Shannon's Entropy Power Inequality via Restricted Minkowski Sums

S.J. Szarek[1] and D. Voiculescu[2]

[1] Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106-7058, U.S.A. and Université Pierre et Marie Curie, Equipe d'Analyse, Bte 186, 4, Place Jussieu, 75252 Paris, France

[2] Department of Mathematics, University of California, Berkeley, CA 94720-3840, U.S.A.

1 Introduction and Preliminaries

If X is an $\mathbb{R}^n$-valued random variable whose distribution μ_X is absolutely continuous with respect to the Lebesgue measure λ_n and f is the corresponding density, the *entropy* of X is defined via $h(X) := \int_{\mathbb{R}^n} f \log \frac{1}{f} d\lambda_n$. One of the fundamental results of Information Theory (see, e.g., [SW]) is the Shannon's *Entropy Power Inequality*, which affirms that if X, Y are two such variables which are independent, then

$$\exp(2h(X)/n) + \exp(2h(Y)/n) \leq \exp(2h(X+Y)/n)\,. \tag{1}$$

Shannon's original variational argument seems incomplete, but there exist (at least) two other proofs of (1) due to Stam ([S], 1959) and Lieb ([L], 1978); see [CT] or [DCT] for more background and history. The purpose of this note is to present a new proof of the classical Entropy Power Inequality in the spirit of [SV], where its noncommutative (free) analogue was shown. Our proof is conceptually related to Lieb's argument as it uses a rearrangement inequality from [BLL], belonging to the same circle of ideas as [L].

While having one more proof of a classical fact may be perceived as being of limited value, the present argument appears to have the advantage of being much more direct than the other ones. Additionally, we hope that the more geometric approach may shed some new light on the noncommutative theory, where even the most appropriate definitions of concepts, particularly in the multivariate case, haven't been determined, cf. the series of papers [V] and their references.

As in [SV], our argument is based on a geometric result resembling formally the classical Brunn-Minkowski inequality. Let A, B be subsets of a

Both authors were supported in part by grants from the National Science Foundation. This research was partially carried out by the second named author for the Clay Mathematics Institute.

vector space and $\Theta \subset A \times B$. We will call

$$A +_\Theta B := \{x + y \,:\, (x, y) \in \Theta\}$$

the restricted (to Θ) sum of A and B. We then have (below and in what follows, all sets and functions are assumed to be measurable and $\lambda = \lambda_n$ is the Lebesgue measure in the appropriate dimension which may vary between occurrences)

Lemma 1. *For any $\varepsilon > 0$ there exists $\delta > 0$ such that if $n \in \mathbb{N}$, $A, B \subset \mathbb{R}^n$ and $\Theta \subset A \times B \subset \mathbb{R}^{2n}$ verify*

$$\lambda(\Theta) \geq (1 - \delta)^n \lambda(A)\lambda(B)\,,$$

then

$$\lambda(A +_\Theta B)^{2/n} \geq (1 - \varepsilon)\left(\lambda(A)^{2/n} + \lambda(B)^{2/n}\right)\,. \tag{2}$$

Remarks.

(i) The Lemma above is slightly different from the version stated in [SV]. Its analogue in that paper (Theorem 1.2) asserts a stronger inequality $\lambda(A +_\Theta B)^{2/n} \geq \lambda(A)^{2/n} + \lambda(B)^{2/n}$ under a stronger hypothesis: $(1-\delta)^n$ replaced by $1 - \delta$, where – in cases of interest – $\delta > 0$ can be chosen independently of n, A, B and Θ. However, the present variant follows easily, with a "nearly" optimal dependence $\varepsilon = O(\delta^{1/2})$, from Corollary 1.5 and Remark 1.6 in [SV], or can be directly derived from the rearrangement inequality of [BLL].

(ii) A formally stronger "restricted Prékopa-Leindler inequality" was proved in [B]; it is quite likely that it can be used to prove (1) in an even more direct way.

We will also need the following elementary fact, closely related to the traditional information-theoretic definition of entropy as a measure of the volume of the "effective support" of a "large" sample of X. To make the exposition more clear, we shall concentrate on the scalar case, which contains all the ingredients of the general setting (see the remark at the end of this section).

Let $X_1, X_2, \ldots, X_N, \ldots$ be a sequence of independent copies of a real random variable X (with density f, as before; only variables with density need to be considered in this context) and denote by $\mathcal{P}$ the underlying probability measure. Given $N \in \mathbb{N}$, let $F = F_N : \mathbb{R}^N \to \mathbb{R}_+$ be the joint density of $X_1, X_2, \ldots, X_N$ with respect to λ_N; of course $F(x_1, x_2, \ldots, x_N) = f(x_1)f(x_2)\ldots f(x_N)$. We then have

Proposition 2. *There exist two positive sequences* (α_k) *and* (β_k) *(depending on* X*) converging to* 0 *such that if we set, for* $N \in \mathbb{N}$,

$$V_N = V_N(X) := \left\{ e^{N(-h(X)-\alpha_N)} \leq F_N \leq e^{N(-h(X)+\alpha_N)} \right\},$$

then

$$\mathcal{P}\left((X_1, X_2, \ldots, X_N) \in V_N\right) \equiv \int_{V_N} F_N d\lambda_N \ > 1 - \beta_N \, .$$

V_N is closely related to the set of "typical sequences" from Information Theory (cf. [SW], sections 7 and 21). For completeness, we include a simple proof of the Proposition in the Appendix, even though neither the result nor the gist of the argument are new. Of course we could have simplified the statement by requiring that $(\alpha_k) = (\beta_k)$; however, as shall be clear from what follows, the roles of these two sequences are quite different. In particular, for our purposes it would have been enough to have (β_k) just "sufficiently small", e.g, $\leq 1/4$ for large k.

An immediate consequence of the Proposition is that

$$(1 - \beta_N) e^{N(h(X)-\alpha_N)} < \lambda(V_N) \leq e^{N(h(X)+\alpha_N)} \tag{3}$$

and hence, as $N \to \infty$,

$$\frac{\log \lambda(V_N(X))}{N} \to h(X) \quad \textit{and} \quad \lambda\big(V_N(X)\big)^{2/N} \to \exp\big(2h(X)\big) \, . \tag{4}$$

2 The Proof

The idea of the rest of the proof is now as follows. Let (X_k) and (Y_k) be sequences of *jointly* independent copies of X and Y respectively. Given N, the set $V_N(X+Y)$ of typical sequences $X_1+Y_1, X_2+Y_2, \ldots, X_N+Y_N$ is *roughly* the same as the Minkowski sum of $V_N(X)$ and $V_N(Y)$, the sets of typical sequences $X_1, X_2, \ldots, X_N$ and $Y_1, Y_2, \ldots, Y_N$ respectively. Accordingly, by the inequality (2) for restricted Minkowski sums ("restricted" because of the qualification "roughly" above) we have *approximately* $\lambda(V_N(X+Y))^{2/N} \geq \lambda(V_N(X))^{2/N} + \lambda(V_N(Y))^{2/N}$ ("approximately" because of the $1-\varepsilon$ factor in (2)) and (1) follows by letting $N \to \infty$ and using the second limit relation in (4).

To make this sketch precise, we apply Lemma 1 with $A := V_N(X)$, $B := V_N(Y)$ and

$$\Theta := \left\{ (a,b) \in A \times B \, : \, a+b \in V_N(X+Y) \right\}.$$

Since, by definition, $A +_\Theta B \subset C := V_N(X+Y)$, leaving for a moment aside the issue of the exact values of $\varepsilon = \varepsilon_N$ and $\delta = \delta_N$ that intervene, we get from Lemma 1 that

$$\lambda\big(V_N(X+Y)\big)^{2/N} \geq (1-\varepsilon_N)\Big(\lambda\big(V_N(X)\big)^{2/N} + \lambda\big(V_N(Y)\big)^{2/N}\Big). \tag{5}$$

We claim that with our choice of A, B and Θ one has

$$\left(\frac{\lambda_{2N}(\Theta)}{\lambda_N(A)\cdot\lambda_N(B)}\right)^{1/N} \to 1$$

as $N \to \infty$ and so, when applying Lemma 1, one may have choose $\delta = \delta_N$ so that (as $N \to \infty$) $\delta_N \to 0$, hence $\varepsilon_N \to 0$, and so, by the argument sketched earlier, the inequality (5) becomes in the limit (1).

To prove our claim $\delta_N \to 0$ we denote $\mathbf{X} := (X_1, X_2, \ldots, X_N)$, $\mathbf{Y} := (Y_1, Y_2, \ldots, Y_N)$ and observe that

$$\begin{aligned} 1-\beta_N &\le \mathcal{P}(\mathbf{X}+\mathbf{Y}\in C) \\ &\le \mathcal{P}(\mathbf{X}\in A, \mathbf{Y}\in B \& \mathbf{X}+\mathbf{Y}\in C) + \mathcal{P}\big((\mathbf{X},\mathbf{Y})\notin A\times B\big) \\ &= \mathcal{P}\big((\mathbf{X},\mathbf{Y})\in\Theta\big) + \big(1-\mathcal{P}(\mathbf{X}\in A)\mathcal{P}(\mathbf{Y}\in B)\big) \\ &\le \mathcal{P}\big((\mathbf{X},\mathbf{Y})\in\Theta\big) + 1 - (1-\beta_N)^2\,. \end{aligned}$$

by the definitions of A, B, C and the estimates on the corresponding probabilities given by Proposition 2. (We note that even though *a priori* the sequences (α_k), (β_k) depend on the random variable in question, they can be chosen to verify the Proposition for the random variables X, Y and $X+Y$ *simultaneously*.) By simple calculation, the above implies

$$1 - 3\beta_N + \beta_N^2 \le \mathcal{P}\big((\mathbf{X},\mathbf{Y})\in\Theta\big) \equiv \int_\Theta G_N d\lambda\,,$$

where G_N is the density of $(\mathbf{X},\mathbf{Y})$ on $\mathbb{R}^{2N}$, necessarily equal to the product of the densities of $\mathbf{X}$ and $\mathbf{Y}$. Using the upper bounds on the latter densities implicit in the hypothesis of Proposition 2 (where the sets $V_N(\,\cdot\,)$ were defined; recall that $\Theta \subset A\times B \equiv V_N(X)\times V_N(Y)$) we deduce

$$\int_\Theta G_N d\lambda \le \lambda(\Theta)\cdot e^{N(-h(X)+\alpha_N)}\cdot e^{N(-h(Y)+\alpha_N)}\,.$$

On the other hand, by (3),

$$\lambda(A)\cdot\lambda(B) \le e^{N(h(X)+\alpha_N)}\cdot e^{N(h(Y)+\alpha_N)}\,.$$

Combining the last three inequalities we obtain

$$\begin{aligned} \frac{\lambda(\Theta)}{\lambda(A)\cdot\lambda(B)} &\ge \frac{(1-3\beta_N+\beta_N^2)\cdot e^{N(h(X)-\alpha_N)}\cdot e^{N(h(Y)-\alpha_N)}}{e^{N(h(X)+\alpha_N)}\cdot e^{N(h(Y)+\alpha_N)}} \\ &= (1-3\beta_N+\beta_N^2)e^{-4N\alpha_N}\,, \end{aligned}$$

whence

$$\left(\frac{\lambda(\Theta)}{\lambda(A)\cdot\lambda(B)}\right)^{1/N} \ge (1-3\beta_N+\beta_N^2)^{1/N}e^{-4\alpha_N} \to 1$$

when $N \to \infty$, as required.

Remark. The proof extends immediately to the multivariate case. Note that if X is $\mathbb{R}^n$-valued, then the corresponding density F_N "lives" on $\mathbb{R}^{nN}$, and so an application of Lemma 1 yields exponents $2/(Nn)$ and results in an additional n in $\exp(2h(X)/n)$.

3 Appendix: The Proof of Proposition 2

As mentioned earlier, Proposition 2 and its immediate consequences are closely related to the traditional information-theoretic definition of entropy, which is as follows. In the notation of the Proposition, let $N \in \mathbb{N}$ and let $V' = V'_N \subset \mathbb{R}^N$ be any smallest (volumewise) set verifying $\int_{V'_N} F_N = 1/2$; we then set $h(X) := \lim_{n\to\infty} \log \lambda_N(V'_N)/N$. The fact that the limit exists, that it remains unchanged if we replace $1/2$ by some other $\beta \in (0,1)$ as well as the equivalence of the two definitions can be easily deduced, e.g., ¿from the argument below, or formally from the assertion of the Proposition.

To prove the Proposition, consider the expression

$$\frac{\log F_N}{N} \equiv \frac{\sum_{j=1}^N \log f(x_j)}{N}$$

and think of it as a random variable on $\mathbb{R}^\infty$ endowed with the product measure $\mathcal{P} := \otimes_{j=1}^\infty \mu_X$ (where $d\mu_X \equiv f d\lambda_1$ is the law of X); it becomes then $\frac{1}{N}\sum_{j=1}^N \log f(X_j)$. By the law of large numbers, as $N \to \infty$, the last sequence converges ($\mathcal{P}$-a.e., or in probability with respect to $\mathcal{P}$), to the expected value of $\log f(X)$ which in turn equals $\int_{\mathbf{R}} \log f d\mu_X = \int_{\mathbf{R}} f \log f d\lambda_1 \equiv -h(X)$. In particular, there exists a positive sequence (α_k) converging to 0 (depending on X) such that,

$$\mathcal{P}\left(-h(X) - \alpha_N \le \frac{1}{N}\sum_{j=1}^N \log f(X_j) \le -h(X) + \alpha_N \right) \to 1$$

as $N \to \infty$, which is just a rephrasing of the assertion of the Proposition.

References

[B] Barthe F. (1999) Restricted Prékopa-Leindler inequality. Pacific J. Math. 189(2):211–222

[BLL] Brascamp H.J., Lieb E.H., Luttinger J.M. (1974) A general rearrangement inequality for multiple integrals. J. Funct. Analysis 17:227–237

[CT] Cover, T. M., Thomas, J. A. (1991) Elements of information theory. Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York.

[DCT] Dembo A., Cover T.M., Thomas J.A. (1991) Information theoretic inequalities. IEEE Transactions on Information Theory 37(6):1501–1518

[L] Lieb E.H. (1978) Proof of an entropy conjecture of Wehrl. Commun. Math. Phys. 62:35–41

[SW] Shannon C.E., Weaver W. (1963) The Mathematical Theory of Communications. University of Illinois Press

[S] Stam A.J. (1959) Some inequalities satisfied by the quantities of information of Fisher and Shannon. Information and Control 2:101–112

[SV] Szarek S.J., Voiculescu D. (1996) Volumes of restricted Minkowski sums and the free analogue of the entropy power inequality. Commun. Math. Phys. 178:563–570

[V] Voiculescu D. (1993) The analogues of entropy and of Fisher's information measure in free probability theory, I. Commun. Math. Phys. 155:71–92; (1994) *ibidem*, II. Invent. Math. 118:411–440; (1996) *ibidem*, III, The absence of Cartan subalgebras. Geom. Funct. Anal. 6(1):172–199; (1997) *ibidem*, IV, Maximum entropy and freeness. In: Free Probability Theory (Waterloo, ON, 1995), Fields Inst. Commun., 12, Amer. Math. Soc., 293–302; (1998) *ibidem*, V, Noncommutative Hilbert transforms. Invent. Math. 132(1):189–227