Monotonic Decrease of the Non-Gaussianness of the Sum of Independent Random Variables: A Simple Proof

Antonia M. Tulino, *Senior Member, IEEE*, and
Sergio Verdú, *Fellow, IEEE*

Abstract—Artstein, Ball, Barthe, and Naor have recently shown that the non-Gaussianness (divergence with respect to a Gaussian random variable with identical first and second moments) of the sum of independent and identically distributed (i.i.d.) random variables is monotonically nonincreasing. We give a simplified proof using the relationship between non-Gaussianness and minimum mean-square error (MMSE) in Gaussian channels. As Artstein *et al.*, we also deal with the more general setting of nonidentically distributed random variables.

Index Terms—Central limit theorem, differential entropy, divergence, entropy power inequality, minimum mean-square error (MMSE), non-Gaussianness, relative entropy.

I. INTRODUCTION

Versions of the central limit theorem for continuous random variables have been shown in the sense that the divergence between an n-fold convolution and the Gaussian density with identical first and second moments (non-Gaussianness) vanishes as $n \to \infty$ ([1] and references therein). Although long suspected that the non-Gaussianness decreases at each convolution, it was not shown until 2004 (in the equivalent version of increasing differential entropy) by Artstein, Ball, Barthe, and Naor [2] by means of a tour-de-force in functional analysis. They showed the result for independent and identically distributed (i.i.d.) random variables, as well as generalizations of the result where nonidentical distributions are allowed. As a consequence, [2] shows a new entropy power inequality: For independent (not necessarily i.i.d.) random variables,

$$n \exp(2h(X_1 + \ldots + X_{n+1})) \geq \sum_{\ell=1}^{n+1} \exp\left(2h\left(\sum_{j \neq \ell} X_j\right)\right) \quad (1)$$

where $h(X) = -\int f_X(x) \log f_X(x) \mathrm{d}x$ denotes the differential entropy of the density f_X. Note that (1) implies, but is not implied by, Shannon's entropy power inequality [3]

$$\exp\left(2h(X_1 + \ldots + X_n)\right) \geq \sum_{i=1}^{n} \exp\left(2h(X_i)\right). \quad (2)$$

The simplest proof of (2), given in [4], hinges on an elementary estimation argument: it is better to observe two noisy measurements than just their sum. Unlike the standard proof of (2) in which Fisher's information takes center stage, [4] works exclusively with minimum mean-square error (MMSE) using the results in [5]. Among them, the result that is relevant to this correspondence is the following.

Lemma 1 ([5, Theorem 14]): For every random variable X with $\sigma_X^2 < \infty$, its non-Gaussianness[1] is given by

$$D(X) \triangleq \mathsf{D}(P_X \| \mathcal{N}(\mathsf{E}X, \sigma_X^2)) = \frac{1}{2} \log 2\pi e \sigma_X^2 - h(X)$$

$$= \frac{1}{2} \int_0^\infty \frac{\sigma_X^2}{1 + \gamma\sigma_X^2} - \mathsf{mmse}(X, \gamma) \, \mathrm{d}\gamma \quad (3)$$

where

$$\mathsf{mmse}(X, \gamma) \triangleq \mathsf{E}[(X - \mathsf{E}[X \mid \sqrt{\gamma}X + N])^2] \quad (4)$$

and N is a unit-variance Gaussian random variable.

Thus, not only is the Gaussian the hardest random variable to estimate (in MMSE sense) at any signal-to-noise ratio but if $\mathsf{mmse}(X, \gamma) \geq \mathsf{mmse}(Y, \gamma)$ for every $\gamma > 0$ then $D(X) \leq D(Y)$.

II. RESULT

Theorem 1 Let $X_1, X_2, \ldots X_n$ be independent random variables with variances $\sigma_i^2, i = 1, \ldots, n$

$$D\left(\sum_{i=1}^{n} X_i\right) \leq \sum_{i=1}^{n} \lambda_i D\left(\sum_{\substack{j=1 \\ j \neq i}}^{n} X_j\right) + \frac{1}{2} D(\underline{\lambda} \| \underline{\alpha}) \quad (5)$$

where $\underline{\lambda} = \{\lambda_i\}$ is an arbitrary probability mass function on $\{1, \ldots n\}$, and the probability mass function $\underline{\alpha} = \{\alpha_i\}$ is defined by

$$(n - 1)\alpha_i = 1 - \frac{\sigma_i^2}{\sigma^2} \quad (6)$$

with

$$\sigma^2 = \sigma_1^2 + \ldots + \sigma_n^2 \quad (7)$$

Several particular cases of Theorem 1 are of interest.

1) **Theorem 2** If $X_1, X_2, \ldots X_n$ are i.i.d. random variables, then

$$D\left(\sum_{i=1}^{n} X_i\right) \leq D\left(\sum_{i=1}^{n-1} X_i\right). \quad (8)$$

2) Particularizing to $\underline{\lambda} = \underline{\alpha}$, we see that on average, non-Gaussianness increases by omitting one of the random variables from the sum, where the weight depends on the relative variance of the omitted random variable.

3) The main result of [2] is as follows.

Theorem 3 [2, Theorem 2] Let $V_1, V_2, \ldots V_n$ be independent random variables, and let $\boldsymbol{a} = (a_1, \ldots, a_n)$ with $\|\boldsymbol{a}\| = 1$. Then

$$h\left(\sum_{i=1}^{n} a_i V_i\right) \geq \sum_{i=1}^{n} \frac{1 - a_i^2}{n - 1} h\left(\sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{a_j V_j}{\sqrt{1 - a_i^2}}\right) \quad (9)$$

[1]For convenience, all logarithms are natural; thus, we measure non-Gaussianness in nats. We use divergence both for continuous distributions

$$D(f \| g) = \int \log \frac{f(x)}{g(x)} f(x) \mathrm{d}x$$

as well as discrete distributions

$$D(\underline{\lambda} \| \underline{\alpha}) = \sum_{i=1}^{n} \lambda_i \log \frac{\lambda_i}{\alpha_i}.$$

It can be checked that Theorems 3 and 1 are equivalent by identifying $X_i = a_i V_i, (n-1)\lambda_i = 1 - a_i^2$, and using the fact that $D(aX)$ is invariant to nonzero a. Note that in the case where all the random variables are Gaussian, Theorem 3 is tight if and only if $V_1, V_2, \ldots V_n$ have equal variance.

4) The special case of Theorem 1 for two random variables $X_1 \cos\alpha$ and $X_2 \sin\alpha$ with $\lambda_1 = \sin^2\alpha$ yields Lieb's inequality which is equivalent to Shannon's entropy power inequality (cf. [4, Lemma 1]).

It is easy to generalize Theorem 1 to consider sums of subsets of $\{X_1, \ldots X_n\}$ of smaller cardinality than $n-1$. In such case, α_i becomes a normalized version of the sum of the variances of the random variables in the ith such subset.

III. PROOF OF THEOREM 1

Proof: Without loss of generality we assume all random variables to have zero mean. For convenience we use the following notation:

$$Y_i = \sqrt{\gamma} \sum_{i=1}^n X_i + N_i \qquad (10)$$

$$Y = \sqrt{\gamma} \sum_{i=1}^n X_i + N \qquad (11)$$

$$X_{\backslash i} = \sum_{\substack{j=1 \\ j \neq i}}^n X_j \qquad (12)$$

$$Y_{\backslash i} = \sqrt{\gamma} X_{\backslash i} + N_{\backslash i} \qquad (13)$$

- $\{N_i\}$ are independent Gaussian random variables with variance $1 - \beta_i$, with

$$\beta_i = \lambda_i(n-1).$$

- $N = \sum_{i=1}^n N_i$ is Gaussian with unit variance.
- $N_{\backslash i} = N - N_i$ is Gaussian with variance β_i.

Note for future use that

$$\sum_{i=1}^n X_i = \frac{1}{n-1} \sum_{i=1}^n X_{\backslash i} \qquad (14)$$

$$Y = \frac{1}{n-1} \sum_{i=1}^n Y_{\backslash i} \qquad (15)$$

and

$$\text{var}\left\{\beta_i^{-\frac{1}{2}} X_{\backslash i}\right\} = \frac{1}{\beta_i} \sum_{\substack{j=1 \\ j \neq i}}^n \sigma_j \qquad (16)$$

$$= \frac{\sigma^2 - \sigma_i^2}{\lambda_i(n-1)} \qquad (17)$$

$$= \frac{\alpha_i \sigma^2}{\lambda_i}. \qquad (18)$$

Making use of Lemma 1 and the invariance of $D(aX)$ to nonzero a, we can write

$$\sum_{i=1}^n \lambda_i D\left(\sum_{\substack{j=1 \\ j \neq i}}^n X_j\right) - D\left(\sum_{i=1}^n X_i\right) = \frac{A}{2} + \frac{B}{2} \qquad (19)$$

where

$$A = \sum_{i=1}^n \lambda_i \int_0^\infty \frac{\text{var}\left\{\beta_i^{-\frac{1}{2}} X_{\backslash i}\right\}}{1 + \gamma \text{var}\left\{\beta_i^{-\frac{1}{2}} X_{\backslash i}\right\}} - \frac{\sigma^2}{1 + \sigma^2 \gamma} d\gamma \qquad (20)$$

$$= \sum_{i=1}^n \lambda_i \log\left(\frac{\alpha_i}{\lambda_i}\right) \qquad (21)$$

$$= -D(\underline{\lambda} \| \underline{\alpha}) \qquad (22)$$

and

$$B = \int_0^\infty \text{mmse}\left(\sum_{i=1}^n X_i, \gamma\right) - \sum_{i=1}^n \lambda_i \text{mmse}\left(\beta_i^{-\frac{1}{2}} X_{\backslash i}, \gamma\right) d\gamma \qquad (23)$$

where in (21) we used

$$\log\frac{a}{b} = \int_0^\infty \frac{a}{1 + a\gamma} - \frac{b}{1 + b\gamma} d\gamma. \qquad (24)$$

Thus, (5) follows from the property of MMSE in Lemma 2 below. $\square$

Lemma 2: Let $X_1, \ldots X_n$ be independent. For any $\gamma > 0$ and probability mass function $\underline{\lambda}$

$$\text{mmse}\left(\sum_{i=1}^n X_i, \gamma\right) \geq \sum_{i=1}^n \lambda_i \text{mmse}\left(\beta_i^{-\frac{1}{2}} X_{\backslash i}, \gamma\right) \qquad (25)$$

where $\beta_i = \lambda_i(n-1)$.

Proof: Writing

$$\text{mmse}(X, \gamma) = \mathsf{E}[X^2] - \mathsf{E}\left[\mathsf{E}^2[X \mid \sqrt{\gamma} X + N]\right]$$

with unit-variance Gaussian N, and since

$$\sum_{i=1}^n \lambda_i \text{var}\{\beta_i^{-\frac{1}{2}} X_{\backslash i}\} = \text{var}\{\sum_{i=1}^n X_i\}$$

the difference between the left- and right-hand sides of (25) becomes

$$\sum_{i=1}^n \lambda_i \mathsf{E}\left[\mathsf{E}^2\left[\beta_i^{-\frac{1}{2}} X_{\backslash i} \;\middle|\; \frac{\sqrt{\gamma}}{\sqrt{\beta_i}} X_{\backslash i} + \frac{1}{\sqrt{\beta_i}} N_{\backslash i}\right]\right]$$
$$- \mathsf{E}\left[\mathsf{E}^2\left[\sum_{i=1}^n X_i \mid Y\right]\right]$$

$$= \frac{1}{n-1} \sum_{i=1}^n \mathsf{E}\left[\mathsf{E}^2[X_{\backslash i} \mid Y_{\backslash i}]\right] - \mathsf{E}\left[\mathsf{E}^2\left[\sum_{i=1}^n X_i \mid Y\right]\right] \qquad (26)$$

$$\geq 0 \qquad (27)$$

All that remains is to justify (27)

$$\mathsf{E}\left[\mathsf{E}^2\left[\sum_{i=1}^n X_i \mid Y\right]\right]$$
$$= \frac{1}{(n-1)^2} \mathsf{E}\left[\mathsf{E}^2\left[\sum_{i=1}^n X_{\backslash i} \mid Y\right]\right] \qquad (28)$$

$$= \frac{1}{(n-1)^2} \mathsf{E}\left[\left(\mathsf{E}\left[\sum_{i=1}^n \mathsf{E}[X_{\backslash i} \mid Y_i, Y_{\backslash i}] \mid Y\right]\right)^2\right] \qquad (29)$$

$$= \frac{1}{(n-1)^2} \mathsf{E}\left[\left(\mathsf{E}\left[\sum_{i=1}^{n} \mathsf{E}[X_{\backslash i} \,|\, Y_{\backslash i}] \,|\, Y\right]\right)^2\right] \quad (30)$$

$$\leq \frac{1}{(n-1)^2} \mathsf{E}\left[\left(\sum_{i=1}^{n} \mathsf{E}[X_{\backslash i} \,|\, Y_{\backslash i}]\right)^2\right] \quad (31)$$

$$\leq \frac{1}{n-1} \sum_{i=1}^{n} \mathsf{E}\left[\mathsf{E}^2\left[X_{\backslash i} \,|\, Y_{\backslash i}\right]\right] \quad (32)$$

where

(28) $\Leftarrow$ (14);

(29) $\Leftarrow$ Y is a function (sum) of Y_i and $Y_{\backslash i}$;

(30) $\Leftarrow$ $X_{\backslash i}$ is independent of Y_i (conditioned on $Y_{\backslash i}$);

(31) $\Leftarrow$ $\mathsf{E}[V^2] \geq \mathsf{E}[\mathsf{E}^2[V|Y]]$ for any V;

(32) $\Leftarrow$ Lemma 3 (Appendix). $\qquad\square$

The main result (Theorem 1) is seen to be a corollary of an estimation-theoretic property (Lemma 2.) Note that the only two inequalities used in the proof of Lemma 2 are the nonnegativity of MMSE and the "strengthened Cauchy–Schwarz" inequality of [2] (Lemma 2). In contrast, the proof of ([2, Theorem 2]) is considerably more complex involving a far-from-elementary variational representation of Fisher's information, whose explicit solution is cumbersome enough to be sidestepped in [2]. As pointed out in [5], its MMSE-mutual information formula (from which Lemma 1 is derived) is essentially equivalent to de Bruijn's identity involving the differential entropy and Fisher's information. However, starting with the proofs of the identities themselves, continuing with the proof of the entropy power inequality [4], up to the present correspondence, the evidence so far is rather convincing that using MMSE in lieu of Fisher's information yields much crisper proofs in addition to a wide variety of new results (e.g., [5] and [6]). Once more, engineering intuition provides a powerful guide to get to the core of the mathematics.

APPENDIX

The following result can be found in [2]. We give a more concrete proof for completeness.

Lemma 3: Let $W_1, \ldots, W_n$ be independent. Let Z_i be a deterministic function of $W_1, W_2, \ldots, W_{i-1}, W_{i+1}, \ldots, W_n$, such that $\mathsf{E}[Z_i] = 0$. Then,

$$\mathsf{E}\left[\left(\sum_{i=1}^{n} Z_i\right)^2\right] \leq (n-1) \sum_{i=1}^{n} \mathsf{E}[Z_i^2]. \quad (33)$$

Proof: Denoting

$$\Delta_{ij} = \mathsf{E}[Z_i \,|\, W_j, \ldots, W_n] - \mathsf{E}[Z_i \,|\, W_{j+1}, \ldots, W_n]$$

and noticing that for $j = n$, $\mathsf{E}[Z_i \,|\, W_{j+1}, \ldots, W_n] = \mathsf{E}[Z_i] = 0$, we can write

$$Z_i = \sum_{j=1}^{n} \Delta_{ij} \quad (34)$$

Without loss of generality, let us assume that $j < k$. In view of the independence of the $\{W_i\}_{i=1}^n$,

$$\mathsf{E}\left[\mathsf{E}[Z_i \,|\, W_j, \ldots, W_n]\mathsf{E}[Z_\ell \,|\, W_k, \ldots, W_n]\right]$$
$$= \mathsf{E}\left[\mathsf{E}[Z_i \,|\, W_k, \ldots, W_n]\mathsf{E}[Z_\ell \,|\, W_k, \ldots, W_n]\right] \quad (35)$$

from which it immediately follows that

$$\mathsf{E}[\Delta_{ij}\Delta_{\ell k}] = 0, \qquad \text{for } j \neq k.$$

Thus,

$$\mathsf{E}[Z_i^2] = \sum_{j=1}^{n} \mathsf{E}\left[\Delta_{ij}^2\right] \quad (36)$$

and

$$\mathsf{E}\left[\left(\sum_{i=1}^{n} Z_i\right)^2\right] = \mathsf{E}\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij}\right)^2\right] \quad (37)$$

$$= \sum_{j=1}^{n} \mathsf{E}\left[\left(\sum_{i=1}^{n} \Delta_{ij}\right)^2\right] \quad (38)$$

$$\leq (n-1) \sum_{j=1}^{n}\sum_{i=1}^{n} \mathsf{E}\left[\Delta_{ij}^2\right] \quad (39)$$

$$= (n-1) \sum_{i=1}^{n} \mathsf{E}\left[Z_i^2\right] \quad (40)$$

where (40) is equivalent to (36); (39) is simply because

$$(a_1 + a_2 + \ldots + a_{n-1})^2 \leq (n-1) \sum_{i=1}^{n-1} a_i^2$$

and $\sum_{i=1}^{n} \Delta_{ij}$ contains at most $n-1$ nonzero terms since $\Delta_{ii} = 0$. $\qquad\square$

REFERENCES

[1] O. Johnson, *Information Theory And The Central Limit Theorem*. London, U.K.: Imperial College Press, 2004.

[2] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, "Solution of Shannon's problem on the monotonicity of entropy," *J. Amer. Math. Soc.*, vol. 17, no. 4, pp. 975–982, May. 12, 2004.

[3] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, Jul. and Oct. 1948.

[4] S. Verdú and D. Guo, "A simple proof of the entropy power inequality," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2165–2166, May 2006.

[5] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[6] A. Lozano, A. M. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3033–3051, Jul. 2006.