Jaynes & Shannon's Constrained Ignorance

TomQTX

Sant Job Skolaj-Lise, 42 Kerguestenen Straed,
56100 BroAnOriant, Breizh[*]
(Dated: June 30, 2021)

In this simple article, based solely on what is written in Edwin T. Jaynes' book *Probability Theory: The Logic of Science* [1], we look at a way to derive the expression for Shannon's entropy from a purely variational approach, using constraints. The results are therefore not new, but the context, however, may give a coherent formalism, where the maximum entropy principle appears naturally. In the first part, we define general "ignorance" and, as we re-set the context, express how the maximum entropy principle would fit correctly. Then in the second part, we derive the somehow general expected expression for the entropy, but using two approaches : the first one where we, biased, know a vague shape for the entropy, and the second one where we just look at the general case, knowing nothing about its expression. The first approach, less appealing in our opinion, leads to some difficulties in order to "reattach" correctly the pieces of the puzzle, therefore being more a training model. The second approach, however, seems to lead to better results.

Contents

I. INTRODUCTION

In the following, we try to recover the expression of the entropy from a low level approach, with only few assumptions on the context, in the same way as done in Jaynes' book [1]. We define a quantity we call *ignorance* instead of *incertitude* as it seems to get better with the constraints, and knowing that it should be continuous, symmetrical and be expressed similarly in any sub-situation, we derive[5] results leading to the expression of the entropy. We are undeniably biased as we know what we should expect, but keeping it general, we look at some subtleties which express our (real) "ignorance" on how to set the calculations.

The resulting formalism seems appealing and if not yet done, may give some other insights.

This has been done trying to achieve a certain point of rigor. However, errors (in the notation, in the calculations, ..) may have been done (if so, we would be glad to be aware of it). Moreover, we present a work which reports suite of ideas, therefore, this may lack of conciseness, when parts, as the following one, may be too quick, and we encourage the reader to read [1] first, even if it is not necessary.

II. IGNORANCE

Suppose a variable x can take on n different discrete values $(x_1, .., x_n)$, which correspond to n different propositions $(A_1, .., A_n)$. As Jaynes, we are asking

What probabilities $(p_1, .., p_n)$ will Jaynes'robot assign to the possibilities $(x_1, .., x_n)$? .

A. What are the knowledge at this point ?

- The sum of all probabilities is equal to one

$$\sum_{i=1}^{n} p_i = 1 \tag{1}$$

therefore, the *"ignorance of the knowns"* associated to this information is ... 0 and could be expressed simply as

$$H[\lambda, p_1, .., p_n] = \lambda_0(x) \left(\sum_{i=1}^{n} p_i - 1 \right) \tag{2}$$

where $h[p_1, .., p_n] = \left(\sum_{i=1}^{n} p_i - 1 \right)$ is a constraint obtained after derivation with respect to $\lambda_0(x)$, a general Lagrange multiplier.

1. About this Lagrange multiplier : so far x is just a set of yet-to-be-determined variables, and in this actual work, we will consider it as a constant,or a least, independent of the probabilities p_i. However, it may be interesting to put some dependencies, as

[*]Electronic address: XX

it may lead to possible ways to express problems in this formalism which would not be described in the maximization of the entropy method derived as "usual"[6]

2. Expression of the ignorance of the knowns in Eq.(2) takes a simple form which, depending on the context, may not give satisfying results (the final ignorance is constant for instance), leading to the following statement : a more general way of expressing such ignorance would be from the general expression

$$H[\lambda, p_1, .., p_n] = \frac{1}{m}\lambda_0(x)\left(\sum_{i=1}^{n} p_i - 1\right)^m \tag{3}$$

for all $m \in \mathbb{N}^*$, where the factor $\frac{1}{m}$ being here to not have to rescale after derivation. m is necessarily a positive integer as, if one set it negative, the ignorance would give and infinity due to the division by the constraint. However, as suggested by [4] and explained later, one could do the calculation and rescale it by the infinity factor, expressed for instance in the term $ln(x + y - 1)$ as $x + y \to 1$.

However, whatever the choice of $m > 0$, due to constraint, this ignorance will always give 0 in the final expression and we expect that, in this formalism, it will change nothing on the results : two robots doing the same calculations but with different m, are expected at the end to derive the same results about the expression of the ignorance/probabilities. This will be confirmed at the end, with some settings better than others (in fact, the simplest it stays, the better it is).

- Let us say that we know another set of k constraints about the probabilities, $k \leq n$,

$$f_i[p_1, ...p_n] = 0, \qquad \forall 0 < i \leq k \tag{4}$$

the associated ignorance, also null, would be in the same way as previously

$$H[\lambda_i, p_1, .., p_n] = \sum_{i=1}^{k} \lambda_i(x) f_i[p_1, ...p_n] \tag{5}$$

For instance, it could be that then $p_1 = 2p_2$. What are the consequences in this formalism ?

B. Requirements on the ignorance

So far we have dealt with known notions, leading to null ignorance, but ignorance here is taken in its literal form *"lack of knowledge or information"* and H as defined before is a way to put a "degree" about the global situation at the end. Therefore, the requirement we think of would be [3]

1. **Continuity** : H should be continuous, so that changing the values of the probabilities by a very small amount should only change the ignorance by a small amount.

2. **Symmetry** : H should be unchanged if the outcomes p_i are re-ordered.

As we see the derivation of Shannon entropy in [1] p 347, originally express in the work of Shannon [2], one could think about ignorance/uncertainties as the total *expected*/average ignorance, having put all the information we know at the beginning of the calculations(this principle of *not omitting any information* seems important in Jaynes mind and would have been respected in any case). Moreover, in his book [1], Jaynes was arguing that we should obtain at some point a "variational approach"[7] and this is exactly what is done here.

Therefore, one would expect the total ignorance to be in this framework such that

$$H_{tot} = H_{knowns} + H_{unknowns}. \tag{6}$$

whith $H_{knowns} = 0$ and $H_{unknowns} = \sum_i p_i H_i$, the average of the ignorances. However, at the end, we will not assume this expression, rather setting only $H_{unknowns} = \sum_i H_i$, leading to .. better results.

1. *At first,* $H_{unknowns} = \sum_i p_i H_i$

As somehow explained by Shannon and Jaynes, let us imagine that at first the robot is aware of three propositions (A_1, A_2, A_3) of unknown probabilities p_1, p_2 and p_3. The ignorance of the robots would therefore be

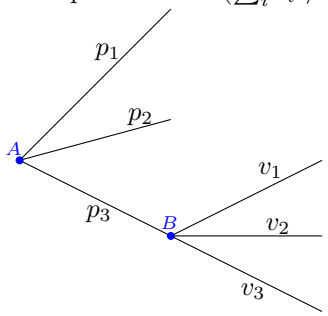
$$H[p_1, p_2, p_3] = \frac{\lambda(x)}{m}\left(\sum_{i=1}^{3} p_i - 1\right)^m + \sum_{i=1}^{3} p_i H_i[p_i]. \tag{7}$$

In the case $n = 1$, it is just

$$H[p_1, p_2, p_3] = \\ \lambda(p_1 + p_2 + p_3 - 1) + p_1 H_1[p_1] + p_2 H_2[p_2] + p_3 H_3[p_3]$$

2. *After an update*

Then, as illustrated below, the robot learns that the third propositions may in fact be a combination of three (or less or more) sub-propositions of probabilities (v_1, v_2, v_3) with $v_1 + v_2 + v_3 = p_3$, with thus $v_i = p(A'_i|A_3)$. The situation is represented by the tree below, which is not here a tree diagram of probabilities ($\sum_i v_i \neq 1$).



The "sub-ignorance" for the proposition A_3 would therefore be written somehow as

$$p_3 \times H_3[p_3] = \frac{\mu(x)}{m}\left(\sum_{i=1}^{3} v_i - p_3\right)^m + \sum_{i=1}^{3} v_i H_i[v_i] \tag{8}$$

leading to an update of the previous ignorance,

$$H[p_1, p_2, p_3] = \\ \lambda(p_1 + p_2 + p_3 - 1) + p_1 H_1[p_1] + p_2 H_2[p_2] \\ +\mu(v_1 + v_2 + v_3 - p_3) + v_1 H'_1[v_1] + v_2 H'_2[v_2] + v_2 H'_2[v_2]. \tag{9}$$

In this case we are dealing with another constraint $\mu(x)$, illustrating what said previously for the ignorance in Eq.(5).

After the update, the robot is having now five propositions A_i, $1 \leq i \leq 5$ in total, of possibilities p_i, and so has an updated expected ignorance

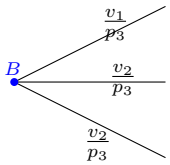
$$H[p_1, p_2, p_3, p_4, p_5] = \frac{\lambda(x)}{n}\left(\sum_{i=1}^{5} p_i - 1\right)^n + \sum_{i=1}^{5} p_i H_i[p_i] \quad (10)$$

3. Commentaries

Moreover, after dividing the ignorance in Eq.(8) by p_3, setting $w_i = \frac{v_i}{p_3}$, therefore $\sum_i \frac{v_i}{p_3} = \sum_i w_i = 1$, and rescaling the Lagrange multiplier $\mu(x) \to \frac{\mu(x)}{p_3}$, one "get back" probabilities such that

$$H_3\left[\frac{v_1}{p_3}, \frac{v_2}{p_3}, \frac{v_3}{p_3}\right] = \frac{\mu(x)}{m}\left(\sum_{i=1}^{3} w_i - 1\right)^m + \sum_{i=1}^{3} w_i H_i[w_i] \quad (11)$$

where p_3 is considered here as a constant parameter. Ignorance in Eq.(11) is simply the one the robot would have if it does not know about the previous propositions other than A_3, thus its state of knowledge starting at the node B in the figure before.



4. In a nutshell ..

Taking into account the possible updates we yet do not know, we see that the general expression of the "Ignorance" we are dealing with so far is therefore, as we are biased,

$$H_{tot}[p_1, .., p_n] = H[\lambda_\mu, p_1, .., p_n] + \sum_{i=1}^{n} p_i H_i\left[\frac{v_1}{p_i}, .., \frac{v_r}{p_i}\right], \quad (12)$$

but could also be set generally in the same way as

$$H_{tot}[p_1, .., p_n] = H[\lambda_\mu, p_1, .., p_n] + \sum_{i=1}^{n} H_i\left[\frac{v_1}{p_i}, .., \frac{v_r}{p_i}\right], \quad (13)$$

with at least in our case

$$H[\lambda_\mu, p_1, .., p_n] = \frac{1}{m}\lambda(u)\left(\sum_{i=1}^{n} p_i - 1\right)^m = 0 \quad (14)$$

is the constraint dealing with the information that we know, therefore with no ignorance about it, and $H_i[p_i]$ the *local* ignorance about the situation on each branch.

Consequently, we could ask *"what are the probabilities that minimize/maximize our ignorance ?"* which, at first, would lead us to an expression for the ignorance which should be at least similar to the one of the information entropy, and then, to the (usual) expression of the probabilities : as a consequence, this is just the maximum entropy principle, which makes also sense in this formalism.

III. EXTREMIZATION OF THE IGNORANCE

In the following we study the case where we express the ignorance for different values of m, first as a training and then to demonstrate in the general case what gives the need to minimize the ignorance in both approaches.

A. Knowing part of the expression of H

From Eq.(14), we consider the constraint

$$H[\lambda_u, p_1, .., p_n] = \lambda(u)\frac{1}{m}\left(\sum_{i=1}^{n} p_i - 1\right)^m = 0 \quad (15)$$

for $m \geq 1$, whose variation with respect to p_i gives simply

$$\delta H[\lambda_u, p_1, .., p_n] = \sum_{w=p_i} \delta w \lambda(u)\left(\sum_{i=1}^{n} p_i - 1\right)^{m-1}$$

The variation of the total ignorance would leads then to

$$\begin{aligned}\delta H \;=\; & \sum_{w=p_i} \delta w \left[\lambda(u)\left(\sum_{i=1}^{n} p_i - 1\right)^{m-1} + H_w(x_1, .., x_n)\right. \quad (16)\\ & \left. - \sum_{i=1}^{m} x_i \times \frac{\partial H_w}{\partial x_i}(x_1, .., x_n)\right].\end{aligned}$$

where v_i has been extracted from H_w via $x_i = \frac{v_i}{p_j}$, as it will be clear in the following. However, we know that $\sum_j x_j = 1$, therefore this information should appear at some point. In fact, we should have thought about it even in the previous part where $m = 1$.

Will it change something ? To see it, we will set this information on the x_i at two places, with factor p and q taken values in $\{0; 1\}$ in Eq.(12) and consider the sub-propositions of probabilities x_i

1. The general equations

Considering the expression

$$H_{tot}[p_1, .., p_n] = H[\lambda_\mu, p_1, .., p_n] + \sum_{i=1}^{n} p_i H_i\left[\frac{v_1}{p_i}, .., \frac{v_r}{p_i}\right], \quad (17)$$

from the reasoning in Eq.(11), we could set

$$\begin{aligned} H_i[x_1, .., x_r] \;&=\; \frac{\mu(u)}{b}\left(\sum_{j=1}^{r} x_j - 1\right)^b + \sum_{j=1}^{r} x_j H_j[x_j] \quad (18)\\ &=\; \frac{\mu(u)}{b}\left(\sum_{j=1}^{r} x_j - 1\right)^b + f[x_j] \quad (19)\end{aligned}$$

as the unknown variable in our calculations is the H_k. Moreover, we could have continued and express again the sub-ignorance $H_j[x_j]$ further, in terms of the sub-sub-ignorance, but it would have been redundant as we would process to the same calculations at each node of the probability tree, again and again. Therefore here, our unknown variable is simply $f[x_j]$ which represents the situation from the probability tree of nodes A and B_i for each p_i, shown before.

$$\delta H = \sum_{w=p_i} \delta w \left[\lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} + H_w \left[\frac{v_1}{w}, \frac{v_2}{w}, ..., \frac{v_m}{w} \right] + w \times \sum_{i=1}^{m} \frac{\partial \left(\frac{v_i}{w} \right)}{\partial w} \times \frac{\partial H_w}{\partial \left(\frac{v_i}{w} \right)} \left[\frac{v_1}{w}, \frac{v_2}{w}, ..., \frac{v_m}{w} \right] \right] \quad (20)$$

$$= \sum_{w=p_i} \delta w \left[\lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} + H_w \left[x_i \right] - \sum_{i=1}^{m} x_i \times \frac{\partial H_w}{\partial x_i} \left[x_1, .., x_n \right] \right] \quad (21)$$

$$= \sum_{w=p_i} \delta w \left[\lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} + \left(p \frac{\mu(u)}{b} \left(\sum_{j=1}^{r} x_j - 1 \right)^{b} + f[x_j] \right) - \sum_{i=1}^{m} x_i \frac{\partial}{\partial x_i} \left(q \times \frac{\mu(u)}{b} \left(\sum_{j=1}^{r} x_j - 1 \right)^{b} + f[x_j] \right) \right] \quad (22)$$

$$\Leftrightarrow \; 0 = \lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} + p \frac{\mu(u)}{b} \left(\sum_{j=1}^{r} x_j - 1 \right)^{b} + f[x_i] - \left(\sum_{i=1}^{r} x_i \right) \left[q \times \mu(u) \left(\sum_{j=1}^{r} x_j - 1 \right)^{b-1} + \frac{\partial f[x_i]}{\partial x_i} \right] \quad \forall \; \delta w \quad (23)$$

$$\Leftrightarrow \; 0 = \lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} + p \frac{\mu(u)}{b} \left(\sum_{j=1}^{r} x_j - 1 \right)^{b} - q \times \mu(u) \left(\sum_{i=1}^{r} x_i \right) \left[\left(\sum_{j=1}^{r} x_j - 1 \right)^{b-1} \right] + f[x_i] - \sum_{i=1}^{r} x_i \frac{\partial f[x_i]}{\partial x_i} \quad (24)$$

$$\Leftrightarrow \; 0 = \lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} + \frac{\mu(u)}{b} \left(\sum_{j=1}^{r} x_j - 1 \right)^{b-1} \left[p(\sum_{j=1}^{r} x_j - 1) - bq \left(\sum_{i=1}^{r} x_i \right) \right] + f[x_i] - \sum_{i=1}^{r} x_i \frac{\partial f[x_i]}{\partial x_i} \quad (25)$$

leading us therefore to solve in the general case

$$0 = \lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} + \frac{\mu(u)}{b} \left(\sum_{j=1}^{r} x_j - 1 \right)^{b-1} \left[(p - bq) \left(\sum_{i=1}^{r} x_i \right) - p \right] + f[x_i] - \sum_{i=1}^{r} x_i \frac{\partial f[x_i]}{\partial x_i} \quad (26)$$

Commentaries and assumptions at this point :

- we put p and q in order to distinguish from where the information that $\sum_{i=1}^{r} x_j - 1 = 0$ comes from : as it appears twice in the calculation, it may be redundant to do so and we may consider that one of the two terms could be superfluous. However, it could also play a major role in the expression of the solution when derived, and therefore we keep the q in front, as such.
- $b = m$: not really an assumption as by redundancy it has to be true [assuming that Ignorance at each node has the same expression]
- For simplicity, we are looking only at two sub-propositions of possibilities x and y such that $x + y = 1$. We guess that any sub-situation can be seen as : A_x "something happens", A_y "something does not", and by recurrence at each node it should be true. For instance, from proposition A_3 we could have sub-proposition $A_{3,1} = A_3$ of probability $x = 1$.
- Regarding $\alpha(p_i, m) = \lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1}$, it should be "just" a constant for $f[x_j]$, that is, x_i are considered now as independent of p_j.

Therefore, in the following, we will try to solve

$$0 = \alpha(m) + \frac{\mu}{b} (x + y - 1)^{b-1} \left[(p - bq)(x + y) - p \right] + f[x_i] - x \frac{\partial f[x, y]}{\partial x} - y \frac{\partial f[x, y]}{\partial y} \quad (27)$$

This equation is a linear first order PDE we can rewrite as

$$a(x, y) u_x + b(x, y) u_y = f(x, y, u) \quad (28)$$

with $u = f(x, y)$ such that $u_z = \partial_z u = \frac{\partial u}{\partial z}(x, y)$ for $z = x, y$, and

$$\begin{aligned} a(x, y) &= x, \\ b(x, y) &= y, \\ f(x, y, u) &= \alpha(m) + \frac{\mu}{b} (x + y - 1)^{b-1} \left[(p - bq)(x + y) - p \right] + u \end{aligned}$$

Using the method of characteristics, we have to solve

$$\frac{dx}{a} = \frac{dy}{b} = \frac{du}{f} \quad (29)$$

that is

$$\frac{dx}{x} = \frac{dy}{y} = \frac{du}{\alpha + \frac{\mu}{b} (x + y - 1)^{b-1} \left[(p - bq)(x + y) - p \right] + u} \quad (30)$$

1. From the two first ones, we get

$$\frac{dx}{x} = \frac{dy}{y} \Leftrightarrow y = c_1 x, \quad (31)$$

and so $c_1 = \frac{y}{x}$

2. From the second ones, setting $(1+c_1)x = cx$,

$$\frac{dx}{x} = \frac{du}{\alpha(m) + \frac{\mu}{b}(cx-1)^{b-1}[cx(p-bq)-p] + u} \quad (32)$$

$$\Leftrightarrow \frac{dx}{x} = \frac{du}{\alpha(m) + \frac{\mu}{b}(cx-1)^{b-1}[(p-bq)cx-1] + u} \quad (33)$$

$$\Leftrightarrow \frac{du}{dx} = \frac{\alpha(m) + \frac{\mu}{b}(cx-1)^{b-1}[(p-bq)cx-1] + u}{x} \quad (34)$$

$$\Leftrightarrow \frac{du}{dx} - \frac{u}{x} = \frac{\alpha(m) + \frac{\mu}{b}(cx-1)^{b-1}[(p-bq)cx-p]}{x} \quad (35)$$

3. Multiplying both side by $\frac{1}{x}$, we get

$$\frac{1}{x}\frac{du}{dx} - \frac{u}{x^2} = \frac{d}{dx}\left(\frac{u}{x}\right) = \frac{\alpha + \frac{\mu}{b}(cx-1)^{b-1}[(p-bq)cx-p]}{x^2} \quad (36)$$

and we have therefore to solve

$$\frac{u(x,y)}{x} = \int_a^x \frac{\alpha + \frac{\mu}{b}(c\xi-1)^{b-1}[(p-bq)c\xi-p]}{\xi^2} d\xi + c_2\left(\frac{y}{x}\right) \quad (37)$$

$$= \beta(x,y) - \frac{\alpha}{x} + \frac{\mu}{b}c\int_{ca}^{cx} \frac{dw}{w^2}(w-1)^{b-1}((p-bq)w-p) \quad (38)$$

where $\beta(x,y) = c_2\left(\frac{y}{x}\right) + constants$, a is a constant and we set $w = c\xi$ for more simplicity.

We can therefore express the "solution" as

$$u(x,y) = \beta\left(\frac{y}{x}\right)x - \alpha(m) + \frac{\mu(cx)}{b}(pI_1 - bqI_2) \quad (39)$$

where

$$I_1 = \int_{ca}^{cx} \frac{dw}{w^2}(w-1)^b \quad (40)$$

$$I_2 = \int_{ca}^{cx} \frac{dw}{w}(w-1)^{b-1} \quad (41)$$

2. What are the results of $b=1$ or $b=2$?

a. case where $b=1$

In this case, setting $m = b = 1$, we have $\alpha(m) = \lambda$ the Lagrange multiplier (here considered as constant). Regarding the integrals,

$$I_1 = \int_{ca}^{cx} \frac{dw}{w^2}(w-1) = \int_{ca}^{cx} dw\left(\frac{1}{w} - \frac{1}{w^2}\right) \quad (42)$$

$$= ln(cx) + \frac{1}{cx} + const \quad (43)$$

$$I_2 = \int_{ca}^{cx} \frac{dw}{w} = ln(cx) + const \quad (44)$$

and therefore

$$u(x,y) = \beta\left(\frac{y}{x}\right)x - \lambda + \mu(cx)\left(p \times \left(ln(cx) + \frac{1}{cx}\right) - q \cdot ln(cx)\right)$$

$$= \beta\left(\frac{y}{x}\right)x - \lambda + \mu p + \mu(p-q)(cx)ln(cx) \quad (45)$$

from which we could say that

- if $u(x,y) = constant \times (x+y)$, as the constraints $x+y = cx \to 1$ will be applied at the end, this will lead $u(x,y)$ to be only a constant, and we could rescale it in order to absorb it. However, the drawback of this formulation is also that .. $ln(cx) \to 1$ as we will talk later.
- as $p=1$, then we have a $\alpha - \mu$ term, which corresponds to, as $\alpha_{b=1} = \lambda(u)$, $\lambda(u) - \mu(u)$. Our guess would be that at each node and sub-nodes, we have the same "kind of information", and therefore we would put $\lambda(u) = \mu(u)$, leading $\lambda(u) - \mu(u)$ to be zero. In the other way around, we would just have either to rescale by removing the constants, or either express any quantity in terms of $H[p] - H_0[P]$ where $H_0[p]$ is a reference value (the minimum, maximum, .. of the ignorance).
- if $p=q$, the logarithm term will disappear, at least for the case $b=1$. As we would like ignorance to decrease when the probabilities are known to be 0 or 1, either
  - we set $p=0$ and $q=1$, and we have with the choice of $\mu > 0$ the kind of expression we need (after rescaling the expression due to terms as α),
  - or we set $p=1$, $q=0$ and choosing $\mu < 0$ (equivalent to $\mu\left(1 - \sum_i x_i\right)$) would give us

$$u(x) = \beta x + \mu(cx)ln(cx) \quad (46)$$

$$u(x,y) = \beta\left(\frac{y}{x}\right)x + \beta\left(\frac{x}{y}\right)y + \mu(x+y)ln(x+y) \quad (47)$$

b. case where $b=2$

In this case

$$I_1 = \int_{ca}^{cx} \frac{dw}{w^2}(w-1)^2 = \int_{ca}^{cx} dw\left(1 - \frac{2}{w} + \frac{1}{w^2}\right) \quad (48)$$

$$= cx - 2ln(cx) - \frac{1}{cx} + const \quad (49)$$

$$= \frac{(cx-1)(cx+1)}{cx} - 2ln(cx) \quad (50)$$

$$I_2 = \int_{ca}^{cx} \frac{dw}{w}(w-1) = \int_{ca}^{cx} dw\left(1 - \frac{1}{w}\right) \quad (51)$$

$$= cx - ln(cx) \quad (52)$$

and therefore, the ignorance would be

$$u(x,y) = \beta x - \alpha + \frac{\mu(cx)}{2} \times \left[p\frac{(cx-1)(cx+1)}{cx} - 2q(cx)\right]$$

$$+\frac{\mu(cx)}{2} \times [-2pln(cx) - 2q(-ln(cx))] \quad (53)$$

thus

$$u(x,y) = \beta x - \alpha + \frac{\mu}{2} \times \left[p(cx-1)(cx+1) - 2q(cx)^2\right]$$

$$-\mu(p-q)(cx)ln(cx). \quad (54)$$

We could say also that

- regarding $-\alpha+\frac{\mu}{2}p(cx-1)(cx+1)$, as we would expect that $m=b=2$, then

$$\alpha(p_i)=\lambda(u)\left(\sum_{i=1}^{n}p_i-1\right)\sim\lambda\times(cp-1) \tag{55}$$

both terms are constraints in $c\xi-1$ and so will vanish.

- Now, with this in mind, comparing Eq.(45) et Eq.(54), as we expect the ignorance to remain the same whatever the choice of the power $m=b$ of the constraint, we would expect no terms in $q(cx)^s$ for different values of s depending on b. So, if this has to be true, then, we should set $q=0$ for the theory to remain coherent. However, setting $p=0$ and $q=1$ gives a $\mu(cx)ln(cx)$ term as for the case where $b=1$. However, as such, we would have to consider $\mu<0$ in order for the ignorance to behave correctly.

- The case $q=0$, $p=1$ and $\mu>0$ is of interest as it leads to the expression for the ignorance, after the constraint being applied, to be similar to Eq.(47), that is

$$u(x,y)=\beta\left(\frac{y}{x}\right)x+\beta\left(\frac{x}{y}\right)y-\mu(x+y)ln(x+y) \tag{56}$$

As a consequences of the choices before, the expression for the total ignorance in Eq.(12) would be somehow

$$H=\frac{\lambda(u)}{2}\left(\sum_{i=1}^{n}p_i-1\right)^2 \tag{57}$$

$$+\sum_{i=1}^{n}p_i\times\left[\beta\left[\frac{v_1}{p_i},\frac{v_2}{p_i}\right]+\lambda\left(\frac{v_1}{p_i}+\frac{v_2}{p_i}\right)ln\left(\frac{v_1}{p_i}+\frac{v_2}{p_i}\right)\right] \tag{58}$$

with $x=\frac{v_1}{p_i}$, $y=\frac{v_2}{p_i}$ and $v_1+v_2=p_i$, $\mu=\lambda$.
When we will apply it to a situation, the constraint will be fulfilled and so H will reduce roughly to

$$H=[0]+[1]+\lambda\sum_{i=1}^{n}p_i\times ln\left(\frac{v_1}{p_i}+\frac{v_2}{p_i}\right) \tag{59}$$

as for the case where $b=1$. Here $[X]$ means terms linear in X, and so having no consequences as the constrained are applied, and after rescaling.

Commentaries :

- the case $b=2$ is appealing in the sens that for a variational problem in physics, H would be similar to a Lagrangian/Hamiltonian where velocities of potential energies are globally in ξ^2. However, here it seems to be independent of the power, therefore this analogy is just to say.

- **More importanlty**, in order to apply the same logic at each node of the tree diagram, from Eq.(8) with $v_jH_j[v_j]$, to Eq.(11) with $\frac{v_j}{p_i}H_j\left[\frac{v_j}{p_i}\right]$, we did a mixed-up change of variables which, even if it was logic regarding Eq.(10), was also done in H_j. Consequently, due to the $\frac{1}{p_i}$ factor, differentiating with respect to p_i, we obtained a negative sign which leads to a logarithm solution for the ignorance (not obtained by a plus sign). But we artificially pass from $\sum_j v_j=p_i$ to $\sum_j x_j=1$, *i.e.* from $v_1+v_2=p_3$ to $x+y=1$, and so to $p_i\times ln\left(\frac{v_1}{p_i}+\frac{v_2}{p_i}\right)$ instead of $p_i\times ln(p_i)$ as expected. One way to cure it would have to look at $H=...+p_iH_i[p_i]=...+v_jH_j[v_j]\rightarrow ..+p_i\frac{v_j}{p_i}H_j\left[p_i\times\frac{v_j}{p_i}\right]$ but differentiating w.r.t p_i would give much more complicated equations, and this would have been a patch to an artificially ill defined solution, as the next part shows a better way of doing it.

  Relately, as shown in Eq.(56), we see a logarithm term which should

  a) go to zero as $x+y=1$ (except if we multiply it by p_i as said just before)

  b) at this sub-node where a proposition is separated in more sub-propositions of possibilities x and y, also give us the relation

$$\beta(x,y)+(x+y)ln(x+y)\quad\rightarrow\quad xln(x)+yln(y). \tag{60}$$

  As $\beta(x,y)$ has not yet specified, we could take a specific value to remove the unwanted term, but this is again an artificial way of doing.

As a consequence, as this first approach seems unsatisfying in our opinion, and as we expect similar expression for the entropy for all value of $m=b$, we will stop here and look at a more general and promising way at this point.

B. Specifying nothing about H_i

1. *Derivation of the solution*

Starting from the general expression

$$H[p_1,p_2,...,p_n]=\frac{\lambda(x)}{m}\left(\sum_{i=1}^{n}p_i-1\right)^m+\sum_{i=1}^{n}H_i\left[\frac{v_1}{p_i},..,\frac{v_r}{p_i}\right] \tag{61}$$

where we only require H_i on the r "sub"-probabilities at each sub-node for each p_i (to recall, this is more coherent as each sub-tree is a probability tree, and as usual, the probabilities are multiplied from branch to branch the more we know about sub-situations, *i.e.* sub-propositions). As before,

$$\delta H=\sum_{w=p_i}\delta w\left[\lambda(u)\left(\sum_{i=1}^{n}p_i-1\right)^{m-1}+\frac{\partial}{\partial w}\left(H_w\left[\frac{v_1}{w},..,\frac{v_r}{w}\right]\right)\right] \tag{62}$$

$$=\sum_{w=p_i}\delta w\left[\lambda(u)\left(\sum_{i=1}^{n}p_i-1\right)^{m-1}+p\frac{\mu(u)}{b}\left(\sum_{j=1}^{r}\left(\frac{v_j}{w}\right)-1\right)^b\right.$$

$$\left.+\sum_{j=1}^{r}\frac{\partial\left(\frac{v_j}{w}\right)}{\partial w}\frac{\partial}{\partial\left(\frac{v_j}{w}\right)}\left(q\frac{\mu(u)}{b}\left(\sum_{j=1}^{r}\left(\frac{v_j}{w}\right)-1\right)^b+H_w\left[\frac{v_i}{w}\right]\right)\right] \tag{63}$$

where we put the constraints on $\frac{v_j}{w}$ inside (with q) or outside (p) the derivation in order to keep it general and see how they impact the results.

$$\delta H = \sum_{w=p_i} \delta w \left[\lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} + p \frac{\mu(u)}{b} \left(\sum_{j=1}^{r} \left(\frac{v_j}{w} \right) - 1 \right)^{b} \right.$$
$$\left. - \frac{1}{w} \sum_{j=1}^{r} \left(\frac{v_j}{w} \right) \left(q\mu(u) \left(\sum_{j=1}^{r} \left(\frac{v_j}{w} \right) - 1 \right)^{b-1} + \frac{\partial H_w[v/w]}{\partial \left(\frac{v_j}{w} \right)} \right) \right] \quad (64)$$
$$= 0 \qquad \Leftrightarrow \forall w \qquad [..] = 0, \quad (65)$$

that is, setting $x_j = \frac{v_j}{w}$

$$\sum_{j=1}^{r} x_j \frac{\partial H[x_j]}{\partial x_j} = w \times \lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1} \quad (66)$$
$$+ w \frac{\mu(u)}{b} \left(\sum_{j=1}^{r} x_j - 1 \right)^{b-1} \times \left(\left(p - \frac{bq}{w} \right) \sum_{l=1}^{r} x_l - p \right).$$

Looking again at two sub-propositions $A_{w,1}$ and $A_{w,2}$, with $x = \frac{v_1}{w}$ and $y = \frac{v_2}{w}$ s.t $x + y = 1$, we derive the solution.

Using for short $\alpha(p,m) = \lambda(u) \left(\sum_{i=1}^{n} p_i - 1 \right)^{m-1}$, and the method of characteristics as Eq.(31) giving $y = c_1 x \Leftrightarrow x + y = (1 + c_1)x = cx$, we have to solve, as for Eq.(32),

$$\frac{dx}{x} = \frac{du}{w \times \alpha(p,m) + w \frac{\mu}{b} (cx - 1)^{b-1} \left(\left(p - \frac{bq}{w} \right) cx - p \right)} \quad (67)$$

$$\frac{du}{dx} = \frac{w\alpha}{x} + \frac{cw\mu}{b} \left(p - \frac{bq}{w} \right) (cx - 1)^{b-1} - \frac{w\mu p}{b} \frac{1}{x} (cx - 1)^{b-1} \quad (68)$$

and so, a being a constant, $\beta(x,y)$ having also constants (like the ones from a), we have

$$u(x,y) = c_2 \left(\frac{y}{x} \right) + w \times \alpha(p,m) \int_a^x d\xi \frac{1}{\xi} \quad (69)$$
$$+ \frac{cw\mu}{b} \left(p - \frac{bq}{w} \right) \int_a^x d\xi (c\xi - 1)^{b-1} - \frac{w\mu p}{b} \int_a^x d\xi \frac{(c\xi - 1)^{b-1}}{\xi} \quad (70)$$

that is,

$$u(x,y) = \beta \left(\frac{y}{x} \right) + w\alpha(m,p) ln(x) + \frac{cw\mu}{b} \left(p - \frac{bq}{w} \right) I_1 - \frac{w\mu p}{b} I_2 \quad (71)$$

Again, if $m \geq 2$, then, as a constraint we will have $\alpha = \lambda(u)(cp - 1) \to 0$ and so this term with a logarithm vanishes when we consider the constraint in the final expression of the Ignorance. However, as the expression of I_2 shows, there is another logarithm term which should appear.

2. *case where $b = 1$*

A really interesting case because it is the simplest one which leads to what we expect, and even more, in what we think a coherent way.

$$I_1 = \int_a^x d\xi (c\xi - 1)^{b-1} = \int_a^x d\xi = [\xi]^x \to x \quad (72)$$

$$I_2 = \int_a^x d\xi \frac{(c\xi - 1)^{b-1}}{\xi} = \int_a^x d\xi \frac{(1}{\xi} = [ln(\xi)]^x \to ln(x) \quad (73)$$

our solution is now

$$u(x,y) = \beta \left(\frac{y}{x} \right) + w[\alpha(m) - p\mu] ln(x) + w\mu \left(p - \frac{bq}{w} \right) (cx) \quad (74)$$

Commentaries

- At the end, in the ignorance, constraints will play no major role as they do not influence it. However, we see that they appear here within the solution $u(x,y)$ via their Lagrange multiplier, and also via cx for the last term. For this term in cx, as the constraint are satisfied when applying the solution, we have $cx \to 1$, but not $ln(x) \to 0$! As a consequence, this leads to the constant $w\mu p - \mu q$ in the expression of the ignorance.

- In fact, at the end, this expression $w\mu p - \mu q$ will play no role, as it leads in Eq.(61) to the term

$$\sum_{i=1}^{n} (p_i \times \mu p - \mu q) = \mu p \sum_{i=1}^{n} p_i - \mu q \sum_{i=1}^{n} 1 \to \mu p - \mu q n \quad (75)$$

  after applying the constraint and setting back $w \equiv p_i$. Giving always n propositions at start, this former term is just a constant. In fact, all term linear in w will be considered at the end as a constant due to the summation and the constraint.

- Moreover, assuming that all constraints are implemented in a same way, we would set $m = b = 1$, leading to $\alpha(1) = \lambda(u)$, but also that $\lambda(..) = \mu(..)$. As a consequence, the remaining term, the logarithm one, becomes $w\lambda(..)(1 - p) ln(x)$. As $x = \frac{v_1}{w}$,

$$w\lambda(..)(1 - p) ln(x) = \lambda(..)(1 - p) w (ln(v_1) - ln(w)) \quad (76)$$
$$= (1 - p)\lambda(..)(-w \times ln(w) + w \times ln(v_1)) \quad (77)$$

  1. As we considered in our derivation that v_j are independent of w, the last term is linear in w and therefore, as for $\mu p w$, will lead to a constant in the final expression of the ignorance when constraints are applied.

     Moreover, as $w = v_1 + v_2$, terms like $w \times ln(v_1)$ will have mixed terms as $(v_1 + v_2) ln(v_1)$. This is again linked to Eq.(60) where we encountered a similar problem, which is a consequence of the form $p \times ln(p)$.

  2. Then, dealing with the last term (except for $\beta(x,y)$ which condense the constants and help to restore the symmetry of the ignorance as $u(x,y) = u(y,x)$), we see a factor $1 - p$. As $p = 0$ or 1, the only way to keep the logarithm of w is to set $p = 0$: this is interesting because it was set artificially to consider the constraint on v_i from outside (v_i make sens only in $H_i[p_i]$), and therefore it is better as this leads to no consequences on what we expect.

In fact, with what we said previously, we see that whatever the value of $q \in \{0; 1\}$, it has also no consequences on the expression of the ignorance which varies : At the end, it is like the obtained solution is given w.r.t p_i but we used its consequences on sub-proposition to solve the equation w.r.t them. We could therefore have solved two equations from Eq.(66) where $q = 0$ or $q = 1$, leading to similar solutions in

$wln(w)$ but it makes sens to consider $q = 1$ as it considers the situation on the sub-node. Therefore

$$0 = w\lambda - q\lambda \times (x+y) - x\frac{\partial H}{\partial x} - y\frac{\partial H}{\partial y} \quad (78)$$

$$\Leftrightarrow \quad 0 = \lambda(w - q(x+y)) - x\frac{\partial H}{\partial x} - y\frac{\partial H}{\partial y} \quad (79)$$

$$\Leftrightarrow \quad u(x,y) = \beta\left(\frac{y}{x}\right) + \lambda w ln(x) - q\lambda(cx) \quad (80)$$

$$u(x,y) \to -2\lambda w \times ln(w) + [w] + [cx \to 1] \quad (81)$$

as we restore the symmetry by setting $\beta\left(\frac{y}{x}\right) \sim \lambda w ln(y)$ and as $ln(x) + ln(y) = ln(v_1) + ln(v_2) - 2ln(w)$.

However, if we generalize it with more than 2 sub-propositions, as r sub-propositions, we get $u(x,y,..) \sim -r\lambda w ln(w)$, and so, from Eq.(61), we obtain

$$H[p_1,..,p_n] = \lambda(u)\left(\sum_{i=1}^{n} p_i - 1\right) - \sum_{i=1}^{n} \lambda(r)(p_i ln(p_i)) + \lambda[1] \quad (82)$$

where [1] condense all the constants.

Commentaries about $\lambda(r)$: As we said before, we used the sub-propositions to get the equation we need to solve. In our case, we "knew" that it exists r sub-propositions, but someone may have known that only $r-1$ sub-propositions in the same case, thus leading to a factor $r-1$ instead of r. We could "cure" this reasoning saying that *a priori* we do not know the r sub-propositions, except that a proposition A_w has at least two sub-propositions which are a sub-proposition and its contrary ($A_w = A + \bar{A}$) of probability x and y such that $x + y = 1$. We could say that $A_w = A$ of probability $x = 1$, and so $r = 1$. However, a concern comes from that A_j, $j \neq 1$ constitute $\bar{A}_1$, so it would lead to a mix between the probabilities.

However, just saying that there is one sub-proposition which is the proposition (of probability $x = 1$), leads simply to $r = 1$ in general (but also $ln(x) = 1$ ...) , leading to the solution

$$H[p_1,..,p_n] = \lambda(u)\left[\left(\sum_{i=1}^{n} p_i - 1\right) - \sum_{i=1}^{n}(p_i ln(p_i)) + [1]\right] \quad (83)$$

- We can always rescale $H[p]$ and deal with $h[p]$ such that $H[p] = \lambda h[p]$ as λ is an arbitrary choice and the ignorance has to be the same for every individu with same knowledge on the situation : this expression has therefore to be invariant as such.
- We can also take care of the constants in [1] by always expressing information in terms of $H[p] - [1]$, or ignorance relatively to maximum/minimum ignorance as $H[p] - H_m[p_m]$ for instance.

When constraints are applied, the ignorance, also known as the *information entropy* would therefore correspond to

$$h[p_1,..,p_n] \equiv -\sum_{w=p_i} w \times ln(w) \quad (84)$$

as expected.

3. *cases where $b \geq 2$*

As said previously, in these cases, we would have, roughly speaking,

$$\alpha(b,p) = \lambda(u)\left(\sum_{i=1}^{n} p_i - 1\right)^{b-1} \to [0] \quad (85)$$

when applying the constraint. Moreover, I_1 will have the general form

$$I_1 = \sum_{v=1}^{b} \gamma_1(v,c) \times (cx)^v \quad (86)$$

where $\gamma_1(v)$ are numerical coefficients obtained after integrating ($b = 2$, $\gamma_1(1) = -1$ and $\gamma_1(2) = -\frac{1}{2c}$). And for I_2, we obtain

$$I_2 = \sum_{v=1}^{b-1} \gamma_2(v) \times (cx)^v - ln(x) \quad (87)$$

The solution can thus be expressed as

$$u(x,y) = \beta\left(\frac{y}{x}\right) + [0]_\lambda + \sum_{v=1}^{b} \Gamma(v,c)(cx)^b + \frac{w\mu p}{b} lnx \quad (88)$$

which becomes when applying the constraints $(cx - 1)$ and restoring the symmetry

$$u(x,y) = [0]_\lambda + [1]_\mu + \frac{w\mu p}{b}(lnx + lny) \quad (89)$$

or in general, doing the same simplifications as the case $b = 1$,

$$u(x,y) = [0]_\lambda + [1]_{\lambda,\mu} + \frac{p\mu r}{b} \times wlnw \quad (90)$$

Commentaries : from this last equation, we could say that

- we can also rescale this expression in order to absorb the $[0]_\lambda + [1]_{\lambda,\mu}$ terms, and considering $\lambda = \frac{\mu}{b}$ we could also rescale as in Eq.(83),
- λ does not play a role at all, except to add constants via the I_1 term. Instead, it's really μ : $b = m$ and $\lambda = \mu$ seem to be irrelevant in the final expression,
- $p = 1$ is important, that is the constraint we add "outside", artificially. The constraint "inside", with q, which would make more sens in our opinion as it represents the sub-nodes, makes no effect (except for adding a constant) as in the first approach in this case.
- we need to take $\bar{\mu} = -\mu$, or the constraint to be as $\mu(1 - \sum_i v_i)$, in order for the ignorance to behave correctly,
- and then, with these modifications, in these cases too, we obtain the expected expression for the ignorance to correspond to the *information entropy*, for any value of b (but in a less appealing way).

IV. CONCLUSION

1. We have included constraints not as $\lambda \sum_i p_i$ as done for instance in [1], but as $\lambda(\sum_i p_i - 1)$. This allows us to define what we call *Ignorance* H, where

$$H = H_{knowns} + H_{unknowns} \tag{91}$$

where H_{knowns} encodes the ignorance due to the constraints, therefore of zero ignorance.

2. In the first approach, we dealt with a quasi-known expression of the expression, *i.e.* with the factor p_i in front of H_i. In this case, it was like maximizing/minimizing the expected value of "local" sub-ignorance (at each branch of p_i) but leading to a final expression not really convincing as the logarithm term has to vanish when the constraints are applied. This was due, in our opinion, to the ill way of defining what happens at each sub-node such that,roughly speaking, $p_i H_i[p_i] \to p_i \times \frac{v_j}{p_i} H_i \left[\frac{v_j}{p_i}\right]$. But we may have set it wrong and a more coherent way is possible.

3. However, we found way to cure this, starting from even before, not knowing at all the expression for the ignorance but just that it has also to apply in the same way at each node. Then we were able to get the expected expression for the Shannon entropy, but still with some interrogations linked to the same ones in the first approach.

4. Mathematically, we have started from A but included sub-nodes as B in order to implement the fact that it has to be similar at each node. This helped us to obtain the correct expression for the differential expressions with the differentiation of the $\frac{1}{p_i}$ factors, leading to an expression in $w\ ln(w)$ primitive of $ln(w) + 1$ and so the role of the exponential.

5. Moreover, we have also seen (at least partially) that the expression of the ignorance was somehow independent of the power taken for the constraints. In fact, the simplest case of power 1 seems in our opinion even better as we were able to obtain Eq.(84) in a coherent way, the higher power needing some adjustments.

6. In this way, the *Maximization Entropy Principle* makes naturally sense as it is just the procedure to minimize our ignorance. It helped us to derive first the expression of the ignorance one has to obtain in order to be coherent, and secondly, knowing the expression but not the probabilities inside, to obtain these probabilities as usual and shown for instance in [1].

7. Regarding the Lagrange multiplier, we were able to incorporate their subjectivity in an invariant way as the final expression of the Ignorance has to be the same whatever the choice of the multipliers. However, due to the presence of constants [1], it would be better to express any quantity with respect to a reference value (as for temperature), that is, using $H[p] - H_m[p_m]$ for instance, in order to keep only the meaningful parts of the ignorance.

To summarize :

Having knowledge on what we should have expected, we were biased but this helped us to start from zero and look at the situation from another perspective : having some notions about constraints and variational problems, reading the nice construction of the theory [1] and on the maxmization entropy principle, gave us thoughts about including constraints on the probability in such a way that it could make sens.

As a consequence, we have defined general what we call "ignorance" and the procedure was "only" to try to minimize it (at least) and see if we could get back the correct expression for Shannon entropy: this is just the application of the maxmization entropy principle which appears naturally in this framework.

To conclude, an extension of this work, at least in the way it has been done, may be helpful for instance in decision theory where one would define a quantity like the average risk, and try to minimize it as done here. This, however, will be kept for further researches.

V. ACKNOWLEDGMENTS

The author would like to express his deepest gratitude to Abhay, Aurelien, Martin, .. for time and space spend together. Thanks also to Lê Nguyên Hoang for its pedagogical work which leads to look deeper to the Bayesian approach. Wolframalpha was used to check the calculations, and Geogebra to plot figures using tikz in LateX.

---

[1] Jaynes, E. T. (2003) "Probability Theory: The Logic of Science", Cambridge University Press, p. 351-355. ISBN 978-0521592710
[2] Shannon, Claude Elwood (July 1948). "A Mathematical Theory of Communication" . Bell System Technical Journal. 27 (3): 379–423.
[3] https://en.wikipedia.org/wiki/Entropy_(information_theory)
[4] https://www.wolframalpha.com/input/?i=0+%3D+A*%28x%2By-1%29**%28-2%29%2B+f%28x%2Cy%29+-+x+df%2Fdx+-+y+df%2Fdy
[5] If this has been done before, we would be sorry for having missing this. We are beginner in this field, therefore, please, let us know, we will change the article correspondingly.
[6] Actually, it's just a guess for now on.
[7] In what I have seen yet, no derivatives where used in order to get the expression of Shannon's entropy.