A mutual information inequality and some applications

Ken Lau *Chandra Nair* *David Ng*

Abstract

In this paper we derive an inequality relating linear combinations of mutual information between subsets of mutually independent random variables and an auxiliary random variable. As corollaries of this inequality, we obtain new results and generalizations and new proofs of known results.

1 Introduction

In this paper we obtain an information inequality relating linear combinations of mutual information between subsets of mutually independent random variables and an auxiliary random variable. Our main result is a rather elementary inequality which surprisingly implies a variety of non-trivial inequalities and yields new inequalities. We are directly motivated by the work of Balister and Bollobás [1] who present generalizations of Shearer's lemma [2, 3], Han's inequality [4], and the Madiman–Tetali inequality [5]. We obtain a compression type inequality similar to Theorem 4.2 of [1], generalizing the work in [6]. We are also motivated by the work of Courtade [7] who presents an elementary proof of monotonicity of entropy power and Fisher information which was originally established by Artstein, Ball, Barthe and Naor [8]. Using a certain perturbative auxiliary, we recover the generalized Stam's inequality [9], which extends Stam's inequality for Fisher information [10] and the Artstein–Ball–Barthe–Naor inequality [8], as a corollary of our main result. We also extend the results involving maximal correlation by Dembo–Kagan–Shepp [11], strong data processing constants in [6], and obtain new Kullback–Leibler (KL) divergence convexity results.

1.1 Main results

Throughout this article we adapt the following notations. We denote by $[a : b]$ the set of integers $\geq a$ and $\leq b$. We denote by $|T|$ the cardinality of a set T. For random variables $X_1, \dots, X_n$ and for $T \subseteq [1 : n]$, we write $X_T := \{X_i\}_{i \in T}$, the tuple consisting of X_i where $i \in T$.

We define the notion of *compression* of multisets, as defined in [1], whereas we view a multiset as the finite sequence of its multiplicities. The relation "is a compression of" defines a partial order on the collection of multisets in $[1 : n]$. A finite sequence of non-negative real numbers $\{\beta_T\}_T$ indexed by $T \subseteq [1 : n]$ is minimal under this partial order (i.e. cannot be further compressed) if and only if the set $\{T \subseteq [1 : n] : \beta_T \neq 0\}$ is totally ordered under set inclusion.

Definition 1. Let n be a positive integer and let $\{\alpha_T\}_T, \{\beta_T\}_T$ be two finite sequences of non-negative real numbers indexed by $T \subseteq [1 : n]$. We call $\{\beta_T\}_T$ an *elementary compression* of $\{\alpha_T\}_T$ if there exist $A, B \subseteq [1 : n]$ with $A \not\subseteq B$ and $B \not\subseteq A$, and $0 < \delta \leq \min\{\alpha_A, \alpha_B\}$ such that for all $T \subseteq [1 : n]$ we have

$$\beta_T = \begin{cases} \alpha_T - \delta & \text{if } T = A \text{ or } T = B, \\ \alpha_T + \delta & \text{if } T = A \cup B \text{ or } T = A \cap B, \\ \alpha_T & \text{otherwise.} \end{cases}$$

The result of a finite sequence of elementary compressions of $\{\alpha_T\}_T$ is called a *compression* of $\{\alpha_T\}_T$.

The following lemma is rather immediate but forms the basis of most of the results in this paper.

Lemma 1. *Let $X_1, \dots, X_n$ be random variables that are mutually independent conditioned on a random variable $S_\emptyset$, and let U be any auxiliary random variable. Then the following hold:*

(i) $I(U; S_\emptyset, X_A) + I(U; S_\emptyset, X_B) \leq I(U; S_\emptyset, X_{A\cup B}) + I(U; S_\emptyset, X_{A\cap B})$ *for all* $A, B \subseteq [1 : n]$.

(ii) $\sum_{T\subseteq[1:n]} \alpha_T I(U; S_\emptyset, X_T) \leq \sum_{T\subseteq[1:n]} \beta_T I(U; S_\emptyset, X_T)$, *for any non-negative real numbers* α_T, β_T $(T \subseteq [1 : n])$ *such that* $\{\beta_T\}$ *is a compression of* $\{\alpha_T\}$.

(iii) $\sum_{T\subseteq[1:n]}\beta_T I(U;S_\emptyset,X_T)\le I(U;S_\emptyset,X_{[1:n]})+(c-1)I(U;S_\emptyset)$, *where* β_T $(T\subseteq[1:n])$ *are non-negative real numbers satisfying* $\sum_{T\subseteq[1:n]:T\ni i}\beta_T\le 1$ *for all* $i=1,\ldots,n$, *and* $c:=\sum_{T\subseteq[1:n]}\beta_T$.

Proof. Suppose $A,B\subseteq[1:n]$. Then

$$\begin{aligned}
I(U;S_\emptyset,X_B)-I(U;S_\emptyset,X_{A\cap B}) &= I(U;X_{B\setminus A}|S_\emptyset,X_{A\cap B})\\
&\le I(U,X_{A\setminus B};X_{B\setminus A}|S_\emptyset,X_{A\cap B})\\
&\overset{\text{(a)}}{=} I(U,X_{A\setminus B};X_{B\setminus A}|S_\emptyset,X_{A\cap B})-I(X_{A\setminus B};X_{B\setminus A}|S_\emptyset,X_{A\cap B})\\
&= I(U;X_{B\setminus A}|S_\emptyset,X_A)\\
&= I(U;S_\emptyset,X_{A\cup B})-I(U;S_\emptyset,X_A),
\end{aligned}$$

where (a) holds by the mutual independence of the X_i's conditioned on $S_\emptyset$. Rearranging gives

$$I(U;S_\emptyset,X_A)+I(U;S_\emptyset,X_B)\le I(U;S_\emptyset,X_{A\cup B})+I(U;S_\emptyset,X_{A\cap B}),$$

which is (i).

If $\{\beta_T\}$ is an elementary compression of $\{\alpha_T\}$, then the inequality in (ii) follows from (i) by canceling like terms on both sizes. Since a compression is obtained as a sequence of elementary compressions, (ii) follows.

We will show (iii) by induction on n. Indeed the base case $n=1$ is trivial. Note that (i) gives

$$I(U;S_\emptyset,X_{[1:n-1]})+I(U;S_\emptyset,X_{T\cup\{n\}})\le I(U;S_\emptyset,X_{[1:n]})+I(U;S_\emptyset,X_T)$$

for all $T\subseteq[1:n-1]$. Suppose β_T $(T\subseteq[1:n])$ are non-negative real numbers satisfying $\sum_{T\subseteq[1:n]:T\ni i}\beta_T\le 1$ for all $i=1,\ldots,n$. Then

$$\begin{aligned}
\sum_{T\subseteq[1:n]}\beta_T I(U;S_\emptyset,X_T) &= \sum_{T\subseteq[1:n-1]}\big(\beta_T I(U;S_\emptyset,X_T)+\beta_{T\cup\{n\}}I(U;S_\emptyset,X_{T\cup\{n\}})\big)\\
&\le \sum_{T\subseteq[1:n-1]}\big(\beta_T I(U;S_\emptyset,X_T)+\beta_{T\cup\{n\}}(I(U;S_\emptyset,X_{[1:n]})-I(U;S_\emptyset,X_{[1:n-1]})+I(U;S_\emptyset,X_T))\big)\\
&\overset{\text{(a)}}{\le} I(U;S_\emptyset,X_{[1:n]})-I(U;S_\emptyset,X_{[1:n-1]})+\sum_{T\subseteq[1:n-1]}(\beta_T+\beta_{T\cup\{n\}})I(U;S_\emptyset,X_T)\\
&\overset{\text{(b)}}{\le} I(U;S_\emptyset,X_{[1:n]})-I(U;S_\emptyset,X_{[1:n-1]})+I(U;S_\emptyset,X_{[1:n-1]})+(c-1)I(U;S_\emptyset)\\
&= I(U;S_\emptyset,X_{[1:n]})+(c-1)I(U;S_\emptyset),
\end{aligned}$$

where (a) holds since $\sum_{T\subseteq[1:n-1]}\beta_{T\cup\{n\}}\le 1$, and (b) follows by applying the induction hypothesis to the non-negative real numbers $\{\beta_T+\beta_{T\cup\{n\}}\}_{T\subseteq[1:n-1]}$. □

Definition 2. Let X_i $(i=1,\ldots,n)$ and S_T $(T\subseteq[1:n])$ be random variables. We call $\{S_T\}_T$ a *layered function family* on $X_1,\ldots,X_n$ if $S_\emptyset$ is independent of $X_{[1:n]}$, and for every non-empty $T\subseteq[1:n]$ and $i\in T$ there is a function $g_{T,i}$ such that $S_T=g_{T,i}(S_{T\setminus\{i\}},X_i)$.

Remark 1. Clearly a trivial example of a layered function family is given by $S_T:=(S_\emptyset,X_T)$. A canonical example of a layered function family is given by $S_T:=S_\emptyset+\sum_{i\in T}f_i(X_i)$, where f_i's are functions taking values in some Abelian monoid. In particular,

(i) $S_T:=S_\emptyset+\sum_{i\in T}X_i$, where $S_\emptyset,X_i\in\mathbb{R}^d$;

(ii) $S_T:=\max(\{S_\emptyset\}\cup\{X_i\}_{i\in T})$, where $S_\emptyset,X_i\in\mathbb{R}$;

are examples of layered function families.

Remark 2. Layered function families play a similar role as that of *partition-determined functions* in [12] and it may be possible that they are intrinsically trying to capture a similar behaviour and dependence structure. For our results, we prefer to stick with the definition of layered function families. Note that [12] deals with dependent random variables while here our main focus is on mutually independent random variables.

Lemma 2. *Let* $\{S_T\}_T$ *be a layered function family on mutually independent random variables* $X_1,\ldots,X_n$. *Suppose* $U\to S_{[1:n]}\to(S_\emptyset,X_{[1:n]})$ *forms a Markov chain. Then the following hold:*

(i) $U \to S_T \to (S_\emptyset, X_T)$ *forms a Markov chain for all* $T \subseteq [1:n]$.

(ii) $I(U; S_T) = I(U; S_\emptyset, X_T)$ *for all* $T \subseteq [1:n]$.

Proof. Suppose $T \subseteq [1:n]$. Consider

$$\begin{aligned}
0 &\overset{\text{(a)}}{=} I(U; S_\emptyset, X_{[1:n]} | S_{[1:n]}) \\
&= I(U; S_\emptyset, X_T, X_{[1:n]\setminus T} | S_{[1:n]}) \\
&\overset{\text{(b)}}{=} I(U; S_\emptyset, X_T, X_{[1:n]\setminus T}, S_T | S_{[1:n]}) \\
&\geq I(U; S_\emptyset, X_T | S_{[1:n]}, X_{[1:n]\setminus T}, S_T) \\
&\overset{\text{(c)}}{=} I(U; S_\emptyset, X_T | X_{[1:n]\setminus T}, S_T) \\
&\overset{\text{(d)}}{=} I(U; S_\emptyset, X_T | X_{[1:n]\setminus T}, S_T) + I(X_{[1:n]\setminus T}; S_\emptyset, X_T | S_T) \\
&= I(U, X_{[1:n]\setminus T}; S_\emptyset, X_T | S_T) \\
&\geq I(U; S_\emptyset, X_T | S_T) \\
&\geq 0,
\end{aligned}$$

where (a) holds since $U \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain, (b) holds since S_T is a function of $(S_\emptyset, X_T)$, (c) holds since $S_{[1:n]}$ is a function of $(S_T, X_{[1:n]\setminus T})$, and (d) holds since $X_{[1:n]\setminus T}$ and $(S_\emptyset, X_T, S_T)$ are independent. This shows (i). Furthermore,

$$\begin{aligned}
I(U; S_T) &\overset{\text{(a)}}{=} I(U; S_T, S_\emptyset, X_T) \\
&\overset{\text{(b)}}{=} I(U; S_\emptyset, X_T),
\end{aligned}$$

where (a) holds since $U \to S_T \to (S_\emptyset, X_T)$ forms a Markov chain, and (b) holds since S_T is a function of $(S_\emptyset, X_T)$. This shows (ii). □

We now state the main theorem. The proof is an immediate application of Lemma 2 to Lemma 1.

Theorem 1. *Let* $\{S_T\}_T$ *be a layered function family on mutually independent random variables* $X_1, \ldots, X_n$. *Suppose* $U \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ *forms a Markov chain. Then the following hold:*

(i) $I(U; S_A) + I(U; S_B) \leq I(U; S_{A\cup B}) + I(U; S_{A\cap B})$ *for all* $A, B \subseteq [1:n]$.

(ii) $\sum_{T\subseteq[1:n]} \alpha_T I(U; S_T) \leq \sum_{T\subseteq[1:n]} \beta_T I(U; S_T)$, *for any non-negative real numbers* α_T, β_T $(T \subseteq [1:n])$ *such that* $\{\beta_T\}$ *is a compression of* $\{\alpha_T\}$.

(iii) $\sum_{T\subseteq[1:n]} \beta_T I(U; S_T) \leq I(U; S_{[1:n]}) + (c-1) I(U; S_\emptyset)$, *where* β_T $(T \subseteq [1:n])$ *are non-negative real numbers satisfying* $\sum_{T\subseteq[1:n]: T\ni i} \beta_T \leq 1$ *for all* $i = 1, \ldots, n$, *and* $c := \sum_{T\subseteq[1:n]} \beta_T$.

It turns out the freedom in choosing the auxiliary random variable U plays a rather important role in the development of the inequalities.

1.2 Two families of perturbative auxiliaries

In this section we will present two families of auxiliaries that will turn out to be useful for obtaining corollaries to Theorem 1.

Lemma 3. *Let* $\{S_T\}_T$ *be a layered function family on mutually independent random variables* $X_1, \ldots, X_n$. *Suppose* f *is an* $\mathbb{R}^d$*-valued bounded measurable function, defined on the set of values of* $S_{[1:n]}$, *such that* $\mathrm{E}[f(S_{[1:n]})] = 0$. *Then there exists a family of random variables* $\{U^{(\epsilon)}\}_\epsilon$, *indexed by small enough* $\epsilon > 0$, *such that* $U^{(\epsilon)} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ *forms a Markov chain and*

$$I(U^{(\epsilon)}; S_T) = \frac{1}{2}\epsilon^2 \, \mathrm{E}[\| \mathrm{E}[f(S_{[1:n]}) | S_T] \|^2] + O(\epsilon^3)$$

for all $T \subseteq [1:n]$.

Proof. Let $\tilde{p}(\cdot)$ be the probability mass function of the uniform distribution on the Boolean hypercube $\{\pm 1\}^d$. For small enough $\epsilon > 0$, define the random variable $U^{(\epsilon)}$ taking values in $\{\pm 1\}^d$, satisfying the Markov chain $U^{(\epsilon)} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$, according to

$$p_{U^{(\epsilon)}|S_{[1:n]}}(u|s) := \tilde{p}(u)(1 + \epsilon\langle f(s), u\rangle).$$

Note that $p_{U^{(\epsilon)}}(u) = \tilde{p}(u)$ (which follows from $\mathrm{E}[f(S_{[1:n]})] = 0$), $\mathrm{E}[U^{(\epsilon)}] = 0$ and $\mathrm{E}[U^{(\epsilon)}U^{(\epsilon)\intercal}] = I$. For any $T \subseteq [1:n]$, since $U^{(\epsilon)} \to S_{[1:n]} \to S_T$ forms a Markov chain,

$$\begin{aligned} p_{U^{(\epsilon)}|S_T}(u|S_T) &= \mathrm{E}[p_{U^{(\epsilon)}|S_{[1:n]}}(u|S_{[1:n]})|S_T] \\ &= \tilde{p}(u)(1 + \epsilon\langle \mathrm{E}[f(S_{[1:n]})|S_T], u\rangle). \end{aligned}$$

Then we have

$$\begin{aligned} I(U^{(\epsilon)}; S_T) &= \mathrm{E}_{U^{(\epsilon)},S_T}\left[\log \frac{p(U^{(\epsilon)}|S_T)}{p(U^{(\epsilon)})}\right] \\ &= \mathrm{E}_{U^{(\epsilon)},S_T}\left[\log(1 + \epsilon\langle \mathrm{E}[f(S_{[1:n]})|S_T], U^{(\epsilon)}\rangle)\right] \\ &= \mathrm{E}_{S_T}\left[\sum_u \tilde{p}(u)(1 + \epsilon\langle \mathrm{E}[f(S_{[1:n]})|S_T], u\rangle)\log(1 + \epsilon\langle \mathrm{E}[f(S_{[1:n]})|S_T], u\rangle)\right] \\ &= \mathrm{E}_{S_T}\left[\sum_u \tilde{p}(u)\left(\epsilon\langle \mathrm{E}[f(S_{[1:n]})|S_T], u\rangle + \frac{1}{2}\epsilon^2\langle \mathrm{E}[f(S_{[1:n]})|S_T], u\rangle^2 + O(\epsilon^3)\right)\right] \\ &= \frac{1}{2}\epsilon^2 \operatorname{tr}\left(\mathrm{E}[\mathrm{E}[f(S_{[1:n]})|S_T]\,\mathrm{E}[f(S_{[1:n]})|S_T]^\intercal] \cdot \sum_u \tilde{p}(u)uu^\intercal\right) + O(\epsilon^3) \\ &= \frac{1}{2}\epsilon^2\,\mathrm{E}[\|\,\mathrm{E}[f(S_{[1:n]})|S_T]\|^2] + O(\epsilon^3). \end{aligned}$$

□

Lemma 4. *Let $\{S_T\}_T$ be a layered function family on mutually independent random variables $X_1, \ldots, X_n$. Suppose $q(\cdot)$ is a distribution that is absolutely continuous and has a bounded Radon–Nikodym derivative with respect to the distribution of $S_{[1:n]}$. Then there exists a family of random variables $\{U^{(\epsilon)}\}_\epsilon$, indexed by small enough $\epsilon > 0$, such that $U^{(\epsilon)} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain and*

$$I(U^{(\epsilon)}; S_T) = \epsilon D_{\mathrm{KL}}(p_{\tilde{S}_T} \| p_{S_T}) + O(\epsilon^2)$$

for all $T \subseteq [1:n]$, where the random variable $\tilde{S}_T$ is defined by

$$p_{\tilde{S}_T}(\tilde{s}) := \sum_s p_{S_T|S_{[1:n]}}(\tilde{s}|s)q(s).$$

Proof. Let $f(s) := q(s)/p_{S_{[1:n]}}(s)$ be the Radon–Nikodym derivative. For small enough $\epsilon > 0$, define the random variable $U^{(\epsilon)}$ taking values in $\{0, 1\}$, satisfying the Markov chain $U^{(\epsilon)} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$, according to

$$p_{U^{(\epsilon)}|S_{[1:n]}}(u|s) := \begin{cases} 1 - \epsilon f(s) & \text{if } u = 0, \\ \epsilon f(s) & \text{if } u = 1. \end{cases}$$

Note that $\mathrm{E}[f(S_{[1:n]})] = 1$ and

$$p_{U^{(\epsilon)}}(u) = \begin{cases} 1 - \epsilon & \text{if } u = 0, \\ \epsilon & \text{if } u = 1. \end{cases}$$

For any $T \subseteq [1:n]$, since $U^{(\epsilon)} \to S_{[1:n]} \to S_T$ forms a Markov chain,

$$\begin{aligned} p_{U^{(\epsilon)}|S_T}(u|S_T) &= \mathrm{E}[p_{U^{(\epsilon)}|S_{[1:n]}}(u|S_{[1:n]})|S_T] \\ &= \begin{cases} 1 - \epsilon\,\mathrm{E}[f(S_{[1:n]})|S_T] & \text{if } u = 0, \\ \epsilon\,\mathrm{E}[f(S_{[1:n]})|S_T] & \text{if } u = 1. \end{cases} \end{aligned}$$

Then we have

$$
\begin{aligned}
&I(U^{(\epsilon)}; S_T) \\
&= \mathrm{E}_{U^{(\epsilon)}, S_T}\left[\log \frac{p(U^{(\epsilon)}|S_T)}{p(U^{(\epsilon)})}\right] \\
&= \mathrm{E}_{S_T}\left[\epsilon\, \mathrm{E}[f(S_{[1:n]})|S_T] \log \mathrm{E}[f(S_{[1:n]})|S_T] + (1-\epsilon\, \mathrm{E}[f(S_{[1:n]})|S_T]) \log \frac{1-\epsilon\, \mathrm{E}[f(S_{[1:n]})|S_T]}{1-\epsilon}\right] \\
&= \epsilon\, \mathrm{E}_{S_T}\left[\frac{p_{\tilde{S}_T}(S_T)}{p_{S_T}(S_T)} \log \frac{p_{\tilde{S}_T}(S_T)}{p_{S_T}(S_T)}\right] + \mathrm{E}_{S_T}\left[(1-\epsilon\, \mathrm{E}[f(S_{[1:n]})|S_T])(\epsilon(1-\mathrm{E}[f(S_{[1:n]})|S_T]) + O(\epsilon^2))\right] \\
&= \epsilon D_{\mathrm{KL}}(p_{\tilde{S}_T} \| p_{S_T}) + O(\epsilon^2).
\end{aligned}
$$

□

Remark 3. These two families of perturbative auxiliaries are not new here and have been used extensively in [13, 14] and references therein.

2 Some consequences of the main result

In this section we will outline some existing results, extensions of existing results, as well as the new ones that we obtain as consequences of Theorem 1.

2.1 Entropy power inequalities and Fisher information inequalities

2.1.1 Historical remark

The celebrated *entropy power inequality* (EPI) as originally postulated by Shannon [15] states that if X, Y are independent random variables in $\mathbb{R}^d$ then

$$e^{\frac{2}{d}h(X+Y)} \geq e^{\frac{2}{d}h(X)} + e^{\frac{2}{d}h(Y)},$$

and equality holds if and only if both X, Y are Gaussian with proportional covariance matrices. Stam [10] showed that the EPI is a consequence of

$$\frac{1}{J(X+Y)} \geq \frac{1}{J(X)} + \frac{1}{J(Y)}.$$

Lieb [16] showed the following two (respectively) equivalent forms of the above two inequalities,

$$
\begin{aligned}
h(\sqrt{t}X + \sqrt{1-t}Y) &\geq th(X) + (1-t)h(Y), \\
J(\sqrt{t}X + \sqrt{1-t}Y) &\leq tJ(X) + (1-t)J(Y),
\end{aligned}
$$

for any $t \in (0,1)$, and equality holds if and only if both X, Y are Gaussian with the same covariance matrix. Several other proofs for the EPI were discoverd by Guo–Shamai–Verdu [17] (via MMSE), Rioul [18], and Courtade [19].

Lieb's form of the EPI implies that

$$h\left(\frac{X+Y}{\sqrt{2}}\right) \geq \frac{1}{2}\left(h(X) + h(Y)\right).$$

Lieb [16] conjectured that if $X_1, \ldots, X_n$ are mutually independent and identically distributed real-valued random variables, then $h\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)$ is non-decreasing in n. This conjecture was resolved by Artstein–Ball–Barthe–Naor [8] who showed the following inequality: If $a_1, \ldots, a_{n+1} \geq 0$ satisfies $\sum_{i=1}^{n+1} a_i^2 = 1$ then

$$h\left(\sum_{i=1}^{n+1} a_i X_i\right) \geq \sum_{i=1}^{n+1} \frac{1-a_i^2}{n} h\left(\frac{1}{\sqrt{1-a_i^2}} \sum_{\substack{j=1 \\ j \neq i}}^{n+1} a_j X_j\right),$$

and in particular,

$$h\left(\frac{1}{\sqrt{n+1}}\sum_{i=1}^{n+1} X_i\right) \geq \frac{1}{n+1}\sum_{i=1}^{n+1} h\left(\frac{1}{\sqrt{n}}\sum_{\substack{j=1\\ j\neq i}}^{n+1} X_j\right).$$

Their proof was simplified and extended in a series of works, e.g. Madiman–Barron [9] and Madiman–Ghassemi [20]. The best known version is the *fractional partition* form of the EPI:

$$e^{\frac{2}{d}h\left(\sum_{i=1}^n X_i\right)} \geq \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} \beta_T e^{\frac{2}{d}h\left(\sum_{i\in T} X_i\right)},$$

for any mutually independent random variables $X_1 \ldots, X_n$ in $\mathbb{R}^d$ with densities, and *fractional partition* $\{\beta_T\}_T$, i.e. a finite collection indexed by $T \subseteq [1:n]$, $T \neq \emptyset$, of non-negative real numbers satisfying $\sum_{T\subseteq[1:n]:T\ni i} \beta_T = 1$ for every $i \in [1:n]$. This was derived as a consequence of the following Fisher information inequality, that we shall refer to as the *generalized Stam's inequality*:

$$\frac{1}{J(S_{[1:n]})} \geq \sum_{T\subseteq[1:n]} \beta_T \frac{1}{J(S_T)},$$

where $S_T := \sum_{i\in T} X_i$.

Remark 4. Unlike the $n = 2$ setting, the implication that the generalized Stam's inequality implies the fractional partition form of the EPI did not have a straightforward proof. In this article, we use convex duality to show a straightforward proof of this implication.

2.1.2 Alternate proof of generalized Stam's inequality

In this subsection, we derive the generalized Stam's inequality involving Fisher information as an immediate consequence of our mutual information inequality. While a similar proof technique that we employ has been used by Courtade in [7] for the case of mutually independent and identically distributed random variables, as noted in [21] (Future work, item 4), the extension of the ideas to independent random variables is of independent interest.

Remark 5. To avoid technical issues, we will deal with random variables X with density function f_X that is smooth and rapidly decaying such that $|\log f_X|$ has at most polynomial growth at infinity.

Definition 3. Let X be a random variable in $\mathbb{R}^d$ with density f_X. The *score function* ρ_X of X is defined by

$$\rho_X := \frac{\nabla f_X}{f_X} = \nabla \log f_X.$$

The *Fisher information* $J(X)$ of X is defined by

$$J(X) := \mathrm{E}[\|\rho_X(X)\|^2].$$

Remark 6. Let X, Z be independent random variables in $\mathbb{R}^d$ such that $Z \sim \mathcal{N}(0, I)$. We have the following basic properties of Fisher information:

(i) $J(aX) = a^{-2}J(X)$ for all $a > 0$.

(ii) $\frac{1}{2}J(X + \sqrt{t}Z) = \frac{\partial}{\partial t}h(X + \sqrt{t}Z)$ for all $t \geq 0$.

(iii) If X has a (finite) covariance matrix then

$$h(X) = \frac{d}{2}\log 2\pi e - \frac{1}{2}\int_0^\infty \left(J(X + \sqrt{t}Z) - \frac{d}{1+t}\right) dt.$$

Property (ii) is also called de Bruijn's identity (e.g. [10]). Property (iii) is a consequence of (ii) and is originally shown by Barron [22] (cf. Lemma 3 of [9]).

Our proof employs the following theorem.

Theorem 2 (Stam [10])**.** *Suppose $X_1, \dots, X_n$ are mutually independent random variables in $\mathbb{R}^d$ with densities, and write $S_k := X_1 + \cdots + X_k$. Then*

$$\rho_{S_n}(S_n) = \mathrm{E}[\rho_{S_k}(S_k)|S_n]$$

for all $k = 1, \dots, n$.

Consequently we have

$$\mathrm{E}[\| \mathrm{E}[\rho_{S_k}(S_k)|S_n]\|^2] = J(S_n).$$

We now use Cauchy–Schwarz inequality to obtain an upper bound on the squared norm of the reversed conditional expectation.

Lemma 5. *Let $X_1, \dots, X_n$ be mutually independent random variables in $\mathbb{R}^d$ with densities. For $k = 1, \dots, n$ we write $S_k := X_1 + \cdots + X_k$. Then*

$$\mathrm{E}[\| \mathrm{E}[\rho_{S_n}(S_n)|S_k]\|^2] \geq \frac{J(S_n)^2}{J(S_k)}$$

for all $k = 1, \dots, n$.

Proof. Consider

$$\begin{aligned}
J(S_n) &= \mathrm{E}[\|\rho_{S_n}(S_n)\|^2] \\
&= \mathrm{E}[\langle \rho_{S_n}(S_n), \mathrm{E}[\rho_{S_k}(S_k)|S_n]\rangle] \\
&= \mathrm{E}[\mathrm{E}[\langle \rho_{S_n}(S_n), \rho_{S_k}(S_k)\rangle|S_n]] \\
&= \mathrm{E}[\langle \rho_{S_n}(S_n), \rho_{S_k}(S_k)\rangle] \\
&= \mathrm{E}[\mathrm{E}[\langle \rho_{S_n}(S_n), \rho_{S_k}(S_k)\rangle|S_k]] \\
&= \mathrm{E}[\langle \mathrm{E}[\rho_{S_n}(S_n)|S_k], \rho_{S_k}(S_k)\rangle] \\
&\overset{\text{(a)}}{\leq} \mathrm{E}[\| \mathrm{E}[\rho_{S_n}(S_n)|S_k]\|^2]^{1/2} \, \mathrm{E}[\|\rho_{S_k}(S_k)\|^2]^{1/2} \\
&= \mathrm{E}[\| \mathrm{E}[\rho_{S_n}(S_n)|S_k]\|^2]^{1/2} J(S_k)^{1/2},
\end{aligned}$$

where (a) follows from the Cauchy–Schwarz inequality. This gives the result. □

Proposition 1 (Generalized Stam's inequality, Theorem 2 of [9])**.** *Let $X_1, \dots, X_n$ be mutually independent random variables in $\mathbb{R}^d$ with densities. Suppose β_T ($T \subseteq [1:n]$) are non-negative real numbers satisfying $\sum_{T \subseteq [1:n] : T \ni i} \beta_T \leq 1$ for all $i = 1, \dots, n$. If $\mathrm{E}[\|\rho_{S_{[1:n]}}\|^2] < \infty$, then*

$$\frac{1}{J(S_{[1:n]})} \geq \sum_{T \subseteq [1:n]} \beta_T \frac{1}{J(S_T)},$$

where $S_T := \sum_{i \in T} X_i$.

Proof. Note that $S_\emptyset = 0$. Let us first assume that $\|\rho_{S_{[1:n]}}\|$ is bounded. An application of Lemma 3 (with $f = \rho_{S_{[1:n]}}$) gives the existence of a family of random variables $\{U^{(\epsilon)}\}_\epsilon$, indexed by small enough $\epsilon > 0$, such that $U^{(\epsilon)} \to S_{[1:n]} \to X_{[1:n]}$ forms a Markov chain and

$$I(U^{(\epsilon)}; S_T) = \frac{1}{2}\epsilon^2 \, \mathrm{E}[\| \mathrm{E}[\rho_{S_{[1:n]}}(S_{[1:n]})|S_T]\|^2] + O(\epsilon^3) \tag{1}$$

for all $T \subseteq [1:n]$. Then Theorem 1 (iii) implies

$$\sum_{T \subseteq [1:n]} \beta_T I(U^{(\epsilon)}; S_T) \leq I(U^{(\epsilon)}; S_{[1:n]}). \tag{2}$$

Now consider

$$\begin{aligned}
J(S_{[1:n]}) &= \mathrm{E}[\|\rho_{S_{[1:n]}}(S_{[1:n]})\|^2] \\
&\overset{\text{(a)}}{\geq} \sum_{T \subseteq [1:n]} \beta_T \, \mathrm{E}[\| \mathrm{E}[\rho_{S_{[1:n]}}(S_{[1:n]})|S_T]\|^2] \\
&\overset{\text{(b)}}{\geq} \sum_{T \subseteq [1:n]} \beta_T \frac{J(S_{[1:n]})^2}{J(S_T)},
\end{aligned}$$

where (a) is obtained by putting (1) into (2), dividing by $\frac{1}{2}\epsilon^2$ and then taking $\epsilon \to 0$, and (b) follows from Lemma 5. The result then follows from rearranging.

If $\|\rho_{S_{[1:n]}}\|$ is not bounded, then we define $f_B := \min\left\{1, \frac{B}{\|\rho_{S_{[1:n]}}\|}\right\}\rho_{S_{[1:n]}}$ and the proof proceeds as before with $\rho_{S_{[1:n]}}$ replaced by f_B till inequality (a). Now, we use dominated convergence theorem (since $\mathrm{E}[\|\rho_{S_{[1:n]}}\|^2] < \infty$ and let $B \to \infty$) to recover the form as above with the score functions. □

2.1.3 From generalized Stam's inequality to fractional entropy power inequality

The first two lemmas that we present below are well-known and are the "Lieb-type-equivalent" forms of the fractional EPI and the generalized Stam's inequality. We present a proof here for completeness.

Lemma 6. *Let $X_1, \ldots, X_n$ be mutually independent random variables in $\mathbb{R}^d$. Let $S_T := \sum_{i \in T} X_i$. Suppose β_T ($T \subseteq [1:n]$, $T \neq \emptyset$) are non-negative real numbers satisfying $\sum_{T \subseteq [1:n]: T \ni i} \beta_T \leq 1$ for all $i \in [1:n]$. Then the following are equivalent.*

(i) It holds that

$$e^{\frac{2}{d}h(S_{[1:n]})} \geq \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} \beta_T e^{\frac{2}{d}h(S_T)}.$$

(ii) For all non-negative real numbers w_T ($T \subseteq [1:n]$, $T \neq \emptyset$) with $\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T = 1$, it holds that

$$h(S_{[1:n]}) \geq \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T h\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right).$$

Proof. We first show (i) implies (ii). Indeed,

$$\begin{aligned}
\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T h\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right) &\overset{(a)}{\leq} \frac{d}{2} \log\left(\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T e^{\frac{2}{d}h\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right)}\right) \\
&= \frac{d}{2} \log\left(\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} \beta_T e^{\frac{2}{d}h(S_T)}\right) \\
&\overset{(b)}{\leq} h(S_{[1:n]}),
\end{aligned}$$

where (a) follows from concavity of $\log(\cdot)$ and (b) follows from (i).

Now we show (ii) implies (i). Set $w_T := \beta_T e^{\frac{2}{d}h(S_T)} \left(\sum_{\substack{\tilde{T} \subseteq [1:n] \\ \tilde{T} \neq \emptyset}} \beta_{\tilde{T}} e^{\frac{2}{d}h(S_{\tilde{T}})}\right)^{-1}$. Note that

$$h\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right) = \frac{d}{2} \log \frac{\beta_T e^{\frac{2}{d}h(S_T)}}{w_T} = \frac{d}{2} \log\left(\sum_{\substack{\tilde{T} \subseteq [1:n] \\ \tilde{T} \neq \emptyset}} \beta_{\tilde{T}} e^{\frac{2}{d}h(S_{\tilde{T}})}\right)$$

is independent of choice of T, and hence (i) follows immediately from (ii). □

Lemma 7. *Let $X_1, \ldots, X_n$ be mutually independent random variables in $\mathbb{R}^d$. Let $S_T := \sum_{i \in T} X_i$. Suppose β_T ($T \subseteq [1:n]$, $T \neq \emptyset$) are non-negative real numbers satisfying $\sum_{T \subseteq [1:n]: T \ni i} \beta_T \leq 1$ for all $i \in [1:n]$. Then the following are equivalent.*

(i) It holds that

$$\frac{1}{J(S_{[1:n]})} \geq \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} \beta_T \frac{1}{J(S_T)}.$$

(ii) For all non-negative real numbers w_T ($T \subseteq [1:n]$, $T \neq \emptyset$) with $\sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} w_T = 1$, it holds that

$$J(S_{[1:n]}) \leq \sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right).$$

Proof. We first show (i) implies (ii). Indeed,

$$\begin{aligned}
\sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right) &\overset{(a)}{\geq} \left(\sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} w_T \frac{1}{J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right)}\right)^{-1} \\
&= \left(\sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} \beta_T \frac{1}{J(S_T)}\right)^{-1} \\
&\overset{(b)}{\geq} J(S_{[1:n]}),
\end{aligned}$$

where (a) follows from convexity of $(\cdot)^{-1}$ and (b) follows from (i).

Now we show (ii) implies (i). Set $w_T := \beta_T \frac{1}{J(S_T)} \left(\sum_{\substack{\tilde{T}\subseteq[1:n]\\\tilde{T}\neq\emptyset}} \beta_{\tilde{T}} \frac{1}{J(S_{\tilde{T}})}\right)^{-1}$. Note that

$$J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right) = \frac{w_T}{\beta_T} J(S_T) = \left(\sum_{\substack{\tilde{T}\subseteq[1:n]\\\tilde{T}\neq\emptyset}} \beta_{\tilde{T}} \frac{1}{J(S_{\tilde{T}})}\right)^{-1}$$

is independent of choice of T, and hence (i) follows immediately from (ii). □

Lemma 8 below is the crucial component of our proof. The lemma is used to show that by restricting our attention to optimal fractional partitions, we can essentially extend the proof for $n = 2$ to larger values of n.

Lemma 8. *Let w_T ($T \subseteq [1:n]$, $T \neq \emptyset$) be non-negative real numbers. Then the maximization*

$$\max_{\substack{\beta_T\geq 0\\ \sum_{T\ni i}\beta_T\leq 1}} \sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} w_T \log \beta_T$$

is attained at $\beta_T = \frac{w_T}{\sum_{i\in T}\lambda_i}$, for some $\lambda_i > 0$ ($i \in [1:n]$), with $\sum_{T\subseteq[1:n]:T\ni i} \beta_T = 1$ for all $i \in [1:n]$.

Proof. Consider

$$\begin{aligned}
\max_{\substack{\beta_T\geq 0\\ \sum_{T\ni i}\beta_T\leq 1}} \sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} w_T \log \beta_T &\overset{(a)}{=} \min_{\lambda_i\geq 0} \max_{\beta_T\geq 0} \left(\sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} w_T \log \beta_T + \sum_{i=1}^{n} \lambda_i \left(1 - \sum_{T\ni i} \beta_T\right)\right) \\
&= \min_{\lambda_i\geq 0} \left(\sum_{i=1}^{n} \lambda_i + \max_{\beta_T\geq 0} \sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} \left(w_T \log \beta_T - \beta_T \sum_{i\in T} \lambda_i\right)\right) \\
&\overset{(b)}{=} \min_{\lambda_i\geq 0} \left(\sum_{i=1}^{n} \lambda_i + \sum_{\substack{T\subseteq[1:n]\\T\neq\emptyset}} \left(w_T \log \frac{w_T}{\sum_{i\in T}\lambda_i} - w_T\right)\right),
\end{aligned}$$

where (a) holds by strong duality since Slater's condition (see Theorem 3.2.8 in [23] for instance) is satisfied for the maximization on the left hand side, and (b) holds since the maximum is attained at $\beta_T = \frac{w_T}{\sum_{i\in T}\lambda_i}$.

The minimization on the last line is a convex problem and is attained at some λ_i^*'s satisfying the first-order condition $\sum_{T\ni i} \frac{w_T}{\sum_{j\in T}\lambda_j^*} = 1$ $(i \in [1:n])$. Let $\beta_T^* := \frac{w_T}{\sum_{i\in T}\lambda_i^*}$. Then

$$\begin{aligned}\max_{\substack{\beta_T\geq 0\\ \sum_{T\ni i}\beta_T\leq 1}} \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T\log\beta_T &\leq \sum_{i=1}^n \lambda_i^* + \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}}\left(w_T\log\beta_T^* - \beta_T^*\sum_{i\in T}\lambda_i^*\right)\\ &= \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T\log\beta_T^* + \sum_{i=1}^n\lambda_i^* - \sum_{i=1}^n\left(\lambda_i^*\sum_{T\ni i}\beta_T^*\right)\\ &= \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T\log\beta_T^*,\end{aligned}$$

hence the maximization on the left hand side of the first line is attained at $\beta_T = \beta_T^*$. □

The following lemma shows that the dual variables λ_i in the proof of Lemma 8 represent the variances of the Gaussians while extending the proof from $n = 2$ to larger n using an approach of calculus of variations.

Lemma 9. *Let $X_1, \ldots, X_n$ be mutually independent random variables in $\mathbb{R}^d$. Let $S_T := \sum_{i\in T} X_i$. Let w_T $(T \subseteq [1:n],\ T \neq \emptyset)$ be non-negative real numbers satisfying $\sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T = 1$. Let β_T $(T \subseteq [1:n],\ T \neq \emptyset)$ be non-negative real numbers satisfying $\sum_{T\subseteq[1:n]:T\ni i}\beta_T \leq 1$ for all $i \in [1:n]$. Then (i) implies (ii).*

(i) For all $X_1, \ldots, X_n$, $\{w_T\}$ and $\{\beta_T\}$ it holds that

$$J(S_{[1:n]}) \leq \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}}S_T\right).$$

(ii) For all $X_1, \ldots, X_n$, $\{w_T\}$ and $\{\beta_T\}$ it holds that

$$h(S_{[1:n]}) \geq \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T h\left(\sqrt{\frac{\beta_T}{w_T}}S_T\right).$$

Proof. It suffices to show that (ii) holds for the β_T's that maximize the right hand side. In view of Lemma 8 we can write $\beta_T = \frac{w_T}{\sum_{i\in T}\lambda_i}$ for some $\lambda_i > 0$ $(i \in [1:n])$ such that $\sum_{T\subseteq[1:n]:T\ni i}\beta_T = 1$ is satisfied for all $i \in [1:n]$. Consequently, we have

$$\sum_{i=1}^n\lambda_i = \sum_{i=1}^n\left(\lambda_i\sum_{T\ni i}\beta_T\right) = \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}}\left(\beta_T\sum_{i\in T}\lambda_i\right) = \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T = 1.$$

Now for $t \in [0,1]$ define

$$f(t) := h\left(\sqrt{1-t}S_{[1:n]} + \sqrt{t}Z\right) - \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T h\left(\sqrt{\frac{\beta_T}{w_T}}\sqrt{1-t}S_T + \sqrt{t}Z\right),$$

where $Z \sim \mathcal{N}(0,1)$. Note that $f(1) = 0$ and hence it suffices to show $f'(t) \leq 0$ for all $0 \leq t \leq 1$. Indeed

$$\begin{aligned}
f'(t) &= \frac{1}{2}\frac{1}{1-t}\left(J\left(\sqrt{1-t}S_{[1:n]} + \sqrt{t}Z\right) - \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}}\sqrt{1-t}S_T + \sqrt{t}Z\right)\right) \\
&= \frac{1}{2}\frac{1}{1-t}\left(J\left(\sqrt{1-t}S_{[1:n]} + \sqrt{\sum_{i=1}^n \lambda_i}\sqrt{t}Z\right) - \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}}\sqrt{1-t}S_T + \sqrt{\frac{\beta_T}{w_T}\sum_{i\in T}\lambda_i}\sqrt{t}Z\right)\right) \\
&= \frac{1}{2}\frac{1}{1-t}\left(J\left(\sum_{i=1}^n X_{i,t}\right) - \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}}\sum_{i\in T} X_{i,t}\right)\right) \\
&\overset{(a)}{\leq} 0,
\end{aligned}$$

where we have set $X_{i,t} := \sqrt{1-t}X_i + \sqrt{\lambda_i t}Z_i$, where $Z_i \sim \mathcal{N}(0,1)$, and (a) follows from (i). □

2.2 Discrete convexity, strong data processing constant and maximal correlation

In this subsection, we establish some discrete convexity results and consequently some results about strong data processing constants and maximal correlations of joint distributions generalizing results in [6] and [11].

The following is a subclass of layered function families that we will also be considering in this section.

Definition 4. Let $\{S_T\}_T$ be a layered function family on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. We call the layered function family $\{S_T\}_T$ *symmetric* if for all permutations π of $[1:n]$ the distributions of $(S_{[1:n]}, S_\emptyset, X_1, \ldots, X_n)$ and $(S_{[1:n]}, S_\emptyset, X_{\pi(1)}, \ldots, X_{\pi(n)})$ are the same.

Remark 7. If $X_1, \ldots, X_n$ are mutually independent and identically distributed random variables, Remark 1 (i) and (ii) are examples of symmetric layered function families.

Lemma 10 (Discrete convexity). *Suppose φ_k $(k = 0, 1, \ldots, n)$ are real numbers satisfying*

$$\varphi_{k-1} + \varphi_{k+1} \geq 2\varphi_k \tag{3}$$

for all $k = 1, \ldots, n-1$. Then

$$\varphi_k \leq \frac{n-k}{n-l}\varphi_l + \frac{k-l}{n-l}\varphi_n$$

for all $l = 0, 1, \ldots, n-1$, and k satisfying $l \leq k \leq n$.

Proof. Note that $k = n$ and $l = k$ are immediate, so we assume $l < k < n$. Observe that $\varphi_k - \varphi_{k-1}$ is nondecreasing in k. Then

$$\begin{aligned}
\varphi_n - \varphi_k &= (\varphi_n - \varphi_{n-1}) + (\varphi_{n-1} - \varphi_{n-2}) + \cdots + (\varphi_{k+1} - \varphi_k) \\
&\geq (n-k)(\varphi_{k+1} - \varphi_k) \\
&\geq (n-k)(\varphi_k - \varphi_{k-1}) \\
&\geq \frac{n-k}{k-l}((\varphi_k - \varphi_{k-1}) + (\varphi_{k-1} - \varphi_{k-2}) + \cdots + (\varphi_{l+1} - \varphi_l)) \\
&= \frac{n-k}{k-l}(\varphi_k - \varphi_l).
\end{aligned}$$

The result follows by rearranging. □

Proposition 2. *Let $\{S_T\}_T$ be a symmetric layered function family on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. Suppose U is a random variable such that $U \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain. Then $I(U; S_T)$ is a function of $|T|$, and we have*

$$I(U; S_T) + I(U; S_{T\cup\{i,j\}}) \geq I(U; S_{T\cup\{i\}}) + I(U; S_{T\cup\{j\}})$$

for all $T \subseteq [1:n]$ *and distinct elements* i, j *in* $[1:n] \setminus T$*. Furthermore,*

$$I(U; S_T) \leq \frac{n-|T|}{n} I(U; S_\emptyset) + \frac{|T|}{n} I(U; S_{[1:n]})$$

for all $T \subseteq [1:n]$*.*

Proof. We first show that $I(U; S_T)$ is a function of $|T|$. It suffices to establish $I(U; S_T) = I(U; S_{[1:|T|]})$ for all $T \subseteq [1:n]$. Take a permutation π of $[1:n]$, that is increasing on $[1:|T|]$, such that $T = \{\pi(i)\}_{i=1,\ldots,|T|}$. From the definition of symmetric layered function family and the Markov chain $U \to S_{[1:n]} \to (S_\emptyset, X_1, \ldots, X_n)$, we have that the distributions of $(U, S_\emptyset, X_1, \ldots, X_n)$ and $\big(U, S_\emptyset, X_{\pi(1)}, \ldots, X_{\pi(n)}\big)$ are the same. In particular, the distributions of $\big(U, S_\emptyset, X_{[1:|T|]}\big)$ and $(U, S_\emptyset, X_T)$ are the same. Hence Lemma 2 (ii) gives

$$\begin{aligned} I(U; S_T) &= I(U; S_\emptyset, X_T) \\ &= I(U; S_\emptyset, X_{[1:|T|]}) \\ &= I(U; S_{[1:|T|]}). \end{aligned}$$

Now we show that $\varphi_k := I(U; S_T)$, where T is any subset of $[1:n]$ of cardinality k, satisfies (3). For any $k = 1, \ldots, n-1$, take any $T \subseteq [1:n]$ with $|T| = k-1$ and distinct elements i, j in $[1:n] \setminus T$, and we have

$$\begin{aligned} \varphi_{k-1} + \varphi_{k+1} &= I(U; S_T) + I(U; S_{T\cup\{i,j\}}) \\ &\overset{\text{(a)}}{\geq} I(U; S_{T\cup\{i\}}) + I(U; S_{T\cup\{j\}}) \\ &= 2\varphi_k, \end{aligned}$$

where (a) follows from (i) of Theorem 1. Hence (3) is satisfied. Then an application of Lemma 10 (with $l = 0$) yields

$$\varphi_k \leq \frac{n-k}{n}\varphi_0 + \frac{k}{n}\varphi_n,$$

or equivalently,

$$I(U; S_T) \leq \frac{n-|T|}{n} I(U; S_\emptyset) + \frac{|T|}{n} I(U; S_{[1:n]})$$

for all $T \subseteq [1:n]$. □

Corollary 1. *Let* $\{S_T\}_T$ *be a symmetric layered function family on mutually independent and identically distributed random variables* $X_1, \ldots, X_n$*. Then the following hold:*

(i) Suppose f *is an* $\mathbb{R}^d$*-valued bounded measurable function, defined on the set of values of* $S_{[1:n]}$*, such that* $\mathrm{E}[f(S_{[1:n]})] = 0$*. Then*

$$\mathrm{E}[\|\,\mathrm{E}[f(S_{[1:n]})|S_T]\|^2] \leq \frac{n-|T|}{n}\,\mathrm{E}[\|\,\mathrm{E}[f(S_{[1:n]})|S_\emptyset]\|^2] + \frac{|T|}{n}\,\mathrm{E}[\|f(S_{[1:n]})\|^2]$$

for all $T \subseteq [1:n]$*.*

(ii) Suppose $q(\cdot)$ *is a distribution absolutely continuous and with bounded Radon–Nikodym derivative with respect to the distribution of* $S_{[1:n]}$*. For* $T \subseteq [1:n]$ *let the random variable* $\tilde{S}_T$ *be defined by*

$$p_{\tilde{S}_T}(\tilde{s}) := \sum_s p_{S_T|S_{[1:n]}}(\tilde{s}|s)q(s).$$

Then

$$D_{\mathrm{KL}}(p_{\tilde{S}_T}\|p_{S_T}) + D_{\mathrm{KL}}(p_{\tilde{S}_{T\cup\{i,j\}}}\|p_{S_{T\cup\{i,j\}}}) \geq D_{\mathrm{KL}}(p_{\tilde{S}_{T\cup\{i\}}}\|p_{S_{T\cup\{i\}}}) + D_{\mathrm{KL}}(p_{\tilde{S}_{T\cup\{j\}}}\|p_{S_{T\cup\{j\}}})$$

for all $T \subseteq [1:n]$ *and distinct elements* i, j *in* $[1:n] \setminus T$*. Furthermore,*

$$D_{\mathrm{KL}}(p_{\tilde{S}_T}\|p_{S_T}) \leq \frac{n-|T|}{n} D_{\mathrm{KL}}(p_{\tilde{S}_\emptyset}\|p_{S_\emptyset}) + \frac{|T|}{n} D_{\mathrm{KL}}(p_{\tilde{S}_{[1:n]}}\|p_{S_{[1:n]}})$$

for all $T \subseteq [1:n]$*.*

Proof. (i) and (ii) are direct applications of Lemma 3 and 4, respectively, to Proposition 2. □

Definition 5. Let S be a function on mutually independent and identically distributed random variables $X_1, \dots, X_n$. We call S *cyclically symmetric* if for all cyclic shifts π of $[1:n]$ the distributions of $(S, X_1, \dots, X_n)$ and $(S, X_{\pi(1)}, \dots, X_{\pi(n)})$ are the same.

Remark 8. The function $S := \sum_{i=1}^{n} X_i X_{i+1}$ (with $X_{n+1} := X_1$), where X_i's are mutually independent and identically distributed random variables in $\mathbb{R}$, is an example of cyclically symmetric function.

Proposition 3. *Let S be a cyclically symmetric function on mutually independent and identically distributed random variables $X_1, \dots, X_n$. Suppose U is a random variable such that $U \to S \to X_{[1:n]}$ forms a Markov chain. Then for all $k = 1, \dots, n-1$ we have*

$$I(U; X_{[1:k-1]}) + I(U; X_{[1:k+1]}) \geq 2I(U; X_{[1:k]}).$$

Furthermore,

$$I(U; X_{[1:k]}) \leq \frac{k}{n} I(U; S)$$

for all $k = 0, 1, \dots, n$.

Proof. Since $U \to S \to X_{[1:n]}$ forms a Markov chain and S is a function of $X_{[1:n]}$, we have $I(U; S) = I(U; X_{[1:n]})$. Further from the cyclic symmetry of S and the Markov chain $U \to S \to X_{[1:n]}$, we have that the distributions of $(U, S, X_1, X_2, \dots, X_n)$ and $(U, S, X_n, X_1, \dots, X_{n-1})$ are the same. Consequently, for all $k = 0, \dots, n-1$ we have $I(U; X_{[1:k+1]}) = I(U; X_{[1:k] \cup \{n\}})$. Hence for $k = 1, \dots, n-1$,

$$\begin{aligned}
I(U; X_{[1:k+1]}) - I(U; X_{[1:k]}) &= I(U; X_{[1:k] \cup \{n\}}) - I(U; X_{[1:k]}) \\
&= I(U; X_n | X_{[1:k]}) \\
&\overset{\text{(a)}}{=} I(U; X_n | X_{[1:k]}) + I(X_k; X_n | X_{[1:k-1]}) \\
&= I(U, X_k; X_n | X_{[1:k-1]}) \\
&\geq I(U; X_n | X_{[1:k-1]}) \\
&= I(U; X_{[1:k-1] \cup \{n\}}) - I(U; X_{[1:k-1]}) \\
&= I(U; X_{[1:k]}) - I(U; X_{[1:k-1]}),
\end{aligned}$$

where (a) holds since X_k is independent of $X_{[1:k-1] \cup \{n\}}$. Now $\varphi_k := I(U; X_{[1:k]})$ satisfies (3) and hence by Lemma 10 (with $l = 0$) we have

$$\begin{aligned}
I(U; X_{[1:k]}) &\leq \frac{k}{n} I(U; X_{[1:n]}) \\
&= \frac{k}{n} I(U; S)
\end{aligned}$$

as required. □

2.2.1 Strong data processing constant

Definition 6. The *strong data processing constant* $s_*(X; Y)$ of two random variables X, Y is defined by

$$s_*(X; Y) := \sup_{\substack{p(u|x) \\ I(U;X) \neq 0}} \frac{I(U; Y)}{I(U; X)}.$$

Corollary 2. *Let $\{S_T\}_T$ be a symmetric layered function family on mutually independent and identically distributed random variables $X_1, \dots, X_n$. Then*

$$s_*(S_{[1:n]}; S_T) \leq \frac{n - |T|}{n} s_*(S_{[1:n]}; S_\emptyset) + \frac{|T|}{n}$$

for all $T \subseteq [1:n]$.

Proof. Fix any U satisfying the Markov chain $U \to S_{[1:n]} \to S_T$. Define a random variable $\tilde{U}$, satisfying the Markov chain $\tilde{U} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$, according to

$$p_{\tilde{U}|S_{[1:n]}}(u|s) := p_{U|S_{[1:n]}}(u|s).$$

Indeed $\tilde{U}$ also satisfies the Markov chain $\tilde{U} \to S_{[1:n]} \to S_T$ since S_T is a function of $(S_\emptyset, X_{[1:n]})$. Hence the distributions of $(U, S_{[1:n]}, S_T)$ and $(\tilde{U}, S_{[1:n]}, S_T)$ are the same. Therefore,

$$\begin{aligned} \frac{I(U;S_T)}{I(U;S_{[1:n]})} &= \frac{I(\tilde{U};S_T)}{I(\tilde{U};S_{[1:n]})} \\ &\overset{(a)}{\leq} \frac{n-|T|}{n}\frac{I(\tilde{U};S_\emptyset)}{I(\tilde{U};S_{[1:n]})} + \frac{|T|}{n} \\ &\leq \frac{n-|T|}{n} s_*(S_{[1:n]};S_\emptyset) + \frac{|T|}{n}, \end{aligned}$$

where (a) is an application of Proposition 2. □

Remark 9. Observe that this result generalizes the one in [6] from sums of mutually independent and identically distributed random variables to the more general symmetric layered function families. The proof technique used here is clearly motivated by the arguments in [6].

Corollary 3. *Let S be a cyclically symmetric function on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. Then $s_*(S;X_{[1:k]}) \leq \frac{k}{n}$ for all $k = 1, \ldots, n$.*

Proof. This is immediate from Proposition 3. □

2.2.2 Maximal correlation

The Hirschfeld–Gebelein–Rényi maximal correlation measures the dependence between two random variables in a general probability space. This quantity is first introduced by Hirschfeld [24] and Gebelein [25] and then studied by Rényi [26].

Definition 7. The *Hirschfeld–Gebelein–Rényi maximal correlation* $\rho_m(X;Y)$ of two random variables X, Y is defined by

$$\rho_m(X;Y) := \sup_{\substack{f,g \text{ real-valued measurable} \\ \mathrm{E}[f(X)]=\mathrm{E}[g(Y)]=0 \\ \mathrm{E}[f(X)^2]=\mathrm{E}[g(X)^2]=1}} \mathrm{E}[f(X)g(Y)].$$

An alternative expression for the quantity is formulated by Rényi [26] as follows.

Proposition 4 (Rényi [26])**.** *Let X, Y be random variables. Then*

$$\rho_m(X;Y) = \sup_{\substack{f \text{ real-valued measurable} \\ \mathrm{E}[f(X)]=0 \\ \mathrm{E}[f(X)^2]=1}} \mathrm{E}[\mathrm{E}[f(X)|Y]^2]^{1/2}.$$

Corollary 4. *Let $\{S_T\}_T$ be a symmetric layered function family on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. Then*

$$\rho_m(S_{[1:n]};S_T)^2 \leq \frac{n-|T|}{n}\rho_m(S_{[1:n]};S_\emptyset)^2 + \frac{|T|}{n}$$

for all $T \subseteq [1:n]$.

Proof. By Corollary 1 (i), for any bounded real-valued measurable function f such that $\mathrm{E}[f(S_{[1:n]})] = 0$ and $\mathrm{E}[f(S_{[1:n]})^2] = 1$ we have

$$\begin{aligned} \mathrm{E}[\mathrm{E}[f(S_{[1:n]})|S_T]^2] &\leq \frac{n-|T|}{n}\mathrm{E}[\mathrm{E}[f(S_{[1:n]})|S_\emptyset]^2] + \frac{|T|}{n}\mathrm{E}[f(S_{[1:n]})^2] \\ &\leq \frac{n-|T|}{n}\rho_m(S_{[1:n]};S_\emptyset)^2 + \frac{|T|}{n}. \end{aligned}$$

Taking supremum over f yields the result. □

2.2.3 KL divergence inequality

The KL divergence inequalities obtained in Corollary 1 (ii) imply, by choosing $X_1, \ldots, X_n$ to follow Poisson distribution, certain new convexity results concerning the KL divergence of binomial distribution given a Poisson distribution. Our results have a similar flavor to a conjecture of Yu (Conjecture 1 of [27]) who conjectured that $N \mapsto D_{\mathrm{KL}}\left(\mathrm{Binomial}\left(N, \frac{\lambda}{N}\right) \middle\| \mathrm{Poisson}\left(\lambda\right)\right)$ is completely monotonic. Even the convexity of this function is yet to be proven.

Lemma 11. *Suppose $X_1 \sim \mathrm{Poisson}(\lambda_1)$ and $X_2 \sim \mathrm{Poisson}(\lambda_2)$ are independent and $Y \sim \mathrm{Binomial}(N, \mu)$. Then the random variable $\tilde{Y}$ defined by*

$$p_{\tilde{Y}}(\tilde{y}) := \sum_y p_{X_1|X_1+X_2}(\tilde{y}|y) p_Y(y)$$

satisfies $\tilde{Y} \sim \mathrm{Binomial}\left(N, \frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)$.

Proof. We first compute

$$\begin{aligned} p_{X_1|X_1+X_2}(\tilde{y}|y) &= \frac{p_{X_1}(\tilde{y}) p_{X_2}(y-\tilde{y})}{p_{X_1+X_2}(y)} \\ &= \binom{y}{\tilde{y}} \frac{\lambda_1^{\tilde{y}} \lambda_2^{y-\tilde{y}}}{(\lambda_1+\lambda_2)^y}. \end{aligned}$$

Then

$$\begin{aligned} p_{\tilde{Y}}(\tilde{y}) &= \sum_y p_{X_1|X_1+X_2}(\tilde{y}|y) p_Y(y) \\ &= \sum_{y=\tilde{y}}^{N} \binom{y}{\tilde{y}} \frac{\lambda_1^{\tilde{y}} \lambda_2^{y-\tilde{y}}}{(\lambda_1+\lambda_2)^y} \binom{N}{y} \mu^y (1-\mu)^{N-y} \\ &= \binom{N}{\tilde{y}} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)^{\tilde{y}} \sum_{y=\tilde{y}}^{N} \binom{N-\tilde{y}}{y-\tilde{y}} \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\mu\right)^{y-\tilde{y}} (1-\mu)^{N-y} \\ &= \binom{N}{\tilde{y}} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)^{\tilde{y}} \left(1-\mu+\frac{\lambda_2}{\lambda_1+\lambda_2}\mu\right)^{N-\tilde{y}} \\ &= \binom{N}{\tilde{y}} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)^{\tilde{y}} \left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)^{N-\tilde{y}} \end{aligned}$$

as required. □

Corollary 5. *Let $N \geq 0$, $\tilde{\lambda}, \lambda \geq 0$ and $0 \leq \mu \leq 1$. For $k = 0, 1, \ldots, n$ let*

$$\varphi_k := D_{\mathrm{KL}}\left(\mathrm{Binomial}\left(N, \frac{\tilde{\lambda}+\lambda k}{\tilde{\lambda}+\lambda n}\mu\right) \middle\| \mathrm{Poisson}\left(\tilde{\lambda}+\lambda k\right)\right).$$

Then

$$\varphi_{k-1} + \varphi_{k+1} \geq 2\varphi_k$$

for all $k = 1, \ldots, n-1$, and

$$\varphi_k \leq \frac{n-k}{n}\varphi_0 + \frac{k}{n}\varphi_n$$

for all $k = 0, 1, \ldots, n$.

Proof. Let $S_\emptyset \sim \mathrm{Poisson}(\tilde{\lambda})$ and $X_1, \ldots, X_n \sim \mathrm{Poisson}(\lambda)$ be mutually independent random variables. Let $S_T := S_\emptyset + \sum_{i \in T} X_i$ for non-empty $T \subseteq [1:n]$. Note that $\{S_T\}_T$ forms a symmetric layered function family on $X_1, \ldots, X_n$. Also note that $S_T \sim \mathrm{Poisson}(\tilde{\lambda} + \lambda|T|)$ and $S_{[1:n]} - S_T \sim \mathrm{Poisson}(\lambda(n-|T|))$ are independent. Let $\tilde{S}_T$ be defined as in Corollary 1 (ii) (with $q(\cdot) \sim \mathrm{Binomial}(N, \mu)$). Applying Lemma 11, we have $\tilde{S}_T \sim \mathrm{Binomial}\left(N, \frac{\tilde{\lambda}+\lambda|T|}{\tilde{\lambda}+\lambda n}\mu\right)$. The result then follows from Corollary 1 (ii). □

Corollary 6. *For all $N \geq 0$ and $\lambda \geq 0$, the function*

$$t \mapsto D_{\mathrm{KL}}\left(\mathrm{Binomial}\left(N,t\right) \| \, \mathrm{Poisson}\left(\lambda t\right)\right)$$

is convex on $[0,1]$.

Proof. This is immediate from Corollary 5 (with $\tilde{\lambda} = 0$ and $\mu = 1$) and continuity. □

3 Conclusion and future work

One possible application of our main result is to discover possible connections between sumset inequalities in combinatorics and entropic inequalities in information theory. Sumset inequalities have been playing an important role in additive combinatorics, for instance [12, 28]. Several sumset inequalities have been shown to have entropic equivalents, and for some of these equivalent formulations, one can establish the combinatorial version from the entropic version and vice-versa.

Ruzsa has conjectured the sumset inequality (Conjecture 3.13 of [12]) that, if A_1, A_2, A_3, A_4 are finite subsets of some (possibly non-Abelian) group then

$$\max_{a_2 \in A_2,\, a_3 \in A_3} |A_1 \circ A_2 \circ A_3||A_1 \circ a_2 \circ A_3 \circ A_4||A_1 \circ A_2 \circ a_3 \circ A_4||A_2 \circ A_3 \circ A_4| \geq |A_1 \circ A_2 \circ A_3 \circ A_4|^3.$$

From our main result, however, the entropic analogue of this sumset inequality can be shown. Via an application of Theorem 1 (iii) (with $U := X_1 \circ X_2 \circ X_3 \circ X_4$ and $S_T := X_T$), we have that if X_1, X_2, X_3, X_4 are mutually independent random variables taking value in some (possibly non-Abelian) group then

$$\begin{aligned} &H(X_1 \circ X_2 \circ X_3) + H(X_1 \circ X_2 \circ X_3 \circ X_4 | X_2) + H(X_1 \circ X_2 \circ X_3 \circ X_4 | X_3) + H(X_2 \circ X_3 \circ X_4) \\ &\quad \geq 3H(X_1 \circ X_2 \circ X_3 \circ X_4). \end{aligned}$$

Note that as [12] dealt with dependent random variables, they were not able to establish an entropic inequality that mimicked the previous conjecture (see the paragraph after Conjecture 3.13 in [12]).

In general, the sumset inequality that for subsets $A_1, \ldots, A_n$ of some group,

$$\prod_{i=1}^{n} \max_{a_i \in A_i} |A_1 \circ \cdots \circ A_{i-1} \circ a_i \circ A_{i+1} \circ \cdots \circ A_n| \geq |A_1 \circ \cdots \circ A_n|^{n-1},$$

is known to be true for Abelian groups (Theorem 9.3, Chapter 1 of [29]). For non-Abelian groups it is known to be true for $n \leq 3$ (Corollary 3.12 of [12]) while the other cases remain open (Problem 9.4, Chapter 1 of [29]). On the other hand, the corresponding entropic inequality that for mutually independent random variables $X_1, \ldots, X_n$,

$$\sum_{i=1}^{n} H(X_1 \circ \cdots \circ X_n | X_i) \geq (n-1) H(X_1 \circ \cdots \circ X_n),$$

can be deduced from our main result for all n and (possibly non-Abelian) groups.

Acknowledgements

Chandra Nair wishes to thank Prof. Venkat Anantharam who brought to his attention the compression approach in [1]. The authors also wish to thank Qinghua Ding and Jinpei Zhao for interesting discussion related to this problem.

References

[1] P. Balister and B. Bollobás, "Projections, entropy and sumsets," *Combinatorica*, vol. 32, no. 2, pp. 125–141, Mar 2012. [Online]. Available: https://doi.org/10.1007/s00493-012-2453-1

[2] F. Chung, R. Graham, P. Frankl, and J. Shearer, "Some intersection theorems for ordered sets and graphs," *Journal of Combinatorial Theory, Series A*, vol. 43, no. 1, pp. 23–37, 1986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0097316586900191

[3] J. Radhakrishnan, “Entropy and counting,” *Computational mathematics, modelling and algorithms*, vol. 146, 2003.

[4] T. S. Han, “Nonnegative entropy measures of multivariate symmetric correlations,” *Information and Control*, vol. 36, no. 2, pp. 133–156, 1978. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019995878902759

[5] M. Madiman and P. Tetali, “Information inequalities for joint distributions, with interpretations and applications,” *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2699–2713, 2010.

[6] S. Kamath and C. Nair, “The strong data processing constant for sums of i.i.d. random variables,” in *Information Theory (ISIT), 2015 IEEE International Symposium on*, June 2015, pp. 2550–2552.

[7] T. A. Courtade, “Monotonicity of entropy and Fisher information: a quick proof via maximal correlation,” *Commun. Inf. Syst.*, vol. 16, no. 2, pp. 111–115, 2016. [Online]. Available: https://doi.org/10.4310/cis.2016.v16.n2.a2

[8] S. Artstein, K. Ball, F. Barthe, and A. Naor, “Solution of Shannon’s problem on the monotonicity of entropy,” *Journal of the American Mathematical Society*, vol. 17, no. 4, pp. 975–982, 2004.

[9] M. M. Madiman and A. R. Barron, “Generalized entropy power inequalities and monotonicity properties of information,” *IEEE Trans. Inf. Theory*, vol. 53, no. 7, pp. 2317–2329, 2007. [Online]. Available: https://doi.org/10.1109/TIT.2007.899484

[10] A. Stam, “Some inequalities satisfied by the quantities of information of Fisher and Shannon,” *Information and Control*, vol. 2, no. 2, pp. 101–112, 1959. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019995859903481

[11] A. Dembo, A. Kagan, and L. A. Shepp, “Remarks on the maximum correlation coefficient,” *Bernoulli*, vol. 7, no. 2, pp. 343–350, 04 2001. [Online]. Available: https://projecteuclid.org:443/euclid.bj/1080222081

[12] M. Madiman, A. Marcus, and P. Tetali, “Entropy and set cardinality inequalities for partition-determined functions,” *Random Structures & Algorithms*, vol. 40, no. 4, pp. 399–424, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20385

[13] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, “On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover,” *CoRR*, vol. abs/1304.6133, 2013.

[14] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, “On hypercontractivity and a data processing inequality,” in *2014 IEEE International Symposium on Information Theory (ISIT’2014)*, Honolulu, USA, Jun. 2014, pp. 3022–3026.

[15] C. E. Shannon, “A mathermatical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October, 1948.

[16] E. H. Lieb, “Proof of an entropy conjecture of wehrl,” *Communications in Mathematical Physics*, vol. 62, no. 1, pp. 35–41, 1978. [Online]. Available: https://doi.org/10.1007/BF01940328

[17] D. Guo, S. Shamai, and S. Verdu, “Proof of entropy power inequalities via mmse,” in *2006 IEEE International Symposium on Information Theory*, 2006, pp. 1011–1015.

[18] O. Rioul, “Information theoretic proofs of entropy power inequalities,” *Information Theory, IEEE Transactions on*, vol. 57, no. 1, pp. 33–55, Jan 2011.

[19] T. A. Courtade, “Strengthening the entropy power inequality,” *CoRR*, vol. abs/1602.03033, 2016. [Online]. Available: http://arxiv.org/abs/1602.03033

[20] M. Madiman and F. Ghassemi, “The entropy power of a sum is fractionally superadditive,” in *2009 IEEE International Symposium on Information Theory*, 2009, pp. 295–298.

[21] O. Johnson, “Maximal correlation and the rate of Fisher information convergence in the central limit theorem,” *CoRR*, vol. abs/1905.11913, 2019. [Online]. Available: http://arxiv.org/abs/1905.11913

[22] A. R. Barron, “Entropy and the central limit theorem,” *The Annals of Probability*, vol. 14, no. 1, pp. 336–342, 1986. [Online]. Available: http://www.jstor.org/stable/2244098

[23] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, ser. CMS Books in Mathematics. Springer New York, 2005. [Online]. Available: https://books.google.com.hk/books?id=TXWzqEkAa7IC

[24] O. Hirschfeld, "A connection between correlation and contingency," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, pp. 520–524, 1935.

[25] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwert-problem und sein zusammenhang mit der ausgleichungsrechnung," *Zeitschrift für angew. Math. und Mech.*, vol. 21, pp. 364–379, 1941.

[26] A. Rényi, "On measures of dependence," *Acta. Math. Acad. Sci. Hung.*, vol. 10, pp. 441–451, 1959.

[27] Y. Yu, "Monotonic convergence in an information-theoretic law of small numbers," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5412–5422, 2009.

[28] T. Tao, "Sumset and inverse sumset theory for shannon entropy," *Combinatorics, Probability and Computing*, vol. 19, no. 4, pp. 603–639, jan 2010. [Online]. Available: https://doi.org/10.1017/s0963548309990642

[29] I. Ruzsa, "Sumsets and structure," *Combinatorial Number Theory and Additive Group Theory*, 2009.