The Monotonicity of Information in the Central Limit Theorem and Entropy Power Inequalities

Mokshay Madiman and Andrew Barron
Department of Statistics
Yale University
Email: mokshay.madiman , andrew.barron @yale.edu

***Abstract*— We provide a simple proof of the monotonicity of information in the Central Limit Theorem for i.i.d. summands. Extensions to the more general case of independent, not identically distributed summands are also presented. New families of Fisher information and entropy power inequalities are discussed.**

I. Introduction

Let $X_1, X_2, \ldots, X_n$ be independent random variables with densities and finite variances, and let H denote the (differential) entropy. The classical entropy power inequality of Shannon [1] and Stam [2] states

$$e^{2H(X_1+\ldots+X_n)} \geq \sum_{j=1}^{n} e^{2H(X_j)}. \tag{1}$$

Recently, Artstein, Ball, Barthe and Naor [3] proved a new entropy power inequality

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{n-1} \sum_{i=1}^{n} e^{2H\left(\sum_{j\neq i} X_j\right)}, \tag{2}$$

where each term involves the entropy of the sum of $n-1$ of the variables excluding the i-th, which is an improvement over (1). Indeed, repeated application of (2) for a succession of values of n yields not only (1) but also a whole family of intermediate inequalities

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{\binom{n-1}{m-1}} \sum_{s\in\Omega_m} e^{2H\left(\sum_{j\in s} X_j\right)}, \tag{3}$$

where we write Ω_m for the collection of all subsets of $\{1, 2, \ldots, n\}$ of size m. Below, we give a simplified and direct proof of (3) (and, in particular, of (2)), and also show that equality holds if and only if the X_i are normally distributed (in which case it becomes an identity for sums of variances).

In fact, all these inequalities are particular cases of a generalized entropy power inequality, which we develop in [4]. Let $\mathcal{S}$ be an arbitrary collection of subsets of $\{1, \ldots, n\}$ and let $r = r(\mathcal{S}, n)$ be the maximum number of subsets in $\mathcal{S}$ in which any one index i can appear, for $i = 1, \ldots, n$. Then

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{r} \sum_{s\in\mathcal{S}} e^{2H\left(\sum_{j\in s} X_j\right)}. \tag{4}$$

For example, if $\mathcal{S}$ consists of subsets s whose elements are m consecutive indices in $\{1, \ldots, n\}$, then $r = m$, whereas if $\mathcal{S} = \Omega_m$, then $r = \binom{n-1}{m-1}$. So (4) extends (3). Likewise for general collections we have a corresponding inequality for inverse Fisher information. Details of these results can be found in [4].

These inequalities are relevant for the examination of monotonicity in central limit theorems. Indeed, if X_1 and X_2 are independent and identically distributed (i.i.d.), then (1) is equivalent to

$$H\left(\frac{X_1+X_2}{\sqrt{2}}\right) \geq H(X_1). \tag{5}$$

This fact implies that the entropy of the standardized sums $Y_n = \frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}$ increases along the powers-of-2 subsequence, i.e., $H(Y_{2^k})$ is non-decreasing in k. Characterization of the increase in entropy in (5) was used in proofs of central limit theorems by Shimizu [5], Barron [6] and Johnson and Barron [7]. In particular, Barron [6] showed that the sequence $\{H(Y_n)\}$ of entropies of the normalized sums converges to the entropy of the normal; this, incidentally, is equivalent to the convergence to 0 of the relative entropy (Kullback divergence) from a normal distribution when the X_i have zero mean.

In 2004, Artstein, Ball, Barthe and Naor [3] (hereafter denoted by ABBN [3]) showed that $H(Y_n)$ is in fact a non-decreasing sequence for every n, solving a long-standing conjecture. In fact, (2) is equivalent in the i.i.d. case to the monotonicity property

$$H\left(\frac{X_1+\ldots+X_n}{\sqrt{n}}\right) \geq H\left(\frac{X_1+\ldots+X_{n-1}}{\sqrt{n-1}}\right). \tag{6}$$

Note that the presence of the factor $n-1$ (rather than n) in the denominator of (2) is crucial for this monotonicity.

Likewise, for sums of independent random variables, our inequality (3) is equivalent to "monotonicity on average" properties for certain standardizations; for instance,

$$\exp\left\{2H\left(\frac{X_1+\ldots+X_n}{\sqrt{n}}\right)\right\} \geq \frac{1}{\binom{n}{m}} \sum_{s\in\Omega_m} \exp\left\{2H\left(\frac{\sum_{i\in s} X_i}{\sqrt{m}}\right)\right\}.$$

A similar monotonicity also holds, as we shall show, when the sums are standardized by their variances. Here again the $\binom{n-1}{m-1}$ rather than $\binom{n}{m}$ in the denominator of (3) for the unstandardized version is crucial.

We find that all the above inequalities (including (2)) as well as corresponding inequalities for Fisher information can be proved by simple tools. Two of these tools, a convolution identity for score functions and the relationship between Fisher

information and entropy (discussed in Section II), are familiar in past work on entropy power inequalities. An additional trick is needed to obtain the denominators of $n-1$ and $\binom{n-1}{m-1}$ in (2) and (3) respectively. This is a simple variance drop inequality for statistics expressible via sums of functions of m out of n variables, which is familiar in other statistical contexts (as we shall discuss). It was first used for information inequality development in ABBN [3]. The variational characterization of Fisher information that is an essential ingredient of ABBN [3] is not needed in our proofs.

For clarity of presentation, we find it convenient to first outline the proof of (6) for i.i.d. random variables. Thus, in Section III, we establish the monotonicity result (6) in a simple and revealing manner that boils down to the geometry of projections (conditional expectations). Whereas ABBN [3] requires that X_1 has a C^2 density for monotonicity of Fisher divergence, absolute continuity of the density suffices in our approach. Furthermore, whereas the recent preprint of Shlyakhtenko [8] proves the analogue of the monotonicity fact for non-commutative or "free" probability theory, his method implies a proof for the classical case only assuming finiteness of all moments, while our direct proof requires only finite variance assumptions. Our proof also reveals in a simple manner the cases of equality in (6) (c.f., Schultz [9]). Although we do not write it out for brevity, the monotonicity of entropy for standardized sums of d-dimensional random vectors has an identical proof.

We recall that for a random variable X with density f, the entropy is $H(X) = -E[\log f(X)]$. For a differentiable density, the score function is $\rho_X(x) = \frac{\partial}{\partial x} \log f(x)$, and the Fisher information is $I(X) = E[\rho_X^2(X)]$. They are linked by an integral form of the de Bruijn identity due to Barron [6], which permits certain convolution inequalities for I to translate into corresponding inequalities for H.

Underlying our inequalities is the demonstration for independent, not necessarily identically distributed (i.n.i.d.) random variables with absolutely continuous densities that

$$I(X_1 + \ldots + X_n) \leq \binom{n-1}{m-1} \sum_{s \in \Omega_m} w_s^2 I\Big(\sum_{i \in s} X_i\Big) \qquad (7)$$

for any non-negative weights w_s that add to 1 over all subsets $s \subset \{1, \ldots, n\}$ of size m. Optimizing over w yields an inequality for inverse Fisher information that extends the original inequality of Stam:

$$\frac{1}{I(X_1 + \ldots + X_n)} \geq \frac{1}{\binom{n-1}{m-1}} \sum_{s \in \Omega_m} \frac{1}{I(\sum_{i \in s} X_i)}. \qquad (8)$$

Alternatively, using a scaling property of Fisher information to re-express our core inequality (7), we see that the Fisher information of the sum is bounded by a convex combination of Fisher informations of scaled partial sums:

$$I(X_1 + \ldots + X_n) \leq \sum_{s \in \Omega_m} w_s I\Big(\frac{\sum_{i \in s} X_i}{\sqrt{w_s \binom{n-1}{m-1}}}\Big). \qquad (9)$$

This integrates to give an inequality for entropy that is an extension of the "linear form of the entropy power inequality" developed by Dembo et al [10]. Specifically we obtain

$$H(X_1 + \ldots + X_n) \leq \sum_{s \in \Omega_m} w_s H\Big(\frac{\sum_{i \in s} X_i}{\sqrt{w_s \binom{n-1}{m-1}}}\Big). \qquad (10)$$

Likewise using the scaling property of entropy on (10) and optimizing over w yields our extension of the entropy power inequality

$$\exp\{2H(X_1 + \ldots + X_n)\} \geq \frac{1}{\binom{n-1}{m-1}} \sum_{s \in \Omega_m} \exp\Big\{2H\big(\sum_{j \in s} X_j\big)\Big\}.$$

Thus both inverse Fisher information and entropy power satisfy an inequality of the form

$$\binom{n-1}{m-1} \psi\big(X_1 + \ldots + X_n\big) \geq \sum_{s \in \Omega_m} \psi\Big(\sum_{i \in s} X_i\Big). \qquad (11)$$

We motivate the form (11) using the following almost trivial fact, which is proved in the Appendix for the reader's convenience. Let $[n] = \{1, 2, \ldots, n\}$.

Fact I: For arbitrary non-negative numbers $\{a_i^2 : i \in [n]\}$,

$$\sum_{s \in \Omega_m} \sum_{i \in s} a_i^2 = \binom{n-1}{m-1} \sum_{i \in [n]} a_i^2, \qquad (12)$$

where the first sum on the left is taken over the collection $\Omega_m = \{s \subset [n] : |s| = m\}$ of sets containing m indices.

If Fact I is thought of as (m, n)-additivity, then (8) and (3) represent the (m, n)-superadditivity of inverse Fisher information and entropy power respectively. In the case of normal random variables, the inverse Fisher information and the entropy power equal the variance. Thus in that case (8) and (3) become Fact I with a_i^2 equal to the variance of X_i.

II. Score Functions and Projections

The first tool we need is a projection property of score functions of sums of independent random variables, which is well-known for smooth densities (c.f., Blachman [11]). For completeness, we give the proof. As shown by Johnson and Barron [7], it is sufficient that the densities are absolutely continuous; see [7][Appendix 1] for an explanation of why this is so.

Lemma I:[Convolution identity for score functions] If V_1 and V_2 are independent random variables, and V_1 has an absolutely continuous density with score ρ_{V_1}, then $V_1 + V_2$ has the score

$$\rho_{V_1+V_2}(v) = E[\rho_{V_1}(V_1) | V_1 + V_2 = v] \qquad (13)$$

Proof: Let f_{V_1} and f_V be the densities of V_1 and $V = V_1 + V_2$ respectively. Then, either bringing the derivative

inside the integral for the smooth case, or via the more general formalism in [7],

$$\begin{aligned} f'_V(v) &= \frac{\partial}{\partial v} E[f_{V_1}(v - V_2)] \\ &= E[f'_{V_1}(v - V_2)] \\ &= E[f_{V_1}(v - V_2)\rho_{V_1}(v - V_2)] \end{aligned} \tag{14}$$

so that

$$\begin{aligned} \rho_V(v) = \frac{f'_V(v)}{f_V(v)} &= E\left[\frac{f_{V_1}(v - V_2)}{f_V(v)}\rho_{V_1}(v - V_2)\right] \\ &= E[\rho_{V_1}(V_1)|V_1 + V_2 = v]. \end{aligned} \tag{15}$$

■

The second tool we need is a "variance drop lemma", which goes back at least to Hoeffding's seminal work [12] on U-statistics (see his Theorem 5.2). An equivalent statement of the variance drop lemma was formulated in ABBN [3]. In [4], we prove and use a more general result to study the i.n.i.d. case.

First we need to recall a decomposition of functions in $L^2(\mathbb{R}^n)$, which is nothing but the Analysis of Variance (ANOVA) decomposition of a statistic. The following conventions are useful. $[n]$ is the index set $\{1, 2, \ldots, n\}$. For any $s \subset [n]$, X_s stands for the collection of random variables $\{X_i : i \in s\}$. For any $j \in [n]$, $E_j\psi$ denotes the conditional expectation of ψ, given all random variables other than X_j, i.e.,

$$E_j\psi(x_1, \ldots, x_n) = E[\psi(X_1, \ldots, X_n)|X_i = x_i \quad \forall i \neq j] \tag{16}$$

averages out the dependence on the j-th coordinate.

Fact II:[ANOVA DECOMPOSITION] Suppose $\psi : \mathbb{R}^n \to \mathbb{R}$ satisfies $E\psi^2(X_1, \ldots, X_n) < \infty$, i.e., $\psi \in L^2$, for independent random variables $X_1, X_2, \ldots, X_n$. For $s \subset [n]$, define the orthogonal linear subspaces

$$\mathcal{H}_s = \{\psi \in L^2 : E_j\psi = \psi 1_{\{j \notin s\}} \ \forall j \in [n]\} \tag{17}$$

of functions depending only on the variables indexed by s. Then L^2 is the orthogonal direct sum of this family of subspaces, i.e., any $\psi \in L^2$ can be written in the form

$$\psi = \sum_{s \subset [n]} \psi_s, \tag{18}$$

where $\psi_s \in \mathcal{H}_s$.

Remark: In the language of ANOVA familiar to statisticians, when ϕ is the empty set, ψ_ϕ is the mean; $\psi_{\{1\}}, \psi_{\{2\}}, \ldots, \psi_{\{n\}}$ are the main effects; $\{\psi_s : |s| = 2\}$ are the pairwise interactions, and so on. Fact II implies that for any subset $s \subset [n]$, the function $\sum_{\{R:R\subset s\}} \psi_R$ is the best approximation (in mean square) to ψ that depends only on the collection X_s of random variables.

Remark: The historical roots of this decomposition lie in the work of von Mises [13] and Hoeffding [12]. For various refinements and interpretations, see Kurkjian and Zelen [14], Jacobsen [15], Rubin and Vitale [16], Efron and Stein [17], and Takemura [18]; these works include applications of such decompositions to experimental design, linear models, U-statistics, and jackknife theory. The Appendix contains a brief proof of Fact II for the convenience of the reader.

We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is an additive function if there exist functions $f_i : \mathbb{R} \to \mathbb{R}$ such that $f(x_1, \ldots, x_d) = \sum_{i\in[d]} f_i(x_i)$.

Lemma II:[VARIANCE DROP]Let $\psi : \mathbb{R}^{n-1} \to \mathbb{R}$. Suppose, for each $j \in [n]$, $\psi_j = \psi(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$ has mean 0. Then

$$E\left[\sum_{j=1}^n \psi_j\right]^2 \leq (n-1)\sum_{j\in[n]} E[\psi_j]^2. \tag{19}$$

Equality can hold only if ψ is an additive function.

Proof: By the Cauchy-Schwartz inequality, for any V_j,

$$\left[\frac{1}{n}\sum_{j\in[n]} V_j\right]^2 \leq \frac{1}{n}\sum_{j\in[n]} [V_j]^2, \tag{20}$$

so that

$$E\left[\sum_{j\in[n]} V_j\right]^2 \leq n\sum_{j\in[n]} E[V_j]^2. \tag{21}$$

Let $\bar{E}_s$ be the operation that produces the component $\bar{E}_s\psi = \psi_s$ (see the appendix for a further characterization of it); then

$$\begin{aligned} E\left[\sum_{j\in[n]} \psi_j\right]^2 &= E\left[\sum_{s\subset[n]}\sum_{j\in[n]} \bar{E}_s\psi_j\right]^2 \\ &\overset{(a)}{=} \sum_{s\subset[n]} E\left[\sum_{j\notin s} \bar{E}_s\psi_j\right]^2 \\ &\overset{(b)}{\leq} \sum_{s\subset[n]} (n-1)\sum_{j\notin s} E\left[\bar{E}_s\psi_j\right]^2 \\ &\overset{(c)}{=} (n-1)\sum_{j\in[n]} E[\psi_j]^2. \end{aligned} \tag{22}$$

Here, (a) and (c) employ the orthogonal decomposition of Fact II and Parseval's theorem. The inequality (b) is based on two facts: firstly, $E_j\psi_j = \psi_j$ since ψ_j is independent of X_j, and hence $\bar{E}_s\psi_j = 0$ if $j \in s$; secondly, we can ignore the null set ϕ in the outer sum since the mean of a score function is 0, and therefore $\{j : j \notin s\}$ in the inner sum has at most $n-1$ elements. For equality to hold, $\bar{E}_s\psi_j$ can only be non-zero when s has exactly 1 element, i.e., each ψ_j must consist only of main effects and no interactions, so that it must be additive. ■

III. MONOTONICITY IN THE IID CASE

For i.i.d. random variables, inequalities (2) and (3) reduce to the monotonicity $H(Y_n) \geq H(Y_m)$ for $n > m$, where

$$Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i. \tag{23}$$

For clarity of presentation of ideas, we focus first on the i.i.d. case, beginning with Fisher information.

Proposition I:[MONOTONICITY OF FISHER INFORMATION]For i.i.d. random variables with absolutely continuous densities,

$$I(Y_n) \leq I(Y_{n-1}), \tag{24}$$

with equality iff X_1 is normal or $I(Y_n) = \infty$.

Proof: We use the following notational conventions: The (unnormalized) sum is $V_n = \sum_{i\in[n]} X_i$, the leave-one-out sum leaving out X_j is $V^{(j)} = \sum_{i\neq j} X_i$, and the normalized leave-one-out sum is $Y^{(j)} = \frac{1}{\sqrt{n-1}} \sum_{i\neq j} X_i$.

If $X' = aX$, then $\rho_{X'}(X') = \frac{1}{a}\rho_X(X)$; hence

$$\begin{aligned}
\rho_{Y_n}(Y_n) &= \sqrt{n}\rho_{V_n}(V_n) \\
&\overset{(a)}{=} \sqrt{n}E[\rho_{V^{(j)}}(V^{(j)})|V_n] \\
&= \sqrt{\frac{n}{n-1}}E[\rho_{Y^{(j)}}(Y^{(j)})|V_n] \\
&\overset{(b)}{=} \frac{1}{\sqrt{n(n-1)}} \sum_{j=1}^{n} E[\rho_{Y^{(j)}}(Y^{(j)})|Y_n].
\end{aligned} \tag{25}$$

Here, (a) follows from application of Lemma I to $V_n = V^{(j)} + X_j$, keeping in mind that Y_{n-1} (hence $V^{(j)}$) has an absolutely continuous density, while (b) follows from symmetry. Set $\rho_j = \rho_{Y^{(j)}}(Y^{(j)})$; then we have

$$\rho_{Y_n}(Y_n) = \frac{1}{\sqrt{n(n-1)}} E\left[\sum_{j=1}^{n} \rho_j \middle| Y_n\right]. \tag{26}$$

Since the length of a vector is not less than the length of its projection (i.e., by Cauchy-Schwartz inequality),

$$I(Y_n) = E[\rho_{Y_n}(Y_n)]^2 \leq \frac{1}{n(n-1)} E\left[\sum_{j=1}^{n} \rho_j\right]^2. \tag{27}$$

Lemma II yields

$$E\left[\sum_{j=1}^{n} \rho_j\right]^2 \leq (n-1) \sum_{j\in[n]} E[\rho_j]^2 = (n-1)nI(Y_{n-1}), \tag{28}$$

which gives the inequality of Proposition I on substitution into (27). The inequality implied by Lemma II can be tight only if each ρ_j is an additive function, but we already know that ρ_j is a function of the sum. The only functions that are both additive and functions of the sum are linear functions of the sum; hence the two sides of (24) can be finite and equal only if the score ρ_j is linear, i.e., if all the X_i are normal. It is trivial to check that X_1 normal or $I(Y_n) = \infty$ imply equality. ∎

We can now prove the monotonicity result for entropy in the i.i.d. case.

Theorem I: Suppose X_i are i.i.d. random variables with densities. Suppose X_1 has mean 0 and finite variance, and

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \tag{29}$$

Then

$$H(Y_n) \geq H(Y_{n-1}). \tag{30}$$

The two sides are finite and equal iff X_1 is normal.

Proof: Recall the integral form of the de Bruijn identity, which is now a standard method to "lift" results from Fisher divergence to relative entropy. This identity was first stated in its differential form by Stam [2] (and attributed by him to de Bruijn), and proved in its integral form by Barron [6]: if X_t is equal in distribution to $X + \sqrt{t}Z$, where Z is normally distributed independent of X, then

$$H(X) = \tfrac{1}{2}\log(2\pi ev) - \tfrac{1}{2}\int_0^\infty \left[I(X_t) - \frac{1}{v+t}\right] dt \tag{31}$$

is valid in the case that the variances of Z and X are both v. This has the advantage of positivity of the integrand but the disadvantage that is seems to depend on v. One can use

$$\log v = \int_0^\infty \left[\frac{1}{1+t} - \frac{1}{v+t}\right] \tag{32}$$

to re-express it in the form

$$H(X) = \tfrac{1}{2}\log(2\pi e) - \tfrac{1}{2}\int_0^\infty \left[I(X_t) - \frac{1}{1+t}\right] dt. \tag{33}$$

Combining this with Proposition I, the proof is finished. ∎

IV. EXTENSIONS

For the case of independent, non-identically distributed (i.n.i.d.) summands, we need a general version of the "variance drop" lemma.

Lemma III:[VARIANCE DROP: GENERAL VERSION]Suppose we are given a class of functions $\psi^{(s)} : \mathbb{R}^{|s|} \to \mathbb{R}$ for any $s \in \Omega_m$, and $E\psi^{(s)}(X_1, \ldots, X_m) = 0$ for each s. Let w be any probability distribution on Ω_m. Define

$$U(X_1, \ldots, X_n) = \sum_{s\in\Omega_m} w_s \psi^{(s)}(X_s), \tag{34}$$

where we write $\psi^{(s)}(X_s)$ for a function of X_s.Then

$$EU^2 \leq \binom{n-1}{m-1} \sum_{s\in\Omega_m} w_s^2 E[\psi^{(s)}(X_s)]^2, \tag{35}$$

and equality can hold only if each $\psi^{(s)}$ is an additive function (in the sense defined earlier).

Remark: When $\psi^{(s)} = \psi$ (i.e., all the $\psi^{(s)}$ are the same), ψ is symmetric in its arguments, and w is uniform, then U defined above is a U-statistic of degree m with symmetric, mean zero kernel ψ. Lemma III then becomes the well-known bound for

the variance of a U-statistic shown by Hoeffding [12], namely $EU^2 \leq \frac{m}{n} E\psi^2$.

This gives our core inequality (7).

Proposition II: Let $\{X_i\}$ be independent random variables with densities and finite variances. Define

$$T_n = \sum_{i\in[n]} X_i \qquad \text{and} \qquad T_m^{(s)} = T^{(s)} = \sum_{i\in s} X_i, \quad (36)$$

where $s \in \Omega_m = \{s \subset [n] : |s| = m\}$. Let w be any probability distribution on Ω_m. If each $T_m^{(s)}$ has an absolute continuous density, then

$$I(T_n) \leq \binom{n-1}{m-1} \sum_{s\in\Omega_m} w_s^2 I(T_m^{(s)}), \quad (37)$$

where $w_s = w(\{s\})$. Both sides can be finite and equal only if each X_i is normal.

Proof: In the sequel, for convenience, we abuse notation by using ρ to denote several different score functions; $\rho(Y)$ always means $\rho_Y(Y)$. For each j, Lemma I and the fact that $T_m^{(s)}$ has an absolutely continuous density imply

$$\rho(T_n) = E\left[\rho\Big(\sum_{i\in s} X_i\Big)\Big|T_n\right]. \quad (38)$$

Taking a convex combinations of these identities gives, for any $\{w_s\}$ such that $\sum_{s\in\Omega_m} w_s = 1$,

$$\begin{aligned}\rho(T_n) &= \sum_{s\in\Omega_m} w_s E\left[\rho\Big(\sum_{i\in s} X_i\Big)\Big|T_n\right] \\ &= E\Big[\sum_{s\in\Omega_m} w_s \rho(T^{(s)})\Big|T_n\Big].\end{aligned} \quad (39)$$

By applying the Cauchy-Schwartz inequality and Lemma III in succession, we get

$$\begin{aligned}I(T_n) &\leq E\Big[\sum_{s\in\Omega_m} w_s \rho(T^{(s)})\Big]^2 \\ &\leq \binom{n-1}{m-1} \sum_{s\in\Omega_m} E[w_s\rho(T^{(s)})]^2 \\ &= \binom{n-1}{m-1} \sum_{s\in\Omega_m} w_s^2 I(T^{(s)}).\end{aligned} \quad (40)$$

The application of Lemma III can yield equality only if each $\rho(T^{(s)})$ is additive; since the score $\rho(T^{(s)})$ is already a function of the sum $T^{(s)}$, it must in fact be a linear function, so that each X_i must be normal. ∎

APPENDIX

A. Proof of Fact I

$$\begin{aligned}\sum_{s\in\Omega_m} \bar{a}_s^2 &= \sum_{s\in\Omega_m}\sum_{i\in s} a_i^2 = \sum_{i\in[n]}\sum_{S\ni i,|s|=m} a_i^2 \\ &= \sum_{i\in[n]} \binom{n-1}{m-1} a_i^2 = \binom{n-1}{m-1}\sum_{i\in[n]} a_i^2.\end{aligned} \quad (41)$$

B. Proof of Fact II

Let E_s denote the integrating out of the variables in s, so that $E_j = E_{\{j\}}$. Keeping in mind that the order of integrating out independent variables does not matter (i.e., the E_j are commuting projection operators in L^2), we can write

$$\begin{aligned}\phi &= \prod_{j=1}^n [E_j + (I - E_j)]\phi \\ &= \sum_{s\subset[n]} \prod_{j\notin s} E_j \prod_{j\in s} (I - E_j)\phi \\ &= \sum_{s\subset[n]} \phi_s,\end{aligned} \quad (42)$$

where

$$\phi_s = \bar{E}_s\phi \equiv E_{s^c} \prod_{j\notin s} (I - E_j)\phi. \quad (43)$$

In order to show that the subspaces $\mathcal{H}_s$ are orthogonal, observe that for any s_1 and s_2, there is at least one j such that s_1 is contained in the image of E_j and s_2 is contained in the image of $(I - E_j)$; hence every vector in s_1 is orthogonal to every vector in s_2.

REFERENCES

[1] C. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[2] A. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, pp. 101–112, 1959.

[3] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, "Solution of Shannon's problem on the monotonicity of entropy," *J. Amer. Math. Soc.*, vol. 17, no. 4, pp. 975–982 (electronic), 2004.

[4] M. Madiman and A. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *Submitted*, 2006.

[5] R. Shimizu, "On Fisher's amount of information for location family," in *Statistical Distributions in Scientific Work*, G. et al, Ed. Reidel, 1975, vol. 3, pp. 305–312.

[6] A. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, pp. 336–342, 1986.

[7] O. Johnson and A. Barron, "Fisher information inequalities and the central limit theorem," *Probab. Theory Related Fields*, vol. 129, no. 3, pp. 391–409, 2004.

[8] D. Shlyakhtenko, "A free analogue of Shannon's problem on monotonicity of entropy," *Preprint*, 2005. [Online]. Available: arxiv:math.OA/0510103

[9] H. Schultz, "Semicircularity, gaussianity and monotonicity of entropy," *Preprint*, 2005. [Online]. Available: arxiv: math.OA/0512492

[10] A. Dembo, T. Cover, and J. Thomas, "Information-theoretic inequalities," *IEEE Trans. Inform. Theory*, vol. 37, no. 6, pp. 1501–1518, 1991.

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

[12] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *Ann. Math. Stat.*, vol. 19, no. 3, pp. 293–325, 1948.

[13] R. von Mises, "On the asymptotic distribution of differentiable statistical functions," *Ann. Math. Stat.*, vol. 18, no. 3, pp. 309–348, 1947.

[14] B. Kurkjian and M. Zelen, "A calculus for factorial arrangements," *Ann. Math. Statist.*, vol. 33, pp. 600–619, 1962.

[15] R. L. Jacobsen, "Linear algebra and ANOVA," Ph.D. dissertation, Cornell University, 1968.

[16] H. Rubin and R. A. Vitale, "Asymptotic distribution of symmetric statistics," *Ann. Statist.*, vol. 8, no. 1, pp. 165–170, 1980.

[17] B. Efron and C. Stein, "The jackknife estimate of variance," *Ann. Stat.*, vol. 9, no. 3, pp. 586–596, 1981.

[18] A. Takemura, "Tensor analysis of ANOVA decomposition," *J. Amer. Statist. Assoc.*, vol. 78, no. 384, pp. 894–900, 1983.