Fighting with entropy jumps

Fei Cao

July 10, 2023

Contents

Abstract

In the study of the so-called *uniform reshuffling model* arising from econo-physics, we are particularly interested in getting an explicit decay rate in terms of relative entropy from the solution of the uniform reshuffling PDE to the exponential equilibrium. For this purpose, it would be sufficient if we can prove a result indicating a jump in entropy under some transformation of random variables. To be more precise, we want to establish something like

$$\mathrm{H}(X) - \mathrm{H}(U(X+Y)) \ge c\left(\mathrm{H}(X) - \mathrm{H}(\mathcal{E})\right), \tag{1}$$

where X is a random variable with mean $m > 0$ whose density f is supported on $[0,\infty)$, Y is an i.i.d. copy of X, $\mathcal{E} \sim \text{Exponential}(1/m)$, and $U \sim \text{Uniform}[0,1]$ is independent of X, Y and $\mathcal{E}$. Ideally, one hope that the constant $c \in [0,1)$ appearing in (1) is independent of X (or its law f), but this might be too good to be true.

1 Introduction: Entropy jumps in the classical setting

Before we dive into the fundamental question on the validity of (1), it is necessary to recall what is known in the more classical setting. Therefore, we decide to briefly discuss about the entropy jumps associated with re-scaled sum of i.i.d. random variables. It is shown in [1] that if X is a mean-zero $\mathbb{R}$-valued random variable with unit variance and finite entropy, whose density f satisfies a Poincaré inequality, i.e., if for any smooth test function φ,

$$\lambda \mathrm{Var}[\varphi(X)] \leq \mathbb{E}[(\varphi'(X))^2] \tag{2}$$

for some universal constant $\lambda > 0$, then

$$\mathrm{H}(X) - \mathrm{H}\left(\frac{X+Y}{\sqrt{2}}\right) \geq \frac{\lambda}{2+2\lambda}(\mathrm{H}(X) - \mathrm{H}(\mathcal{G})), \tag{3}$$

in which Y is an i.i.d. copy of X and $\mathcal{G} \sim \mathcal{N}(0,1)$. In a nutshell, the entropy of X is greater than that of $(X+Y)/\sqrt{2}$ by a fixed fraction of the entropy gap between X and the Gaussian of the same variance. We remark here that whether the assumption (2) can be removed (or relaxed somehow) is not yet known to our best knowledge. The derivation of (3) relies on a pretty standard strategy: instead of working directly with entropy, we study the Fisher information, which is defined by

$$\mathrm{I}(X) = \int_{\mathbb{R}} \frac{|f'|^2}{f}\,\mathrm{d}x$$

if the law of X is f. In fact, I is much simpler than H in several respects, in particular because it involves quadratic quantities (quoted from [3]). The Gaussian $\mathcal{G}$ is special because not only it has the least entropy among random variables having unit variance, but also it has the smallest Fisher information among random variables having unit variance (see pp 7 of [1] for a short proof). Before we continue our discussion, we mention that the Fisher information of a general density decreases with repeated convolution, which is a simple consequence of the famous Blachman-Stam inequality [2].

Lemma 1. *(Blachman-Stam inequality, three equivalent formulations [3]) Suppose that $\eta \in [0,1]$ and X, Y are independent random variables, then*

(a) $\mathrm{I}(\sqrt{\eta}\,X + \sqrt{1-\eta}\,Y) \le \eta\,\mathrm{I}(X) + (1-\eta)\,\mathrm{I}(Y)$

(b) $\mathrm{I}(X+Y) \le \eta^2\,\mathrm{I}(X) + (1-\eta)^2\,\mathrm{I}(Y)$

(c) $\frac{1}{\mathrm{I}(X+Y)} \ge \frac{1}{\mathrm{I}(X)} + \frac{1}{\mathrm{I}(Y)}$.

Proof We provide a proof in dimension 1 for reader's convenience, the proof presented here is extracted from [2]. Suppose that X and Y has probability density f and g, respectively. Let $Z = X + Y$, then the probability density of Z is $\rho_Z(z) = \int_{\mathbb{R}} f(x)\,g(z-x)\,\mathrm{d}x$. We make a detour to find the law of X given Z, which is denoted by $\rho_{X|Z}(x|z)$. For this goal, we need to determine the joint probability density for X and Z (denoted by $\rho_{X,Z}(x,z)$), then $\rho_{X|Z}(x|z) = \frac{\rho_{X,Z}(x,z)}{\rho_Z(z)}$. For each test function h, we have

$$\mathbb{E}[h(X,Z)] = \int_{\mathbb{R}^2} h(x,x+y)\,f(x)\,g(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbb{R}^2} h(x,z)\,\rho_{X,Z}(x,z)\,\mathrm{d}x\,\mathrm{d}z,$$

from which we deduce that $\rho_{X,Z}(x,z) = f(x)\,g(z-x)$. Now, $\mathrm{d}\rho_Z/\mathrm{d}z$ is

$$\rho_Z'(z) = \int_{\mathbb{R}} f'(x)\,g(z-x)\,\mathrm{d}x. \tag{4}$$

Thus,

$$\frac{\rho_Z'(z)}{\rho_Z(z)} = \int_{\mathbb{R}} \frac{f(x)\,g(z-x)}{\rho_Z(z)} \cdot \frac{f'(x)}{f(x)}\,\mathrm{d}x = \mathbb{E}\left[\frac{f'(X)}{f(X)}\,\middle|\, Z=z\right].$$

Likewise, one has

$$\frac{\rho_Z'(z)}{\rho_Z(z)} = \mathbb{E}\left[\frac{g'(Y)}{g(Y)}\,\middle|\, Z=z\right].$$

Therefore, for any constants a and b,

$$(a+b)\,\frac{\rho_Z'(Z)}{\rho_Z(Z)} = \mathbb{E}\left[a\,\frac{f'(X)}{f(X)} + b\,\frac{g'(Y)}{g(Y)}\,\middle|\, Z\right].$$

Hence, the conditional Jensen's inequality ensures that

$$(a+b)^2\,\left[\frac{\rho_Z'(Z)}{\rho_Z(Z)}\right]^2 \le \mathbb{E}\left[\left(a\,\frac{f'(X)}{f(X)} + b\,\frac{g'(Y)}{g(Y)}\right)^2\,\middle|\, Z\right]. \tag{5}$$

Averaging both sides of (5) over the distribution of Z gives us

$$\begin{aligned}(a+b)^2\,\mathbb{E}\left[\left(\frac{\rho_Z'(Z)}{\rho_Z(Z)}\right)^2\right] &\le \mathbb{E}\left[\left(a\,\frac{f'(X)}{f(X)}+b\,\frac{g'(Y)}{g(Y)}\right)^2\right]\\ &= a^2\,\mathbb{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]+b^2\,\mathbb{E}\left[\left(\frac{g'(Y)}{g(Y)}\right)^2\right],\end{aligned} \tag{6}$$

or equivalently, $(a+b)^2\,\mathrm{I}(X+Y)\le a^2\,\mathrm{I}(X)+b^2\,\mathrm{I}(Y)$. This proves the part (b). Optimizing part (b) over $\eta\in[0,1]$ immediately yields part (c). To justify part (a), we simply set $\sqrt{\eta}\,X$ and $\sqrt{1-\eta}\,Y$ in place of X and Y (respectively) in part (b), and use the elementary observation that the Fisher information I is homogeneous of degree -2 (i.e., $\mathrm{I}(\eta\,X)=\eta^{-2}\,\mathrm{I}(X)$). □

Remark 1. *If X and Y are i.i.d., then part (a) of Lemma 1 yields* $\mathrm{I}\left(\frac{X+Y}{\sqrt{2}}\right)\le \mathrm{I}(X)$.

Coming back to our discussion, it is well-known (among people who know it!) that a remarkable connection between entropy H and information I exists and is provided by the Ornstein-Uhlenbeck semigroup. This auxiliary diffusion semigroup is generated by the following (linear) SDE

$$\mathrm{d}X_t=-X_t\,\mathrm{d}t+\sqrt{2}\,\mathrm{d}B_t, \tag{7}$$

with $X_0:=X$ being a mean-zero random variable having law f and unit variance. The explicit solution of (7) is given by

$$X_t=X\mathrm{e}^{-t}+\sqrt{2}\int_0^t \mathrm{e}^{-(t-s)}\,\mathrm{d}B_s\overset{\mathrm{d}}{=}\sqrt{\mathrm{e}^{-2t}}\,X+\sqrt{1-\mathrm{e}^{-2t}}\,\mathcal{G}, \tag{8}$$

where $\overset{\mathrm{d}}{=}$ means equal in the sense of distribution. If we denote the density of X_t by f_t (hence $f_0=f$), then f_t satisfies the Kolmogorov forward equation $\partial_t f_t=\mathcal{L}^*[f_t]$, with

$$\mathcal{L}^*[\rho](x):=\rho''+(x\rho)'$$

for a probability density ρ. The following relation, which links the entropy H with the information I, is usually referred to as the *de Bruijn's identity*.

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{H}(X_t)=1-\mathrm{I}(X_t). \tag{9}$$

The proof of (9) is merely a simple calculation, which is skipped. Since f_t converges as $t \to \infty$ to the density of $\mathcal{G}$, upon integration the relation (9) is translated to

$$\mathrm{H}(X) - \mathrm{H}(\mathcal{G}) = \int_0^\infty \big(\mathrm{I}(X_t) - 1\big)\, \mathrm{d}t, \tag{10}$$

i.e., the entropy gap between X and $\mathcal{G}$ is the integral of the corresponding information gap. At this point, it should be clear that a information gap can be integrated along the semigroup to recover a entropy gap (a typical example exemplifying this principle lies the derivation of the famous Shannon-Stam inequality from the Blachman-Stam inequality). It is shown in [1] that, if X is a zero-mean random variable with variance 1 and finite Fisher information, whose density f satisfies a Poincaré inequality (2), then

$$\mathrm{I}(X) - \mathrm{I}\left(\frac{X+Y}{\sqrt{2}}\right) \geq \frac{\lambda}{2+2\lambda}(\mathrm{I}(X) - \mathrm{I}(\mathcal{G})), \tag{11}$$

in which Y is an i.i.d. copy of X and $\mathrm{I}(\mathcal{G}) = 1$.

Remark 2. *The proof of (11) is the hardest part of [1], which involves modern transportation argument as well as a tricky application of the calculus of variations.*

Once (11) is established, we want to integrate this relation along the Ornstein-Uhlenbeck semigroup to obtain (3), but we need to ensure that the Poincaré constant does not deteriorate. That being said, if the law of $X_0 := X$ satisfies a Poincaré inequality (2), we want to show that the law of X_t, which is defined by (8), satisfies the same Poincaré inequality as well for each $t > 0$. Since the Gaussian $\mathcal{G}$ has "the best" Poincaré constant (see for instance [4]), it can be shown that the Poincaré constant is strictly improved along the Ornstein-Uhlenbeck semigroup (we do not plan to prove such a semigroup-induced improvement of the Poincaré inequality here, and a short proof based on the law of total variance is given in pp 9 of [1]. The essence of such improvement lies in the fact the Gaussian $\mathcal{G}$ has "the best" Poincaré constant, say λ_*, so if we denote by λ the Poincaré constant of X_0, one can show "by interpolation" that the Poincaré constant of X_t, denoted by λ_t, satisfies $\lambda \leq \lambda_t \leq \lambda_*$ for all $t \geq 0$).

2 Entropy jump in the uniform reshuffling model?

Now we want to fight against the entropy jumps arising from the uniform reshuffling model, and to make our life simpler we assume throughout this section that parameter of the exponential equilibrium is 1, i.e., $m = 1$. At first glance, our problem is much more difficult than what is encountered in the classical setting mentioned in the previous section, because we are now considering multiplication of random variables $U(X+Y)$ instead of the simple re-scaled sum of i.i.d. random variables $\frac{X+Y}{\sqrt{2}}$. As we will see soon, in an attempt to apply a similar strategy as sketched in the previous section to our new problem, one will have to face several crucial difficulties.

To warm up the discussion, we introduce the following (nonlinear) SDE, known as the Cox-Ingersoll-Ross process

$$\mathrm{d}R_t = (1 - R_t)\,\mathrm{d}t + \sqrt{2R_t}\,\mathrm{d}B_t. \tag{12}$$

The (semi-)explicit solution of (12) is given by

$$R_t = 1 + (R_0 - 1)\,\mathrm{e}^{-t} + \sqrt{2}\int_0^t \sqrt{R_s}\mathrm{e}^{-(t-s)}\,\mathrm{d}B_s, \tag{13}$$

in which R_0 is non-negative a.s. having law ρ with unit mean. We remark that $R_t \xrightarrow{t\to\infty} \mathcal{E} \sim$ Exponential(1).

Remark 3. *In wikipedia, it is implicitly suggested that one can view R_t as a squared version of X_t (recall (8)) once we prescribe the initial condition associated with (12) to be $R_0 := X_0^2 = X^2$. However, this is misleading! In fact, $R_t := X_t^2$ does not satisfy the SDE (12) (https://quant.stackexchange.com/questions/31863/cir-process-from-ornstein-uhlenbeck-process, most importantly, the Brownian motions involved in such interpretations are different!) .*

If we denote the density of R_t by ρ_t (with $\rho_0 := \rho$), then ρ_t satisfies the Kolmogorov forward equation $\partial_t \rho_t = Q^*[\rho_t]$, with

$$Q^*[q](x) := (xq)'' - ((1-x)q)' = x(q'' + q') + q' + q$$

for a probability density q. Next, if we introduce the space $\mathcal{M}_1$ as the collection of probability densities having unit mean value whose supported is

contained in $[0,\infty)$, and define the J-information associated with a random variable V whose law g is an element of $\mathcal{M}_1$ by

$$\mathrm{J}(V) = \int_0^\infty x\,\frac{(g')^2}{g}\,\mathrm{d}x.$$

Parallel to the fact that the standard Gaussian $\mathcal{G}$ has the least Fisher information among $\mathbb{R}$-valued random variables having zero mean and unit variance, it is not hard to check that the exponential random variable $\mathcal{E} \sim \text{Exponential}(1)$ has the least J-information among $\mathbb{R}_+$-valued random variables having unit mean. Indeed, the J-information of $\mathcal{E}$ is 1. Suppose that V has (a differentiable) density $g \in \mathcal{M}_1$ and a finite J-information, then

$$\begin{aligned} 0 \le \int_0^\infty \left(\frac{g'}{g}+1\right)^2 x\,g\,\mathrm{d}x &= \int_0^\infty \left(x\frac{(g')^2}{g} + 2\,x\,g' + x\,g\right)\,\mathrm{d}x \\ &= \mathrm{J}(V) - 1 = \mathrm{J}(V) - \mathrm{J}(\mathcal{E}). \end{aligned}$$

The following relation, which links the entropy H with the J-information J, is a consequence of a straightforward computation.

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{H}(R_t) = 1 - \mathrm{J}(R_t). \tag{14}$$

Since ρ_t converges as $t \to \infty$ to the density of $\mathcal{E}$, upon integration the relation (14) is translated to

$$\mathrm{H}(R_0) - \mathrm{H}(\mathcal{E}) = \int_0^\infty \big(\mathrm{J}(R_t) - 1\big)\,\mathrm{d}t, \tag{15}$$

i.e., the entropy gap between R_0 and $\mathcal{E}$ is the integral of the corresponding J-information gap. We now make two (natural) conjectures related to the J-information.

Conjecture (1) For each $\mathbb{R}_+$-valued random variable X with unit mean, let Y be an i.i.d. copy of X, $\mathcal{E} \sim \text{Exponential}(1)$, and $U \sim \text{Uniform}[0,1]$ is independent of X, Y and $\mathcal{E}$. We have

$$\mathrm{J}(X) - \mathrm{J}(U(X+Y)) \ge 0. \tag{16}$$

Conjecture (2) Under the setting of **Conjecture (1)**,

$$\mathrm{J}(X) - \mathrm{J}(U(X+Y)) \ge c\,\big(\mathrm{J}(X) - \mathrm{J}(\mathcal{E})\big), \tag{17}$$

for some universal constant $c \in (0,1)$ (this might be too good to be true if we take into account of the discussions in the previous section, but if this can be proved, then "game over"!).

We emphasize here that (16) is an analog of a consequence of the classical Blachman-Stam inequality (recall Remark 1). But if we try to prove it using the same technique as those used in the proof of Lemma 1, we will have to deal with several difficulties.

Proof of Conjecture (1), Attempt Only! Assume that the density of X is $f \in \mathcal{M}_1$ and define $Z := U(X+Y)$, then the probability density of Z is

$$\rho_Z(z) = \int_{x \geq z} \frac{(f * f)(x)}{x} \, \mathrm{d}x = \int_{\mathbb{R}_+^2} \frac{\mathbb{1}_{[0,x+y]}(z)}{x+y} \, f(x) \, f(y) \, \mathrm{d}x \, \mathrm{d}y. \tag{18}$$

Also, the joint probability density for X and Z, denoted by $\rho_{X,Z}(x,z)$, is given by $\rho_{X,Z}(x,z) = \int_{\mathbb{R}_+} \frac{\mathbb{1}_{[0,x+y]}(z)}{x+y} \, f(x) \, f(y) \, \mathrm{d}y$. However, we can no longer express $\mathrm{d}\rho_Z/\mathrm{d}z$ in a similar form as (4). In fact, (4) uses integration by parts to shift a derivative in z to the derivative in x, and in the setting of the previous section, such integration by parts will not produce any "boundary terms". Here, $\mathrm{d}\rho_Z/\mathrm{d}z$ equals to

$$\rho_Z'(z) = -\frac{(f * f)(z)}{z}. \tag{19}$$

We don't know how to proceed from here... Also, the identity in (6) relies on the fact that $\mathbb{E}\left[\frac{f'(X)}{f(X)}\right] = \int_{\mathbb{R}} f'(x) \, \mathrm{d}x = 0$, but in our case the underlying space is $\mathbb{R}_+$ instead of $\mathbb{R}$, so it seems hard to avoid/cancel "boundary terms" when we evaluate certain integrals...

Unfortunately, **Conjecture (1)** is certainly not true. To see this, simply take f to be the uniform distribution, i.e., $f(x) = \frac{1}{2}\mathbb{1}_{[0,2]}(x)$. However, this is not really a bad news. As the uniform distribution does not have a differentiable density and its support is not the whole half line $[0,\infty)$, so this "counter example" will also make the statement "the standard Gaussian has the smallest Fisher information among zero-mean random variables with unit variance" questionable. So we should care about the validity of **Conjecture**

(1) among differentiable densities with full support on $\mathbb{R}_+$. I came to realize these thanks to a reading of [5].

Actually, from a different point of view, the J-information is a very natural quantity in order to have a quantitative measure of "closeness" of a $\mathbb{R}_+$-supported differentiable density ρ_X (we use ρ_X to denote the law of X) with unit mean to the exponential density $\mathrm{e}^{-x}\mathbb{1}_{x\geq 0}$. This "observation" is inspired from [7] in which the quantitative convergence property of a sum of (possibly dependent) Bernoulli random variables to a Poisson random variable has been investigated. Indeed, let us define our new J-information by

$$\mathrm{J}(X) = \mathbb{E}\left[X\left(\frac{\rho_X'(X)}{\rho_X(X)}+1\right)^2\right].$$

For any differentiable density ρ_X having full support on $\mathbb{R}_+$ and unit mean, define f by $f(x) = \rho_X(x)/\mathrm{e}^{-x}$ for all $x \geq 0$, a logarithmic Sobolev inequality for the exponential distribution (see Exercise 6.11 of [6]) yields that

$$\begin{aligned}\mathrm{D_{KL}}\left(\rho_X \,||\, \text{Exponential}(1)\right) = \mathrm{Ent}(f) &\leq 4\int_0^\infty \mathrm{e}^{-x}\, x\, \frac{\left(\rho_X'(x)\right)^2}{4\,\rho_X(x)}\,\mathrm{d}x \\ &= \mathbb{E}\left[X\left(\frac{\rho_X'(X)}{\rho_X(X)}+1\right)^2\right] = \mathrm{J}(X).\end{aligned}$$

Thus the smaller the $\mathrm{J}(X)$, the closer the random variable X will be to the exponential random variable $\mathcal{E}$. In fact, we have already encountered this new J-information at the beginning of page 7, it can be readily seen that this new definition of J-information differs from our old J-information only by a additive constant (equal to 1).

3 Appendix

3.1 Proof of $\mathrm{H}\left(U(X+Y)\right) \leq \mathrm{H}(X)$

We give a relatively terse proof of the somewhat non-trivial observation that $\mathrm{H}\left(U(X+Y)\right) \leq \mathrm{H}(X)$, based on [8]. Let us denote $X_1 = X$ and X_2 to be an independent copy of X_1. We also set $Z_1 = U(X_1+X_2)$ and $Z_2 = (1-U)(X_1+X_2)$. The following chain of relations lead us to the advertised conclusion:

$$2\,\mathrm{H}(Z_1) = \mathrm{H}(Z_1) + \mathrm{H}(Z_2) \leq \mathrm{H}((Z_1,Z_2)) \leq \mathrm{H}((X_1,X_2)) = 2\,\mathrm{H}(X_1), \quad (20)$$

where the first inequality follows from the well-known super-additivity of H (in fact, the quantity $\mathrm{I}(Z_1; Z_2) = \mathrm{H}((Z_1, Z_2)) - \mathrm{H}(Z_1) - \mathrm{H}(Z_2) \geq 0$ is called the *mutual information* between Z_1 and Z_2) and we only need to prove the validity of the second inequality in (20). We denote by $\rho(x, y)$ the joint density of the pair (Z_1, Z_2), a simple calculation yields that

$$\rho(x, y) = \int_0^{x+y} \frac{1}{x+y} f(z)\, f(x+y-z)\, \mathrm{d}z = \phi(x+y), \tag{21}$$

where

$$\phi(m) := \frac{1}{m} \int_0^m f(z)\, f(m-z)\, \mathrm{d}z = \mathbb{E}\left[f(S)\, f(m-S)\right] \text{ with } S \sim \mathrm{Uniform}[0, m].$$

Therefore, by the convexity of $x \mapsto x\, \ln(x)$ and Jensen's inequality,

$$\begin{aligned}
\mathrm{H}((Z_1, Z_2)) &= \int_{x=0}^{\infty} \int_{y=0}^{\infty} \phi(x+y)\, \ln \phi(x+y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_{m=0}^{\infty} \int_{z=0}^{m} \phi(m)\, \ln \phi(m)\, \mathrm{d}z\, \mathrm{d}m = \int_{m=0}^{\infty} m\, \phi(m)\, \ln \phi(m)\, \mathrm{d}m \\
&\leq \int_{m=0}^{\infty} m\, \mathbb{E}\left[f(S)\, f(m-S)\, \ln\left[f(S)\, f(m-S)\right]\right] \mathrm{d}m \\
&= \int_{m=0}^{\infty} \int_{s=0}^{m} f(s)\, f(m-s)\, \ln\left[f(s)\, f(m-s)\right] \mathrm{d}s\, \mathrm{d}m \\
&= \mathrm{H}((X_1, X_2)).
\end{aligned}$$

Remark 4. *In fact, with a little bit of extra work, one can show that*

$$\mathrm{H}\left((Z_1, Z_2), (X_1, X_2)\right) \leq \mathrm{H}((Z_1, Z_2)) = \mathrm{H}\left((X_1, X_2), (Z_1, Z_2)\right) \leq \mathrm{H}((X_1, X_2)).$$

References

[1] Keith Ball, Franck Barthe, and Assaf Naor. *Entropy Jumps in the Presence of a Spectral Gap*, 2003.

[2] N. M. Blachma. *The Convolution Inequality for Entropy Powers*, 1965.

[3] C. Villani. *Entropy Methods for the Boltzmann Equation*, 2008.

[4] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, 2013.

[5] Elke Uhrmann-Klingen. *Minimal Fisher Information Distributions with Compact-Supports*, 1995.

[6] Stephane Boucheron, Gabor Lugosi, and Olivier Bousquet. *Concentration Inequalities: A Nonasymptotic Theory of Independence*, 2013.

[7] I. Kontoyiannis, P. Harremoes, and O. Johnson. *Entropy and The Law of Small Numbers*, 2005.

[8] S. M. Apenko. *Monotonic entropy growth for a nonlinear model of random exchanges*, 2013.