The strong data processing constant for sums of i.i.d. random variables

Sudeep Kamath
ECE Department,
Princeton University
Princeton, NJ USA
sukamath@princeton.edu

Chandra Nair
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, NT, Hong Kong S.A.R
chandra@ie.cuhk.edu.hk

***Abstract*—We obtain the strong data processing constant for sums of real-valued i.i.d. random variables by means of a very simple information-theoretic proof. As a corollary, we recover a classical result concerning the maximal correlation between sums of a sequence of real-valued i.i.d. random variables.**

***Index Terms*—data-processing inequality**

I. Introduction

Data processing inequalities are fundamental tools that are often used in information theoretic arguments. The classical data processing inequality for mutual information states that if three random variables $U - X - Y$ form a Markov chain, then

$$I(U;Y) \leq I(U;X).$$

A finer question would be to determine the best constant $\mathsf{s}^*(X;Y)$ that depends on the joint law of (X,Y), such that the inequality

$$I(U;Y) \leq \mathsf{s}^*(X;Y) I(U;X),$$

holds for every U such that $U - X - Y$ is Markov. We define $\mathsf{s}^*(X;Y)$ to be *the strong data processing constant* associated with the pair (X,Y).

Remarks: Note that $\mathsf{s}^*(X;Y)$ is not in general symmetric in X and Y. The readers may refer to [1] for some relevant background related to $\mathsf{s}^*(X;Y)$. In particular, computation of $\mathsf{s}^*(X;Y)$ is not an easy task, and an explicit formula for $\mathsf{s}^*(X;Y)$ is known only for some special classes of joint distributions, see [2] and references therein.

The strong data processing constant has important operational meanings in the efficiencies of investment information [3], and common randomness generation [4]. It has also been used to provide outer bounds on multiterminal source coding problems [5].

In this paper, we use information theoretic ideas and tools to obtain the strong data processing constant for sums of real-valued independent and identically distributed (i.i.d.) random variables by means of a very simple proof. As a consequence of this result, we recover two classical results in statistics: the maximal correlation of sums of real-valued i.i.d. random variables [6] and the maximal correlation of jointly Gaussian random variables [7], [8].

All the random variables considered in this paper are real valued and Borel measurable.

II. Maximal Correlation and Strong Data Processing Constant

Given random variables (X,Y), the following two notions of correlation between them are considered in this paper.

The Hirschfeld-Gebelein-Rényi maximal correlation [7], [9], [10] between X and Y is defined as:

$$\rho_{\mathsf{m}}(X;Y) = \sup \mathbb{E} f(X) g(Y), \tag{1}$$

where the supremum is taken over the set of measurable functions $f : \mathcal{X} \to \mathbb{R}, g : \mathcal{Y} \to \mathbb{R}$ satisfying $\mathbb{E}f(X) = \mathbb{E}g(Y) = 0, \mathbb{E}f(X)^2 \leq 1, \mathbb{E}g(Y)^2 \leq 1$.

The strong data processing constant corresponding to X and Y is defined as [3], [11]:

$$\mathsf{s}^*(X;Y) = \sup_{\substack{U: U-X-Y \\ 0<I(U;X)<\infty}} \frac{I(U;Y)}{I(U;X)} \ . \tag{2}$$

Note that $\rho_{\mathsf{m}}(X;Y)$ is symmetric in its arguments whereas $\mathsf{s}^*(X;Y)$ is in general not symmetric. There are several equivalent characterizations of $\mathsf{s}^*(X;Y)$ presented in [1]. In particular, [1] shows that it suffices to restrict U to be binary valued when computing the supremum in (2), thus we may assume U to be real-valued and Borel measurable without loss of generality.

The following relation was established in [11] between one of the equivalent definitions of $\mathsf{s}^*(X;Y)$ and $\rho^2_{\mathsf{m}}(X;Y)$.

Theorem 1. *[11] For any pair of random variables (X,Y), we have*

$$\mathsf{s}^*(X;Y) \geq \rho^2_{\mathsf{m}}(X;Y). \tag{3}$$

III. Main Result

Let $X_1, X_2, \ldots,$ be i.i.d. copies of a non-constant random variable X. Let $S_j := X_1 + X_2 + \ldots + X_j$. If $\mathbb{E}X^2 < \infty$, we get

$$\rho_{\mathsf{m}}^2(S_n; S_m) \geq \frac{m}{n}, \quad \text{for } 1 \leq m \leq n, \tag{4}$$

by considering the standard Pearson correlation between S_n and S_m. Indeed, by a clever argument involving characteristic functions, [12] showed that (4) holds even when $\mathbb{E}X^2 = \infty$.

The following theorem is the main result of this paper.

Theorem 2. *For $1 \leq m \leq n$, and for any U such that $U - S_n - S_m$ is Markov,*

$$I(U; S_m) \leq \frac{m}{n} I(U; S_n). \tag{5}$$

Proof. Let U be any random variable satisfying the Markov chain $U - S_n - S_m$ with $I(U; S_n) < \infty$. Further let $p_{U,S_n,S_m}(u, s_n, s_m) = p_{U|S_n}(u|s_n) q_{S_m|S_n}(s_m|s_n) q_{S_n}(s_n)$.

Observe that the distributions q_{S_n} and $q_{S_m|S_n}$ are fixed by the distribution of X_i and the independence of $(X_1, ..., X_n)$. Now consider $(U', X_1', ..., X_n')$ distributed according to $p_{U|S_n}(u'|s_n') q_{X_1,..,X_n|S_n}(x_1', ..., x_n'|s_n') q_{S_n}(s_n')$. As before, the distribution $q_{X_1,..,X_n|S_n}$ is fixed by the distribution of X_i and the independence of $(X_1, ..., X_n)$. Thus we induce a Markov chain $U' - S_n' - (X_1', ..., X_n')$, while noting that (U, S_n, S_m) has the same distribution as (U', S_n', S_m'). Since we are interested only in $I(U; S_n)$ and $I(U; S_m)$ we may as well replace them by $I(U'; S_n')$ and $I(U'; S_m')$.

Remark: In the rest of the paper, we will drop the primes on the random variables, and assume that we have $U - S_n - (X_1, ..., X_n)$ to be Markov.

For any subset $\mathcal{A} \subseteq \{1, 2, \ldots, n\}$, let $X_{\mathcal{A}}$ denote $\{X_i : i \in \mathcal{A}\}$ and $S_{\mathcal{A}} := \sum_{i \in \mathcal{A}} X_i$. Since $X_1, X_2, \ldots, X_n$ is an i.i.d. sequence and the Markov chain $U - S_n - (X_1, ..., X_n)$ holds, the distribution of $(U, X_{\mathcal{A}})$ depends only on the size of the set $\mathcal{A}$. So, we may define

$$\Phi(i) := I(U; X_{\mathcal{A}}), \text{ for } |\mathcal{A}| = i.$$

Note that for any set $\mathcal{A}$, we have

$$\begin{aligned} I(U; S_n) &\overset{(a)}{=} I(U; S_n, S_{\mathcal{A}}, S_{\mathcal{A}^c}, X_{\mathcal{A}}) \\ &= I(U; S_{\mathcal{A}}, S_{\mathcal{A}^c}, S_n) + I(U; X_{\mathcal{A}} | S_{\mathcal{A}}, S_{\mathcal{A}^c}, S_n) \\ &\overset{(b)}{\geq} I(U; S_n) + I(U; X_{\mathcal{A}} | S_{\mathcal{A}}, S_{\mathcal{A}^c}). \end{aligned}$$

Here (a) follows from the Markov chain $U - S_n - (X_1, ..., X_n)$, (b) follows since $S_n = S_{\mathcal{A}} + S_{\mathcal{A}^c}$.

The last inequality gives

$$\begin{aligned} 0 &\geq I(U; X_{\mathcal{A}} | S_{\mathcal{A}}, S_{\mathcal{A}^c}) \\ &\overset{(c)}{=} I(U, S_{\mathcal{A}^c}; X_{\mathcal{A}} | S_{\mathcal{A}}) \\ &\geq I(U; X_{\mathcal{A}} | S_{\mathcal{A}}) \geq 0, \end{aligned}$$

where (c) uses the independence of $(S_{\mathcal{A}}, X_{\mathcal{A}})$ and $S_{\mathcal{A}^c}$. Thus, we have that $U - S_{\mathcal{A}} - X_{\mathcal{A}}$ is Markov and so, $\Phi(i) = I(U; S_i)$.

Proving $I(U; S_m) \leq \frac{m}{n} I(U; S_n)$ for any $1 \leq m \leq n$ is equivalent to showing that $\Phi(m) \leq \frac{m}{n} \Phi(n)$, or that $\frac{\Phi(i)}{i}$ is non-decreasing over $1 \leq i \leq n$. We will be done if we show that for $1 \leq i \leq n-1$, we have

$$\Phi(i+1) - \Phi(i) \geq \Phi(i) - \Phi(i-1). \tag{6}$$

To see why (6) suffices, note the condition is same as convexity in discrete time. Or alternatively, by induction if $\frac{\Phi(i)}{i} \geq \frac{\Phi(i-1)}{i-1}$ then

$$\begin{aligned} \Phi(i+1) - \Phi(i) &\geq \Phi(i) - \Phi(i-1) \\ &\geq \Phi(i) - \frac{i-1}{i} \Phi(i) = \frac{1}{i} \Phi(i). \end{aligned}$$

Thus $\frac{\Phi(i+1)}{i+1} \geq \frac{\Phi(i)}{i}$. The base case $\frac{\Phi(2)}{2} \geq \Phi(1)$ is immediate from (6) and $\Phi(0) = 0$.

To establish (6) observe that

$$\begin{aligned} &\Phi(i+1) - 2\Phi(i) + \Phi(i-1) \\ &\overset{(d)}{=} I(U; X_1, \ldots, X_{i+1}) - I(U; X_1, \ldots, X_{i-1}, X_{i+1}) \\ &\quad - I(U; X_1, \ldots, X_i) + I(U; X_1, \ldots, X_{i-1}) \\ &= I(U; X_i | X_1, \ldots, X_{i-1}, X_{i+1}) \\ &\quad - I(U; X_i | X_1, \ldots, X_{i-1}) \\ &\overset{(e)}{=} I(U, X_1, \ldots, X_{i-1}, X_{i+1}; X_i) \\ &\quad - I(U, X_1, \ldots, X_{i-1}; X_i) \\ &= I(X_{i+1}; X_i | U, X_1, \ldots, X_{i-1}) \\ &\geq 0, \end{aligned}$$

where (d) follows from the observation that $(U, X_{\mathcal{A}})$ has the same distribution for all sets $\mathcal{A}$ of the same size, and (e) follows from independence of the sequence $X_1, X_2, \ldots, X_n$. This completes the proof. □

Corollary 1. *For $1 \leq m \leq n$, we have*

$$\mathsf{s}^*(S_n; S_m) = \frac{m}{n}.$$

Proof. The corollary is an immediate consequence of Theorem 2, Theorem 1, and the lower bound on maximal correlation established in (4). □

Remark 1. We noted earlier that s^* is not in general symmetric in its arguments. It can be shown in contrast to Corollary 1 that $\mathsf{s}^*(S_m; S_n) > \frac{m}{n}$ in general by choosing, for example, $X \sim \text{Ber}(\epsilon), \epsilon \neq 0, \frac{1}{2}, 1,$ and

$m = 1, n = 2$. This follows from a simple observation regarding $\mathsf{s}^*(X;Y)$ when X is binary valued, i.e. $\mathsf{s}^*(X;Y) = \rho_m^2(X;Y) = \lambda$ at a particular input distribution μ_X and channel $W(y|x)$, only if the function $H(Y) - \lambda H(X)$ is a convex function of the input distribution (for the fixed channel law $W(y|x)$). The details are left as an exercise to the interested reader. The argument uses an alternate characterization of $\mathsf{s}^*(X;Y)$ using convex envelopes that can be found in [1].

As a corollary of our main result, along with Theorem 1 and the lower bound (4), we obtain the following classical result that characterizes the maximal correlation of sums of i.i.d random variables. The proof of Thm. 3 in [6] uses the Efron-Stein decomposition [13] of symmetric functions.

Theorem 3. *(Dembo-Kagan-Shepp Theorem [6])*
For $1 \leq m \leq n$,

$$\rho_{\mathsf{m}}^2(S_n; S_m) = \frac{m}{n}. \tag{7}$$

Another corollary of our main result is the well known fact [7], [8] that if (W, Z) are joint Gaussian with α as the usual correlation coefficient, then

$$\mathsf{s}^*(W;Z) = \rho_{\mathsf{m}}^2(W;Z) = \alpha^2.$$

This can be shown as follows. First, the Pearson correlation lower bound on maximal correlation and Theorem 1 give

$$\mathsf{s}^*(W;Z) \geq \rho_{\mathsf{m}}^2(W;Z) \geq \alpha^2.$$

From our main result with choosing X to be Gaussian with zero mean and variance 1, so that

$$\mathsf{s}^*\left(\frac{S_n}{\sqrt{n}}; \frac{S_m}{\sqrt{m}}\right) = \mathsf{s}^*(S_n; S_m) = \frac{m}{n},$$

which proves the desired result for all positive α such that α^2 is rational. The simple observation that whenever $A - B - C - D$ is Markov, we have

$$\mathsf{s}^*(A;D) \leq \mathsf{s}^*(B;C),$$

which implies that $\mathsf{s}^*(W;Z)$ as a function of α^2 is monotonically increasing. This extends the proof to all real $\alpha \in [-1, 1]$ (by a sandwich argument), since rationals form a dense set in the reals.

Remark 2. The proof of Theorem 2 goes through even if some of the assumptions are suitably relaxed: for instance, $X_1, X_2, \ldots, X_n$ may take values in a finite Abelian group. Note however that in this case, the lower bound (4) may not hold, so Theorem 2 does not help provide a complete characterization of $\mathsf{s}^*(S_n; S_m)$. For example, suppose X_1, X_2 lie in the finite field $\mathbb{F}_p$ for some prime p, and are equiprobable and independent. Then, $S_1 = X_1$ and $S_2 = X_1 + X_2$ are independent, so that $\mathsf{s}^*(S_2; S_1) = 0$.

IV. Conclusion

We computed the strong data processing constant for sums of real-valued i.i.d. random variables using elementary information theoretic tools. On the other hand, this result implies, as a corollary, a non-trivial classical result the computes the maximal correlation between sums of real-valued i.i.d. random variables.

Acknowledgements

Sudeep Kamath would like to acknowledge support from the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

Chandra Nair would like to acknowledge support from the following grants: GRF 2150743, 2150785, 2150829 and the area of excellence grant AoE/E-02/08.

References

[1] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and a data processing inequality," in *Proc. of International Symposium on Information Theory*, Honolulu, Hawaii, USA, July 2014.

[2] ——, "On hypercontractivity and the mutual information between boolean functions," in *51st Annual Allerton Conference on Communication, Control, and Computing*, 2013.

[3] E. Erkip and T. Cover, "The efficiency of investment information," *IEEE Transactions On Information Theory*, vol. 44, pp. 1026–1040, May 1998.

[4] L. Zhao and Y.-K. Chia, "The efficiency of common randomness generation," in *49th Annual Allerton Conference on Communication, Control, and Computing*, 2011.

[5] T. Courtade, "Outer bounds for multiterminal source coding via a strong data processing inequality," in *Proc. of International Symposium on Information Theory*, Istanbul, Turkey, July 2013.

[6] A. Dembo, A. Kagan, and L. Shepp, "Remarks on the maximum correlation coefficient," *Bernoulli*, vol. 7, pp. 343–350, 2001.

[7] 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.

[8] H. Lancaster, "Some properties of the bivariate normal distribution considered in the form of a contingency table," *Biometrika*, vol. 44, pp. 289–292, 1957.

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

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

[11] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *Annals of Probability*, vol. 4, pp. 925–939, 1976.

[12] W. Bryc, A. Dembo, and A. Kagan, "On the maximum correlation coefficient," *Theory Probab. Appl.*, vol. 49, pp. 132–138, 2005.

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