# T'aurais pas une entropie?

by jfb & co ,

**Abstract**

Where we show that it is possible to derive new entropies yielding a particular specified maximum entropy distribution. There are (probably) many errors –I hope not fundamental but is is possible; (certainly many) approximations, typos, maths and language mistakes.Suggestions and improvements will be much appreciated.

## 1. Maximum entropy distributions

Let $S[f] = -\int f(x) \log f(x) \mathrm{d}\mu(x))$ be the Shannon entropy. Subject to $n$ moment constraints such as $\mathbb{E}\left[[T_i(x)\right] = t_i,\ i = 1, \ldots, n$ and to normalization, it is well known that the maximum entropy distribution lies within the exponential family

$$f_X(x) = \exp\left(\sum_{i=1}^{n} \lambda_i T_i(x) + \lambda_0\right).$$

In order to recover known probability distributions (that must belong to the exponential family), it is then sufficient to specify a set of functions $T_i$, i.e., a function $T : \mathbb{R} \mapsto \mathbb{R}^n$ where $n$ is the number of moment constraints. This has been used by many authors. For instance, the gamma distribution can be viewed as a maximum entropy distribution if one knows the moments $\mathbb{E}\left[[X\right]$ and $\mathbb{E}\left[\log(X)\right]$. In order to find maximum entropy distributions with simpler constraints or distributions outside of the exponential family, it is possible to consider other entropies. This is discussed below.

## 2. Maximum $(h, \phi)$-entropy distributions

### *2.1. Definition and maximum $(h, \phi)$-entropy solution*

**Definition 1.** Let $\phi : \Omega \subset \mathbb{R}_+ \mapsto \mathbb{R}$ be a stricly convex differentiable function defined on a closed convex set $\Omega$. Then, if $f$ is a probability distribution defined with respect to a general measure $\mu(x)$ on a set $\mathcal{X}$,

$$H_\phi[f] = -\int_{\mathcal{X}} \phi(f(x)) \mathrm{d}\mu(()x) \tag{1}$$

is the $\phi$-entropy of $f$.

Since $\phi(x)$ is convex, then the entropy functional $H_\phi[f]$ is concave. Also note that the composition of a concave function with a nondecreasing concave function preserves concavity, and that composition of a convex function with a nonincreasing convex function yields a concave functional.

**Definition 1.** *With the same assumption in definition 1,*

$$H_{h,\phi}[f] = h\left(-\int_{\mathcal{X}} \phi(f(x)) \mathrm{d}\mu(()x)\right) \tag{2}$$

*is called $(h, \phi)$-entropy of $f$, where*

- *either $\phi$ is convex and $h$ concave nondecreasing*
- *or $\phi$ is concave and $h$ convex nonincreasing*

These $(h, \phi)$-entropies have been studied in **? ?** ] for instance. In these works neither concavity (resp. convexity) of $h$, nor the differentiability of $\phi$ are imposed.

A useful related quantity to these entropies is the Bregman divergence associated with $\phi$:

**Definition 2.** *With the same assumption in definition 1, the Bregman divergence associated with $\phi$ defined on a closed convex set $\Omega$, is given by*

$$D_\phi(x_1, x_2) = \phi(x_1) - \phi(x_2) - \phi'(x_2)\,(x_1 - x_2)\,. \tag{3}$$

A direct consequence of the strict convexity of $\phi$ is the nonnegativity of the Bregman divergence: $D_\phi(x_1, x_2) \geq 0$ with equality if and only if $x_1 = x_2$.

Consider the problem of maximizing entropy (2) subject to constraints on some moments $\mathbb{E}\left[[T(X)]\right]$ where the normalization constraint is now included in $T$ (namely $T_0(x) = 1$ and $t_0 = 1$). Since $h$ is monotone, it is enough to look for the maximum of the $\phi$-entropy (1),

$$\begin{cases} \max_f & -\displaystyle\int \phi(f(x))\mathrm{d}\mu(()x) \\ \text{s.t.} & \mathbb{E}\left[[T(X)]\right] = t \end{cases} \tag{4}$$

**Proposition 2.** *The probability distribution $f_X$ solution of the Maximum entropy problem* (4) *satisfies the equation*

$$\phi'\big(f_X(x;t)\big) = \lambda^t\, T(x). \tag{5}$$

*where vector $\lambda$ is such that $\mathbb{E}\left[T(X)\right] = t$.*

*Proof.* The maximization problem being concave, the solution exists and is unique. Equation 5 results directly from the classical Lagrange multipliers technique.

An alternative derivation of the result consists in checking that the distribution (5) is effectively a maximum entropy distribution, by showing that $H_\phi[f] > H_\phi[g]$ for all probability distributions with a given (fixed) moment $\mathbb{E}\left[[T(X)]\right]$. To this end, consider the functional Bregman divergence acting on functions defined on a common domain $\mathcal{X}$:

$$D_\phi(f_1, f_2) = \int_{\mathcal{X}} \phi(f_1(x))\mathrm{d}\mu(()x) - \int_{\mathcal{X}} \phi(f_2(x))\mathrm{d}\mu(()x) - \int_{\mathcal{X}} \phi'(f_2(x))\,(f_1(x) - f_2(x))\,\mathrm{d}\mu(()x).$$

From the nonnegativity of the Bregman divergence this functional divergence is nonnegative as well, and zero if and only if $f_1 = f_2$ almost everywhere. Define by

$$C_t = \{f : \mathcal{X} \mapsto \mathbb{R}_+ :\ \mathbb{E}\left[[T(X)]\right] = t\}$$

the set of all probability distributions defined on $\mathcal{X}$ with given moments $t$. Consider now $f_X \in C_t$ such that $\phi'(f_X(x)) = \lambda^t\, T(x)$ and any given function $f \in C_t$. Then

$$\begin{aligned} D_\phi(f, f_X) &= \int_{\mathcal{X}} \phi(f(x))\mathrm{d}\mu(()x) - \int_{\mathcal{X}} \phi(f_X(x))\mathrm{d}\mu(()x) - \int_{\mathcal{X}} \phi'(f_X(x))\,(f(x) - f_X(x))\,\mathrm{d}\mu(()x) \\ &= -H_\phi[f] + H_\phi[f_X] - \int_{\mathcal{X}} \lambda^t\, T(x)\,(f(x) - f_X(x))\,\mathrm{d}\mu(()x) \\ &= H_\phi[f_X] - H_\phi[f] \end{aligned}$$

where we used the fact that $f$ and $f_X$ have the same moments $\mathbb{E}\left[T(X)\right] = t$. By nonegativity of the Bregman functional divergence, we finally get that

$$H_\phi[f_X] \geq H_\phi[f]$$

for all pdf $f$ with the same moments $t$ than $f_X$, with equality if and only if $f = f_X$. In other words, this shows that $f_X$, solution of (5), realizes the minimum of $H_\phi[f]$ over $C_t$. □

*2.2. Defining new entropy functionals*

Given an entropy functional, we thus obtain a maximum entropy distribution. There exists numerous $(h, \phi)$-entropies in the literature. However a few of them lead to explicit forms for the maximum entropy distribution. Therefore, it is of high interest to look for the entropies that lead to a specified distribution as a maximum entropy solution.

Since we will look for the function $\phi$ for a given probability distribution $f_X(x)$ we also see that the corresponding $\lambda$ parameters can be included in the definition of the function.

Let us recall some implicit properties of $\phi(x)$.

- $\phi'(x)$ is defined on a domain included on $f_X(\mathcal{X})$;
- From the strict convexity property of $\phi$, necessarily $\phi'$ is increasing.

The identification of a function $\phi(x)$ such that a given $f_X(x)$ is the associated maximum entropy distribution amounts to solve (5), that is

1. choose $T(x)$,
2. find $\phi'(y)$ such that
$$\lambda^t T(x) + \mu = \phi'\left(f_X(x)\right) = \phi'(y) \tag{6}$$
3. integrate the result to get $\phi(y) = \displaystyle\int \phi'(y)\mathrm{d}y + c$, where $c$ is an integration constant. The entropy being defined by $H_\phi[f] = -\displaystyle\int_{\mathcal{X}} \phi(f(x))\mathrm{d}\mu(()x)$, the constant $c$ will usually be zero.
4. Parameters $\lambda$ may be choosen case by case in order to simplify the expression of $\phi$.

Remind that $\phi'$ must be increasing, thus, necessarily, $\lambda^t\, T(x)$ and $f_X(x)$ must have the same sense of variation.

Observe that since we want $\phi(x)$ to be convex, which means $\phi''(x) \geq 0$ for a twice differeentiable function, it is thus necessary that $\phi'(x)$ is non decreasing on [0,max($f$)]. By the relation

$$\phi'^{-1}\left(\lambda T(x) + \mu\right) = f_X(x). \tag{7}$$

we have that

$$f_X'(x) = \lambda T'(x) \frac{1}{\phi''\left(\phi'^{-1}\left(\lambda T(x) + \mu\right)\right)} = \lambda T'(x) \frac{1}{\phi''\left(f_X(x)\right)}.$$

Hence we get that

$$\phi''\left(f_X(x)\right) = \frac{f_X'(x)}{\lambda T'(x)}$$

and we see that $f_x(x)$ and $T(x)$ must have the same or an opposite variation, depending on the sign of $\lambda$.

Examples: if $\lambda$ is negative, then

- for $T(x) = x$, $f_X(x)$ must be non increasing,
- for $T(x) = x^2$ or $T(x) = |x|$, $f_X(x)$ must be unimodal with a maximum at zero.

For instance, for one moment constraint, if $\lambda_1$ is negative, then

- for $T_1(x) = x$, $f_X(x)$ must be decreasing,
- for $T_1(x) = x^2$ or $T_1(x) = |x|$, $f_X(x)$ must be unimodal with a maximum at zero.

Equation (5) may have no solution, when $\lambda^t\, T(x)$ has not the same variations than $f_X$. But it can also have several solutions.

## 3. $\phi$-escort, $\phi$-Fisher information and generalized Cramér-Rao inequality

## 4. Some examples

### *4.1. Normal distribution and second-order moment*

### *4.2. q-exponential distribution and first-order moment*

### *4.3. q-Normal distribution and second-order moment*

### *4.4. Hyperbolic secant distribution and first-order moment*

Let us consider some specific cases.

1. For a normal distribution, $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$ and $T(x) = x^2$, we begin by computing the inverse $y = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, which gives $-\frac{1}{2}x^2 - \log\sqrt{2\pi} = \log(y)$. Choosing $\lambda = -\frac{1}{2}$, $\mu = -\log\sqrt{2\pi}$ and integrating, we obtain
$$\phi(y) = y \log y - y$$
2. For a Tsallis $q$-exponential, $f_X(x) = C_q\,(1-(q-1)\beta x)_+^{\frac{1}{(q-1)}}$ , $x \geq 0$, and $T(x) = x$. We simply have $C_q^{q-1}\,(1-(q-1)\beta x) = y^{q-1}$. With $\lambda = qC_q^{q-1}\beta$ and $\mu = qC_q^{q-1}/(1-q)$, this yields
$$\phi(y) = \frac{y^q}{1-q}.$$
Taking $\mu = \left(qC_q^{q-1} + 1\right)/(1-q)$ gives
$$\phi(y) = \frac{y^q - y}{1-q},$$
and an associated entropy can be
$$H_\phi[f] = \frac{1}{1-q}\left(\int f(x)^q \mathrm{d}\mu(x) - 1\right),$$
which is nothing but Tsallis entropy.[1]
3. The same entropy functional can readily be obtained for the so-called $q$-Gaussian, or Student-t and -r distributions $f_X(x) = C_q\left(1-(q-1)\beta x^2\right)_+^{\frac{1}{(q-1)}}$ . It suffices to follow the very same steps as above with $T(x) = x^2$.

[1]Of course, we can also take the first $\phi(y) = \frac{y^q}{1-q}$, integrate and add any constant, since adding a constant do not modify the actual value af the minimizer (or maximizer if we consider concave entropies).

4. Let $f_X(x)$ be the hyperbolic secant distribution, with density

$$f_X(x) = \frac{1}{2}\text{sech}(\frac{\pi}{2}x) = \frac{1}{2}\cosh^{-1}(\frac{\pi}{2}x).$$

Obviously, $\frac{\pi}{2}x = \cosh(2y) = \phi'(y)$ with $T(x) = x$, $\lambda = \frac{\pi}{2}$, and

$$\phi(y) = \sinh(2y).$$

So doing, we obtain an hyperbolic sine entropy with the hyperbolic secant distribution as the associated maximum entropy distribution.

## 5. Multiform entropies

Of course, the preceeding derivations require that (6) is effectively solvable. In addition, one has also to choose or design a specific $T(x)$ statistic, as well as the parameters $\lambda$ and $\mu$. In the examples above, we used $T(x) = x$ and $T(x) = x^2$. Particular choices such as $T(x) = x^2$ or $T(x) = |x|$ obviously lead to symmetrical densities.

For nonsymmetrical unimodal densities, the situation is more involved. For instance, if we take $T(x) = x$, then the resolution of (6) amounts to compute the inverse relation of $y = f_X(x)$, which is is multi-valued. Indeed, $f_X(x)$ is not injective and to each $y$ correspond two distinct values of $x$. Let us denote $\mathcal{S}_y$ the image of $f_X$, $I_+ \subseteq \mathbb{R}$ the domain where $f_X(x)$ is non decreasing, and $I_- \subseteq \mathbb{R}$ the domain where $f_X(x)$ is non increasing. We thus have two possible inverses defined respectively say $\phi'_+ : \mathcal{S}_y \mapsto I_+$ and $\phi'_- : \mathcal{S}_y \mapsto I_-$ such that $\phi'^{-1}_+(-x) = f_X(x)$ for $x \in I_+$ and $\phi'^{-1}_-(-x) = f_X(x)$ for $x \in I_-$. Furthermore, by the remarks at the end of section 2.2, we see that $\phi_+$ is convex while $\phi_-$is concave. In this context, our proposal is to define a $\phi$-entropy as follows

$$H_\phi[f_X] = -\int_{I_+} \phi_+ (f_X(x))\, \mathrm{d}\mu(x) - \int_{I_-} \phi_- (f_X(x))\, \mathrm{d}\mu(x).$$

It is easy to check that this entropy functional is no more convex nor concave (for the subset of distributions with support on $I_+$the entropy is convex while it its concave for on the subset of distributions on $I_-$. We propose to look for the *extreme entropy* (instead of the maximum entropy as in the classical case). With a moment constraint, the Lagrangian is

$$L(f_X; \lambda_1, \lambda_0) = \int_{I_+} \phi_+ (f_X(x))\, \mathrm{d}\mu(x) + \int_{I_-} \phi_+ (f_X(x))\, \mathrm{d}\mu(x) + \int_{\mathbb{R}} \lambda_1 x\, \mathrm{d}\mu(x) + \int_{\mathbb{R}} \lambda_0\, \mathrm{d}\mu(x)$$

and its first variation is

$$\delta L(f_X; \lambda_1, \lambda_0) = \phi_+ (f_X(x))\, 1_{I_+} + \phi_- (f_X(x))\, 1_{I_-} + \lambda_1 x + \lambda_0.$$

Thus the critical points are defined by $\phi_+ (f_X(x)) + \lambda_1 x + \lambda_0 = 0$ for $x \in I_+$and $\phi_- (f_X(x)) + \lambda_1 x + \lambda_0 = 0$ for $x \in I_-$, which actually define the extreme entropy distribution $f_X(x)$ as the inverse relation of a multiform entropy. Obviously, this formulation includes the classical maximum entropy approach as a particular case.

Observe that it is still possible to get a maximum or a minimum entropy solution, but on subsets. Thus, it will still be possible to use these entropies in testing problems. For such goal, define $m_+$ to be the moment computed on the subset $I_+ : m_+ = \int_{I_+} x\, f_X(x) \mathrm{d}\mu(x)$, and similarly for a moment $m_-$ computed on $I_-$. By the very same reasoning and proof as in the classical case (see the proof of proposition 2), we have that

- (a) $H_{\phi_+}[f_X] \geq H_{\phi_-}[f_1]$ for all distributions $f_1$ with a fixed moment $m_+$,
- (a) $H_{\phi_-}[f_X] \leq H_{\phi_-}[f_2]$ for all distributions $f_2$ with a fixed moment $m_-$

where $f_X(x) = \phi_+'^{-1}(-x)$ for $x \in I_+$ and $f_X(x) = \phi_-'^{-1}(-x)$ for $x \in I_-$. Hence we will be able to use these entropies for distribution testing, provided that we are able to compute empirical values for $m_+$ and $m_-$ from data, which is quite easy.

*5.1. Example 1. The logistic distribution*

The pdf of the logistic distribution is given by

$$f_X(x) = \frac{e^{-\frac{x}{s}}}{s\left(1+e^{-\frac{x}{s}}\right)^2}.$$

This distribution, which resembles the normal distribution but has heavier tails, has been used in many applications. By direct calculations, we obtain

$$\begin{cases} \phi_-'(y) = & s\ln\left(\frac{1}{2}\,\frac{-2\,ys+1+\sqrt{-4\,ys+1}}{ys}\right), \\ \phi_+'(y) = & s\ln\left(-\frac{1}{2}\,\frac{2\,ys-1+\sqrt{-4\,ys+1}}{ys}\right). \end{cases}$$

The associated entropy is then

$$\begin{cases} \phi_-(y) = -\frac{1}{2}\sqrt{-4\,ys+1} + \frac{1}{2} + ys\,\ln\left(-\frac{\sqrt{-4\,ys+1}-1}{\sqrt{-4\,ys+1}+1}\right) \\ \phi_+(y) = \frac{1}{2}\sqrt{-4\,ys+1} + \frac{1}{2} + ys\,\ln\left(-\frac{\sqrt{-4\,ys+1}-1}{\sqrt{-4\,ys+1}+1}\right) \end{cases}$$

for $y \in [0, \frac{1}{4s}]$, and where we have introduced a integration constant such that $\min_y \phi_+(y) = 0$. For $y > \frac{1}{4s}$, we extend the function and let $\phi_+(y) = +\infty$. Figure 1 gives a representation of this entropy for $s = 1$.

*5.2. Example 2. The gamma distribution*

The probability density function of the gamma distribution is given by

$$f_X(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta\,x}}{\Gamma(\alpha)}$$

We obtain

$$\phi'(y) = -\mathrm{e}^{\frac{1}{\alpha-1}\left(-W\left(-\frac{\beta\left(y\Gamma(\alpha)\beta^{-\alpha}\right)^{(\alpha-1)^{-1}}}{\alpha-1}\right)\alpha + W\left(-\frac{\beta\left(y\Gamma(\alpha)\beta^{-\alpha}\right)^{(\alpha-1)^{-1}}}{\alpha-1}\right) + \ln\left(y\Gamma(\alpha)\beta^{-\alpha}\right)\right)},$$

where $W$ is the Lambert W multivalued 'function' defined by $z = W(z)e^{W(z)}$ (ie the inverse relation of $f(w) = we^w$). Unfortunately, in the general case, we do not have a closed form for $\phi(y)$ as the integral of $\phi'(y)$.[2] Restricting us to the case $\alpha = 2$, we have

$$\phi(y) = \frac{\left(1 - W\left(-\frac{y}{\beta}\right) + y\left(W\left(-\frac{y}{\beta}\right)\right)^2\right)}{\beta\,W\left(-\frac{y}{\beta}\right)} + \frac{\beta}{e},$$

which is convex if we choose the -1 branch of the Lambert function and concave for the 0 branch. An example with $\alpha = 2$ and $\beta = 3$ is given on Figure 2.

[2] This might not be completely unacceptable. Indeed, it is really not difficult to compute numerically the values of $\phi(y)$.

*5.3. Example 3. The arcsine distribution*

As a further example, we consider the case of the arcsine distribution (see wiki) which also yields a multiform entropy. This distribution, defined for $x \in (0, 1)$, is a special case of the Beta distribution with parameters $\alpha = \beta = 1/2$. It has the following pdf:

$$f_X(x) = \frac{1}{\pi\sqrt{x(1-x)}}.$$

Observe that $\min_x f_X(x) = 2/\pi$. Doing our now usual calculations, we obtain

$$\begin{cases} \phi'_-(y) = & -\frac{y\pi+\sqrt{y^2\pi^2-4}}{2y\pi}, \\ \phi'_+(y) = & -\frac{y\pi-\sqrt{y^2\pi^2-4}}{2y\pi}. \end{cases}$$

and the expression of the entropy is

$$\phi_\pm(y) = \frac{1}{2}\frac{\sqrt{y^2\pi^2-4}}{\pi} \pm \frac{1}{\pi}\arctan\left(2\frac{1}{\sqrt{y^2\pi^2-4}}\right) - \frac{1}{2}y,$$

for $y \geq 1/\pi$. The entropy is shown on 3.

*5.4. Example 4. The Chi-squared distribution*

Let us now consider the case of a chi-squared distribution. The probability density, for $x \geq 0$,is given by

$$f_X(x) = c\,x^{\frac{k}{2}-1}\exp -\frac{x}{2}$$

with $c^{-1} = 2^{\frac{k}{2}}\Gamma(\frac{k}{2})$. Instead of $T(x) = x$, we now take $T(x) = x^2$ and $\lambda = 1$, which means that we look for $\phi$ such that $\phi'^{-1}(x^2) = f_X(x)$. Solving, we get that

$$\phi'(y) = \begin{cases} 4\,(n-1)^2\,W\left(\frac{1}{2(n-1)}\,(-y)^{\frac{1}{n-1}}\right)^2 & \text{for } k\text{=2}n\text{even, }n\text{even} \\ 4\,(n-1)^2\,W\left(\pm\frac{1}{2(n-1)}\,(y)^{\frac{1}{n-1}}\right)^2 & \text{for } k\text{=2}n\text{even, }n\text{odd} \\ (k-2)^2\,W\left(\frac{1}{k-2}\,\left(-y^2\right)^{\frac{1}{k-2}}\right)^2 & \text{for } k\text{odd} \end{cases}$$

Among these solutions, we must discard complex valued solutions. Since $y$ is non negative, we see that we can only keep solutions with $k = 2n$ even with $n$ odd (or $n = 2$). For $n = 2$, the solution reduces to

$$\phi'(y) = 4W\left(-\frac{1}{2}y\right)^2.$$

By integration, we obtain the corresponding entropy, e.g.

$$\phi(y) = 4\,\frac{\left(-4+4\,W\left(-1/2\,y\right)-2\,\left(W\left(-1/2\,y\right)\right)^2+\left(W\left(-1/2\,y\right)\right)^3\right)y}{\left(-1/2\,y\right)} \text{ for n=2}$$