Herm J Brascamp, Elliott H Lieb, and Joaquin Mazdak Luttinger. A general rearrangement inequality for multiple integrals. *Journal of functional analysis*, 17(2):227–237, 1974.

[25] Herm Jan Brascamp and Elliott H Lieb. Best constants in Young's inequality, its converse, and its generalization to more than three functions. *Advances in Mathematics*, 20(2):151–173, 1976.

[26] Hermann Brunn. *Über Ovale und Eiflächen*. Akademische Buchdruckerei von R. Straub, 1887.

[27] Luis A Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[28] Anthony Carbery. The Brascamp–Lieb inequalities: recent developments. *Nonlinear Analysis, Function Spaces and Applications*, pages 9–34, 2007.

[29] Eric A Carlen and Dario Cordero-Erausquin. Subadditivity of the entropy and its relation to Brascamp–Lieb type inequalities. *Geometric and Functional Analysis*, 19 (2):373–405, 2009.

[30] Eric A Carlen and Avraham Soffer. Entropy production by block variable summation and central limit theorems. *Communications in mathematical physics*, 140:339–371, 1991.

[31] Eric A Carlen, Elliott H Lieb, and Michael Loss. A sharp analog of Young's inequality on s_n and related entropy inequalities. *The Journal of Geometric Analysis*, 14:487–520, 2004.

[32] Wei-Kuo Chen, Nikos Dafnis, and Grigoris Paouris. Improved hölder and reverse hölder inequalities for gaussian random vectors. *Advances in Mathematics*, 280:643–689, 2015.

[33] Rudolf Clausius. *Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie: vorgetragen in der naturforsch. Gesellschaft den 24. April 1865.* éditeur inconnu, 1865.

[34] Max Costa. On the Gaussian interference channel. *IEEE transactions on information theory*, 31(5):607–615, 1985.

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

[36] Max Costa and Thomas Cover. On the similarity of the entropy power inequality and the Brunn–Minkowski inequality (corresp.). *IEEE Transactions on Information Theory*, 30(6):837–839, 1984.

[37] Thomas A Courtade. A strong entropy power inequality. *IEEE Transactions on Information Theory*, 64(4):2173–2192, 2017.

[38] Thomas A Courtade. Personal Communication, 2023.

[39] Thomas A Courtade and Jingbo Liu. Euclidean forward–reverse Brascamp–Lieb inequalities: Finiteness, structure, and extremals. *The Journal of Geometric Analysis*, 31(4):3300–3350, 2021.

[40] Thomas A Courtade, Max Fathi, and Ashwin Pananjady. Quantitative stability of the entropy power inequality. *IEEE Transactions on Information Theory*, 64(8):5691–5703, 2018.

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

[42] Thomas M Cover and Zhen Zhang. On the maximum entropy of the sum of two dependent random variables. *IEEE Transactions on Information Theory*, 40(4):1244–1246, 1994.

[43] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[44] A De Moivre. Approximatio ad summam terminorum binomii (a+ b) n in seriem expansi (1733). *An English translation is included in The Doctrine of Chances (second edition, 1738, and third edition, 1756)*, 1733.

[45] Amir Dembo, Thomas M Cover, and Joy A Thomas. Information theoretic inequalities. *IEEE Transactions on Information theory*, 37(6):1501–1518, 1991.

[46] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.

[47] Serge Dubuc. Critères de convexité et inégalités intégrales. In *Annales de l'institut Fourier*, volume 27, pages 135–165, 1977.

[48] Albert Einstein et al. On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat. *Annalen der physik*, 17(549-560):208, 1905.

[49] Ronen Eldan and James R Lee. Regularization under diffusion and anti-concentration of temperature. *preprint*, 2014.

[50] Ronen Eldan and Dan Mikulincer. Stability of the Shannon–Stam inequality via the Föllmer process. *Probability Theory and Related Fields*, 177(3):891–922, 2020.

[51] Max Fathi. A sharp symmetrized form of Talagrand's transport-entropy inequality for the Gaussian measure. *Electronic Communications in Probability*, 23:1–9, 2018.

[52] JBJ Fourier. Theorie de la propagation de la chaleur dans les solides, 234 pp, 1807.

[53] Richard Gardner. The Brunn–Minkowski inequality. *Bulletin of the American Mathematical Society*, 39(3):355–405, 2002.

[54] Ankit Garg, Leonid Gurvits, Rafael Oliveira, and Avi Wigderson. Algorithmic and optimization aspects of Brascamp–Lieb inequalities, via operator scaling. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 397–409, 2017.

[55] Yanlin Geng and Chandra Nair. The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages. *IEEE Transactions on Information Theory*, 60(4):2087–2104, 2014.

[56] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pages 354–375. PMLR, 2013.

[57] Shunsuke Ihara. On the capacity of channels with additive non-Gaussian noise. *Information and Control*, 37(1):34–39, 1978.

[58] Oliver Johnson. A conditional entropy power inequality for dependent variables. *IEEE transactions on information theory*, 50(8):1581–1583, 2004.

[59] Robert König and Graeme Smith. The entropy power inequality for quantum systems. *IEEE Transactions on Information Theory*, 60(3):1536–1548, 2014.

[60] PS Laplace. Mémoire sur les approximations des formules qui sont fonctions de très-grands nombres, et sur leur application aux probabilités. mémoires l'institut 1809, 353-415 and 559-565 (supplement), reproduced in oeuvres complètes de laplace, 1878. *Mémoires de l'Académie Royale des Sciences*, 10:207–291, 1810.

[61] Jimmie D Lawson and Yongdo Lim. The geometric mean, matrices, metrics, and more. *The American Mathematical Monthly*, 108(9):797–812, 2001.

[62] Kuan-Yun Lee and Thomas Courtade. Minimax bounds for generalized linear models. *Advances in Neural Information Processing Systems*, 33:9372–9382, 2020.

[63] Joseph Lehec. Representation formula for the entropy and functional inequalities. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 885–899, 2013.

[64] S Leung-Yan-Cheong and Ma Hellman. The Gaussian wire-tap channel. *IEEE transactions on information theory*, 24(4):451–456, 1978.

[65] Elliott H Lieb. Proof of an entropy conjecture of wehrl. *Communications in Mathematical Physics*, 62(1):35–41, 1978.

[66] Elliott H Lieb. Gaussian kernels have only Gaussian maximizers. In *Inventiones mathematicae*, pages 179–208. 1990.

[67] Jingbo Liu. *Information theory from a functional viewpoint*. PhD thesis, Princeton University, 2018.

[68] Jingbo Liu, Thomas A Courtade, Paul Cuff, and Sergio Verdú. Brascamp–Lieb inequality and its reverse: An information theoretic view. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1048–1052. IEEE, 2016.

[69] Jingbo Liu, Thomas A Courtade, Paul W Cuff, and Sergio Verdú. A forward-reverse Brascamp–Lieb inequality: Entropic duality and Gaussian optimality. *Entropy*, 20(6): 418, 2018.

[70] Tie Liu and Pramod Viswanath. An extremal inequality motivated by multiterminal information-theoretic problems. *IEEE Transactions on Information Theory*, 53(5): 1839–1851, 2007.

[71] Mokshay Madiman, James Melbourne, and Peng Xu. Forward and reverse entropy power inequalities in convex geometry. In *Convexity and concentration*, pages 427–485. Springer, 2017.

[72] Hermann Minkowski. *Geometrie der zahlen*. BG Teubner, 1910.

[73] Mehdi Mohseni and John M Cioffi. A proof of the converse for the capacity of gaussian mimo broadcast channels. In *2006 IEEE International Symposium on Information Theory*, pages 881–885. IEEE, 2006.

[74] Edward Nelson. The free Markoff field. *Journal of Functional Analysis*, 12(2):211–227, 1973.

[75] Aleksandar Nikolov and Mohit Singh. Maximizing determinants under partition constraints. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 192–201, 2016.

[76] Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

[77] Mark S Pinsker. Calculation of the rate of information production by means of stationary random processes and the capacity of stationary channel. In *Dokl. Akad. Nauk USSR*, volume 111, pages 753–756, 1956.

[78] Henri Poincaré. Calcul des probabilités. leçons professées pendant le deuxième semestre 1893-1894, 1896.

[79] Henri Poincaré. *Calcul des probabilités*. Gauthier-Villars, 1912.

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

[81] Olivier Rioul. Information theoretic proofs of entropy power inequalities. *IEEE Transactions on Information Theory*, 57(1):33–55, 2010.

[82] Olivier Rioul and Ram Zamir. Equality in the matrix entropy-power inequality and blind separation of real and complex sources. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1842–1846. IEEE, 2019.

[83] Walter Rudin. Functional analysis, international series in pure and applied mathematics (mcgraw-hill, new york, 1991). 1991.

[84] Luis A Santaló. Un invariante afin para los cuerpos convexos del espacio de n dimensiones. *Portugaliae mathematica*, 8:155–161, 1949.

[85] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[86] Saul Stahl. The evolution of the normal distribution. *Mathematics magazine*, 79(2): 96–113, 2006.

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

[88] J Michael Steele. *The Cauchy–Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

[89] Seiji Takano, S Watanabe, M Fukushima, Y Prohorov, and A Shiryaev. The inequalities of Fisher information and entropy power for dependent variables. In *Proceedings of the 7th Japan-Russia Symposium on Probability Theory and Mathematical Statistics, Tokyo*, pages 460–470. World Scientific, 1995.

[90] Ender Tekin and Aylin Yener. The Gaussian multiple access wire-tap channel with collective secrecy constraints. In *2006 IEEE International Symposium on Information Theory*, pages 1164–1168. IEEE, 2006.

[91] Stefán Ingi Valdimarsson. Optimisers for the Brascamp–Lieb inequality. *Israel Journal of Mathematics*, 168(1):253–274, 2008.

[92] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[93] Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

[94] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[95] Cédric Villani. A short proof of the" concavity of entropy power". *IEEE Transactions on Information Theory*, 46(4):1695–1696, 2000.

[96] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

[97] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

[98] John Von Neumann. *Mathematische grundlagen der quantenmechanik*. Die grundlehren der mathematischen wissenschaften in einzeldarstellungen, bd. XXXVIII. J. Springer, Berlin, 1932.

[99] Liyao Wang and Mokshay Madiman. Beyond the entropy power inequality, via rearrangements. *IEEE Transactions on Information Theory*, 60(9):5116–5137, 2014.

[100] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[101] Hanan Weingarten, Yossef Steinberg, and Shlomo Shitz Shamai. The capacity region of the Gaussian multiple-input multiple-output broadcast channel. *IEEE transactions on information theory*, 52(9):3936–3964, 2006.

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

[103] Peiyuan Zhang, Jingzhao Zhang, and Suvrit Sra. Minimax in geodesic metric spaces: Sion's theorem and algorithms. *arXiv preprint arXiv:2202.06950*, 2022.