Information Theoretic Inequalities

Amir Dembo, Thomas M. Cover, *Fellow, IEEE*, and Joy A. Thomas, *Member, IEEE*

Invited Paper

Abstract—**The role of inequalities in information theory is reviewed and the relationship of these inequalities to inequalities in other branches of mathematics is developed.**

Index Terms—**Information inequalities, entropy power, Fisher information, uncertainty principles.**

I. Preface: Inequalities in Information Theory

INEQUALITIES in information theory have been driven by a desire to solve communication theoretic problems. To solve such problems, especially to prove converses for channel capacity theorems, the algebra of information was developed and chain rules for entropy and mutual information were derived. Fano's inequality, for example, bounds the probability of error by the conditional entropy. Some deeper inequalities were developed as early as Shannon's 1948 paper. For example, Shannon stated the entropy power inequality in order to bound the capacity of non-Gaussian additive noise channels.

Information theory is no longer restricted to the domain of communication theory. For this reason it is interesting to consider the set of known inequalities in information theory and search for other inequalities of the same type. Thus motivated, we will look for natural families of information theoretic inequalities.

For example, the entropy power inequality, which says that the entropy of the sum of two independent random vectors is no less than the entropy of the sum of their independent normal counterparts, has a strong formal resemblance to the Brunn Minkowski inequality, which says that the volume of the set sum of two sets is greater than or equal to the volume of the set sum of their spherical counterparts. Similarly, since the exponentiated entropy is a measure of volume it makes sense to consider the surface area of the volume of the typical set associated with a given probability density. Happily, this turns out to be another information quantity, the Fisher information.

A large number of inequalities can be derived from a strengthened Young's inequality. These inequalities include the entropy power inequality, the Brunn Minkowski inequality and the Heisenberg uncertainty inequality. These inequalities are extreme points of the set of inequalities derivable from a central idea. Logically independent derivations of these inequalities exist and are based on Fisher information inequalities such as the Cramér–Rao inequality.

Turning our attention to simple inequalities for differential entropy, we apply them to the standard multivariate normal to furnish new and simpler proofs of the major determinant inequalities in classical mathematics. In particular Hadamard's inequality, Ky Fan's inequality and others can be derived this way. Indeed we find some new matrix inequalities by this method. Moreover the entropy power inequality, when specialized to matrices, turns out to yield Minkowski's determinant inequality, yet another tangency with the Minkowski of Brunn–Minkowski.

In the process of finding determinant inequalities we derive some new differential entropy inequalities. We restate one of them as follows. Suppose one is looking at ocean waves at a certain subset of points. Then the average entropy per sample of a random subset of samples can be shown to increase as the number of sampling points increases. On the other hand, the per sample conditional entropy of the samples, conditioned on the values of the remaining samples, monotonically decreases. Once again using these entropy inequalities on the standard multivariate normal leads to associated matrix inequalities and in particular to an extension of the sequence of inequalities found by Hadamard and Szasz.

By turning our attention from the historically necessary inequalities to the natural set of inequalities suggested by information theory itself, we find, full circle, that these inequalities turn out to be useful as well. They improve determinant inequalities, lead to overlooked inequalities for the entropy rate of random subsets and demonstrate the unity between physics, mathematics, information theory and statistics (through unified proofs of the Heisenberg, entropy power, Fisher information and Brunn–Minkowski inequalities).

The next section is devoted to differential entropy inequalities for random subsets of samples. These inequalities when specialized to multivariate normal vari-

Manuscript received February 1, 1991. This work was supported in part by the National Science Foundation under Grant NCR-89-14538 and in part by JSEP Contract DAAL 03-91-C-0010. A Dembo was supported in part by the SDIO/IST, managed by the Army Research Office under Contract DAAL 03-90-G-0108 and in part by the Air Force Office of Scientific Research, Air Force Systems Command under Contract AF88-0327. Sections III and IV are based on material presented at the IEEE/CAM Workshop on Information Theory, 1989.

A. Dembo is with the Statistics Department, Stanford University, Stanford, CA 94305.

T. M. Cover is with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.

J. Thomas was with Stanford University. He is now with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598.

IEEE Log Number 9103368.

ables provide the determinant inequalities presented in Section V. Section III focuses on the entropy power inequality (including the related Brunn–Minkowski, Young's and Fisher information inequalities) while Section IV deals with various uncertainty principles and their interrelations.

II. Information Inequalities

A. Basic Inequalities

In this section, we introduce some of the basic information theoretic quantities and a few well-known simple inequalities using convexity. We assume throughout that the vector $X = (X_1, X_2, \cdots, X_n)$ has a probability density $f(x_1, x_2, \cdots, x_n)$. We need the following definitions.

Definition: The *entropy* $h(X_1, X_2, \cdots, X_n)$, sometimes written $h(f)$, is defined by

$$h(X_1, X_2, \cdots, X_n) = -\int f \ln f = E(-\ln f(X)).$$

The entropy may be infinite and it is well defined as long as either $E(\max\{\ln f(X), 0\})$ or $E(\max\{-\ln f(X), 0\})$ are finite.

The entropy is a measure of the number of bits required to describe a random variable to a particular accuracy. Approximately $b + h(X)$ bits suffice to describe X to b-bit accuracy. Also, $e^{(2/n)h(X)}$ can be interpreted as the effective support set size for the random variable X. This point is further explored in Section III.

Definition: The functional

$$D(f\|g) = \int f(x) \ln (f(x)/g(x))\, dx$$

is called the *relative entropy*, where f and g are probability densities.

The relative entropy $D(f\|g)$ is also known as the Kullback Leibler *information number*, *information for discrimination*, and *information distance*. We also note that $D(f\|g)$ is the error exponent in the hypothesis test of density f versus g.

Definition: The *conditional entropy* $h(X|Y)$ of X given Y is defined by

$$h(X|Y) = -\int f(x, y) \ln f(x|y)\, dx\, dy.$$

We now observe certain natural properties of these information quantities.

Lemma 1: $D(f\|g) \geq 0$, with equality iff $f = g$ a.e.

Proof: Let A be the support set of f. Then, by Jensen's inequality,

$$-D(f\|g) = \int_A f \ln (g/f)$$

$$\leq \ln \int_A f(g/f) = \ln \int_A g \leq \ln 1 = 0,$$

with equality only if $g/f = 1$, a.e., by the strict concavity of the logarithm (see [18], [29]). □

Lemma 2: If (X, Y) have a joint density, then $h(X|Y) = h(X, Y) - h(Y)$.

Proof:

$$h(X|Y) = -\int f(x, y) \ln f(x|y)\, dx\, dy$$

$$= -\int f(x, y) \ln (f(x, y)/f(y))\, dx\, dy$$

$$= -\int f(x, y) \ln f(x, y)\, dx\, dy + \int f(y) \ln f(y)\, dy$$

$$= h(X, Y) - h(Y). \quad \square$$

Lemma 3: $h(X|Y) \leq h(X)$, with equality iff X and Y are independent.

Proof:

$$h(X) - h(X|Y) = \int f(x, y) \ln (f(x|y)/f(x))$$

$$= \int f(x, y) \ln (f(x, y)/f(x)f(y)) \geq 0,$$

by $D(f(x, y)\|f(x)f(y)) \geq 0$. Equality implies $f(x, y) = f(x)f(y)$, a.e., by strict concavity of the logarithm. □

Lemma 4 (Chain Rule, Subadditivity of the Entropy):

$$h(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} h(X_i|X_{i-1}, X_{i-2}, \cdots, X_1)$$

$$\leq \sum_{i=1}^{n} h(X_i)$$

with equality iff $X_1, X_2, \cdots, X_n$ are independent.

Proof: The equality is the chain rule for entropies, which we obtain by repeatedly applying Lemma 2. The inequality follows from Lemma 3, and we have equality iff $X_1, X_2, \cdots, X_n$ are independent. □

We will also need the entropy maximizing property of the multivariate normal. Throughout we denote by $\phi_K(x)$ the joint density of the multivariate normal vector with zero-mean and covariance K.

Lemma 5: Let the random vector $X \in \boldsymbol{R}^n$ have zero-mean and covariance $K = EXX^t$, i.e., $K_{ij} = EX_iX_j$, $1 \leq i, j \leq n$. Then $h(X) \leq \frac{1}{2} \ln (2\pi e)^n |K|$, with equality iff $f(x) = \phi_K(x)$.

Proof: Let $g(x)$ be any density satisfying $\int g(x) x_i x_j\, dx = K_{ij}$, for all i, j. Then,

$$0 \leq D(g\|\phi_K)$$

$$= \int g \ln (g/\phi_K)$$

$$= -h(g) - \int g \ln \phi_K$$

$$= -h(g) - \int \phi_K \ln \phi_K$$

$$= -h(g) + h(\phi_K), \quad (1)$$

where the substitution $\int g \ln \phi_K = \int \phi_K \ln \phi_K$ follows from

the fact that g and ϕ_K yield the same expectation of the quadratic form $\ln \phi_K(x)$. □

B. Subset Inequalities for Entropy

Motivated by a desire to prove Szasz's generalization of Hadamard's inequality in Section V, we develop a new inequality on the entropy rates of random subsets of random variables.

Let $X_1, X_2, \cdots, X_n$ be a set of n random variables with an arbitrary joint distribution. Let S be any subset of the indices $\{1,2,\cdots,n\}$. We will use $X(S)$ to denote the subset of random variables with indices in S and S^c to denote the complement of S with respect $\{1,2,\cdots,n\}$. For example, if $S=\{1,3\}$, then $X(S)=\{X_1, X_3\}$ and $X(S^c)=\{X_2, X_4, X_5, \cdots, X_n\}$. Recall that the entropy $h(X)$ of a random vector $X \in R^k$ with density function $f(x)$ is

$$h(X) = -\int f(x) \ln f(x)\, dx.$$

If $S=\{i_1, i_2, \cdots, i_k\}$, let

$$h(X(S)) = h(X_{i_1}, X_{i_2}, \cdots, X_{i_k}).$$

Let

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:\,|S|=k} \frac{h(X(S))}{k}$$

be the entropy rate per element for subsets of size k averaged over all k-element subsets. Here, $h_k^{(n)}$ is the average entropy in bits per symbol of a randomly drawn k-element subset of $\{X_1, X_2, \cdots, X_n\}$. This quantity is monotonically nonincreasing in k as stated in the following theorem (due to Han [27]).

Theorem 1:

$$h_1^{(n)} \geq h_2^{(n)} \geq \cdots \geq h_n^{(n)}.$$

Proof: (Following [16]). We first prove the inequality $h_n^{(n)} \leq h_{n-1}^{(n)}$. We write

$$\begin{aligned}
h(X_1, X_2, \cdots, X_n) &= h(X_1, X_2, \cdots, X_{n-1}) \\
&\quad + h(X_n | X_1, X_2, \cdots, X_{n-1}), \\
h(X_1, X_2, \cdots, X_n) &= h(X_1, X_2, \cdots, X_{n-2}, X_n) \\
&\quad + h(X_{n-1} | X_1, X_2, \cdots, X_{n-2}, X_n) \\
&\leq h(X_1, X_2, \cdots, X_{n-2}, X_n) \\
&\quad + h(X_{n-1} | X_1, X_2, \cdots, X_{n-2}), \\
&\vdots \\
h(X_1, X_2, \cdots, X_n) &\leq h(X_2, X_3, \cdots, X_n) + h(X_1).
\end{aligned}$$

Adding these n inequalities and using the chain rule, we obtain

$$\begin{aligned}
&nh(X_1, X_2, \cdots, X_n) \\
&\quad \leq \sum_{i=1}^{n} h(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n) \\
&\qquad + h(X_1, X_2, \cdots, X_n)
\end{aligned}$$

or

$$\begin{aligned}
&\frac{1}{n} h(X_1, X_2, \cdots, X_n) \\
&\quad \leq \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)}{n-1}, \quad (2)
\end{aligned}$$

which is the desired result $h_n^{(n)} \leq h_{n-1}^{(n)}$.

We now prove that $h_k^{(n)} \leq h_{k-1}^{(n)}$ for all $k \leq n$, by first conditioning on a k-element subset, then taking a uniform choice over its $(k-1)$-element subsets. For each k-element subset, $h_k^{(k)} \leq h_{k-1}^{(k)}$, and hence, the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements. □

Corollary 1: Let $r > 0$, and define

$$s_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:\,|S|=k} e^{(rh(X(S))/k)}. \quad (3)$$

Then,

$$s_1^{(n)} \geq s_2^{(n)} \geq \cdots \geq s_n^{(n)}. \quad (4)$$

Proof: Starting from (2) in Theorem 1, we multiply both sides by r, exponentiate, and then apply the arithmetic-mean, geometric-mean inequality to obtain

$$\begin{aligned}
e^{(1/n)rh(X_1, X_2, \cdots, X_n)} &\leq e^{1/n \sum_{i=1}^{n} (rh(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)/n-1)} \\
&\leq \frac{1}{n} \sum_{i=1}^{n} e^{(rh(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)/n-1)}, \\
&\qquad \text{for all } r \geq 0, \quad (5)
\end{aligned}$$

which is equivalent to $s_n^{(n)} \leq s_{n-1}^{(n)}$. Now we use the same arguments as in Theorem 1, taking an average over all subsets to prove the result that for all $k \leq n$, $s_k^{(n)} \leq s_{k-1}^{(n)}$. □

The conditional entropy rate per element for a k element subset S is $h(X(S)|X(S^c))/k$.

Definition: The *average conditional entropy rate* per element for all subsets of size k is the average of the previous quantities for k-element subsets of $\{1,2,\cdots,n\}$, i.e.,

$$g_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:\,|S|=k} \frac{h(X(S)|X(S^c))}{k}.$$

Here, $g_k(S)$ is the entropy per element of the set S conditional on the elements of the set S^c. When the size of the set S increases, one could expect a greater dependence between the elements of the set S, and expect a decrease in the entropy per element. This explains Theorem 1.

In the case of the conditional entropy per element, as k increases, the size of the conditioning set S^c decreases and the entropy of the set S increases since conditioning reduces entropy. In the conditional case, the increase in entropy per element due to the decrease in conditioning dominates the decrease due to additional dependence between the elements and hence, we have the following

theorem that is a consequence of the general formalism developed by Han [27].

Theorem 2:

$$g_1^{(n)} \le g_2^{(n)} \le \cdots \le g_n^{(n)}.$$

Proof: The proof proceeds on lines very similar to the proof of the theorem for the unconditional entropy per element for a random subset. We will first prove that $g_n^{(n)} \ge g_{n-1}^{(n)}$, and then use this to prove the rest of the inequalities.

By the chain rule, the entropy of a collection of random variables is less than the sum of the entropies, i.e.,

$$h(X_1, X_2, \cdots, X_n) \le \sum_{i=1}^{n} h(X_i).$$

Subtracting both sides of this inequality from $nh(X_1, X_2, \cdots, X_n)$, we have

$$\begin{aligned}(n-1)h(X_1, X_2, \cdots, X_n) &\ge \sum_{i=1}^{n} (h(X_1, X_2, \cdots, X_n) - h(X_i)) \\ &= \sum_{i=1}^{n} h(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n | X_i).\end{aligned}$$

Dividing this by $n(n-1)$, we obtain

$$\frac{h(X_1, X_2, \cdots, X_n)}{n} \ge \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n | X_i)}{n-1},$$

which is equivalent to $g_n^{(n)} \ge g_{n-1}^{(n)}$.

We now prove that $g_k^{(n)} \ge g_{k-1}^{(n)}$ for all $k \le n$, by first conditioning on a k-element subset, then taking a uniform choice over its $(k-1)$-element subsets. For each k-element subset, $g_k^{(k)} \ge g_{k-1}^{(k)}$, and hence, the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements. □

C. Inequalities for Average Mutual Information between Subsets

The previous two theorems can be used to prove the following statement about mutual information.

Corollary 2: Let

$$f_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S: |S| = k} \frac{I(X(S); X(S^c))}{k}.$$

Then,

$$f_1^{(n)} \ge f_2^{(n)} \ge \cdots \ge f_n^{(n)}.$$

Proof: This result follows from the identity $I(X(S); X(S^c)) = h(X(S)) - h(X(S)|X(S^c))$ and Theorems 1 and 2. □

We now prove an inequality for the average mutual information between a subset and its complement, averaged over all subsets of size k in a set of random variables. This inequality will be used to prove yet another determinant inequality along the lines of Szasz's theorem; however, unlike the inequalities in the previous section, there is no normalization by the number of elements in the subset.

Let

$$i_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S: |S| = k} I(X(S); X(S^c))$$

be the average mutual information between a subset and its complement averaged over all subsets of size k. By the symmetry of mutual information and the definition of $i_k^{(n)}$, it is clear that $i_k^{(n)} = i_{n-k}^{(n)}$.

Theorem 3:

$$i_1^{(n)} \le i_2^{(n)} \le \cdots \le i_{\lfloor n/2 \rfloor}^{(n)}.$$

Remark: Note that the dependence between sets and their complements is greatest when they are of equal size.

Proof: Let $k \le \lfloor n/2 \rfloor$. Consider a particular subset S of size k. S has k subsets of size $k-1$. Let S_j denote the subset $S - \{j\}$. Then

$$\begin{aligned}&kI(X(S); X(S^c)) - \sum_{j \in S} I\big(X(S_j); X(S_j^c)\big) \\ &\quad = \sum_{j \in S} I\big(X(S_j), X_j; X(S^c)\big) - I\big(X(S_j); X(S^c), X_j\big) \\ &\quad = \sum_{j \in S} I\big(X(S_j); X(S^c)\big) + I\big(X_j; X(S^c) | X(S_j)\big) \\ &\qquad - I\big(X(S_j); X(S^c)\big) - I\big(X(S_j); X_j | X(S^c)\big) \\ &\quad = \sum_{j \in S} h\big(X_j | X(S_j)\big) - h\big(X_j | X(S_j), X(S^c)\big) \\ &\qquad - h\big(X_j | X(S^c)\big) + h\big(X_j | X(S^c), X(S_j)\big) \\ &\quad = \sum_{j \in S} h\big(X_j | X(S_j)\big) - h\big(X_j | X(S^c)\big).\end{aligned}$$

Summing this over all subsets of size k, we obtain

$$\sum_{S: |S| = k} \left[kI(X(S); X(S^c)) - \sum_{j \in S} I\big(X(S_j); X(S_j^c)\big) \right] = \sum_{S: |S| = k} \sum_{j \in S} h\big(X_j | X(S_j)\big) - h\big(X_j | X(S^c)\big).$$

Reversing the order of summation, we obtain

$$\begin{aligned}&\sum_{S: |S| = k} \left[kI(X(S); X(S^c)) - \sum_{j \in S} I\big(X(S_j); X(S_j^c)\big) \right] \\ &\quad = \sum_{j=1}^{n} \sum_{S: |S| = k,\, S \ni j} h\big(X_j | X(S_j)\big) - h\big(X_j | X(S^c)\big) \\ &\quad = \sum_{j=1}^{n} \sum_{S': |S'| = k-1,\, S' \not\ni j} h\big(X_j | X(S')\big) \\ &\qquad - h\big(X_j | X(\{S' \cup j\}^c)\big) \\ &\quad = \sum_{j=1}^{n} \left[\sum_{S': S' \subset \{j\}^c, |S'| = k-1} h\big(X_j | X(S')\big) \right. \\ &\qquad \left. - \sum_{S'': S'' \subset \{j\}^c, |S''| = n-k} h\big(X_j | X(S'')\big) \right]. \qquad (6)\end{aligned}$$

Since $k \le \lfloor n/2 \rfloor$, $k-1 < n-k$. So we would expect that the second sum in (6) to be less than the first sum, since both sums have the same number of terms but the second sum corresponds to entropies with more conditioning. We will prove this by using a simple symmetry argument.

The set S'' with $n-k$ elements has $\binom{n-k}{k-1}$ subsets of size $k-1$. For each such subset S' of size $k-1$, we have

$$h(X_j|X(S'')) \le h(X_j|X(S')), \tag{7}$$

since conditioning reduces entropy. Since (7) is true for each subset $S' \subset S''$, it is true of the average over subsets. Hence,

$$h(X_j|X(S'')) \le \frac{1}{\binom{n-k}{k-1}} \sum_{S': S' \subset S'', |S'| = k-1} h(X_j|X(S')). \tag{8}$$

Summing (8) over all subsets S'' of size $n-k$, we get

$$\begin{aligned} &\sum_{S'': |S''| = n-k} h(X_j|X(S'')) \\ &\quad \le \sum_{S'': |S''| = n-k} \frac{1}{\binom{n-k}{k-1}} \sum_{S': S' \subset S'', |S'| = k-1} h(X_j|X(S')) \\ &\quad = \sum_{S': |S'| = k-1} h(X_j|X(S')), \end{aligned} \tag{9}$$

since by symmetry, each subset S' occurs in $\binom{n-k}{n-2k+1} = \binom{n-k}{k-1}$ sets S''.

Combining (6) and (9), we get

$$\sum_{S: |S| = k} \left[kI(X(S); X(S^c)) - \sum_{j \in S} I(X(S_j); X(S_j^c)) \right] \ge 0.$$

Since each set of size $k-1$ occurs $n-k+1$ times in the second sum, we have

$$\begin{aligned} &\sum_{S: |S| = k} kI(X(S); X(S^c)) \\ &\quad \ge \sum_{S: |S| = k} \sum_{j \in S} I(X(S_j); X(S_j^c)) \\ &\quad = (n-k+1) \sum_{S': |S'| = k-1} I(X(S'); X(S'^c)). \end{aligned}$$

Dividing this equation by $k\binom{n}{k}$, we have the theorem

$$\begin{aligned} i_k^{(n)} &= \frac{1}{\binom{n}{k}} \sum_{S: |S| = k} I(X(S); X(S^c)) \\ &\ge \frac{1}{\binom{n}{k-1}} \sum_{S': |S'| = k-1} I(X(S'); X(S'^c)) = i_{k-1}^{(n)}. \end{aligned}$$

III. The Entropy Power and Related Analytical Inequalities

The entropy power inequality, which says that the entropy of the sum of two independent random vectors is no less than the entropy of the sum of their independent normal counterparts, has a strong formal resemblance to the Brunn Minkowski inequality, which says that the volume of the set sum of two sets is greater than or equal to the volume of the set sum of their spherical counterparts. Both are interpreted here as convexity inequalities for Rényi entropies that measure the uncertainty associated with a random variable X via the pth norm of its density (see Section III-A). A strengthened version of Young's inequality about the norms of convolutions of functions, due to Beckner [3] and Brascamp and Lieb [8] is equivalent to a more general convexity inequality, with both the entropy power and the Brunn-Minkowski inequality being extreme points (see Section III-B).

This proof of the entropy power inequality (due to Lieb [30]) is different from Stam's [38] proof, which relies upon a convexity inequality for Fisher information. Nevertheless, the interpretation of the entropy power inequality as a convexity inequality for entropy allows for a new, simpler version of Stam's proof, presented here in Section III-C.

Isoperimetric versions of the entropy power and the Fisher information inequalities have derivations that parallel the classical derivation of the isoperimetric inequality as a consequence of the Brunn–Minkowski inequality (see Section III-D following Costa and Cover [14] and Dembo [19]).

A. Entropy Power and Brunn–Minkowski Inequalities

The definition of the entropy power and the associated entropy power inequality stated next are due to Shannon [37]. The entropy power inequality is instrumental in establishing the capacity region of the Gaussian broadcast channel ([5]) and in proving convergence in relative entropy for the central limit theorem ([2]).

Definition: The *entropy power* of a random vector $X \in \boldsymbol{R}^n$ with a density is

$$N(X) = \frac{1}{2\pi e} \exp\left(\frac{2}{n} h(X)\right).$$

In particular, $N(X) = |K|^{1/n}$ when $X \sim \phi_K$.

Theorem 4 (Entropy Power Inequality): If X, Y are two independent random vectors with densities in $\boldsymbol{R}^n$ and both $h(X)$ and $h(Y)$ exist, then,

$$N(X+Y) \ge N(X) + N(Y). \tag{10}$$

Equality holds iff X and Y are both multivariate normal with proportional covariances.

In the sequel (see Section III-C), we shall present a simplified version of Stam's first proof of this inequality (in [38]) as well as a less known proof due to Lieb [30].

The next matrix inequality (Oppenheim [36], Marshall and Olkin [32, p. 475]) follows immediately from the entropy power inequality when specialized to the multivariate normal.

Theorem 5 (Minkowski's Inequality [34]): For any two nonnegative definite matrices K_1, K_2

$$|K_1 + K_2|^{1/n} \geq |K_1|^{1/n} + |K_2|^{1/n},$$

with equality iff K_1 is proportional to K_2.

Proof: Let X_1, X_2 be independent with $X_i \sim \phi_{K_i}$. Noting that $X_1 + X_2 \sim \phi_{K_1+K_2}$ and using the entropy power inequality yields

$$\begin{aligned} |K_1 + K_2|^{1/n} &= N(X_1 + X_2) \\ &\geq N(X_1) + N(X_2) \\ &= |K_1|^{1/n} + |K_2|^{1/n}. \end{aligned}$$ □

The following alternative statement of the entropy power inequality is given in Costa and Cover [14].

Theorem 6: For any two independent random vectors X, Y such that both $h(X)$ and $h(Y)$ exist,

$$h(X + Y) \geq h(\tilde{X} + \tilde{Y}), \tag{11}$$

where $\tilde{X}, \tilde{Y}$ are two independent multivariate normal with proportional covariances, chosen so that $h(\tilde{X}) = h(X)$ and $h(\tilde{Y}) = h(Y)$.

Proof: For $\tilde{X}$ and $\tilde{Y}$ multivariate normal, Minkowski's inequality and the entropy power inequality (10), hold with equality. Furthermore, $\tilde{X}$ and $\tilde{Y}$ are chosen so that

$$N(\tilde{X}+\tilde{Y}) = N(\tilde{X}) + N(\tilde{Y}) = N(X) + N(Y) \leq N(X+Y),$$

where the last inequality follows from (10). Thus (10) and (11) are equivalent. □

Alternatively, the entropy power inequality also amounts to the convexity of the entropy under the "covariance preserving transformation" $\sqrt{\lambda}X + \sqrt{1-\lambda}Y$ as follows.

Theorem 7: For any $0 \leq \lambda \leq 1$,

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \geq 0. \tag{12}$$

Proof: For $\tilde{X}$ and $\tilde{Y}$ the inequality (12) holds trivially with equality. Therefore, (12) is equivalent to

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

The latter inequality is merely (11) with $\sqrt{\lambda}X$ substituted for X and $\sqrt{1-\lambda}Y$ substituted for Y. □

Remark: Theorem 7 parallels part of Lieb's proof of Theorem 4 (in [30]).

In parallel with the above derivation of Minkowski's inequality, the following theorem due to Ky Fan [22] results from specializing (12) to the multivariate normal.

Theorem 8 (K. Fan [22]): $\ln |K|$ is concave.

Proof: Consider (12) for $X \sim \phi_{K_1}$ and $Y \sim \phi_{K_2}$. Then, $\sqrt{\lambda}X + \sqrt{1-\lambda}Y$ is also multivariate normal with covariance $\lambda K_1 + (1-\lambda)K_2$, and (12) becomes

$$\ln|\lambda K_1 + (1-\lambda)K_2| \geq \lambda \ln|K_1| + (1-\lambda)\ln|K_2|. \quad \square \tag{13}$$

Remark: See Section V-A for an alternative information theoretic proof of both Theorems 5 and 8, which avoids the entropy power inequality.

The entropy power inequality has a strong formal resemblance of the Brunn–Minkowski inequality. For defining the latter, let μ denote Lebesgue measure in $\boldsymbol{R}^n$ (i.e., set volume in $\boldsymbol{R}^n$) and $A + B$ denote the Minkowski sum (in $\boldsymbol{R}^n$) of the (measurable) sets A and B, that is

$$A + B = \{x + y : x \in A, y \in B\}.$$

Theorem 9 (Brunn-Minkowski Inequality [24]):

$$\mu(A + B)^{1/n} \geq \mu(A)^{1/n} + \mu(B)^{1/n}. \tag{14}$$

Proof: For a very simply geometric proof, see [24]. An alternative proof of this inequality as an extreme point of Young's inequality (which is due to Brascamp and Lieb, see [7] and [9]) is presented in Section III-B.

The entropy power is a measure of the effective variance of a random vector while $\mu(A)^{1/n}$ measures the effective radius of a set A. Thus, the entropy power inequality, which says that the effective variance of the sum of two independent random vectors is no less than the sum of the effective variances of these vectors, is the dual of the Brunn–Minkowski inequality, which says that the effective radius of the set sum of two sets is no less than the sum of the effective radii of these sets. In this formal duality normal random variables are the analog of balls (being the equality cases for the previously mentioned inequalities), and the sum of two independent random vectors is the analog of the Minkowski sum of sets. This analogy is suggested in [14], where the existence of a family of intermediate inequalities is conjectured.

We shall further develop this issue here and show in Section III-B that Young's inequality is the bridge between the entropy power and the Brunn–Minkowski inequalities. The following family of Rényi entropies helps in illustrating these relationships.

Definition: The pth *Rényi entropy* $h_p(X)$ of a random variable X with density f in $\boldsymbol{R}^n$ is defined by

$$h_p(X) = \frac{1}{1-p} \ln E\left[f(X)^{(p-1)}\right] = \frac{p}{1-p} \ln\left(\|f\|_p\right), \tag{15}$$

for $0 < p < \infty$, $p \neq 1$, where $\|f\|_p = [\int f(x)^p \, dx]^{1/p}$. The Rényi entropies for $p = 0$ and $p = 1$ are defined as the limits of $h_p(X)$ as $p \to 0$ and $p \to 1$, respectively. It follows directly from the previous definition that

$$h_0(X) = \lim_{p \to 0} h_p(X) = \ln \mu(\{x : f(x) > 0\}), \tag{16}$$

and

$$h_1(X) = \lim_{p \to 1} h_p(X) = h(X). \tag{17}$$

Therefore, the (Shannon) entropy is identified with the Rényi entropy of index $p = 1$, while the logarithm of the essential support of the density is identified with the Rényi entropy of index $p = 0$.

A convexity inequality for Rényi entropies of index $p = 0$, which is the dual of (12), is the following.

Theorem 10: For any $0 \le \lambda \le 1$ and any two independent random vectors $\boldsymbol{X}, \boldsymbol{Y}$,

$$h_0(\lambda \boldsymbol{X} + (1-\lambda)\boldsymbol{Y}) - \lambda h_0(\boldsymbol{X}) - (1-\lambda) h_0(\boldsymbol{Y}) \ge 0. \quad (18)$$

Remarks:

a) While Theorem 7 deals with convexity under the "variance preserving transformation" $\sqrt{\lambda}\,\boldsymbol{X} + \sqrt{1-\lambda}\,\boldsymbol{Y}$, this theorem deals with convexity under the "support size preserving transformation" $\lambda \boldsymbol{X} + (1-\lambda)\boldsymbol{Y}$.

b) The proof of Theorem 10 is deferred to Section III-B. A family of convexity inequalities for Rényi entropies is derived there as consequences of Young's inequality and both Theorems 7 and 10 are obtained as extreme (limit) points. Here we derive only the Brunn–Minkowski inequality as a consequence of Theorem 10.

Proof of Theorem 9: Choose a pair of independent random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ in $\boldsymbol{R}^n$ such that the support of the density of $\lambda \boldsymbol{X}$ is the set A and the support of the density of $(1-\lambda)\boldsymbol{Y}$ is B. Clearly, the support of the density of $\lambda \boldsymbol{X} + (1-\lambda)\boldsymbol{Y}$ is the (essential) Minkowski sum $A + B$, while $(1/\lambda)A$ and $(1/(1-\lambda))B$ are the support sets of the densities of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Therefore, taking (16) into account, the inequality (18) specializes for these random vectors to

$$\ln \mu(A+B) \ge \lambda \ln \mu((1/\lambda)A) + (1-\lambda) \ln \mu((1/(1-\lambda))B). \quad (19)$$

Observing that $\ln \mu((1/\lambda)A) = \ln \mu(A) - n \ln \lambda$ and $\ln \mu((1/(1-\lambda))B) = \ln \mu(B) - n \ln(1-\lambda)$, the Brunn–Minkowski inequality results when rearranging the above inequality for the particular choice of $\lambda = \mu(A)^{1/n}/(\mu(A)^{1/n} + \mu(B)^{1/n})$. □

B. Young's Inequality and Its Consequences

There is a strong formal resemblance between the convexity inequalities (12) and (18) (where the former yields the entropy power inequality while the latter results in the Brunn–Minkowski inequality). This resemblance suggests the existence of a family of intermediate inequalities. Young's inequality, which is presented in the sequel, results after few manipulations with these inequalities (see (21)). In particular, we follow Lieb's (in [30]) and Brascamp and Lieb's (in [9]) approach in regarding and proving Theorems 7 and 10 (respectively) as limits of (21).

For that purpose let $L_p(\boldsymbol{R}^n)$ denote the space of complex valued measurable functions on $\boldsymbol{R}^n$ with $\|f\|_p < \infty$ and let $f \star g(\boldsymbol{x}) = \int f(\boldsymbol{x} - \boldsymbol{y}) g(\boldsymbol{y})\, d\boldsymbol{y}$ denote the convolution operation.

The following sharp version of Young's inequality is due to Beckner [3] and Brascamp and Lieb [8].

Theorem 11 (Young's Inequality): If $1/r + 1 = 1/q + 1/p$, then for $1 \le r, p, q \le \infty$,

$$\sup_{\substack{f \in L_p(\boldsymbol{R}^n) \\ g \in L_q(\boldsymbol{R}^n)}} \left\{ (\|f \star g\|_r)/(\|f\|_p \|g\|_q) \right\} \le (c_p c_q / c_r)^{n/2}. \quad (20)$$

Here,

$$c_p = (p)^{1/p}/|p'|^{1/p'},$$

where p' is the Hölder conjugate of p (i.e., $1/p + 1/p' = 1$) and c_q and c_r are likewise defined. The converse inequality holds for the infimum of $\|f \star g\|_r / \|f\|_p \|g\|_q$ when $0 < r, p, q \le 1$.

Remark: For the multivariate normal densities $f = \phi_{\lambda K_1}$ and $g = \phi_{(1-\lambda)K_2}$ (where $\lambda = (1/p')/(1/r')$, and consequently $1 - \lambda = (1/q')/(1/r')$), Young's inequality reduces to K. Fan's matrix Theorem 8. Actually, (20) is established in [8] by showing that the supremum is achieved by multivariate normal densities, where the constants in the right side of (20) are determined by applying K. Fan's matrix Theorem 8. For a detailed study of cases of equality in this and related inequalities see [31].

The following convexity inequality for Rényi entropies (which is the natural extension of Theorem 7) is a direct consequence of Young's inequality.

Theorem 12: For any $0 < r \le \infty$, $r \ne 1$ and any $0 \le \lambda \le 1$, let p, q be such that $1/p' = \lambda 1/r'$ and $1/q' = (1-\lambda)1/r'$, then for any two independent random vectors $\boldsymbol{X}, \boldsymbol{Y}$ with densities in $\boldsymbol{R}^n$,

$$h_r(\sqrt{\lambda}\,\boldsymbol{X} + \sqrt{1-\lambda}\,\boldsymbol{Y}) - \lambda h_p(\boldsymbol{X}) - (1-\lambda) h_q(\boldsymbol{Y}) \ge h_r(\phi_I) - \lambda h_p(\phi_I) - (1-\lambda) h_q(\phi_I), \quad (21)$$

provided that both $h_p(\boldsymbol{X})$ and $h_q(\boldsymbol{Y})$ exist.

Here, ϕ_I stands for the standard normal density in $\boldsymbol{R}^n$.

In establishing the inequality (21) we use the well-known scaling property of Rényi entropies

$$h_p(\alpha \boldsymbol{X}) = h_p(\boldsymbol{X}) + n \ln |\alpha|. \quad (22)$$

This identity follows from the definition in (15) by a change of variable argument.

Proof: Fix r and λ. We plan to apply Young's inequality for f the density of $\sqrt{\lambda}\,\boldsymbol{X}$ and g the density of $\sqrt{1-\lambda}\,\boldsymbol{Y}$. Since $h_p(\boldsymbol{X})$ and $h_q(\boldsymbol{Y})$ are well defined, so are

$$h_p(\sqrt{\lambda}\,\boldsymbol{X}) = -p' \ln \|f\|_p = h_p(\boldsymbol{X}) + \frac{n}{2} \ln \lambda$$

and

$$h_q(\sqrt{1-\lambda}\,\boldsymbol{Y}) = -q' \ln \|g\|_q = h_q(\boldsymbol{Y}) + \frac{n}{2} \ln(1-\lambda).$$

These identities are applications of (15) and (22), and in particular they imply that $f \in L_p(\boldsymbol{R}^n)$ and $g \in L_q(\boldsymbol{R}^n)$. Further, since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed independent,

$$-r' \ln \|f \star g\|_r = h_r(\sqrt{\lambda}\,\boldsymbol{X} + \sqrt{1-\lambda}\,\boldsymbol{Y}).$$

Observe that p, q in Theorem 12 are such that $1/p' + 1/q' = 1/r'$ (so that $1/r + 1 = 1/q + 1/p$), and $1/r' < 0$

implies $0 < r, p, q < 1$, while $1/r' > 0$ implies $1 < r, p, q$. Therefore, Theorem 11 is applicable for f and g, resulting in

$$-r' \ln\left\{(\|f \star g\|_r)/(\|f\|_p \|g\|_q)\right\} \geq -r' \frac{n}{2} \ln(c_p c_q / c_r). \tag{23}$$

This inequality holds with equality for $f = \phi_{\lambda I}$ and $g = \phi_{(1-\lambda)I}$ (i.e., $X \sim \phi_I$, $Y \sim \phi_I$) since for any $p \neq 0$, $p \neq 1$

$$h_p(\phi_I) = \frac{n}{2}\left[\ln 2\pi + \frac{1}{1-p} \ln \frac{1}{p}\right]. \tag{24}$$

Combining all these identities, the inequality (23) results in (21). □

We now show that the convexity Theorems 7 and 10 (i.e., the inequalities (12) and (18) respectively) are the extreme limit points $r \to 1$ and $r \to 0$ of the Rényi entropy convexity Theorem 12.

Proof of Theorem 7: Fix $0 \leq \lambda \leq 1$, and assume that $h(X)$ and $h(Y)$ are well defined. Further assume that (21) holds for some $r_0 \neq 1$. Then, Theorem 12 holds for any choice of r between r_0 and 1 (i.e., the entropies $h_p(X)$ and $h_q(Y)$ exist for the resulting p and q). It is easily verified that $r \to 1$ with λ fixed implies that $p \to 1$ and $q \to 1$. Therefore, by the continuity of entropies (17) in the limit as $r \to 1$, the inequality (21) reduces to (12), thus completing the proof of Theorem 7. □

Proof of Theorem 10: Again fix $0 \leq \lambda \leq 1$. Now assume that $h_0(X)$ and $h_0(Y)$ are well defined and that (21) holds for some $r_0 < 1$. Then Theorem 12 holds for any choice of r between r_0 and 0 (i.e., the entropies $h_p(X)$ and $h_q(Y)$ exist for the resulting p and q). Further, as $r \to 0$, also $p = 1/(1-\lambda(1-1/r)) \to 0$ and $q = 1/(1-(1-\lambda)(1-1/r)) \to 0$. Thus, in the limit $r \to 0$, the inequality (21) reduces by (16) to

$$\begin{aligned} &h_0(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h_0(X) - (1-\lambda)h_0(Y) \\ &\quad \geq \lim_{r \to 0}\left\{h_r(\phi_I) - \lambda h_p(\phi_I) - (1-\lambda)h_q(\phi_I)\right\} \\ &\quad = \frac{n}{2} \lim_{r \to 0}\left\{\frac{1}{1-r}\ln\frac{1}{r} - \frac{\lambda}{1-p}\ln\frac{1}{p} - \frac{(1-\lambda)}{1-q}\ln\frac{1}{q}\right\}, \end{aligned} \tag{25}$$

where the right-hand equality is in view of (24).

Note that $\lambda/(1-p) + (1-\lambda)/(1-q) = (1+r)/(1-r)$ and $\lim_{r \to 0}(r/p) = \lambda$ while $\lim_{r \to 0}(r/q) = (1-\lambda)$. Therefore,

$$\begin{aligned} &\lim_{r \to 0}\left\{\frac{1}{1-r}\ln\frac{1}{r} - \frac{\lambda}{1-p}\ln\frac{1}{p} - \frac{(1-\lambda)}{1-q}\ln\frac{1}{q}\right\} \\ &\quad = \lim_{r \to 0}\left\{\frac{-r}{1-r}\ln\frac{1}{r} - \frac{\lambda}{1-p}\ln\frac{r}{p} - \frac{(1-\lambda)}{(1-q)}\ln\frac{r}{q}\right\} \\ &\quad = H(\lambda) + \lim_{r \to 0}\frac{r}{1-r}\ln r = H(\lambda), \end{aligned}$$

where $H(\lambda) \triangleq -\lambda \ln \lambda - (1-\lambda)\ln(1-\lambda)$. Combining this limit with (25) yields

$$h_0(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) = \lambda h_0(X) - (1-\lambda)h_0(Y) \geq \frac{n}{2}H(\lambda).$$

Inequality (18) is now obtained by the rescaling $X \leftarrow \sqrt{\lambda}X$ and $Y \leftarrow \sqrt{1-\lambda}Y$ (using the scaling property (22)). This completes the proof of Theorem 10. □

Remarks:

a) The proof of Theorem 7 follows Lieb's proof of the entropy power inequality (see [30]).

b) In [9], Brascamp and Lieb prove the Prékopa–Liendler inequality

$$\int \sup_y \left\{f\left(\frac{x-y}{1-\lambda}\right)^{1-\lambda} g\left(\frac{y}{\lambda}\right)^{\lambda}\right\} dx \geq 1, \tag{26}$$

for every pair of densities f, g in $\mathbf{R}^n$ and any $0 < \lambda < 1$. For $g(\cdot)$ a uniform density on A/λ and $f(\cdot)$ a uniform density on $B/(1-\lambda)$, this inequality reduces to the Brunn–Minkowski inequality (19). The proof of Theorem 10 is a simplified version of Brascamp and Lieb's proof of (26).

c) Theorem 7 of [8] deals with $X_1, \cdots, X_k$, independent random variables with densities in $\mathbf{R}^n$, and $(k-1) \geq l \geq 1$ deterministic linear combinations of these variables $Y_1, \cdots, Y_l$. Let V have the density of Y_1 conditional upon $Y_1 = \cdots = Y_l$, then this theorem implies that the minimum of

$$\left\{h_r(V) - \sum_{j=1}^{k} \lambda_j h_{p_j}(X_j)\right\}$$

is obtained for $X_1, \cdots, X_k$ normal random variables with appropriate diagonal covariance matrices. This theorem holds for any $1 < r \leq \infty$, and any $\lambda_j = r'/p_j' \geq 0$ such that $\sum_{j=1}^{k}\lambda_j = 1 + r'(l-1)$. For $l = 1$, $\sum_{j=1}^{k}\lambda_j = 1$ and $V = Y_1 = X_1 + \cdots + X_k$, this inequality results in Young's inequality. It seems plausible that new entropy inequalities may be derived by considering limits of this more general inequality for $l > 1$.

C. Fisher Information and the Entropy Power Inequality

Stam's proof of the entropy power inequality (see [38]) is based on a simple inequality about Fisher information coupled with a continuous normal perturbation argument. A simplified version of this proof is presented here, where a *simple explicit normal perturbation* yields the convexity inequality (12). As we have seen already, inequality (12) is equivalent to the entropy power inequality (10).

Definition: The *Fisher information* of X with respect to a scalar translation parameter is

$$J(X) = \int \nabla f(x)' \cdot \nabla f(x) \frac{dx}{f(x)}. \tag{27}$$

Equivalent statements of the following convexity inequality about Fisher information are proved in [6], [14], [38]. (For matrix versions see [20]).

Theorem 13 (Fisher Information Inequality): For any two independent random vectors X, Y and any $0 \le \lambda \le 1$,

$$\lambda J(X) + (1-\lambda) J(Y) - J\left(\sqrt{\lambda}\, X + \sqrt{1-\lambda}\, Y\right) \ge 0. \quad (28)$$

This is the first instrumental tool for the proof of the entropy power inequality presented in the sequel. The second tool is DeBruijn's identity, the link between entropy and Fisher information (for proofs consider [6], [14], [38]).

Theorem 14 (DeBruijn's identity [38]): Let X be any random vector in R^n such that $J(X)$ exists and let $Z \sim \phi_I$ be a standard normal, which is independent of X. Then

$$\frac{d}{d\epsilon} h\left(X + \sqrt{\epsilon}\, Z\right)_{|\epsilon = 0} = \frac{1}{2} J(X). \quad (29)$$

We are now ready to present the simplified version of Stam's proof.[1]

Proof of Theorem 7 (by normal perturbations): Consider the continuous family of pairs of independent random vectors

$$X_t = \sqrt{t}\, X + \sqrt{1-t}\, X_0, \qquad 0 \le t \le 1,$$

$$Y_t = \sqrt{t}\, Y + \sqrt{1-t}\, Y_0, \qquad 0 \le t \le 1,$$

where the standard multivariate normals $X_0 \sim \phi_I$ and $Y_0 \sim \phi_I$ are independent of X, Y and of each other. Fix $0 \le \lambda \le 1$ and let $V_t = \sqrt{\lambda}\, X_t + \sqrt{1-\lambda}\, Y_t$. Clearly, $V_0 \sim \phi_I$ is also a standard normal, and $V_t = \sqrt{t}\, V_1 + \sqrt{1-t}\, V_0$ for all $0 \le t \le 1$. We now consider the function

$$s(t) = h(V_t) - \lambda h(X_t) - (1-\lambda) h(Y_t), \qquad \text{for } 0 \le t \le 1.$$

Theorem 7 (i.e., inequality (12)) amounts to $s(1) \ge 0$, and since V_0, X_0, and Y_0 are identically distributed $s(0) = 0$. Therefore, our goal is to establish the differential inequality

$$\frac{d}{dt}\{s(t)\} \ge 0, \qquad 0 \le t \le 1, \quad (30)$$

which clearly implies inequality (12) and thus completes the proof. By virtue of the scaling property (22) (applied here for $p = 1$, $\alpha = \sqrt{1/t}$ and for the variables X_t, Y_t and V_t) the function $s(t)$ may also be expressed as

$$s(t) = h\left(V_1 + \sqrt{\epsilon_t}\, V_0\right) - \lambda h\left(X + \sqrt{\epsilon_t}\, X_0\right) - (1-\lambda) h\left(Y + \sqrt{\epsilon_t}\, Y_0\right),$$

where $\epsilon_t = ((1/t) - 1)$. Therefore, by DeBruijn's identity (29)

$$\frac{d}{dt}\{s(t)\} = \frac{1}{2}\frac{d}{dt}\{\epsilon_t\}\left\{J\left(V_1 + \sqrt{\epsilon_t}\, V_0\right) - \lambda J\left(X + \sqrt{\epsilon_t}\, X_0\right) - (1-\lambda) J\left(Y + \sqrt{\epsilon_t}\, Y_0\right)\right\}.$$

Since $d\epsilon_t / dt = -1/t^2$ we obtain by an application of the well-known scaling property $J(X) = \alpha^2 J(\alpha X)$

$$2t\frac{d}{dt}\{s(t)\} = \lambda J(X_t) + (1-\lambda) J(Y_t) - J(V_t). \quad (31)$$

Since $V_t = \sqrt{\lambda}\, X_t + \sqrt{1-\lambda}\, Y_t$ the Fisher information inequality (28) applies to (31) and thus establishes the differential inequality (30). □

Remarks:

a) The representation (31) is very similar to the one in [1]. Such a representation was also used in [2] for proving a strong version of the central limit theorem.

b) Two independent proofs of the entropy power inequality via the equivalent convexity inequality (12) have been presented. In the first proof, the underlying tool is Young's inequality from mathematical analysis, and results about (Shannon's) entropy are the limit as $r \to 1$ of analogous results about Rényi entropies (i.e., about norms of operators in $L_r(R^n)$). In the second proof, the underlying tool is a sufficient statistic inequality for Fisher information, and results about entropy are obtained by integration over the path of a continuous normal perturbation. This proof also settles the cases of equality that are not determined in the first proof. We will encounter this duality again in Section IV where uncertainty principles are derived by similar arguments.

c) The strong formal resemblance between convexity inequalities (12) and (18) dealing with entropies and the Minkowski sum of sets suggests the following inequality:

$$\frac{\mu(A+B)}{S(A+B)} \ge \frac{\mu(A)}{S(A)} + \frac{\mu(B)}{S(B)}, \quad (32)$$

as the dual of the Fisher information inequality (37). Here, $S(C)$ denotes the outer Minkowski content of the boundary of a set C, which is defined as

$$S(C) = \liminf_{\epsilon \to 0} \frac{1}{\epsilon}\left[\mu(C + \epsilon B_1) - \mu(C)\right],$$

where B_ρ denotes a ball of radius ρ centered at the origin (in particular, when C is a convex set or a set with piecewise smooth boundary then $S(C)$ coincides with the usual surface area of C; see [10], p. 69).

When inequality (32) holds, the Brunn–Minkowski inequality follows by a continuous perturbation (by balls) argument paralleling Stam's proof of the entropy power inequality. However, (32) does not hold in general for nonconvex sets. For example, it is false when A is the unit ball and B is the union of two balls of distance 3 apart (so that $A + B$ is also the union of two disjoint balls).

[1] At the time of the writing of this paper, the same result was independently derived by Carlen and Soffer and will appear in [13].

Does (32) hold when both A and B are compact, convex and nonempty sets? Alternatively, is the ratio of volume-to-surface area increased by Minkowski sums for such sets? When in addition A (or B) is a ball the inequality (32) indeed holds as a direct consequence of the Alexandrov–Fenchel inequality (see [10], p. 143).

d) Consider the functions

$$s_r(t) = h_r(V_t) - \lambda h_p(X_t) - (1-\lambda) h_q(Y_t), \qquad 0 \le t \le 1.$$

Theorem 7 (inequality (12)) amounts to $s_1(1) - s_1(0) \ge 0$ and therefore is a direct consequence of $2t(ds_1(t)/dt) \ge 0$. It can be shown that Young's inequality (20) is in essence equivalent to $s_r(1) - s_r(0) \ge 0$. Therefore, it is tempting to suggest that the stronger inequality $2t(ds_r(t)/dt) \ge 0$, holds for all $0 \le t \le 1$ and for some (or all) $r \ne 1$. The latter inequality holds iff for every $X \sim f$ and $Y \sim g$

$$\{J(V_r)\} \le \lambda^2((1-\lambda)r + \lambda)\{J(X)\} + (1-\lambda)^2(\lambda r + (1-\lambda))\{J(Y)\}, \quad (33)$$

where the density of V_r is proportional to

$$\left[\int f(v-y)^{((1-\lambda)+\lambda/r)} g(y)^{(\lambda+(1-\lambda)/r)} dy\right]^r.$$

Note that for $r = 1$, $V = X + Y$ and (33) is merely the Fisher information inequality (28). In conclusion, if (33) holds for $r \ne 1$ then this remark is the skeleton of a new proof of Young's inequality for these values of r, a proof which is orthogonal to the existing proofs of [3] and [8].

D. Isoperimetric Inequalities

The classical isoperimetric inequality states that balls have the smallest surface area per given volume. Recall that $S(A)$ is the surface area of a set A and that B_1 is the unit ball. So, an alternative statement of the isoperimetric inequality is as follows.

Theorem 15 (The Classical Isoperimetric Inequality):

$$S(A) \ge n \cdot \mu(A)^{(n-1/n)} \mu(B_1)^{(1/n)}$$

with equality if A is a ball in $\boldsymbol{R}^n$.

Proof: Consider the nth power of the Brunn–Minkowski inequality (14) for $B_\epsilon = \epsilon B_1$ (so that $\mu(B_\epsilon)^{1/n} = \epsilon \mu(B_1)^{1/n}$). The isoperimetric inequality results by subtracting $\mu(A)$, dividing by ϵ and considering the limit as $\epsilon \downarrow 0$. □

A dual "isoperimetric inequality" was derived by such an approach out of the entropy power inequality (see [14] following [38]).

Theorem 16 (Isoperimetric Inequality for Entropies): For any random vector X in $\boldsymbol{R}^n$ for which the Fisher information $J(X)$ exists,

$$\frac{1}{n} J(X) N(X) \ge 1. \quad (34)$$

Proof (following [14]): For $Y = \sqrt{\epsilon}\, Z$, where Z is a standard multivariate normal (so $N(Y) = \epsilon$), the entropy power inequality (10) reduces to

$$\frac{1}{\epsilon}\left[N(X + \sqrt{\epsilon}\, Z) - N(X)\right] \ge 1. \quad (35)$$

Clearly,

$$\frac{d}{d\epsilon}\left\{N(X + \sqrt{\epsilon}\, Z)\right\}_{|\epsilon=0} = \frac{2}{n} N(X) \frac{d}{d\epsilon}\left\{h(X + \sqrt{\epsilon}\, Z)\right\}_{|\epsilon=0}.$$

Therefore, in the limit $\epsilon \downarrow 0$, inequality (35) yields the isoperimetric inequality for entropies by DeBruijn's identity (29). □

Remark: Inequality (34) is equivalent to Gross's logarithmic Sobolev inequality (see [25]). This is discussed in [12]. For more literature on this subject see [26].

The same approach is applied in [19] for deriving the following isoperimetric inequality about Fisher information.

Theorem 17 (Fisher Information Isoperimetric Inequality): When the Fisher information $J(X)$ of a random vector X in $\boldsymbol{R}^n$ exists and is differentiable with respect to a small independent normal perturbation then

$$\frac{d}{d\epsilon}\left\{\left[\frac{1}{n}\left\{J(X + \sqrt{\epsilon}\, Z)\right\}\right]^{-1}\right\}_{\epsilon=0} \ge 1. \quad (36)$$

Proof (following [19]): While the Fisher information inequality (28) is the dual of the convexity inequality (12), the inequality

$$J(X+Y)^{-1} - J(X)^{-1} - J(Y)^{-1} \ge 0, \quad (37)$$

where X, Y are any two independent random vectors, is the dual of the entropy power inequality (10). This equivalent statement of the Fisher information inequality is proved for example in [6] (for $n = 1$) and [20] (for $n \ne 1$).

For $Y = \sqrt{\epsilon}\, Z$ (so that $J(Y)^{-1} = \epsilon/n$) and in the limit $\epsilon \downarrow 0$ this inequality yields

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}\left\{J(X + \sqrt{\epsilon}\, Z)^{-1} - J(X)^{-1}\right\} - \frac{1}{n} = \frac{d}{d\epsilon}\left\{J(X + \sqrt{\epsilon}\, Z)^{-1}\right\}_{|\epsilon=0} - \frac{1}{n} \ge 0.$$

Since this is the same inequality as (36) the proof is completed. □

Remark: Inequality (36) is equivalent to the "Γ_2 inequality" of Bakry and Emery (see [1]).

The Fisher information isoperimetric inequality suggests that the sensitivity of the inverse of the Fisher information with respect to a small independent normal perturbation is minimal when the unperturbed variable already possesses a multivariate normal distribution. Note that the inverse of the Fisher information is exactly the

Cramér–Rao lower bound for the error of the estimate of a translation parameter (see also Section IV-B).

The concavity of the entropy power, which is proved directly in great length in [15], is the following immediate corollary of the Fisher information isoperimetric inequality (36).

Corollary 3 (Concavity of the Entropy Power): When the Fisher information of X exists and is differentiable with respect to a small independent normal perturbation then

$$\frac{d^2}{d\epsilon^2}\{N(X+\sqrt{\epsilon}Z)\}_{|\epsilon=0} \le 0. \tag{38}$$

Proof (following [19]): Two applications of DeBruijn's identity (29) yield

$$\frac{d^2}{d\epsilon^2}\{N(X+\sqrt{\epsilon}Z)\}_{|\epsilon=0} = N(X)\left\{\left[\frac{1}{n}\{J(X)\}\right]^2 + \frac{1}{n}\frac{d}{d\epsilon}\{J(X+\sqrt{\epsilon}Z)\}_{|\epsilon=0}\right\}.$$

The isoperimetric Fisher information inequality is clearly equivalent to

$$\left[\frac{1}{n}\{J(X)\}\right]^2 + \frac{1}{n}\frac{d}{d\epsilon}\{J(X+\sqrt{\epsilon}Z)\}_{|\epsilon=0} \le 0$$

and the proof of (38) is thus completed. □

In conclusion, the entropy power of $X_t = X + \sqrt{t}Z$ is concave with respect to the variance t of the additive normal perturbation. Moreover, since DeBruijn's identity holds for any random vector Z whose first four moments coincide with those of the standard multivariate normal, so does the concavity inequality (38).

IV. Uncertainty Principles

In [38], the Weyl–Heisenberg uncertainty principle is derived from a specific version of the Cramér–Rao inequality. This idea is further developed here in Section IV-B, where we rederive the well-known fact that the Cramér–Rao inequality for location parameter is *exactly* the Weyl–Heisenberg uncertainty principle. Strong ties between Young's inequality, the entropy power and the Fisher information inequalities where explored in Section III. Similarly, Hirschman's uncertainty principle, which is presented in Section IV-C, is a consequence of the Hausdorff–Young inequality and it involves entropy powers of conjugate variables. Hausdorff–Young inequalities exist for various groups and result in the corresponding uncertainty principles. One such example, which is presented in Section IV-D, is related to bounds on the sizes of support sets of conjugate variables (see [21] for many other bounds of this type).

A new proof of Wehrl's conjecture about the minimal possible value of the classical entropy associated with certain quantum systems is presented in Section IV-E.[2]

[2]It was brought to our attention by an anonymous referee that this result was obtained independently by Carlen and will appear in [11].

While Lieb's proof of this conjecture (in [30]) is based on Hausdorff–Young and Young inequalities, here a stronger "incremental" result is derived as a direct consequence of the isoperimetric inequality for entropies. This demonstrates once again the close relationship between Fisher information and entropy.

A. Stam's Uncertainty Principle

We adopt the following definition of conjugate variables in quantum mechanics.

Definition: Associate with any complex wave amplitude function ψ in $L_2(R^n)$ a probability density

$$f_\psi(x) = |\psi(x)|^2 / \|\psi\|_2^2.$$

Let $\phi(y) \in L_2(R^n)$ be the Fourier transform of $\psi(x)$, and $g_\phi(y)$ the density similarly associated with ϕ. Then, the random vectors $X \sim f_\psi$ and $Y \sim g_\phi$ are called conjugate variables.

Stam's uncertainty principle relates the Fisher information *matrix* associated with a random vector (defined next) with the covariance of its conjugate variable.

Definition: The Fisher information *matrix* $J(X)$ of a random vector X with a density f is

$$J(X) = \int \nabla f(x)(\nabla f(x))' \frac{dx}{f(x)}.$$

Theorem 18 (Stam's Uncertainty Principle): Let K_X and K_Y be the covariance matrices of the conjugate random variables X and Y. Then

$$16\pi^2 K_Y - J(X) \ge 0, \tag{39}$$

or, by the symmetrical roles of X and Y,

$$16\pi^2 K_X - J(Y) \ge 0. \tag{40}$$

Proof: See [38]. □

Remark: The left side of the matrix inequalities above is a nonnegative definite matrix. This is the interpretation of all matrix inequalities in the sequel.

The following identities, which are important consequences of Stam's proof of Theorem 18, are derived in [20].

Stam's Identities:

$$J(X) = 16\pi^2 K_Y, \quad \text{if } \bar{\psi}(x)/\psi(x) = \exp(i\varphi), \tag{41}$$

where φ is a constant independent of x. Similarly,

$$J(Y) = 16\pi^2 K_X, \quad \text{if } \bar{\phi}(y)/\phi(y) = \exp(i\varphi). \tag{42}$$

B. Heisenberg's Principle and the Cramér – Rao Inequality

Heisenberg's uncertainty principle is often stated as

$$\sigma_X \sigma_Y \ge \frac{h}{2\pi},$$

where h is Planck's constant and σ_X and σ_Y are the standard deviations of a pair X, Y of conjugate variables in R^1. However, the definition of conjugate variables in

Section IV-A corresponds to a proper normalization, in which $h/2\pi$ is replaced by $1/4\pi$. This normalization yields the following multivariate uncertainty relationships.

Theorem 19: The n-dimensional Weyl–Heisenberg uncertainty principle may be stated in any of the following four equivalent forms:

$$16\pi^2 K_Y^{1/2} K_X K_Y^{1/2} - I \geq 0 \quad (43)$$

$$16\pi^2 K_X^{1/2} K_Y K_X^{1/2} - I \geq 0 \quad (44)$$

$$16\pi^2 K_Y - K_X^{-1} \geq 0 \quad (45)$$

$$16\pi^2 K_X - K_Y^{-1} \geq 0, \quad (46)$$

where X, Y are any pair of conjugate vectors (see Section IV-A for definition).

There exists a simple and direct proof of this inequality as a consequence of an appropriate Cauchy–Schwartz inequality. Here we present an alternative proof illustrating the connection of this uncertainty principle with the Cramér–Rao inequality:

Theorem 20 (Cramér – Rao Inequality):

$$J(X) - K_X^{-1} \geq 0, \quad (47)$$

$$J(Y) - K_Y^{-1} \geq 0. \quad (48)$$

Proof of the Weyl – Heisenberg Inequality: Adding Stam's uncertainty principle (39) and the Cramér–Rao inequality (47) yields the Weyl–Heisenberg principle (45). □

We interpret this relationship by suggesting that Stam's uncertainty principle "measures" the fluctuations in the phase of the amplitude wave functions $\psi(x)$ and $\phi(y)$, while the Cramér–Rao inequality "measures" the amount of "nonnormality" of the associated densities $f_\psi(x)$ and $g_\phi(y)$.

Actually, Stam's identities (41), (42) establish the *equivalence* of the Weyl–Heisenberg principle and the specific Cramér–Rao inequality given in Theorem 20. This equivalence is established by proving the Cramér–Rao inequality as a consequence of the Weyl–Heisenberg principle.

Proof of the Cramér – Rao Inequality (47): Suppose that X is a random variable in R^n with a density $f(x)$ for which $J(X) < \infty$. Let $\psi(x) = \sqrt{f(x)}$ be the associated real valued amplitude wave function. Clearly, Stam's identity (41) holds. Substituting this identity into the Weyl–Heisenberg principle (45) yields the Cramér–Rao inequality (47). □

Remark: This equivalence is generalized in [20], and shown there to hold between general families of Weyl–Heisenberg and Cramér–Rao inequalities.

C. Hausdorff – Young Inequality and Hirschman's Uncertainty Principle

An immediate consequence of Stam's uncertainty principle (39) is that

$$16\pi^2 |K_Y|^{1/n} \geq |J(X)|^{1/n},$$

where throughout this section X, Y is any pair of conjugate variables. A seemingly unrelated fact, a stronger version of Theorem 16 (the isoperimetric inequality for entropies) whose detailed derivation is given in [20], states that

$$N(X)|J(X)|^{1/n} \geq 1.$$

By combining the above two inequalities one obtains

$$16\pi^2 |K_Y|^{1/n} N(X) \geq 1.$$

Now, the maximum entropy inequality $N(Y) \leq |K_Y|^{1/n}$ (Lemma 5) suggests the following sharper uncertainty principle.

Theorem 21 (Hirschman's Uncertainty Principle):

$$16\pi^2 N(Y) N(X) \geq 1. \quad (49)$$

This uncertainty principle was conjectured by Hirschman (in [28]) who proved a weaker version with a smaller constant. It follows as a corollary of the following strong version of Hausdorff–Young inequality (due to Beckner [3]).

Theorem 22 (Hausdorff – Young Inequality): Let $\phi(y)$ be the Fourier transform of $\psi(x) \in L_p(R^n)$. Then for any $1 \leq p \leq 2$

$$\|\phi\|_{p'} \leq c_p^{n/2} \|\psi\|_p \quad (50)$$

where $(1/p) + (1/p') = 1$, and $c_p = p^{1/p}/p'^{(1/p')}$.

Remarks:

a) In [40], the time duration of the function $\psi(x)$ is measured via $\tau_p = \exp\{h_{p/2}(X)\}$ and its bandwidth is measured by $\theta_{p'} = \exp\{h_{p'/2}(Y)\}$. In this terminology, Hirschman's uncertainty principle amounts to the following "time-bandwidth" uncertainty relation

$$\tau_p \theta_{p'} \geq \left(\frac{e}{2}\right)^n.$$

b) One can also establish Young's inequality in the range $1 \leq p, q \leq 2 \leq r$ out of the Hausdorff–Young inequality (Theorem 22) and elementary properties of the Fourier transform (see [3]).

c) Cases of equality in (50) are studied in [31].

d) Carlen [12] obtains the isoperimetric inequality (34) as a consequence of Hirschman's uncertainty principle.

D. A Discrete Version of Hirschman's Uncertainty Principle

Hausdorff–Young inequalities exist for Fourier transforms on groups other than R^n. Each of these inequalities yields the corresponding Hirschman's uncertainty principle by considering the limit as $p \to 2$. As an explicit example to demonstrate this idea we show here that any unitary square matrix U (possibly of infinite dimension), with $\sup_{ij}|u_{ij}| = M < 1$, yields a nontrivial Hausdorff–Young inequality and consequently the following uncertainty principle.

Theorem 23: The integer valued random variables X, Y with $P(X=i) = |x_i|^2 / \|x\|_2^2$ and $P(Y=i) = |(Ux)_i|^2 / \|Ux\|_2^2$ are "conjugate" variables, where x is any vector with $\|x\|_2 < \infty$. For any such pair

$$H(X) + H(Y) \geq 2 \ln\left(\frac{1}{M}\right). \tag{51}$$

Proof: The unitary matrix U is an isometry on the appropriate Hilbert space, i.e., for every x, $\|Ux\|_2 = \|x\|_2$. Furthermore, clearly $\|Ux\|_\infty \leq M\|x\|_1$, where $\|x\|_p = [\sum_{i=1} |x_i|^p]^{1/p}$ and $\|x\|_\infty = \sup_{i=1}\{|x_i|\}$. Riesz's interpolation theorem (between the extreme bounds above for $p=1$ and $p=2$) yields the following "Hausdorff–Young" inequality for any vector x and any $1 \leq p \leq 2$

$$\|Ux\|_{p'} \leq M^{(2-p)/p} \|x\|_p, \tag{52}$$

where $1/p' + 1/p = 1$. Consider now a pair of conjugate variables X and Y with distribution functions as previously defined. Then (52) implies an uncertainty principle for the (discrete) Rényi entropies of X and Y. Specifically, let

$$H_{p/2}(X) = \frac{1}{1-(p/2)} \ln \sum_i P(X=i)^{p/2} = \frac{p}{1-p/2} \ln\left(\|x\|_p / \|x\|_2\right),$$

and

$$H_{p'/2}(Y) = \frac{1}{1-(p'/2)} \ln \sum_i P(Y=i)^{p'/2} = \frac{p'}{1-p'/2} \ln\left(\|Ux\|_{p'} / \|Ux\|_2\right),$$

then (52) reads

$$\left(\frac{1}{p} - \frac{1}{2}\right) H_{p/2}(X) + \left(\frac{1}{2} - \frac{1}{p'}\right) H_{p'/2}(Y) \geq \left(\frac{1}{p} - \frac{1}{2}\right) 2 \ln\left(\frac{1}{M}\right).$$

For $(1/p) = (1/2) + \epsilon$, $(1/p') = (1/2) - \epsilon$ and as $\epsilon \downarrow 0$ this inequality (when divided by ϵ) yields the uncertainty principle (51). □

Remarks: This uncertainty principle is nontrivial for $M < 1$. For example, consider the discrete Fourier transform of size n that corresponds to a unitary matrix U for which $M = |u_{ij}| = 1/\sqrt{n}$. Here, Hirschman's uncertainty principle becomes

$$\sum_{k=1}^{n} P(X=k) \log_2 \frac{1}{P(X=k)} + \sum_{k=1}^{n} P(Y=k) \log_2 \frac{1}{P(Y=k)} \geq \log_2 n, \tag{53}$$

where the vector $\sqrt{P(Y=k)}$ is the discrete Fourier transform of the vector $\sqrt{P(X=k)}$.

This inequality is sharp. For example, starting with $P(X=1)=1$ results in a uniform distribution $P(Y=k) = 1/n$ for $k = 1, 2, \cdots, n$, and for this pair of distributions the previous inequality holds with equality.

The discrete entropy is bounded above by the base 2 logarithm of the size of the support set of the distribution. Therefore, the uncertainty principle (53) implies that the product of the support sets of the vector x and its discrete Fourier transform is at least the dimension n of the Fourier transform. This is Theorem 1 of [21] (where similar support-set inequalities are derived also for x such that $(1-\epsilon)$ of $\|x\|_2$ is concentrated over a relatively small index set).

E. Wehrl's Conjecture

Wehrl introduced a new definition of the "classical" entropy corresponding to a quantum system in an attempt to build a bridge between quantum theory and thermodynamics (see [39]). Consider a single particle in $\boldsymbol{R}^n$. The (quantum) state of the particle is characterized by the "density matrix" ρ, a nonnegative definite linear operator on $L_2(\boldsymbol{R}^n)$ of unit trace (i.e., whose eigenvalues are nonnegative real numbers that sum to one). The coherent states are the normalized $L_2(\boldsymbol{R}^n)$ functions

$$\psi(x|p,q) = \left(\frac{1}{2\pi}\right)^{n/2} \left(\frac{1}{\pi}\right)^{n/4} \exp\left\{-\frac{1}{2}(x-q)'(x-q) + ip'x\right\},$$

where $p \in \boldsymbol{R}^n$ and $q \in \boldsymbol{R}^n$ are respectively the momentum and position parameters associated with the coherent state. Note that when the particle is in quantum state $\psi(x|p,q)$ then its associated probability density $|\psi(x|p,q)|^2 / \|\psi\|_2^2$ is $\phi_{I/2}(x-q)$.

For any quantum operator ρ one can associate the following classical probability density function f_ρ on the parameter space $\mathbb{R}^{2n}$

$$f_\rho(p,q) = \int \bar{\psi}(x|p,q) \rho[\psi(x|p,q)]\, dx, \tag{54}$$

where $\bar{\psi}$ denotes the complex conjugate of ψ. Wehrl argued that the proper definition of the "classical" entropy associated with the operator ρ is the normalized (Shannon) entropy of f_ρ, i.e., $h(X_\rho) - n \ln(2\pi)$, where X_ρ is a random variable on $\mathbb{R}^{2n}$ with density f_ρ. Wehrl and others have studied the properties of this classical entropy (see, for example, [39]). One of the appealing properties they demonstrate is that the classical measure of uncertainty $h(X_\rho)$ is an upper bound to the quantum measure of uncertainty, i.e., the discrete quantum entropy $-\operatorname{tr}(\rho \ln \rho)$. As the quantum entropy is always nonnegative they argue that while the differential entropy $h(\cdot)$ may well be negative it is never so for the physically meaningful variables, i.e., for those of the form of X_ρ for some quantum operator ρ.

The quantum entropy is zero on any pure state (i.e., whenever the operator ρ is of rank 1). On the other hand, Wehrl conjectured that the classical entropy is never zero,

i.e., in the "classical" theory there is an inherent minimal level of uncertainty (due to "quantization") the value of which is n. Further, this minimal uncertainty is obtained iff the operator ρ is a projection operator on one of the coherent states.

Wehrl's conjecture, which is restated below as a lower bound on the entropy power of X_ρ, was proved in [30] by an application of the strong versions of Young and Hausdorff–Young inequalities (cases of equality were later determined by Carlen [11]).

Theorem 24 (Wehrl–Lieb): For X_ρ a random variable in $\mathbb{R}^{2n}$ with density f_ρ of the form of (54)

$$N(X_\rho) \geq 1,$$

and equality holds iff ρ is of rank 1 and X_ρ has a standard normal distribution.

Remarks:

a) It is fairly easy to show that the above conditions for equality are equivalent to ρ being a projection operator on exactly one coherent state.

b) Both the previous discussion and statement of Theorem 24 correspond to the normalization under which $h/2\pi$ is replaced by $1/4\pi$. In the real world all levels of uncertainty are to be appropriately restated in terms of multiples of Planck's constant h.

Recall the isoperimetric inequality for entropies (34)

$$\frac{1}{2n} J(X_\rho) N(X_\rho) \geq 1,$$

with equality iff X_ρ has a standard normal distribution. Because of this result, the above theorem (Wehrl's conjecture) is an immediate consequence of the following stronger "incremental" version.

Theorem 25 (Carlen [11], Dembo [20]): For X_ρ, as before,

$$\frac{1}{2n} J(X_\rho) \leq 1,$$

with equality iff ρ is an operator of rank 1.

Remark: Starting with Theorem 25 and applying a perturbation argument similar to the one presented in Section III-C yields the monotonicity of $N(\sqrt{t}\, X_\rho + \sqrt{1-t}\, X_{\rho^*})$, with respect to $t \in [0,1]$, where ρ^* is any projection operator on a coherent state and X_ρ and X_{ρ^*} are independent random vectors. The appropriate interpretation of this result is, however, unclear.

Proof: The operator ρ may be decomposed into $\rho = \sum_{i=1}^{\infty} \lambda_i P_i$ where $\lambda_i \geq 0$, $\sum_{i=1}^{\infty} \lambda_i = 1$, and P_i are rank one projection operators. Therefore, by (54) and the linearity of ρ and P_i,

$$f_\rho(p,q) = \sum_{i=1}^{\infty} \lambda_i f_{P_i}(p,q).$$

The projection operators P_i correspond to densities

$$f_{P_i}(p,q) = \left| \int_{\mathbf{R}^n} e_i(x) \psi(x|p,q)\, dx \right|^2,$$

where $e_i \in L_2(\mathbf{R}^n)$ and $\|e_i\|_2 = 1$. Theorem 25 is thus the immediate consequence of the following two lemmas.

Lemma 6: For any two random vectors X, Y in $\mathbf{R}^{2n}$ and any $0 \leq \lambda \leq 1$, let $Z = B_\lambda X + (1 - B_\lambda) Y$, where B_λ denotes a Bernoulli (λ) random variable, independent of both X and Y. The density of Z is therefore the convex combination $\lambda f + (1-\lambda) g$, where f, g are the densities of X and Y, respectively. Then,

$$\lambda J(X) + (1-\lambda) J(Y) - J(Z) \geq 0.$$

Proof: With this notation, after some manipulations we obtain

$$\begin{aligned} &\lambda J(X) + (1-\lambda) J(Y) - J(Z) \\ &\quad = \lambda(1-\lambda) \int \frac{g(p,q) f(p,q)}{\lambda f(p,q) + (1-\lambda) g(p,q)} \\ &\qquad \cdot \left(\nabla \ln \frac{f(p,q)}{g(p,q)} \right)' \left(\nabla \ln \frac{f(p,q)}{g(p,q)} \right) dp\, dq. \end{aligned} \quad (55)$$

Since $(\nabla \ln f(p,q)/g(p,q))'(\nabla \ln f(p,q)/g(p,q)) \geq 0$, the integral in the right side of (55) is nonnegative and the proof is complete. □

Lemma 7: For any random vector X in $\mathbb{R}^{2n}$ with a density of the form $f(p,q) = |\int e(x) \psi(x|p,q)\, dx|^2$ where $e \in L_2(\mathbf{R}^{2n})$ and $\|e\|_2 = 1$,

$$J(X) = 2n.$$

The proof of this lemma is by direct calculation. (For details see [20]).

V. Determinant Inequalities

A. Basic Inequalities

Throughout we will assume that K is a nonnegative definite symmetric $n \times n$ matrix. Let $|K|$ denote the determinant of K. In Section III, we have seen that the entropy power inequality yields the Minkowski inequality (see Theorem 5) and the concavity of $\ln |K|$ (see Theorem 8).

We now give Hadamard's inequality using the proof in [17]. See also [33] for an alternative proof.

Theorem 26 (Hadamard): $|K| \leq \prod_{i=1}^{n} K_{ii}$, with equality iff $K_{ij} = 0$, $i \neq j$.

Proof: Let $X \sim \phi_K$. Then

$$\begin{aligned} \frac{1}{2} \ln (2\pi e)^n |K| &= h(X_1, X_2, \cdots, X_n) \\ &\leq \sum_{i=1}^{n} h(X_i) = \sum_{i=1}^{n} \frac{1}{2} \ln 2\pi e |K_{ii}|, \end{aligned}$$

with equality iff $X_1, X_2, \cdots, X_n$ are independent, i.e., $K_{ij} = 0$, $i \neq j$. □

We now provide a direct information theoretic proof of Fan's (see [22]) Theorem 8 (which states that $\ln|K|$ is a concave function of K). This proof does not use the entropy power inequality, and provides an alternative to the proof in Section III.

Proof of Theorem 8: Let X_1 and X_2 be normally distributed n-vectors, $X_i \sim \phi_{K_i}(x)$, $i=1,2$. Let the random variable θ have distribution $\Pr\{\theta=1\}=\lambda$, $\Pr\{\theta=2\}=1-\lambda$, $0\le\lambda\le 1$. Let θ, X_1, and X_2 be independent and let $Z=X_\theta$. Then Z has covariance $K_Z=\lambda K_1+(1-\lambda)K_2$. However, Z will not be multivariate normal. By first using Lemma 5, followed by Lemma 3, we have

$$\frac{1}{2}\ln(2\pi e)^n|\lambda K_1+(1-\lambda)K_2|\ge h(Z)\ge h(Z|\theta)$$
$$=\lambda\frac{1}{2}\ln(2\pi e)^n|K_1|+(1-\lambda)\frac{1}{2}\ln(2\pi e)^n|K_2|.$$

Thus,

$$|\lambda K_1+(1-\lambda)K_2|\ge|K_1|^\lambda|K_2|^{1-\lambda}, \tag{56}$$

as desired. □

Taking logarithms and letting $\lambda K_1=A$, $(1-\lambda)K_2=B$, we obtain

$$\log|A+B|\ge\lambda\log\left|\frac{A}{\lambda}\right|+(1-\lambda)\log\left|\frac{B}{(1-\lambda)}\right|$$
$$=\lambda\log|A|+(1-\lambda)\log|B|+nH(\lambda). \tag{57}$$

Maximizing the right-hand side over λ, we obtain the optimum value of λ as $|A|^{1/n}/(|A|^{1/n}+|B|^{1/n})$. Substituting this in (57), we obtain the Minkowski inequality (Theorem 5).

We now prove a property of Toeplitz matrices. A Toeplitz matrix K, which arises as the covariance matrix of a stationary random process, is characterized by the property that $K_{ij}=K_{rs}$ if $|i-j|=|r-s|$. Let K_k denote the principal minor $K(1,2,\cdots,k)$. The following property can be proved easily from the properties of the entropy function.

Theorem 27: If the positive definite $n\times n$ matrix K is Toeplitz, then

$$|K_1|\ge|K_2|^{1/2}\ge\cdots\ge|K_{n-1}|^{1/(n-1)}\ge|K_n|^{1/n}$$

and $|K_k|/|K_{k-1}|$ is decreasing in k.

Proof: Let $(X_1,X_2,\cdots,X_n)\sim\phi_{K_n}$. Then the quantities $h(X_k|X_{k-1},\cdots,X_1)$ are decreasing in k, since

$$h(X_k|X_{k-1},\cdots,X_1)=h(X_{k+1}|X_k,\cdots,X_2)$$
$$\ge h(X_{k+1}|X_k,\cdots,X_2,X_1) \tag{58}$$

where the equality follows from the Toeplitz assumption and the inequality from the fact that conditioning reduces entropy. Thus the running averages

$$\frac{1}{k}h(X_1,\cdots,X_k)=\frac{1}{k}\sum_{i=1}^{k}h(X_i|X_{i-1},\cdots,X_1)$$

are decreasing in k. The theorem then follows from $h(X_1,X_2,\cdots,X_k)=(1/2)\ln(2\pi e)^k|K_k|$. □

Remark: Since $h(X_n|X_{n-1},\cdots,X_1)$ is a decreasing sequence, it has a limit. Hence, by the Cesáro mean limit theorem,

$$\lim_{n\to\infty}\frac{h(X_1,X_2,\cdots,X_n)}{n}$$
$$=\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}h(X_k|X_{k-1},\cdots,X_1)$$
$$=\lim_{n\to\infty}h(X_n|X_{n-1},\cdots,X_1). \tag{59}$$

Translating this to determinants, one obtains the result

$$\lim_{n\to\infty}|K_n|^{1/n}=\lim_{n\to\infty}\frac{|K_n|}{|K_{n-1}|}.$$

B. Inequalities for Ratios of Determinants

We first prove a stronger version of Hadamard's theorem due to K. Fan [23].

Theorem 28: For all $1\le p\le n$,

$$\frac{|K|}{|K(p+1,p+2,\cdots,n)|}\le\prod_{i=1}^{p}\frac{|K(i,p+1,p+2,\cdots,n)|}{|K(p+1,p+2,\cdots,n)|}.$$

Proof: We use the same idea as in Theorem 26, except that we use the conditional form of Lemma 4, to obtain

$$\frac{1}{2}\ln(2\pi e)^p\frac{|K|}{|K(p+1,p+2,\cdots,n)|}$$
$$=h(X_1,X_2,\cdots,X_p|X_{p+1},X_{p+2},\cdots,X_n)$$
$$\le\sum_{i=1}^{p}h(X_i|X_{p+1},X_{p+2},\cdots,X_n)$$
$$=\sum_{i=1}^{p}\frac{1}{2}\ln 2\pi e\frac{|K(i,p+1,p+2,\cdots,n)|}{|K(p+1,p+2,\cdots,n)|}. \quad\square \tag{60}$$

If $(X_1,X_2,\cdots,X_n)\sim\phi_{K_n}$, we know that the conditional density of X_n given $(X_1,X_2,\cdots,X_{n-1})$ is univariate normal with mean linear in $X_1,X_2,\cdots,X_{n-1}$ and conditional variance σ_n^2. Here σ_n^2 is the minimum mean-square error $E(X_n-\hat{X}_n)^2$ over all linear estimators $\hat{X}_n$ based on $X_1,X_2,\cdots,X_{n-1}$.

Lemma 8: $\sigma_n^2=|K_n|/|K_{n-1}|$.

Proof: Using the conditional normality of X_n, Lemma 2 results in

$$\frac{1}{2}\ln 2\pi e\sigma_n^2=h(X_n|X_1,X_2,\cdots,X_{n-1})$$
$$=h(X_1,X_2,\cdots,X_n)-h(X_1,X_2,\cdots,X_{n-1})$$
$$=\frac{1}{2}\ln(2\pi e)^n|K_n|-\frac{1}{2}\ln(2\pi e)^{n-1}|K_{n-1}|$$
$$=\frac{1}{2}\ln 2\pi e|K_n|/|K_{n-1}|. \quad\square \tag{61}$$

Minimization of σ_n^2 over a set of allowed covariance matrices $\{K_n\}$ is aided by the following theorem.

Theorem 29: $\ln(|K_n|/|K_{n-p}|)$ is concave in K_n.

Proof: We remark that Theorem 8 is not applicable because $\ln(|K_n|/|K_{n-p}|)$ is the difference of two concave functions. Let $\mathbf{Z} = X_\theta$, where $X_1 \sim \phi_{S_n}(x)$, $X_2 \sim \phi_{T_n}(x)$, $\Pr\{\theta = 1\} = \lambda = 1 - \Pr\{\theta = 2\}$, and X_1, X_2, θ are independent. The covariance matrix K_n of $\mathbf{Z}$ is given by

$$K_n = \lambda S_n + (1-\lambda) T_n.$$

The following chain of inequalities proves the theorem:

$$\begin{aligned}
&\lambda \frac{1}{2} \ln (2\pi e)^p |S_n|/|S_{n-p}| + (1-\lambda)\frac{1}{2}\ln(2\pi e)^p |T_n|/|T_{n-p}| \\
&\quad \overset{(a)}{=} \lambda h\left(X_{1,n}, X_{1,n-1}, \cdots, X_{1,n-p+1} | X_{1,1}, \cdots, X_{1,n-p}\right) + \\
&\qquad (1-\lambda) h\left(X_{2,n}, X_{2,n-1}, \cdots, X_{2,n-p+1} | X_{2,1}, \cdots, X_{2,n-p}\right) \\
&\quad = h\left(Z_n, Z_{n-1}, \cdots, Z_{n-p+1} | Z_1, \cdots, Z_{n-p}, \theta\right) \\
&\quad \overset{(b)}{\le} h\left(Z_n, Z_{n-1}, \cdots, Z_{n-p+1} | Z_1, \cdots, Z_{n-p}\right) \\
&\quad \overset{(c)}{\le} \frac{1}{2} \ln (2\pi e)^p \frac{|K_n|}{|K_{n-p}|},
\end{aligned} \tag{62}$$

where a) follows from

$$\begin{aligned}
&h\left(X_n, X_{n-1}, \cdots, X_{n-p+1} | X_1, \cdots, X_{n-p}\right) \\
&\qquad = h(X_1, \cdots, X_n) - h(X_1, \cdots, X_{n-p}),
\end{aligned}$$

b) follows from the conditioning lemma, and c) follows from a conditional version of Lemma 5.

Theorem 29 for the case $p = 1$ is due to Bergstrøm [4]. However, for $p - 1$, we can prove an even stronger theorem, also due to Bergstrøm [4]. □

Theorem 30: $|K_n|/|K_{n-1}|$ is concave in K_n.

Proof: Again we use the properties of normal random variables. Let us assume that we have two independent normal random vectors, $X \sim \phi_{A_n}$ and $Y \sim \phi_{B_n}$. Let $\mathbf{Z} = X + Y$.

Then

$$\begin{aligned}
\frac{1}{2}\ln 2\pi e \frac{|A_n + B_n|}{|A_{n-1} + B_{n-1}|} &\overset{(a)}{=} h(Z_n | Z_{n-1}, Z_{n-2}, \cdots, Z_1) \\
&\overset{(b)}{\ge} h(Z_n | Z_{n-1}, Z_{n-2}, \cdots, Z_1, X_{n-1}, X_{n-2}, \cdots, X_1, Y_{n-1}, Y_{n-2}, \cdots, Y_1) \\
&\overset{(c)}{=} h(X_n + Y_n | X_{n-1}, X_{n-2}, \cdots, X_1, Y_{n-1}, Y_{n-2}, \cdots, Y_1) \\
&\overset{(d)}{=} E\frac{1}{2}\ln\left[2\pi e \operatorname{var}(X_n + Y_n | X_{n-1}, X_{n-2}, \cdots, X_1, Y_{n-1}, Y_{n-2}, \cdots, Y_1)\right] \\
&\overset{(e)}{=} E\frac{1}{2}\ln\left[2\pi e\left(\operatorname{var}(X_n | X_{n-1}, X_{n-2}, \cdots, X_1) + \operatorname{var}(Y_n | Y_{n-1}, Y_{n-2}, \cdots, Y_1)\right)\right] \\
&\overset{(f)}{=} E\frac{1}{2}\ln\left(2\pi e\left(\frac{|A_n|}{|A_{n-1}|} + \frac{|B_n|}{|B_{n-1}|}\right)\right) \\
&= \frac{1}{2}\ln\left(2\pi e\left(\frac{|A_n|}{|A_{n-1}|} + \frac{|B_n|}{|B_{n-1}|}\right)\right).
\end{aligned} \tag{63}$$

In this derivation, a) follows from Lemma 8, b) from the fact the conditioning decreases entropy and c) follows from the fact that Z is a function of X and Y. The sum $X_n + Y_n$ is normal conditioned on $X_1, X_2, \cdots, X_{n-1}, Y_1, Y_2, \cdots, Y_{n-1}$, and hence, we can express its entropy in terms of its variance, obtaining equality d). Then e) follows from the independence of X_n and Y_n conditioned on the past $X_1, X_2, \cdots, X_{n-1}, Y_1, Y_2, \cdots, Y_{n-1}$, and f) follows from the fact that for a set of jointly normal random variables, the conditional variance is constant, independent of the conditioning variables (Lemma 8).

In general, by setting $A = \lambda S$ and $B = (1-\lambda)T$, we obtain

$$\frac{|\lambda S_n + (1-\lambda) T_n|}{|\lambda S_{n-1} + (1-\lambda) T_{n-1}|} \ge \lambda \frac{|S_n|}{|S_{n-1}|} + (1-\lambda)\frac{|T_n|}{|T_{n-1}|},$$

i.e., $|K_n|/|K_{n-1}|$ is concave. □

Simple examples show that $|K_n|/|K_{n-p}|$ is not necessarily concave for $p \ge 2$.

C. Subset Inequalities for Determinants

We now prove a generalization of Hadamard's inequality due to Szasz [35]. Let $K(i_1, i_2, \cdots, i_k)$ be the principal submatrix of K formed by the rows and columns with indexes $i_1, i_2, \cdots, i_k$.

Theorem 31 (Szasz): If K is a positive definite $n \times n$ matrix and P_k denotes the product of all the principal k-rowed minors of K, i.e.,

$$P_k = \prod_{1 \le i_1 < i_2 < \cdots < i_k \le n} |K(i_1, i_2, \cdots, i_k)|,$$

then

$$P_1 \geq P_2^{1/\binom{n-1}{2}} \geq P_3^{1/\binom{n-1}{2}} \geq \cdots \geq P_n.$$

Proof: Let $X \sim \phi_K$. Then the theorem follows directly from Theorem 1, with the identification $h_k^{(n)} = (1/n)\ln P_k + (1/2)\ln 2\pi e$. □

We can also prove a related theorem.

Theorem 32: Let K be a positive definite $n \times n$ matrix and let

$$S_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 \leq i_2 < \cdots < i_k \leq n} |K(i_1, i_2, \cdots, i_k)|^{1/k}.$$

Then,

$$\frac{1}{n}\operatorname{tr}(K) = S_1^{(n)} \geq S_2^{(n)} \geq \cdots \geq S_n^{(n)} = |K|^{1/n}.$$

Proof: This follows directly from the corollary to Theorem 1, with the identification $s_k^{(n)} = (2\pi e)S_k^{(n)}$, and $r = 2$ in (3) and (4). □

Define the geometric mean of $(|K|/|K(S^c)|)^{1/k}$ over k-element subsets by

$$Q_k = \left(\prod_{S:|S|=k} \frac{|K|}{|K(S^c)|} \right)^{1/k\binom{n}{k}}.$$

Theorem 33:

$$\prod_{i=1}^{n} \sigma_i^2 = Q_1 \leq Q_2 \leq \cdots \leq Q_{n-1} \leq Q_n = |K|^{1/n}.$$

Proof: The theorem follows immediately from Theorem 2 and the identification

$$h(X(S)|X(S^c)) = \frac{1}{2}\ln(2\pi e)^k \frac{|K|}{|K(S^c)|}.$$

The outermost inequality, $Q_1 \leq Q_n$, can be rewritten as

$$|K| \geq \prod_{i=1}^{n} \sigma_i^2,$$

where

$$\sigma_i^2 = \frac{|K|}{|K(1,2,\cdots,i-1,i+1,\cdots,n)|} \tag{64}$$

is the minimum mean-squared error in the linear prediction of X_i from the remaining X's. It is the conditional variance of X_i given the remaining X_j's if $X_1, X_2, \cdots, X_n$ is jointly normal. Combining this with Hadamard's inequality gives upper and lower bounds on the determinant of a positive definite matrix.

Corollary 4:

$$\prod_i K_{ii} \geq |K| \geq \prod_i \sigma_i^2.$$

Hence, the determinant of a covariance matrix lies between the product of the unconditional variances K_{ii} of the random variables X_i and the product of the conditional variances σ_i^2.

Let

$$R_k = \left(\prod_{S:|S|=k} \frac{|K(S)||K(S^c)|}{|K|} \right)^{1/k\binom{n}{k}}.$$

Theorem 34:

$$R_1 \geq R_2 \geq \cdots \geq R_{n-1} \geq R_n.$$

Proof: The theorem follows immediately from Corollary 2 and the identification

$$I(X(S); X(S^c)) = \frac{1}{2}\ln \frac{|K(S)||K(S^c)|}{|K|}.$$

In particular, the outer inequality $R_1 \geq R_n$ results in

$$\left(\prod_{i=1}^{n} \frac{|K_{ii}||K(\{i\}^c)|}{|K|} \right)^{1/n} \geq 1. \qquad \square \tag{65}$$

Finally, we can convert Theorem 3 into a statement about determinants by considering $X_1, X_2, \cdots, X_n$ to be normally distributed with covariance matrix K.

Let

$$T_k = \left(\prod_{S:|S|=k} \frac{|K(S)||K(S^c)|}{|K|} \right)^{1/\binom{n}{k}}.$$

Theorem 35:

$$T_1 \leq T_2 \leq \cdots \leq T_{\lfloor n/2 \rfloor}.$$

Proof: The theorem follows directly from Theorem 3 and (64). □

Acknowledgment

A. Dembo thanks S. Karlin for pointing attention to [8] and [30] and Y. Peres and G. Kalai for pointing attention to [10]. The authors also thank E. Carlen for providing preprints of [11], [12], and [13].

References

[1] D. Bakry and M. Emery, "Seminaire de Probabilities XIX," in *Lecture Notes in Mathematics, 1123*. New York: Springer, 1985, pp. 179–206.

[2] A. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, no. 1, pp. 336–342, 1986.

[3] W. Beckner, "Inequalities in Fourier analysis," *Ann. Math.*, vol. 102, pp. 159–182, 1975.

[4] R. Bellman, "Notes on matrix theory—IV: An inequality due to Bergstrøm," *Amer. Math. Monthly*, vol. 62, pp. 172–173, 1955.

[5] P. P. Bergmans, "A simple converse for broadcast channels with additive white normal noise," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 279–280, 1974.

[6] N. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 267–271, Apr. 1965.

[7] H. J. Brascamp and E. J. Lieb, "Some inequalities for Gaussian measures and the long range order of the one dimensional plasma," in *Functional Integration and Its Applications*, A. M. Arthors, Ed. Oxford: Clarendon Press, 1975.

[8] ____, "Best constants in Young's inequality, its converse and its generalization to more than three functions," *Adv. Math.*, vol. 20, pp. 151–173, 1976.

[9] ____, "On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation," *J. Functional Anal.*, vol. 22, pp. 366–389, 1976.

[10] Y. D. Burago and V. A. Zalgaller, *Geometric Inequalities*. New York: Springer Verlag, 1980.
[11] E. A. Carlen, "Some integral identities and inequalities for entire functions and their application to the coherent state transform," *J. Functional Anal.*, 1991.
[12] ____, "Superadditivity of Fisher's information and logarithmic Sobolev inequalities," *J. Functional Anal.*, 1991.
[13] E. A. Carlen and A. Soffer, "Entropy production by convolution and central limit theorems with strong rate information," *Commun. Math. Phys.*, 1991.
[14] M. Costa and T. M. Cover, "On the similarity of the entropy power inequality and the Brunn–Minkowski inequality," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 837–839, 1984.
[15] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 751–760, 1985.
[16] T. M. Cover and J. A. Thomas, "Determinant inequalities via information theory," *SIAM J. Matrix Anal. and its Applicat.*, vol. 9, no. 3, pp. 384–392, July 1988.
[17] T. M. Cover and A. El Gamal, "An information theoretic proof of Hadamard's inequality," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 930–931, Nov. 1983.
[18] I. Csiszar, "Informationstheoretische konvegenenzbegriffe im raum der vahrscheinlichkeitsverteilungen," *Publ. Math. Inst., Hungarian Academy of Sci.*, VII, ser. A, pp. 137–157, 1962.
[19] A. Dembo, "A simple proof of the concavity of the entropy power with respect to the variance of additive normal noise," *IEEE Trans. Inform. Theory*, vol. 35, pp. 887–888, July 1989.
[20] ____, "Information inequalities and uncertainty principles," Tech. Rep., Dept. of Statist., Stanford Univ., Stanford, CA, 1990.
[21] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, vol. 49, pp. 906–931, 1989.
[22] K. Fan, "On a theorem of Weyl concerning the eigenvalues of linear transformations II," *Proc. National Acad. Sci. U.S.*, vol. 36, 1950, pp. 31–35.
[23] ____, "Some inequalities concerning positive-definite matrices," *Proc. Cambridge Phil. Soc.*, vol. 51, 1955, pp. 414–421.
[24] H. Federer, *Geometric measure theory, vol. B 153 of Grundl. Math. Wiss.* Berlin: Springer-Verlag, 1969.
[25] L. Gross, "Logarithmic Sobolev inequalities," *Amer. J. Math.*, vol. 97, pp. 1061–1083, 1975.
[26] ____, "Logarithmic Sobolev inequalities for the heat kernel on a Lie group," in *White Noise Analysis*. Singapore: World Scientific, 1990.
[27] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inform. Contr.*, vol. 36, pp. 133–156, 1978.
[28] I. I. Hirschman, "A note on entropy," *Amer. J. Math.*, vol. 79, pp. 152–156, 1957.
[29] S. Kullback, "A lower bound for discrimination information in terms of variation," *IEEE Trans. Inform. Theory*, vol. IT-4, pp. 126–127, 1967.
[30] E. H. Lieb, "Proof of an entropy conjecture of Wehrl," *Commun. Math. Phys.*, vol. 62, pp. 35–41, 1978.
[31] ____, "Gaussian kernels have Gaussian maximizers," *Inventions Math.*, vol. 102, pp. 179–208, 1990.
[32] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*. New York: Academic Press, 1979.
[33] ____, "A convexity proof of Hadamard's inequality," *Amer. Math. Monthly*, vol. 89, pp. 687–688, 1982.
[34] H. Minkowski, "Diskontinuitätsbereich für arithmetische äquivalenz," *J. für Math.*, vol. 129, pp. 220–274, 1950.
[35] L. Mirsky, "On a generalization of Hadamard's determinantal inequality due to Szasz," *Arch. Math.*, vol. 8, pp. 274–275, 1957.
[36] A. Oppenheim, "Inequalities connected with definite Hermitian forms," *J. Lon. Math. Soc.*, vol. 5, pp. 114–119, 1930.
[37] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[38] A. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inform. Contr.*, vol. 2, pp. 101–112, 1959.
[39] A. Wehrl, "General properties of entropy," *Rev. Modern Phys.*, vol. 50, pp. 221–260, 1978.
[40] M. Zakai, "A class of definitions of 'duration' (or 'uncertainty') and the associated uncertainty relations," *Inform. Contr.*, vol. 3, pp. 101–115, 1960.