Solution of Shannon's Problem on the Monotonicity of Entropy

Shiri Artstein* Keith Ball† Franck Barthe‡ Assaf Naor§

Abstract

It is shown that if $X_1, X_2, \ldots$ are independent and identically distributed square-integrable random variables then the entropy of the normalized sum

$$\mathrm{Ent}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)$$

is an increasing function of n. This resolves an old problem which goes back to [6, 7, 5].

The result also has a version for non-identically distributed random variables or random vectors.

1 Introduction

The entropy of a real valued random variable X with density $f : \mathbf{R} \to [0, \infty)$ is defined as

$$\mathrm{Ent}(X) = -\int_{\mathbf{R}} f \log f$$

provided that the integral makes sense. Among random variables with variance 1, the standard Gaussian G has the largest entropy. If X_i are independent copies of a random variable X with variance 1, then the normalized sums

$$Y_n = \frac{1}{\sqrt{n}} \sum_1^n X_i$$

approach the standard Gaussian as n tends to infinity, in a variety of senses.

In the 1940's Shannon (see [6]) proved that $\mathrm{Ent}(Y_2) \geq \mathrm{Ent}(Y_1)$, that is, the entropy of the normalized sum of two independent copies of a random variable is larger than that of the original (In fact, the first rigorous proof of this fact is due to Stam [7]. A shorter proof was obtained by Lieb [5]). Inductively it follows that for the sequence of powers of 2, $\mathrm{Ent}(Y_{2^k}) \geq \mathrm{Ent}(Y_{2^{k-1}})$. It was therefore naturally conjectured that the entire sequence $\mathrm{Ent}_n = \mathrm{Ent}(Y_n)$ increases with n. This problem, though seemingly elementary, remained open: the conjecture was formally stated by Lieb in 1978 [5]. It was even unknown whether the relation $\mathrm{Ent}_2 \leq \mathrm{Ent}_3$ holds true and this special case

*Supported in part by the EU Grant HPMT-CT-2000-00037, The Minkowski center for Geometry and the Israel Science Foundation

†Supported in part by NSF Grant DMS-9796221

‡Supported in part by EPSRC Grant GR/R37210

§Supported in part by the BSF, Clore Foundation and EU Grant HPMT-CT-2000-00037

helps to explain the difficulty: there is no natural way to "build" the sum of three independent copies of X out of the sum of two.

In this article we develop a new understanding of the underlying structure of sums of independent random variables and use it to prove the general statement for entropy:

Theorem 1 (Entropy increases at every step). *Let $X_1, X_2, \dots$ be independent and identically distributed square-integrable random variables. Then*

$$\mathrm{Ent}\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) \le \mathrm{Ent}\left(\frac{X_1+\cdots+X_{n+1}}{\sqrt{n+1}}\right).$$

The main point of the result is that one can now see clearly that convergence in the central limit theorem is driven by an analogue of the second law of thermodynamics. There are versions of the central limit theorem for non-identically distributed random variables and in this context we have the following non-identically distributed version of Theorem 1:

Theorem 2. *Let $X_1, X_2, \dots, X_{n+1}$ be independent random variables and $(a_1, \dots, a_{n+1}) \in S^n$ a unit vector. Then*

$$\mathrm{Ent}\left(\sum_{i=1}^{n+1} a_i X_i\right) \ge \sum_{j=1}^{n+1} \frac{1-a_j^2}{n} \cdot \mathrm{Ent}\left(\frac{1}{\sqrt{1-a_j^2}} \cdot \sum_{i\ne j} a_i X_i\right).$$

In particular,

$$\mathrm{Ent}\left(\frac{X_1+\cdots+X_{n+1}}{\sqrt{n+1}}\right) \ge \frac{1}{n+1}\sum_{j=1}^{n+1} \mathrm{Ent}\left(\frac{1}{\sqrt{n}}\sum_{i\ne j} X_i\right).$$

A by-product of this generalization is the following *entropy power inequality*, the case $n = 2$ of which is a classical fact (see [7, 5]).

Theorem 3 (Entropy power for many summands). *Let $X_1, X_2, \dots, X_{n+1}$ be independent square-integrable random variables. Then*

$$\exp\left[2\,\mathrm{Ent}\left(\sum_{i=1}^{n+1} X_i\right)\right] \ge \frac{1}{n}\sum_{j=1}^{n+1} \exp\left[2\,\mathrm{Ent}\left(\sum_{i\ne j} X_i\right)\right].$$

Theorem 3 can be deduced from Theorem 2 as follows. Denote $E = \exp\left[2\,\mathrm{Ent}\left(\sum_{i=1}^{n+1} X_i\right)\right]$ and $E_j = \exp\left[2\,\mathrm{Ent}\left(\sum_{i\ne j} X_i\right)\right]$. It is easy to check that $\mathrm{Ent}(X+Y) \ge \mathrm{Ent}(X)$ whenever X and Y are independent. It follows that $E \ge E_j$ for all j, so the required inequality holds trivially if $E_j \ge \frac{1}{n}\sum_{i=1}^{n+1} E_i$ for some j. We may therefore assume that $\lambda_j \equiv E_j / \left(\sum_{i=1}^{n+1} E_i\right) < 1/n$ for all j. Setting $a_j = \sqrt{1 - n\lambda_j}$, an application of Theorem 2 shows that

$$\mathrm{Ent}\left(\sum_{i=1}^{n+1} X_i\right) \ge \sum_{j=1}^{n+1} \lambda_j \mathrm{Ent}\left(\frac{1}{\sqrt{n\lambda_j}}\sum_{i\ne j} X_i\right).$$

This inequality simplifies to give the statement of Theorem 3, using the fact that for $\lambda > 0$, $\mathrm{Ent}(\lambda X) = \log\lambda + \mathrm{Ent}(X)$.

2 Proof of Theorem 2

The argument begins with a reduction from entropy to another information-theoretic notion, the Fisher information of a random variable. For sufficiently regular densities, the Fisher information can be written as

$$J(X) = \int_{\mathbf{R}} \frac{(f')^2}{f}.$$

Among random variables with variance 1, the Gaussian has *smallest* information: namely 1. There is a remarkable connection between Fisher information and entropy, provided by the adjoint Ornstein-Uhlenbeck semigroup, which goes back to de Bruijn (see e.g. [7]), Bakry-Emery [1] and Barron [3]. A particularly clear explanation is given in the article of Carlen and Soffer [4]. The point is that the entropy gap $\mathrm{Ent}(G) - \mathrm{Ent}(X)$ can be written as an integral

$$\int_0^\infty \left(J(X^{(t)}) - 1 \right) \, dt$$

where $X^{(t)}$ is the evolute at time t of the random variable X under the semigroup. Since the action of the semigroup commutes with self-convolution, the increase of entropy can be deduced by proving a *decrease* of information $J\left(Y_{n+1}\right) \leq J\left(Y_n\right)$, where now Y_n is the normalized sum of IID copies of an evolute $X^{(t)}$. Each such evolute $X^{(t)}$ has the same distribution as an appropriately weighted sum of X with a standard Gaussian:

$$\sqrt{e^{-2t}} X + \sqrt{1 - e^{-2t}} G.$$

Since we may assume that the density of X has compact support, the density of each $X^{(t)}$ has all the smoothness and integrability properties that we shall need below. (See [4] for details.)

To establish the decrease of information with n, we use a new and flexible formula for the Fisher information, which is described in Theorem 4. A version of this theorem (in the case $n = 2$) was introduced in [2], and was motivated by an analysis of the transportation of measures and a local analogue of the Brunn-Minkowski inequality.

Theorem 4 (Variational characterization of the information). *Let $w : \mathbf{R}^n \to (0, \infty)$ be a continuously twice differentiable density on $\mathbf{R}^n$ with*

$$\int \frac{\|\nabla w\|^2}{w} \, , \, \int \|Hess(w)\| \, < \infty.$$

Let e be a unit vector and h the marginal density in direction e defined by

$$h(t) = \int_{te + e^\perp} w.$$

Then the Fisher information of the density h satisfies

$$J(h) \leq \int_{\mathbf{R}^n} \left(\frac{\mathrm{div}(pw)}{w} \right)^2 w,$$

for any continuously differentiable vector field $p : \mathbf{R}^n \to \mathbf{R}^n$ with the property that for every x,

$$\langle p(x), e \rangle = 1$$

(and, say, $\int \|p\| w < \infty$).

If w satisfies $\int \|x\|^2 w(x) < \infty$, then there is equality for some suitable vector field p.

Remark The condition $\int \|p\|w < \infty$ is not important in applications but makes for a cleaner statement of the theorem. We are indebted to Giovanni Alberti for pointing out this simplified formulation of the result.

Proof. The conditions on w ensure that for each t,

$$h'(t) = \int_{te+e^\perp} \partial_e w. \tag{1}$$

If

$$\int_{\mathbf{R}^n} \left(\frac{\operatorname{div}(pw)}{w} \right)^2 w$$

is finite, then $\operatorname{div}(pw)$ is integrable on $\mathbf{R}^n$ and hence on almost every hyperplane perpendicular to e. If $\int \|p\|w$ is also finite, then on almost all of these hyperplanes the $(n-1)$-dimensional divergence of pw in the hyperplane, integrates to zero by the Gauss-Green Theorem. Since the component of p in the e direction is always 1, we have that for almost every t,

$$h'(t) = \int_{te+e^\perp} \operatorname{div}(pw)$$

Therefore

$$J(h) = \int \frac{h'(t)^2}{h(t)} = \int \frac{(\int \operatorname{div}(pw))^2}{\int w}$$

and by the Cauchy-Schwarz inequality this is at most

$$\int_{\mathbf{R}^n} \frac{(\operatorname{div}(pw))^2}{w}.$$

To find a field p for which there is equality, we want to arrange that

$$\operatorname{div}(pw) = \frac{h'(t)}{h(t)} w$$

since then we have

$$\int_{\mathbf{R}^n} \frac{(\operatorname{div}(pw))^2}{w} = \int_{\mathbf{R}^n} \frac{h'(t)^2}{h(t)^2} w = J(h).$$

Since we also need to ensure that $\langle p, e \rangle = 1$ identically, we construct p separately on each hyperplane perpendicular to e, to be a solution of the equation

$$\operatorname{div}_{e^\perp}(pw) = \frac{h'(t)}{h(t)} w - \partial_e w. \tag{2}$$

Regularity of p on the whole space can be ensured by using a "consistent" method for solving the equation on the separate hyperplanes. Note that for each t, the conditions on w ensure that $\operatorname{div}_{e^\perp}(pw)$ is integrable on $te + e^\perp$, while the "constant" $\frac{h'(t)}{h(t)}$ ensures that the integral is zero.

The hypothesis $\int \|x\|^2 w(x) < \infty$ is needed only if we wish to construct a p for which $\int \|p\|w < \infty$. For each real t and each $y \in e^\perp$ set

$$F(t, y) = \frac{h'(t)}{h(t)} w(te + y) - \partial_e w(te + y).$$

The condition

$$\int \frac{\|\nabla w\|^2}{w} < \infty$$

shows that

$$\int \frac{F^2}{w} < \infty$$

and under the assumption $\int \|x\|^2 w(x) < \infty$ this guarantees that

$$\int |F| \cdot \|x\| < \infty. \tag{3}$$

As a result we can find solutions to our equation

$$\mathrm{div}_{e^\perp}(pw) = F$$

satisfying $\int \|pw\| < \infty$, for example in the following way.

Fix an orthonormal basis of $e^\perp$ and index the coordinates of $y \in e^\perp$ with respect to this basis, $y_1, y_2, \ldots, y_{n-1}$. Let $K(y_2, \ldots, y_{n-1}) = \int F(t, y)\, dy_1$ and choose a rapidly decreasing density g on the line. Then the function

$$F(t, y) - g(y_1)K(y_2, \ldots, y_n)$$

integrates to zero on every line in the y_1 direction so we can find $p_1(y)$ so that $p_1 w$ tends to zero at infinity and its y_1-derivative is $F(t, y) - g(y_1)K(y_2, \ldots, y_n)$. The integrability of $p_1 w$ is immediate from equation (3). Continue to construct the coordinates of p, one by one, by conditioning on subspaces of successively smaller dimension.

$\square$

Remarks Theorem 4 could be stated in a variety of ways. The proof makes it clear that all we ask of the expression $\mathrm{div}(pw)$ is that it be a function for which $\mathrm{div}_{e^\perp}(pw) - \partial_e w$ integrates to zero on each hyperplane perpendicular to e. If pw decays sufficiently rapidly at infinity, then this condition is guaranteed by the Gauss-Green Theorem, and thus there is a certain naturalness about realizing the function as a divergence. However, the real point of the formulation given above is that in each of the applications that we have found for the theorem, it is much easier to focus on the vector field p than on its divergence, even though formally, the fields themselves play no role. This latter fact is reassuring given that there are so many different fields with the same divergence.

In the proof of Theorem 2, below, let f_i be the density of X_i and consider the product density

$$w(x_1, \ldots, x_{n+1}) = f_1(x_1) \cdots f_{n+1}(x_n).$$

The density of $\sum_{i=1}^{n+1} a_i X_i$ is the marginal of w in the direction $(a_1, \ldots, a_{n+1}) \in S^n$. We shall show that if w satisfies the conditions of Theorem 4, then for every unit vector $\hat{a} = (a_1, \ldots, a_{n+1}) \in S^n$ and every $b_1, \ldots, b_{n+1} \in \mathbf{R}$ satisfying $\sum_{j=1}^{n+1} b_j \sqrt{1 - a_j^2} = 1$ we have

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \leq n \sum_{j=1}^{n+1} b_j^2 J\left(\frac{1}{\sqrt{1 - a_j^2}} \sum_{i \neq j} a_i X_i\right). \tag{4}$$

This will imply Theorem 2 since we may choose $b_j = \frac{1}{n}\sqrt{1-a_j^2}$ and apply (4) to the Ornstein-Uhlenbeck evolutes of the X_i's, $X_i^{(t)}$, and then integrate over $t \in (0,\infty)$. It is worth observing that since the Fisher information is homogeneous of order -2, inequality (4) implies the following generalization of the Blachman-Stam inequality [7] to more than two summands:

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

Proof of Theorem 2. In order to prove (4), for every j denote $\hat{a}_j = \frac{1}{\sqrt{1-a_j^2}}(a_1,\ldots,a_{j-1},0,a_{j+1}\ldots,a_n)$, which is also a unit vector. Let $p^j : \mathbf{R}^{n+1} \to \mathbf{R}^{n+1}$ be a vector field which realizes the information of the marginal of w in direction $\hat{a}_j$ as in Theorem 4, that is $\langle p^j, \hat{a}_j\rangle \equiv 1$ and

$$J\left(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i\right) = \int_{\mathbf{R}^n} \left(\frac{\mathrm{div}(wp^j)}{w}\right)^2 w.$$

Moreover, we may assume that p^j does not depend on the coordinate x_j, and that the j^{th} coordinate of p^j is identically 0, since we may restrict to n dimensions, use Theorem 4, and then artificially add the j^{th} coordinate while not violating any of the conclusions.

Consider the vector field $p : \mathbf{R}^{n+1} \to \mathbf{R}^{n+1}$ given by $p = \sum_{j=1}^{n+1} b_j p^j$. Since $\sum_{j=1}^{n+1} b_j\sqrt{1-a_j^2} = 1$, $\langle p, \hat{a}\rangle \equiv 1$. By Theorem 4,

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \leq \int_{\mathbf{R}^n} \left(\frac{\mathrm{div}(wp)}{w}\right)^2 w = \int_{\mathbf{R}^n} \left(\sum_{j=1}^{n+1} b_j \frac{\mathrm{div}(wp^j)}{w}\right)^2 w.$$

Let y_j denote $b_j\frac{\mathrm{div}(wp^j)}{w}$. Our aim is to show that in $L_2(w)$ (the Hilbert space with weight w):

$$\|y_1 + \cdots + y_{n+1}\|^2 \leq n\left(\|y_1\|^2 + \cdots \|y_{n+1}\|^2\right).$$

With no assumptions on the y_i, the Cauchy-Schwarz inequality would give a coefficient of $n+1$ instead of n on the right hand side. However, our y_i have additional properties. Define $T_1 : L_2(w) \to L_2(w)$ by

$$(T_1\phi)(x) = \int \phi(u, x_2, x_3, \ldots, x_{n+1}) f_1(u)du$$

(so that $T_1(\phi)$ is independent of the first coordinate) and similarly for T_i, by integrating out the i^{th} coordinate against f_i. Then the T_i are commuting orthogonal projections on the Hilbert space $L_2(w)$. Moreover for each i, $T_i y_i = y_i$ since y_i is already independent of the i^{th} coordinate, and for each j, $T_1\cdots T_{n+1}y_j = 0$ because we integrate a divergence. These properties ensure the slightly stronger inequality that we need. We summarize it in the following lemma which completes the proof of Theorem 1.

Lemma 5. *Let $T_1,\ldots T_m$ be m commuting orthogonal projections in a Hilbert space H. Assume that we have m vectors $y_1,\ldots,y_m$ such that for every $1 \leq j \leq m$, $T_1\cdots T_m y_j = 0$. Then*

$$\|T_1y_1 + \cdots + T_my_m\|^2 \leq (m-1)\left(\|y_1\|^2 + \cdots \|y_m\|^2\right).$$

Proof. Since the projections are commuting, we can decompose the Hilbert space $H = \bigoplus_{\varepsilon\in\{0,1\}^m}^{\perp} H_\varepsilon$ where $H_\varepsilon = \{x : T_i x = \varepsilon_i x, 1 \le i \le m\}$. This is an orthogonal decomposition, so that for $\phi \in H$, $\|\phi\|^2 = \sum_{\varepsilon\in\{0,1\}^m} \|\phi_\varepsilon\|^2$. We decompose each y_i separately as, $y_i = \sum_{\varepsilon\in\{0,1\}^m} y^i_\varepsilon$. The condition in the statement of the lemma implies that $y^i_{(1,\dots,1)} = 0$ for each i. Therefore

$$T_1 y_1 + \cdots + T_m y_m = \sum_{i=1}^m \sum_{\varepsilon\in\{0,1\}^m} T_i y^i_\varepsilon = \sum_{\varepsilon\in\{0,1\}^m} \sum_{\varepsilon_i=1} y^i_\varepsilon.$$

Now we can compute the norm of the sum as follows

$$\|T_1 y_1 + \cdots + T_m y_m\|^2 = \sum_{\varepsilon\in\{0,1\}^m} \left\| \sum_{\varepsilon_i=1} y^i_\varepsilon \right\|^2.$$

Every vector on the right hand side is a sum of at most $m-1$ summands, since the only vector with m 1's does not contribute anything to the sum. Thus we can complete the proof of the lemma using the Cauchy-Schwarz inequality:

$$\|T_1 y_1 + \cdots + T_n y_m\|^2 \le \sum_{\varepsilon\in\{0,1\}^m} (m-1) \sum_{\varepsilon(i)=1} \|y^i_\varepsilon\|^2 = (m-1) \sum_{i=1}^m \|y_i\|^2.$$

□

Remark The result of this note applies also to the case where X is a random vector. In this case the Fisher information is realized as a matrix, which in the sufficiently smooth case can be written as

$$J_{i,j} = \int \left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_j}\right)\frac{1}{f}.$$

Theorem 4 generalizes in the following way. Let w be a density defined on $\mathbf{R}^n$ and let h be the (vector) marginal of w on the subspace E. For $x \in E$ we define $h(x) = \int_{x+E^\perp} w$. Then for a unit vector e in E

$$\langle Je, e\rangle = \inf \int \left(\frac{\mathrm{div}(wp)}{w}\right)^2 w,$$

where the infimum runs over all $p : \mathbf{R}^n \to \mathbf{R}^n$ for which the orthogonal projection of p into E is constantly e. The rest of the argument is exactly the same as in the one-dimensional case.

References

[1] D. Bakry and M. Emery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX*, number 1123 in Lect. Notes in Math., pages 179–206. Springer, 1985.

[2] K. Ball, F. Barthe, and A. Naor. Entropy jumps in the presence of a spectral gap. *Duke Math. J.*, 119(1):41–63, 2003.

[3] A. R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14:336–342, 1986.

[4] E. A. Carlen and A. Soffer. Entropy production by block variable summation and central limit theorems. *Commun. Math. Phys.*, 140(2):339–371, 1991.

[5] E. H. Lieb. Proof of an entropy conjecture of Wehrl. *Comm. Math. Phys.*, 62 no. 1:35–41, 1978.

[6] C. E. Shannon and W. Weaver. *The mathematical theory of communication.* University of Illinois Press, Urbana, IL, 1949.

[7] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Info. Control*, 2:101–112, 1959.

Shiri Artstein
School of Mathematical Sciences
Tel Aviv University
Ramat Aviv, Tel Aviv 69978
Israel
E-mail artst@post.tau.ac.il

Keith M. Ball,
Department of Mathematics,
University College London,
Gower Street,
London WC1 6BT,
United Kingdom.
E-mail: kmb@math.ucl.ac.uk

Franck Barthe,
Institut de Mathématiques,
Laboratoire de Statistique et Probabilités, CNRS UMR C5583,
Université Paul Sabatier,
31062 Toulouse Cedex 4
France.
E-mail: barthe@math.ups-tlse.fr

Assaf Naor,
Theory Group,
Microsoft Research,
One Microsoft Way,
Redmond WA 98052-6399
USA.
E-mail: anaor@microsoft.com