# 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 in [11], strong data-processing constants in [6], and obtain new relative entropy convexity results.

### 1.1 Main

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 $X_1, \ldots, X_n$ and for $T \subseteq [1:n]$, we write $X_T := \{X_i\}_{i\in T}$, the tuple consisting of $X_i$ where $i \in T$.

**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 $0 \leq \delta \leq \min\{\alpha_A, \alpha_B\}$ 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 $\{\alpha_T\}_T$.

**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.

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

**Definition 3.** Let $\{S_T\}_T$ be a layered function family on mutually independent and identically distributed random variables $X_1, \dots, X_n$. We call the layered function family $\{S_T\}_T$ *symmetric* if for all permutations $\pi$ of $[1:n]$ the distributions of $(S_{[1:n]}, S_\emptyset, X_1, \dots, X_n)$ and $(S_{[1:n]}, S_\emptyset, X_{\pi(1)}, \dots, X_{\pi(n)})$ are the same.

*Remark* 2. If $X_1, \dots, X_n$ are mutually independent and identically distributed random variables, Remark 1 (i) and (ii) are examples of symmetric layered function families.

**Lemma 1.** *Let $\{S_T\}_T$ be a layered function family on mutually independent random variables $X_1, \dots, X_n$. 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

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

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,

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

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. As the proof below shows (and similar to the case in [1]), the main ingredient is an elementary two-point inequality shown in part (i) below.

**Theorem 1.** *Let $\{S_T\}_T$ be a layered function family on mutually independent random variables $X_1, \dots, X_n$. 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) $\sum_{T\subseteq[1:n]} \alpha_T I(U; S_T) \leq \sum_{T\subseteq[1:n]} \beta_T I(U; S_T)$, where $\alpha_T, \beta_T$ ($T \subseteq [1:n]$) are non-negative real numbers such that $\{\beta_T\}_T$ is a compression of $\{\alpha_T\}_T$.*

*(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 $\beta_T$ ($T \subseteq [1:n]$) are non-negative real numbers satisfying $\sum_{T\subseteq[1:n]:T\ni i} \beta_T \leq 1$ for all $i = 1, \dots, n$, and $c := \sum_{T\subseteq[1:n]} \beta_T$.*

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

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

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

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

which, together with part (ii) of Lemma 1, gives (i). Note that (ii) is immediate from (i) as a compression is obtained as a sequence of elementary compressions.

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

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

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

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

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

## 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 2.** *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 $\mathrm{E}[U^{(\epsilon)}{U^{(\epsilon)}}^T] = I$. 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

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

□

**Lemma 3.** *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 absolutely continuous 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 [12, 13] and references therein.

# 2 Some consequences of Theorem 1

## 2.1 Generalized Stam Inequality

In this subsection, we show a generalized Stam inequality involving Fisher information as an immediate consequence of our mutual information inequality. The results established in this section are not new, and a similar proof technique we employ has been used by Courtade in [7] for the case of mutually independent and identically distributed random variables. However, as noted in [14] (Future work, item 4), the extension of the ideas to independent random variables had been of interest. The proof in this section does the extension to independent random variables.

**Definition 4.** Let $X$ be a random variable in $\mathbb{R}^d$ with density $f_X$. The *score function* $\rho_X$ 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* 4. 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 [15] (cf. Lemma 3 of [9]).

*Remark* 5. The Fisher information of sum of independent random variables satisfies a certain property that is first observed by Stam [10]: If $X_1, X_2$ are independent random variables in $\mathbb{R}^d$ with densities $f_1, f_2$, respectively, then

$$\begin{aligned}
\rho_{X_1+X_2}(y) &= \frac{\nabla(f_1 * f_2)(y)}{(f_1 * f_2)(y)} \\
&= \frac{(\nabla f_1 * f_2)(y)}{(f_1 * f_2)(y)} \\
&= \frac{(\rho_{X_1} f_1 * f_2)(y)}{(f_1 * f_2)(y)} \\
&= \frac{\int \rho_{X_1}(x_1) f_1(x_1) f_2(y - x_1)\, dx_1}{\int f_1(x_1) f_2(y - x_1)\, dx_1} \\
&= \mathrm{E}[\rho_{X_1}(X_1) | X_1 + X_2 = y],
\end{aligned}$$

and hence

$$\rho_{X_1+X_2}(X_1 + X_2) = \mathrm{E}[\rho_{X_1}(X_1) | X_1 + X_2].$$

In general, suppose $X_1, \ldots, X_n$ 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 $k = 1, \ldots, n$.

**Lemma 4.** *Let $X_1, \ldots, X_n$ be mutually independent random variables in $\mathbb{R}^d$ with densities. For $k = 1, \ldots, n$ 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 $k = 1, \ldots, n$.*

*Proof.* Consider

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

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

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

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

*Proof.* Note that $S_\emptyset = 0$. An application of Lemma 2 (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

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

Now consider

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

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 4. The result then follows from rearranging. □

*Remark* 6. Proposition 1 implies the fractional superadditivity of entropy power [16] (see also [9]). On the other hand, one immediately obtains the Artstein–Ball–Barthe–Naor inequality [8]: If $X_1, \dots, X_n$ are mutually independent and identically distributed random variables in $\mathbb{R}^d$ with densities, then for all $k = 1, \dots, n$,

(i) $J\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) \le J\left(\frac{X_1+\cdots+X_k}{\sqrt{k}}\right)$;

(ii) $h\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) \ge h\left(\frac{X_1+\cdots+X_k}{\sqrt{k}}\right)$, if $X_i$'s have a finite covariance matrix;

where (i) follows from setting $\beta_T = 0$ for $|T| \ne k$ in Proposition 1, and (ii) is a consequence of (i) and Remark 4 (iii).

## 2.2 Discrete Convexity, Strong Data Processing 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. The results in this section generalize the known results in [6] and [11].

**Lemma 5** (Discrete convexity). *Suppose $\varphi_k$ $(k = 0, 1, \ldots, n)$ 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 $\pi$ 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 $\left(U, S_\emptyset, X_{\pi(1)}, \ldots, X_{\pi(n)}\right)$ are the same. In particular, the distributions of $\left(U, S_\emptyset, X_{[1:|T|]}\right)$ and $(U, S_\emptyset, X_T)$ are the same. Hence Lemma 1 (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

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

where (a) follows from (i) of Theorem 1. Hence (3) is satisfied. Then an application of Lemma 5 (with $l = 0$) yields

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

or equivalently,

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

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 $X_1, \dots, X_n$. 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*

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

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

*(ii) Suppose $q(\cdot)$ is a distribution absolutely continuous 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*

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

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

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

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

*Proof.* (i) and (ii) are direct applications of Lemma 2 and 3, 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 $\pi$ 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* 7. 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*

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

*Furthermore,*

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

*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 $(U, S, X_1, X_2, \ldots, X_n)$ and $(U, S, X_n, X_1, \ldots, X_{n-1})$ are the same. Consequently, for all $k = 0, \ldots, n-1$ we have $I(U;X_{[1:k+1]}) = I(U;X_{[1:k]\cup\{n\}})$. Hence for $k = 1, \ldots, n-1$,

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

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 5 (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 $X_1, \ldots, X_n$. 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* 8. 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 [17] and Gebelein [18] and then studied by Rényi [19].

**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 [19] as follows.

**Proposition 4** (Rényi [19])**.** *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 $X_1, \dots, X_n$. Then*

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

*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

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

Taking supremum over $f$ yields the result. □

### 2.2.3 KL divergence inequality

**Lemma 6.** *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*

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

*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 $S_\emptyset \sim \text{Poisson}(\tilde{\lambda})$ and $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ 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 $S_T \sim \text{Poisson}(\tilde{\lambda} + \lambda|T|)$ and $S_{[1:n]} - S_T \sim \text{Poisson}(\lambda(n-|T|))$ are independent. Let $\tilde{S}_T$ be defined as in Corollary 1 (ii) (with $q(\cdot) \sim \text{Binomial}(N, \mu)$). Applying Lemma 6, we have $\tilde{S}_T \sim \text{Binomial}\left(N, \frac{\tilde{\lambda}+\lambda|T|}{\tilde{\lambda}+\lambda n}\mu\right)$. The result then follows from Corollary 1 (ii). □

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

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

*is convex on $[0, 1]$.*

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

*Remark* 9. The above result has a similar flavor to the open problem listed in https://archive.siam.org/journals/categories/09-001.php where the author makes a complete monotonicity conjecture between binomial and Poisson distributions, but is unable to even prove the convexity. An interested reader may also see Conjecture 1 in [20].

## 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 Zhao Jinpei 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] 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.

[13] 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.

[14] 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

[15] 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

[16] 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.

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

[18] 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.

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

[20] 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.