Generalized Entropy Power Inequalities and Monotonicity Properties of Information

Mokshay Madiman, *Member, IEEE*, and Andrew Barron, *Senior Member, IEEE*

***Abstract*—New families of Fisher information and entropy power inequalities for sums of independent random variables are presented. These inequalities relate the information in the sum of n independent random variables to the information contained in sums over subsets of the random variables, for an arbitrary collection of subsets. As a consequence, a simple proof of the monotonicity of information in central limit theorems is obtained, both in the setting of independent and identically distributed (i.i.d.) summands as well as in the more general setting of independent summands with variance-standardized sums.**

***Index Terms*—Central limit theorem, entropy power, information inequalities.**

I. Introduction

Let $X_1, X_2, \ldots, X_n$ be independent random variables with densities and finite variances. Let H denote the (differential) entropy, i.e., if f is the probability density function of X, then $H(X) = -E[\log f(X)]$. The classical entropy power inequality of Shannon [36] and Stam [39] states

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

In 2004, Artstein, Ball, Barthe, and Naor [1] (hereafter denoted by ABBN [1]) proved a new entropy power inequality

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

where each term involves the entropy of the sum of $n-1$ of the variables excluding the ith, and presented its implications for the monotonicity of entropy in the central limit theorem. It is not hard to see, by repeated application of (2) for a succession of values of n, that (2) in fact implies the inequality (4) and hence (1). We will present below a generalized entropy power inequality for subset sums that subsumes both (2) and (1) and also implies several other interesting inequalities. We provide simple and easily interpretable proofs of all of these inequalities. In particular, this provides a simplified understanding of the monotonicity of entropy in central limit theorems. A similar independent and contemporaneous development of the monotonicity of entropy is given by Tulino and Verdú [42].

Our generalized entropy power inequality for subset sums is as follows: if $\mathcal{C}$ is an arbitrary collection of subsets of $\{1, 2, \ldots, n\}$, then

$$e^{2H(X_1+\cdots+X_n)} \geq \frac{1}{r}\sum_{\boldsymbol{s}\in\mathcal{C}} e^{2H\left(\sum_{j\in\boldsymbol{s}} X_j\right)} \quad (3)$$

where r is the maximum number of sets in $\mathcal{C}$ in which any one index appears. In particular, note the following.

1) Choosing $\mathcal{C}$ to be the class $\mathcal{C}_1$ of all singletons yields $r = 1$ and hence (1).
2) Choosing $\mathcal{C}$ to be the class $\mathcal{C}_{n-1}$ of all sets of $n-1$ elements yields $r = n-1$ and hence (2).
3) Choosing $\mathcal{C}$ to be the class $\mathcal{C}_m$ of all sets of m elements yields $r = \binom{n-1}{m-1}$ and hence the inequality

$$\exp\left\{2H\left(X_1+\cdots+X_n\right)\right\} \geq \frac{1}{\binom{n-1}{m-1}}\sum_{\boldsymbol{s}\in\mathcal{C}_m} \exp\left\{2H\left(\sum_{i\in\boldsymbol{s}} X_i\right)\right\}. \quad (4)$$

4) Choosing $\mathcal{C}$ to be the class of all sets of k consecutive integers yields $r = \min\{k, n+1-k\}$ and hence the inequality

$$\exp\left\{2H\left(X_1+\cdots+X_n\right)\right\} \geq \frac{1}{\min\{k, n+1-k\}}\sum_{\boldsymbol{s}\in\mathcal{C}} \exp\left\{2H\left(\sum_{i\in\boldsymbol{s}} X_i\right)\right\}.$$

In general, the inequality (3) clearly yields a whole family of entropy power inequalities, for arbitrary collections of subsets. Furthermore, equality holds in any of these inequalities if and only if the X_i are normally distributed and the collection $\mathcal{C}$ is "nice" in a sense that will be made precise later.

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

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

by using the scaling $H(aX) = H(X) + \log|a|$. This fact implies that the entropy of the standardized sums $Y_n = \frac{\sum_{i=1}^n X_i}{\sqrt{n}}$ for i.i.d. X_i increases along the powers-of-two subsequence, i.e., $H(Y_{2^k})$ is nondecreasing in k. Characterization of the change in information quantities on doubling of sample size

Manuscript received April 15, 2006; revised March 16, 2007. The material in this paper was presented in part at the Inaugural Workshop, Information Theory and Applications Center, San Diego, CA, February 2006 and at the IEEE International Symposium on Information Theory, Seattle, WA, July 2006.

The authors are with the Department of Statistics, Yale University, New Haven, CT 06511 USA (e-mail: mokshay.madiman@yale.edu; andrew.barron@yale.edu).

Communicated by V. A. Vaishampayan, Associate Editor At Large.
Digital Object Identifier 10.1109/TIT.2007.899484

was used in proofs of central limit theorems by Shimizu [37], Brown [8], Barron [4], Carlen and Soffer [10], Johnson and Barron [21], and ABBN [2]. In particular, Barron [4] showed that the sequence $\{H(Y_n)\}$ 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. ABBN [1] showed that $H(Y_n)$ is in fact a nondecreasing sequence for every n, solving a long-standing conjecture. In fact, (2) is equivalent in the i.i.d. case to the monotonicity property

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

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, the inequality (4) is equivalent to "monotonicity on average" properties for certain standardizations; for instance

$$H\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) \geq \frac{1}{\binom{n}{m}} \sum_{\boldsymbol{s}\in\mathcal{C}_m} H\left(\frac{\sum_{i\in\boldsymbol{s}} X_i}{\sqrt{m}}\right).$$

A similar monotonicity also holds, as we shall show, for arbitrary collections $\mathcal{C}$ and even when the sums are standardized by their variances. Here again the factor r (rather than the cardinality $|\mathcal{C}|$ of the collection) in the denominator of (3) for the unstandardized version is crucial.

Outline of our development.

We find that the main inequality (3) (and hence all of the above inequalities) 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 information inequalities. An additional trick is needed to obtain the denominator r in (3). This is a simple variance drop inequality for statistics expressible via sums of functions of subsets of a collection of variables, particular cases of which are familiar in other statistical contexts (as we shall discuss).

We recall that for a random variable X with differentiable density f, the score function is $\rho(x) = \frac{\partial}{\partial x}\log f(x)$, the score is the random variable $\rho(X)$, and its Fisher information is $I(X) = E[\rho^2(X)]$.

Suppose for the consideration of Fisher information that the independent random variables $X_1,\ldots,X_n$ have absolutely continuous densities. To outline the heart of the matter, the first step boils down to the geometry of projections (conditional expectations). Let ρ_{tot} be the score of the total sum $\sum_{i=1}^n X_i$ and let $\rho_{\boldsymbol{s}}$ be the score of the subset sum $\sum_{i\in\boldsymbol{s}} X_i$. As we recall in Section II, ρ_{tot} is the conditional expectation (or L^2 projection) of each of these subset sum scores given the total sum. Consequently, any convex combination $\sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}}\rho_{\boldsymbol{s}}$ also has projection

$$\rho_{\text{tot}} = E\left[\sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}}\rho_{\boldsymbol{s}} \middle| \sum_{i=1}^n X_i\right]$$

and the Fisher information $I(X_1+\cdots+X_n) = E\left[\rho_{\text{tot}}^2\right]$ has the bound

$$E[\rho_{\text{tot}}^2] \leq E\left[\left(\sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}}\rho_{\boldsymbol{s}}\right)^2\right]. \quad (7)$$

For nonoverlapping subsets, the independence and zero mean properties of the scores provide a direct means to express the right-hand side in terms of the Fisher informations of the subset sums (yielding the traditional Blachman [7] proof of Stam's inequality for Fisher information). In contrast, the case of overlapping subsets requires fresh consideration. Whereas a naive application of Cauchy–Schwarz would yield a loose bound of $|\mathcal{C}|\sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}}^2 E\rho_{\boldsymbol{s}}^2$, instead, a variance drop lemma yields that the right-hand side of (7) is not more than $r\sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}}^2 E\rho_{\boldsymbol{s}}^2$ if each i is in at most r subsets of $\mathcal{C}$. Consequently

$$I(X_1+\cdots+X_n) \leq r\sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}}^2 I\left(\sum_{i\in\boldsymbol{s}} X_i\right) \quad (8)$$

for any weights $w_{\boldsymbol{s}}$ that add to 1 over all subsets $\boldsymbol{s}\subset\{1,\ldots,n\}$ in $\mathcal{C}$. See Sections II and IV for details. Optimizing over w yields an inequality for inverse Fisher information that extends the Fisher information inequalities of Stam and ABBN:

$$\frac{1}{I(X_1+\cdots+X_n)} \geq \frac{1}{r}\sum_{\boldsymbol{s}\in\mathcal{C}} \frac{1}{I\left(\sum_{i\in\boldsymbol{s}} X_i\right)}. \quad (9)$$

Alternatively, using a scaling property of Fisher information to re-express our core inequality (8), we see that the Fisher information of the total sum is bounded by a convex combination of Fisher informations of scaled subset sums:

$$I(X_1+\cdots+X_n) \leq \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} I\left(\frac{\sum_{i\in\boldsymbol{s}} X_i}{\sqrt{w_{\boldsymbol{s}} r}}\right). \quad (10)$$

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.* [15]. Specifically, we obtain

$$H(X_1+\cdots+X_n) \leq \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} H\left(\frac{\sum_{i\in\boldsymbol{s}} X_i}{\sqrt{w_{\boldsymbol{s}} r}}\right). \quad (11)$$

See Section V for details. Likewise, using the scaling property of entropy on (11) and optimizing over w yields our extension of the entropy power inequality

$$\exp\{2H(X_1+\cdots+X_n)\} \geq \frac{1}{r}\sum_{\boldsymbol{s}\in\mathcal{C}} \exp\left\{2H\left(\sum_{j\in\boldsymbol{s}} X_j\right)\right\} \quad (12)$$

described in Section VI.

To relate this chain of ideas to other recent work, we point out that ABBN [1] were the first to show the use of a variance drop lemma for information inequality development (in the leave-one-out case of the collection $\mathcal{C}_{n-1}$). For that case what is new in our presentation is going straight to the projection inequality (7) followed by the variance drop inequality, bypassing any need for the elaborate variational characterization of Fisher information that is an essential ingredient of ABBN [1]. Moreover, whereas ABBN [1] require that the random variables have a C^2 density for monotonicity of Fisher information, absolute continuity of the density (as minimally needed to define the score function) suffices in our approach. Independently and contemporaneously to our work, Tulino and Verdú [42] also found a similar simplified proof of monotonicity properties of entropy and entropy derivative with a clever interpretation (its relationship to our approach is described in Section III after we complete the presentation of our development). Furthermore, a recent preprint of Shlyakhtenko [38] proves an analogue of the information inequalities in the leave-one-out case for noncommutative or "free" probability theory. In that manuscript, he also gives a proof for the classical setting assuming finiteness of all moments, whereas our direct proof requires only finite variance. Our proof also reveals in a simple manner the cases of equality in (6), for which an alternative approach in the free probability setting is in Schultz [35].

While Section III gives a direct proof of the monotonicity of entropy in the central limit theorem for i.i.d. summands, Section VII applies the preceding inequalities to study sums of nonidentically distributed random variables under appropriate scalings. In particular, we show that "entropy is monotone on average" in the setting of variance-standardized sums.

Our subset sum inequalities are tight (with equality in the Gaussian case) for balanced collections of subsets, as will be defined in Section II. In Section VIII, we present refined versions of our inequalities that can even be tight for certain unbalanced collections.

Section IX concludes with some discussion on potential directions of application of our results and methods. In particular, beyond the connection with central limit theorems, we also discuss potential connections of our results with distributed statistical estimation, graph theory and multiuser information theory.

Form of the inequalities.

If $\psi(X)$ represents either the inverse Fisher information or the entropy power of X, then our inequalities above take the form

$$r\psi\left(X_1+\cdots+X_n\right) \geq \sum_{\boldsymbol{s}\in\mathcal{C}} \psi\left(\sum_{i\in\boldsymbol{s}} X_i\right). \tag{13}$$

We motivate the form (13) using the following almost trivial fact.

Fact 1: For arbitrary numbers $\{a_i : i=1,2,\ldots,n\}$

$$\sum_{\boldsymbol{s}\in\mathcal{C}}\sum_{i\in\boldsymbol{s}} a_i = r\sum_{i=1}^{n} a_i \tag{14}$$

if each index i appears in $\mathcal{C}$ the same number of times r.

Indeed

$$\begin{aligned}\sum_{\boldsymbol{s}\in\mathcal{C}}\sum_{i\in\boldsymbol{s}} a_i &= \sum_{i=1}^{n}\sum_{\boldsymbol{s}\ni i,\boldsymbol{s}\in\mathcal{C}} a_i \\ &= \sum_{i=1}^{n} r a_i = r\sum_{i=1}^{n} a_i.\end{aligned}$$

If Fact 1 is thought of as $\mathcal{C}$-additivity of the sum function for real numbers, then (9) and (12) represent the $\mathcal{C}$-*superadditivity* of inverse Fisher information and entropy power functionals, respectively, with respect to convolution of the arguments. In the case of normal random variables, the inverse Fisher information and the entropy power equal the variance. Thus, in that case (9) and (12) become Fact 1 with a_i equal to the variance of X_i.

II. Score Functions and Projections

We use $\rho(x)=\frac{f'(x)}{f(x)}$ to denote the (almost everywhere defined) score function of the random variable X with absolutely continuous probability density function f. The score $\rho(X)$ has zero mean, and its variance is just the Fisher information $I(X)$.

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 (cf., Blachman [7]). For completeness, we give the proof. As shown by Johnson and Barron [21], it is sufficient that the densities are absolutely continuous; see [21, Appendix I] for an explanation of why this is so.

Lemma 1 (Convolution Identity for Scores): If V_1 and V_2 are independent random variables, and V_1 has an absolutely continuous density with score ρ_1, then V_1+V_2 has the score function

$$\rho(v)=E[\rho_1(V_1)|V_1+V_2=v]. \tag{15}$$

Proof: Let f_1 and f 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 [21]

$$\begin{aligned}f'(v) &= \frac{\partial}{\partial v}E[f_1(v-V_2)] \\ &= E\left[f_1'(v-V_2)\right] \\ &= E[f_1(v-V_2)\rho_1(v-V_2)]\end{aligned}$$

so that

$$\begin{aligned}\rho(v) &= \frac{f'(v)}{f(v)} \\ &= E\left[\frac{f_1(v-V_2)}{f(v)}\rho_1(v-V_2)\right] \\ &= E[\rho_1(V_1)|V_1+V_2=v].\end{aligned}$$ □

The second tool we need is a "variance drop lemma," the history of which we discuss in remarks after the proof below. The following conventions are useful.

- $[n]$ is the index set $\{1,2,\ldots,n\}$.

- For any $\boldsymbol{s} \subset [n]$, $X_{\boldsymbol{s}}$ stands for the collection of random variables $(X_i : i \in \boldsymbol{s})$, with the indices taken in their natural (increasing) order.
- For $\psi_{\boldsymbol{s}} : \mathbb{R}^{|\boldsymbol{s}|} \to \mathbb{R}$, we write $\psi_{\boldsymbol{s}}(x_{\boldsymbol{s}})$ for a function of $x_{\boldsymbol{s}}$ for any $\boldsymbol{s} \subset [n]$, so that $\psi_{\boldsymbol{s}}(x_{\boldsymbol{s}}) \equiv \psi_{\boldsymbol{s}}(x_{k_1}, \dots, x_{k_{|\boldsymbol{s}|}})$, where $k_1 < k_2 < \cdots < k_{|\boldsymbol{s}|}$ are the ordered indices in $\boldsymbol{s}$.
- We say that a function $U : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}$*-additive* if it can be expressed in the form $\sum_{\boldsymbol{s}\in\mathcal{C}} \psi_{\boldsymbol{s}}(x_{\boldsymbol{s}})$.

The following notions are not required for the inequalities we present, but help to clarify the cases of equality.

- A collection $\mathcal{C}$ of subsets of $[n]$ is said to be *discriminating* if for any distinct indices i and j in $[n]$, there is a set in $\mathcal{C}$ that contains i but not j. Note that all the collections introduced in Section I were discriminating.
- A collection $\mathcal{C}$ of subsets of $[n]$ is said to be *balanced* if each index i in $[n]$ appears in the same number (namely, r) of sets in $\mathcal{C}$.
- A function $f : \mathbb{R}^d \to \mathbb{R}$ is *additive* if there exist functions $f_i : \mathbb{R} \to \mathbb{R}$ such that $f(x_1, \dots, x_d) = \sum_{i=1}^d f_i(x_i)$, i.e., if it is $\mathcal{C}_1$-additive.

Lemma 2 (Variance Drop): Let

$$U(X_1, \dots, X_n) = \sum_{\boldsymbol{s}\in\mathcal{C}} \psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})$$

be a $\mathcal{C}$-additive function with mean zero components, i.e., $E\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}}) = 0$ for each $\boldsymbol{s} \in \mathcal{C}$. Then

$$EU^2 \le r \sum_{\boldsymbol{s}\in\mathcal{C}} E\left\{\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right\}^2 \tag{16}$$

where r is the maximum number of subsets $\boldsymbol{s} \in \mathcal{C}$ in which any index appears. When $\mathcal{C}$ is a discriminating collection, equality can hold only if each $\psi_{\boldsymbol{s}}$ is an additive function.

Proof: For every subset $\boldsymbol{t}$ of $[n]$, let $\bar{E}_{\boldsymbol{t}}$ be the Analysis of Variance (ANOVA) projection onto the space of functions of $X_{\boldsymbol{t}}$ (see Appendix I for details). By performing the ANOVA decomposition on each $\psi_{\boldsymbol{s}}$, we have

$$\begin{aligned} EU^2 &= E\left(\sum_{\boldsymbol{s}\in\mathcal{C}} \sum_{\boldsymbol{t}\subset\boldsymbol{s}} \bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \\ &= E\left(\sum_{\boldsymbol{t}} \bar{E}_{\boldsymbol{t}} \sum_{\boldsymbol{s}\supset\boldsymbol{t},\boldsymbol{s}\in\mathcal{C}} \psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \\ &= \sum_{\boldsymbol{t}} E\left(\sum_{\boldsymbol{s}\supset\boldsymbol{t},\boldsymbol{s}\in\mathcal{C}} \bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \end{aligned} \tag{17}$$

using the orthogonality of the ANOVA decomposition of U in the last step.

Recall the elementary fact $\left(\sum_{i=1}^m y_i\right)^2 \le m \sum_{i=1}^m y_i^2$, which follows from the Cauchy–Schwarz inequality. In order to apply this observation to the preceding expression, we estimate the number of terms in the inner sum. The outer summation over $\boldsymbol{t}$ can be restricted to nonempty sets $\boldsymbol{t}$, since $\bar{E}_\phi$ has no effect in the summation due to $\psi_{\boldsymbol{s}}$ having zero mean. Thus, any given $\boldsymbol{t}$ in the expression has at least one element, and the sets $\boldsymbol{s} \supset \boldsymbol{t}$ in the collection $\mathcal{C}$ must contain it; so the number of sets $\boldsymbol{s}$ over which the inner sum is taken cannot exceed r. Thus, we have

$$\begin{aligned} EU^2 &\le \sum_{\boldsymbol{t}} r \sum_{\boldsymbol{s}\supset\boldsymbol{t},\boldsymbol{s}\in\mathcal{C}} E\left(\bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \\ &= r \sum_{\boldsymbol{s}\in\mathcal{C}} \sum_{\boldsymbol{t}\subset\boldsymbol{s}} E\left(\bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \\ &= r \sum_{\boldsymbol{s}\in\mathcal{C}} E\left(\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \end{aligned} \tag{18}$$

by rearranging the sums and using the orthogonality of the ANOVA decomposition again. This proves the inequality.

Now suppose $\psi_{\boldsymbol{s}'}$ is not additive. This means that for some set $\boldsymbol{t} \subset \boldsymbol{s}'$ with two elements, $\bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}'}(X_{\boldsymbol{s}'}) \ne 0$. Fix this choice of $\boldsymbol{t}$. Since $\mathcal{C}$ is a discriminating collection, not all of the at most r subsets containing one element of $\boldsymbol{t}$ can contain the other. Consequently, the inner sum in the inequality (18) runs over strictly fewer than r subsets $\boldsymbol{s}$, and the inequality (18) must be strict. Thus, each $\psi_{\boldsymbol{s}}$ must be an additive function if equality holds, i.e., it must be composed only of main effects and no interactions. □

Remark 1: The idea of the variance drop inequality goes back at least to Hoeffding's seminal work [18] on U-statistics. Suppose $\psi : \mathbb{R}^m \to \mathbb{R}$ is symmetric in its arguments, and $E\psi(X_1, \dots, X_m) = 0$. Define

$$U(X_1, \dots, X_n) = \frac{1}{\binom{n}{m}} \sum_{\{\boldsymbol{s}\subset[n]:|\boldsymbol{s}|=m\}} \psi(X_{\boldsymbol{s}}). \tag{19}$$

Then Hoeffding [18] showed

$$EU^2 \le \frac{m}{n} E\psi^2, \tag{20}$$

which is implied by Lemma 2 under the symmetry assumptions. In statistical language, U defined in (19) is a U-statistic of degree m with symmetric, mean zero kernel ψ that is applied to data of sample size n. Thus, (20) quantitatively captures the reduction of variance of a U-statistic when sample size n increases. For $m = 1$, this is the trivial fact that the empirical variance of a function based on i.i.d. samples is the actual variance scaled by n^{-1}. For $m > 1$, the functions $\psi(X_{\boldsymbol{s}})$ are no longer independent, nevertheless, the variance of the U-statistic drops by a factor of $\frac{m}{n}$. Our proof is valid for the more general nonsymmetric case, and also seems to illuminate the underlying statistical idea (the ANOVA decomposition) as well as the underlying geometry (Hilbert space projections) better than Hoeffding's original combinatorial proof. In [16], Efron and Stein assert in their Comment 3 that an ANOVA-like decomposition "yields one-line proofs of Hoeffding's important theorems 5.1 and 5.2"; presumably our proof of Lemma 2 is a generalization of what they had in mind. As mentioned before, the application of such a variance drop lemma to information inequalities was pioneered by ABBN [1]. They proved and used it in the case $\mathcal{C} = \mathcal{C}_{n-1}$ using clear notation that we adapt in developing our generalization above. A further generalization appears when we consider refinements of our main inequalities in Section VIII.

The third key tool in our approach to monotonicity is the well-known link between Fisher information and entropy, whose origin is the de Bruijn identity first described by Stam [39]. This identity, which identifies the Fisher information as the rate of change of the entropy on adding a normal, provides a standard way of obtaining entropy inequalities from Fisher information inequalities. An integral form of the de Bruijn identity was proved by Barron [4]. We express that integral in a form suitable for our purpose (cf., ABBN [1] and Verdú and Guo [43]).

Lemma 3: Let X be a random variable with a density and arbitrary finite variance. Suppose $X_t \overset{(d)}{=} X + \sqrt{t}Z$, where Z is a standard normal independent of X. Then

$$H(X) = \frac{1}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \left[I(X_t) - \frac{1}{1+t}\right]dt. \quad (21)$$

Proof: In the case that the variances of Z and X match, equivalent forms of this identity are given in [3] and [4]. Applying a change of variables using $t = \tau v$ to [3, eq. (2.23)] (which is also equivalent to [4, eq. (2.1)] by another change of variables), one has that

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

if X has variance v and Z has variance 1. This has the advantage of positivity of the integrand but the disadvantage that it seems to depend on v. One can use

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

to re-express it in the form (21), which does not depend on v. □

III. Monotonicity in the i.i.d. Case

For clarity of presentation of ideas, we focus first on the i.i.d. setting. For i.i.d. summands, inequalities (2) and (4) reduce to the monotonicity property $H(Y_n) \geq H(Y_m)$ for $n > m$, where

$$Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i. \quad (22)$$

We exhibit below how our approach provides a simple proof of this monotonicity property, first proved by ABBN [1] using somewhat more elaborate means. We begin by showing the monotonicity of the Fisher information.

Proposition 1 (Monotonicity of Fisher Information): If $\{X_i\}$ are i.i.d. random variables, and Y_{n-1} has an absolutely continuous density, then

$$I(Y_n) \leq I(Y_{n-1}) \quad (23)$$

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

Proof: We use the following notation: The (unnormalized) sum is $S_n = \sum_{i\in[n]} X_i$, and the leave-one-out sum leaving out X_j is $S^{(j)} = \sum_{i\neq j} X_i$. Setting $\rho(S_n)$ to be the score of S_n and ρ_j to be the score of $S^{(j)}$, we have by Lemma 1 that $\rho(S_n) = E[\rho_j|S_n]$ for each j, and hence

$$\rho(S_n) = E\left[\frac{1}{n}\sum_{j\in[n]}\rho_j \middle| S_n\right].$$

Since the norm of the score is not less than that of its projection (i.e., by the Cauchy–Schwarz inequality)

$$I(S_n) = E[\rho^2(S_n)] \leq E\left[\left(\frac{1}{n}\sum_{j\in[n]}\rho_j\right)^2\right].$$

Lemma 2 yields

$$E\left(\frac{1}{n}\sum_{j\in[n]}\rho_j\right)^2 \leq (n-1)\sum_{j\in[n]}\frac{1}{n^2}E\rho_j^2 = \frac{n-1}{n}I(S_{n-1})$$

so that

$$I(S_n) \leq \frac{n-1}{n}I(S_{n-1}).$$

If $X' = aX$, then $\rho_{X'}(X') = \frac{1}{a}\rho_X(X)$ and $a^2 I(X') = I(X)$; hence

$$I(Y_n) = nI(S_n) \leq (n-1)I(S_{n-1}) = I(Y_{n-1}).$$

The inequality implied by Lemma 2 can be tight only if each ρ_j, considered as a function of the random variables $X_i, i \neq j$, is additive. However, we already know that ρ_j is a function of the sum of these random variables. The only functions that are both additive and functions of the sum are linear functions of the sum; hence, the two sides of (23) can be finite and equal only if each of the scores 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. □

The monotonicity result for entropy in the i.i.d. case now follows by combining Proposition 1 and Lemma 3.

Theorem 1 (Monotonicity of Entropy: i.i.d. Case): Suppose $\{X_i\}$ are i.i.d. random variables with densities and finite variance. If the normalized sum Y_n is defined by(22), then

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

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

After the submission of these results to ISIT 2006 [28], we became aware of a contemporaneous and independent development of the simple proof of the monotonicity fact (Theorem 1) by Tulino and Verdú [42]. In their work, they take nice advantage of projection properties through minimum mean-squared error interpretations. It is pertinent to note that the proofs of Theorem 1 (in [42] and in this paper) share essentials, because of the following observations.

Consider estimation of a random variable X from an observation $Y = X + Z$ in which an independent standard normal Z has been added. Then the score function of Y is related to the difference between two predictors of X (maximum likelihood and Bayes), i.e.,

$$-\rho(Y) = Y - E[X|Y], \quad (24)$$

and hence, the Fisher information $I(Y) = E\rho^2(Y)$ is the same as the mean-square difference $E[(Y - E[X|Y])^2]$, or equivalently, by the Pythagorean identity

$$I(Y) = \mathrm{Var}(Z) - E[(X - E[X|Y])^2]. \tag{25}$$

Thus, the Fisher information (entropy derivative) is related to the minimal mean-squared error. These (and more general) identities relating differences between predictors to scores and relating their mean-squared errors to Fisher informations are developed in statistical decision theory in the work of Stein and Brown. These developments are described, for instance, in the point estimation text by Lehmann and Casella [25, Chs. 4.3 and 5.5], in their study of Bayes risk, admissibility and minimaxity of conditional means $E[X|Y]$.

Tulino and Verdú [42] emphasize the minimal mean-squared error property of the entropy derivative and associated projection properies that (along with the variance drop inequality which they note in the leave-one-out case) also give Proposition 1 and Theorem 1. That is a nice idea. Working directly with the minimal mean-squared error as the entropy derivative they bypass the use of Fisher information. In the same manner, Verdú and Guo [43] give an alternative proof of the Shannon–Stam entropy power inequality. If one takes note of the above identities one sees that their proofs and ours are substantially the same, except that the same quantities are given alternative interpretations in the two works, and that we give extensions to arbitrary collections of subsets.

IV. Fisher Information Inequalities

In this section, we demonstrate our core inequality (8).

Proposition 2: Let $\{X_i\}$ be independent random variables with densities and finite variances. Define

$$T_n = \sum_{i \in [n]} X_i \quad \text{and} \quad T^{(\boldsymbol{s})} = \sum_{i \in \boldsymbol{s}} X_i, \tag{26}$$

for each $\boldsymbol{s} \in \mathcal{C}$, where $\mathcal{C}$ is an arbitrary collection of subsets of $[n]$. Let w be any probability distribution on $\mathcal{C}$. If each $T^{(\boldsymbol{s})}$ has an absolute continuous density, then

$$I(T_n) \leq r \sum_{\boldsymbol{s} \in \mathcal{C}} w_{\boldsymbol{s}}^2 I(T^{(\boldsymbol{s})}) \tag{27}$$

where $w_{\boldsymbol{s}} = w(\{\boldsymbol{s}\})$. When $\mathcal{C}$ is discriminating, both sides can be finite and equal only if each X_i is normal.

Proof: Let $\rho_{\boldsymbol{s}}$ be the score of $T^{(\boldsymbol{s})}$. We proceed in accordance with the outline in the Introduction. Indeed, Lemma 1 implies that $\rho(T_n) = E[\rho_{\boldsymbol{s}}|T_n]$ for each $\boldsymbol{s}$. Taking a convex combinations of these identities gives, for any $\{w_{\boldsymbol{s}}\}$ such that $\sum_{\boldsymbol{s} \in \mathcal{C}} w_{\boldsymbol{s}} = 1$

$$\rho(T_n) = \sum_{\boldsymbol{s} \in \mathcal{C}} w_{\boldsymbol{s}} E[\rho_{\boldsymbol{s}}|T_n] = E\left[\sum_{\boldsymbol{s} \in \mathcal{C}} w_{\boldsymbol{s}} \rho_{\boldsymbol{s}} \middle| T_n\right]. \tag{28}$$

Applying the Cauchy–Schwarz inequality

$$\rho^2(T_n) \leq E\left[\left(\sum_{\boldsymbol{s} \in \mathcal{C}} w_{\boldsymbol{s}} \rho_{\boldsymbol{s}}\right)^2 \middle| T_n\right]. \tag{29}$$

Taking the expectation and then applying Lemma 2 in succession, we get

$$E\left[\rho^2(T_n)\right] \leq E\left(\sum\nolimits_{\boldsymbol{s} \in \mathcal{C}} w_{\boldsymbol{s}} \rho_{\boldsymbol{s}}\right)^2 \tag{30}$$

$$\leq r \sum\nolimits_{\boldsymbol{s} \in \mathcal{C}} E(w_{\boldsymbol{s}} \rho_{\boldsymbol{s}})^2 \tag{31}$$

$$= r \sum\nolimits_{\boldsymbol{s} \in \mathcal{C}} w_{\boldsymbol{s}}^2 I(T^{(\boldsymbol{s})}) \tag{32}$$

as desired. The application of Lemma 2 can yield equality only if each $\rho(T^{(\boldsymbol{s})})$ is additive; since the score $\rho(T^{(\boldsymbol{s})})$ is already a function of the sum $T^{(\boldsymbol{s})}$, it must in fact be a linear function, so that each X_i must be normal. □

Naturally, it is of interest to minimize the upper bound of Proposition 2 over the weighting distribution w, which is easily done either by an application of Jensen's inequality for the reciprocal function, or by the method of Lagrange multipliers. Optimization of the bound implies that Proposition 2 is equivalent to the following Fisher information inequalities.

Theorem 2: Let $\{X_i\}$ be independent random variables such that each $T^{(\boldsymbol{s})}$ has an absolutely continuous density. Then

$$\frac{1}{I(T_n)} \geq \frac{1}{r} \sum_{\boldsymbol{s} \in \mathcal{C}} \frac{1}{I(T^{(\boldsymbol{s})})}. \tag{33}$$

When $\mathcal{C}$ is discriminating, the two sides are positive and equal iff each X_i is normal and $\mathcal{C}$ is also balanced.

Remark 2: Theorem 2 for the special case $\mathcal{C} = \mathcal{C}_1$ of singleton sets is sometimes known as the "Stam inequality" and has a long history. Stam [39] was the first to prove Proposition 2 for $\mathcal{C}_1$, and he credited his doctoral advisor de Bruijn with noticing the equivalence to Theorem 2 for $\mathcal{C}_1$. Subsequently, several different proofs have appeared: in Blachman [7] using Lemma 1, in Carlen [9] using another superadditivity property of the Fisher information, and in Kagan [22] as a consequence of an inequality for Pitman estimators. On the other hand, the special case of the leave-one-out sets $\mathcal{C} = \mathcal{C}_{n-1}$ in Theorem 2 was first proved in ABBN [1]. Zamir [46] used data processing properties of the Fisher information to prove some different extensions of the $\mathcal{C}_1$ case, including a multivariate version; see also Liu and Viswanath [27] for some related interpretations. Our result for arbitrary collections of subsets is new; yet our proof of this general result is essentially no harder than the elementary proofs of the original inequality by Stam and Blachman.

Remark 3: While inequality (30) in the proof above uses a Pythagorean inequality, one may use the associated Pythagorean identity to characterize the difference as the mean square of $\sum_{\boldsymbol{s}} w_{\boldsymbol{s}} \rho_{\boldsymbol{s}} - \rho(T_n)$. In the i.i.d. case with $n = 2m$ and $\mathcal{C}$, a disjoint pair of subsets of size m, this drop in Fisher distance from the normal played an essential role in the previously mentioned central limit theorem analyses of [37], [8], [4], [21]. Furthermore, for general $\mathcal{C}$, we have from the variance drop analysis that the gap in inequality (31) is characterized by the nonadditive ANOVA components of the score functions. We point out these observations as an encouragement to examination of the information drop for collections such as $\mathcal{C}_{n-1}$ and $\mathcal{C}_{n/2}$ in refined analysis of central limit theorem rates under more general conditions.

V. Entropy Inequalities

The Fisher information inequality of the previous section yields a corresponding entropy inequality.

Proposition 3 (Entropy of Sums): Let $\{X_i\}$ be independent random variables with densities. Then, for any probability distribution w on $\mathcal{C}$ such that $w_{\boldsymbol{s}} \leq \frac{1}{r}$ for each $\boldsymbol{s}$,

$$H\biggl(\sum_{i\in[n]} X_i\biggr) \geq \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} H\biggl(\sum_{i\in\boldsymbol{s}} X_i\biggr) + \frac{1}{2}H(w) - \frac{1}{2}\log r \quad (34)$$

where $H(w) = \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} \log \frac{1}{w_{\boldsymbol{s}}}$ is the discrete entropy of w. When $\mathcal{C}$ is discriminating, equality can hold only if each X_i is normal.

Proof: As pointed out in the Introduction, Proposition 2 is equivalent to

$$I\biggl(\sum_{i\in[n]} Y_i\biggr) \leq \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} I\biggl(\frac{\sum_{i\in\boldsymbol{s}} Y_i}{\sqrt{w_{\boldsymbol{s}} r}}\biggr)$$

for independent random variables Y_i. For application of the Fisher information inequality to entropy inequalities we also need for an independent standard normal Z that

$$I(T_n + \sqrt{\tau}Z) \leq \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} I\left(\frac{T^{\boldsymbol{s}}}{\sqrt{r w_{\boldsymbol{s}}}} + \sqrt{\tau}Z\right) \quad (35)$$

at least for suitable values of $w_{\boldsymbol{s}}$. We will show below that this holds when $r w_{\boldsymbol{s}} \leq 1$ for each $\boldsymbol{s}$, and thus

$$\begin{aligned} H(T_n) &= \frac{1}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \left(I(T_{[n]} + \sqrt{\tau}Z) - \frac{1}{1+\tau}\right)dt \\ &\geq \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} \biggl[\frac{1}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \biggl\{I\biggl(\frac{T^{\boldsymbol{s}}}{\sqrt{r w_{\boldsymbol{s}}}} + \sqrt{\tau}Z\biggr) \\ &\qquad - \frac{1}{1+\tau}\biggr\}dt\biggr] \\ &= \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} H\biggl(\frac{T^{\boldsymbol{s}}}{\sqrt{r w_{\boldsymbol{s}}}}\biggr) \end{aligned}$$

using (35) for the inequality and Lemma 3 for the equalities. By the scaling property of entropy, this implies

$$H(T_n) \geq \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} H(T^{\boldsymbol{s}}) - \frac{1}{2}\sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} \log w_{\boldsymbol{s}} - \frac{1}{2}\log r,$$

proving the desired result.

The inequality (35) is true though not immediately so (the naive approach of adding an independent normal to each X_i does not work to get our desired inequalities when the subsets have more than one element). What we need is to provide a collection of independent normal random variables Z_j for some set of indices j (possibly many more than n of them). For each $\boldsymbol{s}$ in $\mathcal{C}$ we need an assignment of subset sums of Z_j (called say $Z^{\boldsymbol{s}'}$) which has variance $r w_{\boldsymbol{s}}$, such that no j is in more than r of the subsets $\boldsymbol{s}'$. Then by Proposition 2 (applied to the collection $\mathcal{C}'$ of augmented sets $\boldsymbol{s} \cup \boldsymbol{s}'$ for each $\boldsymbol{s}$ in $\mathcal{C}$) we have

$$I(T_n + \sqrt{t}Z) \leq \sum_{\boldsymbol{s}\in\mathcal{C}} r w_{\boldsymbol{s}}^2 I(T^{\boldsymbol{s}} + \sqrt{t}Z^{\boldsymbol{s}'})$$

from which the desired inequality follows using the fact that $a^2 I(X) = I(X/a)$. Assuming that $r w_{\boldsymbol{s}} \leq 1$ (which will be sufficient for our needs), we provide such a construction of Z_j and their subset sums in the case of rational weights, say $w_{\boldsymbol{s}} = W(\boldsymbol{s})/M$, where the denominator M may be large. Indeed, set $Z_1, \ldots, Z_M$ independent mean-zero normals each of variance $1/M$. For each $\boldsymbol{s}$, we construct a set $\boldsymbol{s}'$ that has precisely $rW(\boldsymbol{s})$ normals and each normal is assigned to precisely r of these sets. This may be done systematically by considering the sets $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots$ in $\mathcal{C}$ in some order. We let $\boldsymbol{s}_1'$ be the first $rW(\boldsymbol{s}_1)$ indices j (not more than M by assumption), we let $\boldsymbol{s}_2'$ be the next $rW(\boldsymbol{s}_2)$ indices (looping back to the first index once we pass M) and so on. This proves the validity of (35) for rational weights; its validity for general weights follows by continuity. ☐

Remark 4: One may re-express the inequality of Proposition 3 as a statement for the relative entropies with respect to normals of the same variance. If X has density f, we write

$$D(X) = E\left[\log\frac{f(X)}{g(X)}\right] = H(Z) - H(X)$$

where Z has the Gaussian density g with the same variance as X. Then, for any probability distribution w on a balanced collection $\mathcal{C}$

$$D\biggl(\sum_{i\in[n]} X_i\biggr) \leq \sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} D\biggl(\sum_{i\in\boldsymbol{s}} X_i\biggr) + \frac{1}{2}D(w\|\eta) \quad (36)$$

where η is the probability distribution on $\mathcal{C}$ given by $\eta_{\boldsymbol{s}} = \frac{\mathrm{Var}(T^{(\boldsymbol{s})})}{r\mathrm{Var}(T_n)}$, and $D(w\|\eta)$ is the (discrete) relative entropy of w with respect to η. When $\mathcal{C}$ is also discriminating, equality holds iff each X_i is normal. Theorem 1 of Tulino and Verdú [42] is the special case of inequality (36) for the collection of leave-one-out sets. Inequality (36) can be further extended to the case where $\mathcal{C}$ is not balanced, but in that case η is a subprobability distribution. The conclusions (34) and (36) are equivalent, and as seen in the next section, are equivalent to our subset sum entropy power inequality.

Remark 5: It will become evident in the next section that the condition $r w_{\boldsymbol{s}} \leq 1$ in Proposition 3 is not needed for the validity of the conclusions (34) and (36) (see Remark 8).

VI. Entropy Power Inequalities

Proposition 3 is equivalent to a subset sum entropy power inequality. Recall that the entropy power $\frac{e^{2H(X)}}{2\pi e}$ is the variance of the normal with the same entropy as X. The term entropy power is also used for the quantity

$$\mathcal{N}(X) = e^{2H(X)} \quad (37)$$

even when the constant factor of $2\pi e$ is excluded.

Theorem 3: For independent random variables with finite variances

$$\mathcal{N}\Big(\sum_{i\in[n]} X_i\Big) \geq \frac{1}{r}\sum_{\boldsymbol{s}\in\mathcal{C}} \mathcal{N}\Big(\sum_{i\in\boldsymbol{s}} X_i\Big). \tag{38}$$

When $\mathcal{C}$ is discriminating, the two sides are equal iff each X_i is normal and $\mathcal{C}$ is also balanced.

Proof: Let $T^{\boldsymbol{s}}$ be the subset sums as defined in (26). Define $\boldsymbol{Z} = \sum_{\boldsymbol{s}\in\mathcal{C}} \mathcal{N}(T^{\boldsymbol{s}})$ as the normalizing constant producing the weights

$$\mu_{\boldsymbol{s}} = \frac{\mathcal{N}(T^{\boldsymbol{s}})}{\boldsymbol{Z}}. \tag{39}$$

If $N(T^{\boldsymbol{s}}) > \boldsymbol{Z}/r$ for some $\boldsymbol{s} \in \mathcal{C}$ then we trivially have $N(T_n) \geq N(T^{\boldsymbol{s}}) > \boldsymbol{Z}/r$ which is the desired result, so now assume $N(T^{\boldsymbol{s}}) \leq \boldsymbol{Z}/r$ for all $\boldsymbol{s}\in\mathcal{C}$, that is, $r\mu_{\boldsymbol{s}} \leq 1$ for each $\boldsymbol{s}$.

Since

$$H(T^{(\boldsymbol{s})}) = \frac{1}{2}\log \mathcal{N}(T^{\boldsymbol{s}}) = \frac{1}{2}\log[\mu_{\boldsymbol{s}}\boldsymbol{Z}]$$

Proposition 3 implies for any weighting distribution w with $rw_{\boldsymbol{s}} \leq 1$ that

$$\begin{aligned} H(T_n) &\geq \frac{1}{2}\sum_{\boldsymbol{s}\in\mathcal{C}} w_{\boldsymbol{s}} \log[\mu_{\boldsymbol{s}}\boldsymbol{Z}] + \frac{1}{2}\Big[H(w) - \log r\Big] \\ &= \frac{1}{2}\Big[\log\frac{\boldsymbol{Z}}{r} - D(w\|\mu)\Big] \end{aligned} \tag{40}$$

where $D(w\|\mu)$ is the (discrete) relative entropy. Exponentiating gives

$$\mathcal{N}(T_n) \geq r^{-1}e^{-D(w\|\mu)}\boldsymbol{Z}. \tag{41}$$

It remains to optimize the right-hand side over w, or equivalently, to minimize $D(w\|\mu)$ over feasible w. Since $r\mu_{\boldsymbol{s}} \leq 1$ by assumption, $w = \mu$ is a feasible choice, yielding the desired inequality.

The necessary conditions for equality follow from that for Proposition 3, and it is easily checked using Fact 1 that this is also sufficient. The proof is complete. □

Remark 6: To understand this result fully, it is useful to note that if a discriminating collection $\mathcal{C}$ is not balanced, it can always be augmented to a new collection $\mathcal{C}'$ that is balanced in such a way that the inequality (38) for $\mathcal{C}'$ becomes strictly better than that for $\mathcal{C}$. Indeed, if index i appears in $r(i)$ sets of $\mathcal{C}$, one can always find $r - r(i)$ sets of $\mathcal{C}$ not containing i (since $r \leq |\mathcal{C}|$), and add i to each of these sets. The inequality (38) for $\mathcal{C}'$ is strictly better since this collection has the same r and the subset sum entropy powers on the right-hand side are higher due to the addition of independent random variables. While equality in (38) is impossible for the unbalanced collection $\mathcal{C}$, it holds for normals for the augmented, balanced collection $\mathcal{C}'$. This illuminates the conditions for equality in Theorem 3.

Remark 7: The traditional Shannon inequality involving the entropy powers of the summands [36] as well as the inequality of ABBN [1] involving the entropy powers of the "leave-one-out" normalized sums are two special cases of Theorem 3, corresponding to $\mathcal{C} = \mathcal{C}_1$ and $\mathcal{C} = \mathcal{C}_{n-1}$. Proofs of the former subsequent to Shannon's include those of Stam [39], Blachman [7], Lieb [26] (using Young's inequality for convolutions with the sharp constant), Dembo, Cover, and Thomas [15] (building on Costa and Cover [13]), and Verdú and Guo [43]. Note that unlike the previous proofs of these special cases, our proof of the equivalence between the linear form of Proposition 3 and Theorem 3 reduces to the nonnegativity of the relative entropy.

Remark 8: To see that (34) (and hence (36)) holds without any assumption on w, simply note that when the assumption is not satisfied, the entropy power inequality of Theorem 3 implies trivially that

$$\mathcal{N}(T_n) \geq r^{-1}\boldsymbol{Z} \geq r^{-1}e^{-D(w\|\mu)}\boldsymbol{Z}$$

for μ defined by (39), and inverting the steps of (40) yields (34).

VII. Entropy Is Monotone on Average

In this section, we consider the behavior of the entropy of sums of independent but not necessarily identically distributed (i.n.i.d.) random variables under various scalings.

First we look at sums scaled according to the number of summands. Fix the collection $\mathcal{C}_m = \{\boldsymbol{s} \subset [n] : |\boldsymbol{s}| = m\}$. For i.n.i.d. random variables X_i, let

$$Y_n = \frac{\sum_{i\in[n]} X_i}{\sqrt{n}} \quad \text{and} \quad Y_m^{(\boldsymbol{s})} = \frac{\sum_{i\in\boldsymbol{s}} X_i}{\sqrt{m}} \tag{42}$$

for $\boldsymbol{s} \in \mathcal{C}_m$ be the scaled sums. Then Proposition 3 applied to $\mathcal{C}_m$ implies

$$H(Y_n) \geq \sum_{\boldsymbol{s}\in\mathcal{C}_m} w_{\boldsymbol{s}} H\left(Y_m^{(\boldsymbol{s})}\right) - \frac{1}{2}\Big[\log\binom{n}{m} - H(w)\Big].$$

The term on the right-hand indicates that we pay a price for deviations of the weighting distribution w from the uniform. In particular, choosing w to be uniform implies that entropy is "monotone on average" with uniform weights for scaled sums of i.n.i.d. random variables. Applying Theorem 3 to $\mathcal{C}_m$ yields a similar conclusion for entropy power. These observations, which can also be deduced from the results of ABBN [1], are collected in Corollary 1.

Corollary 1: Suppose X_i are independent random variables with densities, and the scaled sums are defined by (42). Then

$$\begin{aligned} H(Y_n) &\geq \frac{1}{\binom{n}{m}} \sum_{\boldsymbol{s}\in\mathcal{C}_m} H\left(Y_m^{(\boldsymbol{s})}\right) \\ \text{and } \mathcal{N}(Y_n) &\geq \frac{1}{\binom{n}{m}} \sum_{\boldsymbol{s}\in\mathcal{C}_m} \mathcal{N}\left(Y_m^{(\boldsymbol{s})}\right). \end{aligned} \tag{43}$$

Remark 9: It is interesting to contrast Corollary 1 with the following results of Han [17] and Dembo, Cover, and Thomas [15]. With no assumptions on $(X_1, \ldots, X_n)$ except that they have a joint density, the above authors show that

$$\frac{H(X_{[n]})}{n} \leq \frac{1}{\binom{n}{m}} \sum_{\boldsymbol{s}\in\mathcal{C}_m} \frac{H(X_{\boldsymbol{s}})}{m} \tag{44}$$

and

$$\left[\mathcal{N}(X_{[n]})\right]^{\frac{1}{n}} \leq \frac{1}{\binom{n}{m}} \sum_{\boldsymbol{s}\in\mathcal{C}_m} \left[\mathcal{N}(X_{\boldsymbol{s}})\right]^{\frac{1}{m}}. \tag{45}$$

where $H(X_{\boldsymbol{s}})$ and $\mathcal{N}(X_{\boldsymbol{s}})$ denote the joint entropy and joint entropy power of $X_{\boldsymbol{s}}$, respectively. These bounds have a form very similar to that of Corollary 1. In fact, such an analogy between inequalities for entropy of sums and joint entropies goes much deeper (so that all of the entropy power inequalities we present here have analogues for joint entropy). More details can be found in Madiman and Tetali [30].

Next, we consider sums of independent random variables standardized by their variances. This is motivated by the following consideration. Consider a sequence of i.n.i.d. random variables $\{X_i : i \in \mathbb{N}\}$ with zero mean and finite variances, $\sigma_i^2 = \mathrm{Var}(X_i)$. The variance of the sum of n variables is denoted $v_n = \sum_{i\in[n]} \sigma_i^2$, and the standardized sum is

$$V_n = \frac{\sum_{i\in[n]} X_i}{\sqrt{v_n}}.$$

The Lindeberg–Feller central limit theorem gives conditions under which $V_n \Rightarrow N(0,1)$. Johnson [20] has proved an entropic version of this classical theorem, showing (under appropriate conditions) that $H(V_n) \to \frac{1}{2}\log(2\pi e)$ and hence the relative entropy from the unit normal tends to 0. Is there an analogue of the monotonicity of information in this setting?

We address this question in the following theorem, and give two proofs. The first proof is based on considering appropriately standardized linear combinations of independent random variables, and generalizes Theorem 2 of ABBN [1]. The second proof is outlined in Remark 11.

Theorem 4 (Monotonicity on Average): Suppose $\{X_i : i \in [n]\}$ are independent random variables with densities, and X_i has finite variance σ_i^2. Set $v_n = \sum_{i\in[n]} \sigma_i^2$ and $v_{\boldsymbol{s}} = \sum_{i\in\boldsymbol{s}} \sigma_i^2$ for sets $\boldsymbol{s}$ in the balanced collection $\mathcal{C}$. Define the standardized sums

$$V_n = \frac{\sum_{i\in[n]} X_i}{\sqrt{v_n}} \tag{46}$$

and

$$V^{(\boldsymbol{s})} = \frac{\sum_{i\in\boldsymbol{s}} X_i}{\sqrt{v_{\boldsymbol{s}}}}. \tag{47}$$

Then

$$H(V_n) \geq \sum_{\boldsymbol{s}\in\mathcal{C}} \eta_{\boldsymbol{s}} H(V^{(\boldsymbol{s})}), \tag{48}$$

where $\eta_{\boldsymbol{s}} = \frac{1}{r}\frac{v_{\boldsymbol{s}}}{v_n}$. Furthermore, if $\mathcal{C}$ is also discriminating, then the inequality is strict unless each X_i is normal.

Proof: Let $a_i, i \in [n]$ be a collection of nonnegative real numbers such that $\sum_{i=1}^n a_i^2 = 1$. Define $\bar{a}_{\boldsymbol{s}} = \left[\sum_{i\in\boldsymbol{s}} a_i^2\right]^{\frac{1}{2}}$ and the weights $\lambda_{\boldsymbol{s}} = \frac{\bar{a}_{\boldsymbol{s}}^2}{r}$ for $\boldsymbol{s} \in \mathcal{C}$. Applying the inequalities of Theorem 2, Proposition 3, and Theorem 3 to independent random variables $a_i X_i'$, and utilizing the scaling properties of the relevant information quantities, one finds that

$$\psi\left(\sum_{i=1}^n a_i X_i'\right) \geq \sum_{\boldsymbol{s}\in\mathcal{C}} \lambda_{\boldsymbol{s}} \psi\left(\frac{1}{\bar{a}_{\boldsymbol{s}}}\sum_{i\in\boldsymbol{s}} a_i X_i'\right) \tag{49}$$

where ψ represents either the inverse Fisher information I^{-1} or the entropy H or the entropy power $\mathcal{N}$.

The conclusion of Theorem 4 is a particular instance of (49). Indeed, we can express the random variables of interest as $X_i = \sigma_i X_i'$, so that each X_i' has variance 1. Choose $a_i = \frac{\sigma_i}{\sqrt{v_n}}$, which is valid since $\sum_{i\in[n]} a_i^2 = 1$. Then $\bar{a}_{\boldsymbol{s}}^2 = \sum_{i\in\boldsymbol{s}} a_i^2 = \frac{v_{\boldsymbol{s}}}{v_n}$ and $\lambda_{\boldsymbol{s}} = \eta_{\boldsymbol{s}}$. Thus

$$V_n = \sum_{i\in[n]} \frac{X_i}{\sqrt{v_n}} = \sum_{i\in[n]} a_i X_i',$$

and

$$V^{(\boldsymbol{s})} = \sum_{i\in\boldsymbol{s}} \frac{X_i}{\sqrt{v_{\boldsymbol{s}}}} = \frac{1}{\bar{a}_{\boldsymbol{s}}}\sum_{i\in\boldsymbol{s}} a_i X_i'.$$

Now an application of (49) gives the desired result, not just for H but also for $\mathcal{N}$ and I^{-1}. □

Remark 10: Since the collection $\mathcal{C}$ is balanced, it follows from Fact 1 that $\eta_{\boldsymbol{s}}$ defines a probability distribution on $\mathcal{C}$. This justifies the interpretation of Theorem 4 as displaying "monotonicity on average." The averaging distribution η is tuned to the random variables of interest, through their variances.

Remark 11: Theorem 4 also follows directly from (36) upon setting $w_{\boldsymbol{s}} = \eta_{\boldsymbol{s}}$ and noting that the definition of $D(X)$ is scale invariant (i.e., $D(aX) = D(X)$ for any real number a).

Let us briefly comment on the interpretation of this result. As discussed before, when the summands are i.i.d., entropic convergence of V_n to the normal was shown in [4], and ABBN [1] showed that this sequence of entropies is monotonically increasing. This completes a long-conjectured intuitive picture of the central limit theorem: forming normalized sums that keep the variance constant yields random variables with increasing entropy, and this sequence of entropies converges to the maximum entropy possible, which is the entropy of the normal with that variance. In this sense, the central limit theorem is a formulation of the "second law of thermodynamics" in physics. Theorem 4 above shows that even in the setting of variance-standardized sums of i.n.i.d. random variables, a general monotonicity on average property holds with respect to an arbitrary collection of normalized subset sums. This strengthens the "second law" interpretation of central limit theorems.

A similar monotonicity on average property also holds for appropriate notions of Fisher information in convergence of sums of discrete random variables to the Poisson and compound Poisson distributions; details may be found in [29].

VIII. A Refined Inequality

Various extensions of the basic inequalities presented above are possible; we present one here. To state it, we find it convenient to recall the notion of a fractional packing from discrete mathematics (see, e.g., Chung, Füredi, Garey, and Graham [11]).

Definition 1: Let $\mathcal{C}$ be a collection of subsets of $[n]$. A collection $\{\beta_{\boldsymbol{s}} : \boldsymbol{s} \in \mathcal{C}\}$ of nonnegative real numbers is called a fractional packing for $\mathcal{C}$ if

$$\sum_{\boldsymbol{s}\ni i, \boldsymbol{s}\in\mathcal{C}} \beta_{\boldsymbol{s}} \leq 1 \tag{50}$$

for each i in $[n]$.

Note that if the numbers $\beta_{\boldsymbol{s}}$ are constrained to only take the values 0 and 1, then the condition above entails that not more than one set in $\mathcal{C}$ can contain i, i.e., that the sets $\boldsymbol{s} \in \mathcal{C}$ are pairwise disjoint, and provide a packing of the set $[n]$. We may interpret a fractional packing as a "packing" of $[n]$ using sets in $\mathcal{C}$, each of which contains only a fractional piece (namely, $\beta_{\boldsymbol{s}}$) of the elements in that set.

We now present a refined version of Lemma 2.

Lemma 4 (Variance Drop: General Version): Suppose $U(X_1, \dots, X_n)$ is a $\mathcal{C}$-additive function with mean zero components, as in Lemma 2. Then

$$EU^2 \le \sum_{\boldsymbol{s}\in\mathcal{C}} \frac{1}{\beta_{\boldsymbol{s}}} E\left(\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \tag{51}$$

for any fractional packing $\{\beta_{\boldsymbol{s}} : \boldsymbol{s} \in \mathcal{C}\}$.

Proof: As in the proof of Lemma 2, we have

$$EU^2 = \sum_{\boldsymbol{t}} E\bigg[\sum_{\boldsymbol{s}\supset\boldsymbol{t}, \boldsymbol{s}\in\mathcal{C}} \bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\bigg]^2.$$

We now proceed to perform a different estimation of this expression, recalling as in the proof of Lemma 2, that the outer summation over $\boldsymbol{t}$ can be restricted to nonempty sets $\boldsymbol{t}$. By the Cauchy–Schwarz inequality

$$\bigg[\sum_{\boldsymbol{s}\supset\boldsymbol{t},\boldsymbol{s}\in\mathcal{C}} \bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\bigg]^2 \le \bigg[\sum_{\boldsymbol{s}\supset\boldsymbol{t},\boldsymbol{s}\in\mathcal{C}} \left(\sqrt{\beta_{\boldsymbol{s}}}\right)^2\bigg]\bigg[\sum_{\boldsymbol{s}\supset\boldsymbol{t},\boldsymbol{s}\in\mathcal{C}} \left(\frac{1}{\sqrt{\beta_{\boldsymbol{s}}}}\bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2\bigg].$$

Since any $\boldsymbol{t}$ of interest has at least one element, the definition of a fractional packing implies that

$$\sum_{\boldsymbol{s}\supset\boldsymbol{t},\boldsymbol{s}\in\mathcal{C}} \beta_{\boldsymbol{s}} \le 1.$$

Thus

$$\begin{aligned} EU^2 &\le \sum_{\boldsymbol{t}} \sum_{\boldsymbol{s}\supset\boldsymbol{t},\boldsymbol{s}\in\mathcal{C}} \frac{1}{\beta_{\boldsymbol{s}}} E\left(\bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \\ &= \sum_{\boldsymbol{s}\in\mathcal{C}} \frac{1}{\beta_{\boldsymbol{s}}} \sum_{\boldsymbol{t}\subset\boldsymbol{s}} E\left(\bar{E}_{\boldsymbol{t}}\psi_{\boldsymbol{s}}(X_{\boldsymbol{s}})\right)^2 \\ &= \sum_{\boldsymbol{s}\in\mathcal{C}} \frac{1}{\beta_{\boldsymbol{s}}} E\left(\psi_{\boldsymbol{s}}\left(X_{\boldsymbol{s}}\right)\right)^2. \end{aligned} \tag{52}$$

□

Exactly as before, one can obtain inequalities for Fisher information, entropy, and entropy power based on this form of the variance drop lemma. The idea of looking for such refinements with coefficients depending on $\boldsymbol{s}$ arose in conversations with Tom Cover and Prasad Tetali at ISIT 2006 in Seattle, after our basic results described in the previous sections were presented. In particular, Prasad's joint work with one of us [30] influenced the development of Theorem 5.

Theorem 5: Let $\{\beta_{\boldsymbol{s}} : \boldsymbol{s} \in \mathcal{C}\}$ be any fractional packing for $\mathcal{C}$. Then

$$I^{-1}(X_1 + \cdots + X_n) \ge \sum_{\boldsymbol{s}\in\mathcal{C}} \beta_{\boldsymbol{s}} I^{-1}\left(\sum_{j\in\boldsymbol{s}} X_j\right). \tag{53}$$

For given subset sum informations, the best such lower bound on the information of the total sum would involve maximizing the right-hand side of (53) subject to the linear constraints (50). This linear programming problem, a version of which is the problem of optimal fractional packing well studied in combinatorics (see, e.g., [11]), does not have an explicit solution in general.

A natural choice of a fractional packing in Theorem 5 leads to the following corollary.

Corollary 2: For any collection $\mathcal{C}$ of subsets of $[n]$, let $r(i)$ denote the number of sets in $\mathcal{C}$ that contain i. In the same setting as Theorem 2, we have

$$\frac{1}{I(X_1 + \cdots + X_n)} \ge \sum_{\boldsymbol{s}\in\mathcal{C}} \frac{1}{r(\boldsymbol{s}) I\left(\sum_{j\in\boldsymbol{s}} X_j\right)} \tag{54}$$

where $r(\boldsymbol{s})$ is the maximum value of $r(i)$ over the indices i in $\boldsymbol{s}$.

We say that $\mathcal{C}$ is quasi-balanced if $r(i) = r(\boldsymbol{s})$ for each $i \in \boldsymbol{s}$ and each $\boldsymbol{s}$. If $\mathcal{C}$ is discriminating, equality holds in (54) if the X_i are normal and $\mathcal{C}$ is quasi-balanced.

Remark 12: For any collection $\mathcal{C}$ and any $\boldsymbol{s} \in \mathcal{C}$, $r(\boldsymbol{s}) \le r$ by definition. Thus, Theorem 5 and Corollary 2 generalize Theorem 2. Furthermore, from the equality conditions in Corollary 2, we see that equality can hold in these more general inequalities even for collections $\mathcal{C}$ that are not balanced, which was not possible with the original formulation in Theorem 2.

Remark 13: One can also give an alternate proof of Theorem 5 using Corollary 2 (which could be proved directly), so that the two results are mathematically equivalent. The key to doing this is the observation that nowhere in our proofs do we actually require that the sets $\boldsymbol{s}$ in $\mathcal{C}$ be distinct. In other words, given a collection $\mathcal{C}$, one may look at an augmented collection that has $k_{\boldsymbol{s}}$ copies of each set $\boldsymbol{s}$ in $\mathcal{C}$. Then the inequality (54) holds for the augmented collection with the counts $r(i)$ and $r(\boldsymbol{s})$ appropriately modified. By considering arbitrary augmentations, one can obtain Theorem 5 for fractional packings with rational coefficients. An approximation argument yields the full version. This method of proof, although picturesque, is somewhat less transparent in the details.

Remark 14: It is straightforward to extend Theorem 5 and Corollary 2 to the multivariate case, where X_i are independent $\mathbb{R}^d$-valued random vectors, and $I(X)$ represents the trace of the Fisher information matrix of X. Similarly, extending Theorem 3, one obtains for independent $\mathbb{R}^d$-valued random vectors $X_1, \dots, X_n$ with densities and finite covariance matrices that

$$e^{\frac{2H(X_1+\cdots+X_n)}{d}} \ge \frac{1}{r}\sum_{\boldsymbol{s}\in\mathcal{C}} e^{\frac{2H\left(\sum_{j\in\boldsymbol{s}} X_j\right)}{d}}$$

which implies the monotonicity of entropy for standardized sums of d-dimensional random vectors. We leave the details to the reader.

Remark 15: It is natural to speculate whether an analogous subset sum entropy power inequality holds with $1/r(\boldsymbol{s})$ inside the sum. For each r between the minimum and the maximum of the $r(\boldsymbol{s})$, we can split $\mathcal{C}$ into the sets $\mathcal{C}_r = \{\boldsymbol{s} \in \mathcal{C} : r(\boldsymbol{s}) = r\}$. Under an assumption that no one set $\boldsymbol{s}$ in $\mathcal{C}_r$ dominates, that is, that there is no $\boldsymbol{s}^* \in \mathcal{C}_r$ with $N(T^{\boldsymbol{s}^*}) > \sum_{\boldsymbol{s}\in\mathcal{C}_r} N(T^{\boldsymbol{s}})/r$, we are able to create suitable normals for perturbation of the Fisher information inequality and integrate (in the same manner as in the proof of Proposition 3) to obtain

$$N(T_n) \geq \sum_{\boldsymbol{s}\in\mathcal{C}} \frac{N(T^{\boldsymbol{s}})}{r(\boldsymbol{s})}. \tag{55}$$

The quasi-balanced case (in which $r(i)$ is the same for each i in $\boldsymbol{s}$) is an interesting special case. Then the unions of sets in $\mathcal{C}_r$ are disjoint for distinct values of r. So for quasi-balanced collections the refined subset sum entropy power inequality (55) always holds by combining our observation above with the Shannon–Stam entropy power inequality.

IX. Concluding Remarks

Our main contributions in this paper are rather general $\mathcal{C}$-super-additivity inequalities for Fisher information and entropy power that hold for arbitrary collections $\mathcal{C}$, and specialize to both the Shannon–Stam inequalities and the inequalities of ABBN [1]. In particular, we prove all these inequalities transparently using only simple projection facts, a variance drop lemma, and classical information-theoretic ideas. A remarkable feature of our proofs is that their main ingredients are rather well known, although our generalizations of the variance drop lemma appear to be new and are perhaps of independent interest. Both our results as well as the proofs lend themselves to intuitive statistical interpretations, several of which we have pointed out in the paper. We now point to potential directions of application.

The inequalities of this paper are relevant to the study of central limit theorems, especially for i.n.i.d. random variables. Indeed, we demonstrated monotonicity on average properties in such settings. Moreover, most approaches to entropic central limit theorems involve a detailed quantification of the gaps associated with monotonicity properties of Fisher information when the summands are nonnormal. Since the gap in our inequality is especially accessible due to our use of a Pythagorean property of projections (see Remark 3), it could be of interest in obtaining transparent proofs of entropic central limit theorems in i.n.i.d. settings, and perhaps rate conclusions under less restrictive assumptions than those imposed in [21] and [2].

The new Fisher information inequalities we present are also of interest, because of the relationship of inverse Fisher information to asymptotically efficient estimation. In this context, the subset sum inequality can be interpreted as a comparison of an asymptotic mean-squared error achieved with use of all $X_1, \ldots, X_n$, and the sum of the mean-squared errors achieved in distributed estimation by sensors that observe $(X_i, i \in \boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{C}$. The parameter of interest can either be a location parameter, or (following [21]) a natural parameter of exponential families for which the minimal sufficient statistics are sums. Furthermore, a nonasymptotic generalization of the new Fisher information inequalities holds (see [5] for details), which sheds light on minimax risks for estimation of a location parameter from sums.

Entropy inequalities involving subsets of random variables (although traditionally not involving sums) have played an important role in understanding some problems of graph theory. Radhakrishnan [33] provides a nice survey, and some recent developments (including joint entropy inequalities analogous to the entropy power inequalities in this paper) are discussed in [30]. The appearance of fractional packings in the refined inequality we present in Section VIII is particularly suggestive of further connections to be explored.

In multiuser information theory, subset sums of rates and information quantities involving subsets of random variables are critical in characterizing rate regions of certain source and channel coding problems (e.g., m-user multiple-access channels). Furthermore, there is a long history of the use of the classical entropy power inequality in the study of rate regions, see, e.g., Shannon [36], Bergmans [6], Ozarow [32], Costa [12], and Oohama [31]. For instance, the classical entropy power inequality was a key tool in Ozarow's solution of the Gaussian multiple description problem for two multiple descriptions, but seems to have been inadequate for problems involving three or more descriptions (see Wang and Viswanath [45] for a recent solution of one such problem without using the entropy power inequality). It seems natural to expand the set of tools available for investigation in these contexts.

Appendix I
The Analysis of Variance Decomposition

In order to prove the variance drop lemma, we use a decomposition of functions in $L^2(\mathbb{R}^n)$, which is nothing but the ANOVA decomposition of a statistic. For any $j \in [n]$, $E_j\psi$ denotes the conditional expectation of ψ, given all random variables other than X_j, i.e.,

$$E_j\psi(x_1, \ldots, x_n) = E[\psi(X_1, \ldots, X_n)|X_i = x_i \; \forall\, i \neq j] \tag{56}$$

averages out the dependence on the jth coordinate.

Fact 2 (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 $\boldsymbol{t} \subset [n]$, define the orthogonal linear subspaces

$$\mathcal{H}_{\boldsymbol{t}} = \{\psi \in L^2 : E_j\psi = \psi 1_{\{j \notin \boldsymbol{t}\}} \; \forall\, j \in [n]\} \tag{57}$$

of functions depending only on the variables indexed by $\boldsymbol{t}$. 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

$$\psi = \sum_{\boldsymbol{t}\subset[n]} \bar{E}_{\boldsymbol{t}}\psi \tag{58}$$

where $\bar{E}_{\boldsymbol{t}}\psi \in \mathcal{H}_{\boldsymbol{t}}$, and the subspaces $\mathcal{H}_{\boldsymbol{t}}$ (for $\boldsymbol{t} \subset [n]$) are orthogonal to each other.

Proof: Let $E_{\boldsymbol{t}}$ denote the integrating out of the variables in $\boldsymbol{t}$, 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

$$\psi = \prod_{j=1}^{n} [E_j + (I - E_j)]\psi$$

$$= \sum_{\boldsymbol{t} \subset [n]} \prod_{j \notin \boldsymbol{t}} E_j \prod_{j \in \boldsymbol{t}} (I - E_j)\psi$$
$$= \sum_{\boldsymbol{t} \subset [n]} \bar{E}_{\boldsymbol{t}}\psi \tag{59}$$

where

$$\bar{E}_{\boldsymbol{t}}\psi \equiv E_{\boldsymbol{t}^c} \prod_{j \in \boldsymbol{t}} (I - E_j)\psi. \tag{60}$$

Note that if j is in $\boldsymbol{t}$, $E_j \bar{E}_{\boldsymbol{t}}\psi = 0$, being in the image of the operator $E_j(I - E_j) = 0$. If j is not in $\boldsymbol{t}$, $\bar{E}_{\boldsymbol{t}}\psi$ is already in the image of E_j, and a further application of the projection E_j has no effect. Thus, $\bar{E}_{\boldsymbol{t}}\psi$ is in $\mathcal{H}_{\boldsymbol{t}}$.

Finally, we wish to show that the subspaces $\mathcal{H}_{\boldsymbol{t}}$ are orthogonal. For any distinct sets $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ in $[n]$, there exists an index j which is in one (say $\boldsymbol{t}_1$), but not the other (say $\boldsymbol{t}_2$). Then, by definition, $\bar{E}_{\boldsymbol{t}_1}\psi$ is contained in the image of E_j and $\bar{E}_{\boldsymbol{t}_2}\psi$ is contained in the image of $(I - E_j)$. Hence, $\bar{E}_{\boldsymbol{t}_1}\psi$ is orthogonal to $\bar{E}_{\boldsymbol{t}_2}\psi$. □

Remark 16: In the language of ANOVA familiar to statisticians, when ϕ is the empty set, $\bar{E}_{\phi}\psi$ is the mean; $\bar{E}_{\{1\}}\psi, \bar{E}_{\{2\}}\psi, \ldots, \bar{E}_{\{n\}}\psi$ are the main effects; $\{\bar{E}_{\boldsymbol{t}}\psi : |\boldsymbol{t}| = 2\}$ are the pairwise interactions, and so on. Fact 2 implies that for any subset $\boldsymbol{s} \subset [n]$, the function $\sum_{\{\boldsymbol{t}:\boldsymbol{t} \subset \boldsymbol{s}\}} \bar{E}_{\boldsymbol{t}}\psi$ is the best approximation (in mean square) to ψ that depends only on the collection $X_{\boldsymbol{s}}$ of random variables.

Remark 17: The historical roots of this decomposition lie in the work of von Mises [44] and Hoeffding [18]. For various interpretations, see Kurkjian and Zelen [24], Jacobsen [19], Rubin and Vitale [34], Efron and Stein [16], Karlin and Rinott [23], and Steele [40]; these works include applications of such decompositions to experimental design, linear models, U-statistics, and jackknife theory. Takemura [41] describes a general unifying framework for ANOVA decompositions.

Acknowledgment

We are indebted to Prasad Tetali for several interesting conversations and for the beneficial influence of Prasad's joint work with the first author on entropy inequalities. We are also very grateful to the anonymous referees and the Associate Editor for careful readings of the paper, and their insightful comments.

References

[1] 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, 2004.
[2] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, "On the rate of convergence in the entropic central limit theorem," *Probab. Theory Related Fields*, vol. 129, no. 3, pp. 381–390, 2004.
[3] A. R. Barron, Monotonic Central Limit Theorem for Densities Dep. Statistics, Stanford Univ., Stanford, CA, 1984, Tech. Rep. 50.
[4] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, pp. 336–342, 1986.
[5] M. Madiman, A. R. Barron, A. M. Kagan, and T. Yu, "On the Minimax Risks for Estimation of a Location Parameter From Sums," 2007, preprint.
[6] P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 279–280, Mar. 1974.
[7] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 2, pp. 267–271, Apr. 1965.
[8] L. D. Brown, "A proof of the central limit theorem motivated by the Cramér-Rao inequality," in *Statistics and Probability: Essays in Honor of C. R. Rao*. Amsterdam, The Netherlands: North-Holland, 1982, pp. 141–148.
[9] E. A. Carlen, "Superadditivity of Fisher's information and logarithmic Sobolev inequalities," *J. Funct. Anal.*, vol. 101, no. 1, pp. 194–211, 1991.
[10] E. A. Carlen and A. Soffer, "Entropy production by block variable summation and central limit theorems," *Commun. Math. Phys.*, vol. 140, 1991.
[11] F. R. K. Chung, Z. Füredi, M. R. Garey, and R. L. Graham, "On the fractional covering number of hypergraphs," *SIAM J. Discr. Math.*, vol. 1, no. 1, pp. 45–49, 1988.
[12] M. H. M. Costa, "On the Gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 5, pp. 607–615, Sep. 1985.
[13] M. H. M. Costa and T. M. Cover, "On the similarity of the entropy power inequality and the Brunn-Minkowski inequality," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 6, pp. 837–839, Nov. 1984.
[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[15] A. Dembo, T. M. Cover, and J. A. Thomas, "Information-theoretic inequalities," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1501–1518, Nov. 1991.
[16] B. Efron and C. Stein, "The jackknife estimate of variance," *Ann. Statist.*, vol. 9, no. 3, pp. 586–596, 1981.
[17] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inf. Contr.*, vol. 36, no. 2, pp. 133–156, 1978.
[18] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *Ann. Math. Statist.*, vol. 19, no. 3, pp. 293–325, 1948.
[19] R. L. Jacobsen, "Linear Algebra and ANOVA," Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1968.
[20] O. Johnson, "Entropy inequalities and the central limit theorem," *Stochastic Process. Appl.*, vol. 88, pp. 291–304, 2000.
[21] O. Johnson and A. R. Barron, "Fisher information inequalities and the central limit theorem," *Probab. Theory Related Fields*, vol. 129, no. 3, pp. 391–409, 2004.
[22] A. Kagan, "An inequality for the Pitman estimators related to the Stam inequality," *Sankhya Ser. A*, vol. 64, pp. 281–292, 2002.
[23] S. Karlin and Y. Rinott, "Applications of ANOVA type decompositions for comparisons of conditional variance statistics including jackknife estimates," *Ann. Statist.*, vol. 10, no. 2, pp. 485–501, 1982.
[24] B. Kurkjian and M. Zelen, "A calculus for factorial arrangements," *Ann. Math. Statist.*, vol. 33, pp. 600–619, 1962.
[25] E. L. Lehmann and G. Casella, "Theory of point estimation," in *Springer Texts in Statistics*, 2nd ed. New York: Springer-Verlag, 1998.
[26] E. H. Lieb, "Proof of an entropy conjecture of Wehrl," *Commun. Math. Phys.*, vol. 62, no. 1, pp. 35–41, 1978.
[27] T. Liu and P. Viswanath, "Two Proofs of the Fisher Information Inequality via Data Processing Arguments," 2005, preprint.
[28] M. Madiman and A. R. Barron, "The monotonicity of information in the central limit theorem and entropy power inequalities," in *Proc. IEEE Intl. Symp. Information Theory*, Seattle, WA, Jul. 2006, pp. 1021–1025.
[29] M. Madiman, O. Johnson, and I. Kontoyiannis, "Fisher information, compound Poisson approximation, and the Poisson channel," in *Proc. IEEE Int. Symp. Information Theory*, Nice, France, Jun. 2007.
[30] M. Madiman and P. Tetali, "Sandwich bounds for joint entropy," in *Proc. IEEE Intl. Symp. Information Theory*, Nice, France, Jun. 2007.
[31] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1057–1070, May 1998.
[32] L. Ozarow, "On a source coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, no. 10, pp. 1909–1921, 1980.
[33] J. Radhakrishnan, "Entropy and counting," in *IIT Kharagpur Golden Jubilee Volume.*, J. C. Misra, Ed. New Delhi, India: Narosa, 2001.
[34] H. Rubin and R. A. Vitale, "Asymptotic distribution of symmetric statistics," *Ann. Statist.*, vol. 8, no. 1, pp. 165–170, 1980.
[35] H. Schultz, "Semicircularity, Gaussianity and Monotonicity of Entropy," 2005, preprint.
[36] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
[37] R. Shimizu, "On Fisher's amount of information for location family. In," in *Statistical Distributions in Scientific Work*, G. P. Patil *et. al.*, Ed. Dordrecht, The Netherlands: Reidel, 1975, vol. 3, pp. 305–312.

[38] D. Shlyakhtenko, "A Free Analogue of Shannon's Problem on Monotonicity of Entropy," 2005, preprint.
[39] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Contr.*, vol. 2, pp. 101–112, 1959.
[40] J. M. Steele, "An Efron-Stein inequality for nonsymmetric statistics," *Ann. Statist.*, vol. 14, no. 2, pp. 753–758, 1986.
[41] A. Takemura, "Tensor analysis of ANOVA decomposition," *J. Amer. Statist. Assoc.*, vol. 78, no. 384, pp. 894–900, 1983.
[42] A. M. Tulino and S. Verdú, "Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4295–4297, Sep. 2006.
[43] S. Verdú and D. Guo, "A simple proof of the entropy-power inequality," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2165–2166, May 2006.
[44] R. van Mises, "On the asymptotic distribution of differentiable statistical functions," *Ann. Math. Statist.*, vol. 18, no. 3, pp. 309–348, 1947.
[45] H. Wang and P. Viswanath, "Vector Gaussian multiple description with individual and central receivers," *IEEE Trans. Inf. Theory*, to be published.
[46] R. Zamir, "A proof of the Fisher information inequality via a data processing argument," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1246–1250, May 1998.