Physica Scripta

PAPER

On the foundations of the maximum entropy principle using Fenchel duality for Shannon and Tsallis entropies

Pierre Maréchal[1], **Yasmín Navarrete**[2] and **Sergio Davis**[3,4]

[1] Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse, France
[2] Instituto de Filosofía y Ciencias de la Complejidad (IFICC), Los Alerces 3024, Ñuñoa, Santiago, Chile
[3] Comisión Chilena de Energía Nuclear, Casilla 188-D, Santiago, Chile
[4] Departamento de Física, Facultad de Ciencias Exactas, Universidad Andres Bello. Sazié 2212, piso 7, 8370136, Santiago, Chile

E-mail: pierre.marechal@math.univ-toulouse.fr

Keywords: maximum entropy, Fenchel duality, escort distribution, tsallis entropy

RECEIVED
28 November 2023

REVISED
27 May 2024

ACCEPTED FOR PUBLICATION
7 June 2024

PUBLISHED
21 June 2024

Abstract

In this work, we address two main objectives. The first one is to provide a rigorous foundation to the maximum entropy principle in statistical physics, by making use of the Fenchel-Rockafellar duality. The second objective is to discuss the well-foundedness of the so-called escort distributions in the context of non-extensive entropy maximization. The duality treatment of maximum entropy confirms the non-rigorous results obtained via the usual variational calculus, however, the use of escort distributions yields undefined behavior when used consistently, and only leads to the desired results when used in an ad-hoc manner.

1. Introduction

The principle of maximum entropy, that is, the idea of constructing the most unbiased probability distribution by maximizing an entropy functional, is a powerful inference tool in statistics which is now widely used in several fields of science as well as engineering [1, 2], ecology [3, 4], astronomy [5], social dynamics [6, 7], and signal and image science [8], among others. Originally introduced by Gibbs [9] in statistical mechanics as the justification for the canonical ensemble for equilibrium systems, and extended outside of physics by Jaynes [10], the principle of maximum entropy is based on the interpretation of Shannon's entropy as a measure of information. In this way, maximizing the entropy should be understood as choosing the model with less information content provided that it agrees with the constraints given.

Even after the proof of uniqueness of the Gibbs-Shannon entropy in the context of inference by Shore and Johnson [11], the use of different entropy functionals has been proposed, in order to justify the existence of power-law distributions in complex, non-equilibrium systems such as space and laboratory plasmas [12, 13], turbulent fluids [14], self-gravitating systems of astrophysical interest [15] and also in open, finite systems [16], in what is known as non-extensive statistical mechanics [17, 18]. However, the use of Tsallis' q-entropy seems to require [19, 20] a different kind of expectation constraint that uses the so-called escort distributions, originally introduced by Beck and Schlögl [21], instead of the target distribution. This modification has been both subject to criticism on formal grounds and found to produce inconsistencies [22–28]. Furthermore, it has been largely shown that power laws can be recovered without the need to invoke generalized entropies [29–33], most notably under the framework of superstatistics [34]. Despite this, the use of non-Shannonian entropies have been supported recently under particular assumptions [35].

Maximizing the entropy under expectation constraints is usually performed by making use of variational calculus. More precisely, the Lagrangian of the problem is derivated *formally*, meaning that the infinite-dimensional probability density that is searched for is manipulated as if it would be a finite-dimensional (discrete) probability distribution, and the integral entropy-functional as if it was a finite sum. The functional analysis so avoided needs, somehow, to be clarified. Dealing with infinite-dimensional convex optimization with linear (expectation) constraints can be tackled using a powerful convex analytic machinery, in particular the so-

called *partially finite convex programming*, as introduced in [36, 37] and extended in [38]. Notice that formal derivation eludes the fact that the entropy takes infinite values, and as so cannot be derivated (formally or not) throughout the space of integrable functions. As an example, consider the function defined on $\mathbb{R}$ by

$$p(x) = \begin{cases} \dfrac{1}{x \ln^2 \frac{e}{x}} & \text{if } x \in (0,\ 1), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then p is nonnegative, integrable with unit integral, and such that all its positive moments

$$\mu_n = \int_0^1 p(x) x^n \, \mathrm{d}x, \qquad n = 0,\ 1,\ 2,\ldots \tag{2}$$

exist. It is therefore a well-behaved probability density, and yet it has infinite entropy.

Moving beyond the use of Shannon's entropy, from the point of view of optimization, introduces additional issues to be aware of. Most important of all, the escort distributions introduce non-linear constraints that make the problem no longer convex. Linearity can be restored via a transformation of variables, although this brings us to the following alternative: either there is no solution or there are infinitely many solutions (see section 4.3).

In this paper, we provide a rigorous derivation of the Maximum Entropy distributions, using Fenchel duality argument. Moreover, we apply this strategy for dealing with non-extensive statistics, and show that the escort avoidance of it yields a somewhat undetermined framework.

The manuscript is organized as follows. In section 2, we recall the historical background of maximum entropy in statistical physics, and we outline the computation of maximum entropy distributions via variational calculus. We emphasize that this approach to the derivation of solutions presents weaknesses, which motivates our treatment via Fenchel-Rockafellar duality. In section 3, we review the main mathematical tools for such a treatment. In section 4, we apply the results of the previous section to provide rigorous justification of maximum entropy solutions, in both the Boltzmann-Shannon entropy case and the Tsallis entropy case. Finally, section 5 presents a discussion and conclusion of this work.

2. Maximum entropy in statistical physics

The maximum entropy principle [10] is a conceptual extension of the Gibbs method for the construction of ensembles in statistical mechanics [9]. In the original argument by Gibbs, the equilibrium states in nature are states of maximum thermodynamic entropy

$$\mathcal{S}[p] \coloneqq -k_B \int_V p(\mathbf{\Gamma}) \ln p(\mathbf{\Gamma}) \, \mathrm{d}\mathbf{\Gamma}, \tag{3}$$

where $\mathbf{\Gamma} = (\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N, \boldsymbol{p}_1,\ldots,\boldsymbol{p}_N)$ are the microstates of the system, and $\mathcal{S}$ is constrained by the external conditions. However, after the seminal work of Shannon establishing information theory, Jaynes recognized the maximization of the entropy functional $\mathcal{S}$ as the search for the most unbiased model for the microstate probability, that is, as a problem of statistical inference.

The maximum entropy ensemble under the constraint of fixed mean energy

$$\int_V p(\mathbf{\Gamma}) \mathcal{H}(\mathbf{\Gamma}) \, \mathrm{d}\mathbf{\Gamma} = U, \tag{4}$$

where U corresponds to the *internal energy*, and normalization,

$$\int_V p(\mathbf{\Gamma}) \, \mathrm{d}\mathbf{\Gamma} = 1, \tag{5}$$

is the well-known *canonical ensemble* representing a system in thermal equilibrium at temperature T,

$$p(\mathbf{\Gamma};\ \beta) = \frac{\exp(-\beta \mathcal{H}(\mathbf{\Gamma}))}{Z(\beta)}, \tag{6}$$

where $\beta = 1/(k_B T)$ is the *inverse temperature* and $Z(\beta)$ the *partition function*,

$$Z(\beta) \coloneqq \int_V \exp(-\beta \mathcal{H}(\mathbf{\Gamma})) \, \mathrm{d}\mathbf{\Gamma}. \tag{7}$$

The appropriate value of β for a given value of U is the one that realizes the constraint in (4), which in thermodynamics leads to the *caloric curve* $U(\beta)$,

$$U(\beta) = -\frac{\partial}{\partial \beta} \ln Z(\beta). \tag{8}$$

By replacing (6) into (3), the maximized value of the Gibbs entropy is then

$$S(\beta) := \mathcal{S}[p(\cdot;\beta)] = k_B(\ln Z(\beta) + \beta U), \tag{9}$$

which, by using the definition of the (Helmholtz) free energy $F := U - TS$, leads to

$$F(\beta) = -\frac{1}{\beta}\ln Z(\beta). \tag{10}$$

In the case of non-extensive statistical mechanics, one replaces $\mathcal{S}$ in (3) by the Tsallis entropy,

$$\mathcal{S}_q[p] := \frac{1}{q-1}\left(1 - \int_V p(\mathbf{\Gamma})^q \, \mathrm{d}\mathbf{\Gamma}\right) \tag{11}$$

and maximizes it subject to the usual normalization constraint in (5) and a generalized expectation constraint, of the form

$$\frac{\int_V p(\mathbf{\Gamma})^q \mathcal{H}(\mathbf{\Gamma}) \, \mathrm{d}\mathbf{\Gamma}}{\int_V p(\mathbf{\Gamma})^q \, \mathrm{d}\mathbf{\Gamma}} = U_q. \tag{12}$$

According to variational calculus, this leads to the well-known q-canonical ensemble of Tsallis statistics,

$$p(\mathbf{\Gamma};\, \beta,\, q) = \frac{1}{Z_q(\beta)}[1 + (q-1)\beta\mathcal{H}(\mathbf{\Gamma})]_+^{\frac{1}{1-q}}. \tag{13}$$

In order to state the derivation of the canonical ensemble in (6) in more rigorous terms, we consider the optimization problem

$$(\mathscr{P}) \quad \left| \begin{array}{ll} \text{Maximize} & \mathscr{S}(p) := -\int_V p(\mathbf{x})\ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ \text{subject to} & p \in L^1(V), \\ & \int_V p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1, \\ & \int_V p(\mathbf{x}) H(\mathbf{x}) \, \mathrm{d}\mathbf{x} = E_\circ. \end{array} \right.$$

We (temporarily) assume that the domain V is a bounded subset of $\mathbb{R}^n$, in which $n = 6N$ with N the number of particles. In the above problem, H denotes the Hamiltonian of the system, which we assume to be bounded on V. The control variable p lies in the infinite dimensional space $L^1(V)$, and the integrals in the constraints are well-defined on $L^1(V)$.

The objective functional is the Boltzmann-Shannon entropy. It can be written as

$$\mathscr{S}(p) = -\int_V h_\circ(p(\mathbf{x})) \, \mathrm{d}\mathbf{x}, \tag{14}$$

in which

$$h_\circ(t) = \begin{cases} t\ln t & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ \infty & \text{if } t < 0. \end{cases} \tag{15}$$

It is concave, and clearly maximizing $\mathscr{S}$ is equivalent to minimizing the convex functional $-\mathscr{S}$. The constraints involve the linear mappings

$$\mathbb{I}p := \int_V p(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \quad p \in L^1(V)$$

and

$$\mathbb{A}p := \int_V p(\mathbf{x}) H(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \quad p \in L^1(V).$$

For convenience, we denote by $\mathbb{A}_\circ$ the linear mapping given by $\mathbb{A}_\circ p = (\mathbb{I}p,\, \mathbb{A}p) \in \mathbb{R}^2$.

3. Review of convex analytic tools

Let L be any real vector space. A function f: $L \to [-\infty, \infty]$ is said to be *convex* if its *epigraph*, the set

$$\mathrm{epi}\, f := \{(x,\, \alpha) \in L \times \mathbb{R} | f(x) \leqslant \alpha\},$$

is a convex subset of $L \times \mathbb{R}$. It is said to be *proper convex* if it never takes the value $-\infty$ and it is not identically equal to ∞. A function g: $L \to [-\infty, \infty]$ is said to be *concave* if $-g$ is convex, and *proper concave* if $-g$ is proper

convex. Notice that g is concave if and only if its *hypograph*

$$\text{hypo}\, g := \{(x, \alpha) \in L \times \mathbb{R} | g(x) \geqslant \alpha\}$$

is convex. The *effective domain* of a convex function f is the set

$$\text{dom}\, f = \{x \in L | f(x) < \infty\}.$$

The *effective domain* of a concave function g is the set

$$\text{dom}\, g = \{x \in L | g(x) > \infty\}.$$

The only functions that are both proper convex and proper concave are the affine functions. The effective domain of each affine function is equal to L, both as a convex and concave function.

In optimization, we use *indicator functions* to encode constraints. The indicator function of a subset $C \subset L$ is the function

$$\delta_C(x) := \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{otherwise.} \end{cases}$$

Let now L and Λ be vector spaces paired by a bilinear mapping

$$\langle \cdot, \cdot \rangle L \times \Lambda \mathbb{R}(x, \xi) \langle x, \xi \rangle.$$

An standard example is $L = \mathbb{R}^d = \Lambda$ with the usual Euclidean scalar product. Another example is obtained by taking $L = L^1(V)$ and $\Lambda = L^\infty(V)$ with V a subset of $\mathbb{R}^n$.

The *convex conjugate* of a function f (convex or not) is defined as the function

$$f^\star(\xi) = \sup\{\langle x, \xi \rangle - f(x) | x \in X\}, \quad \xi \in \Lambda.$$

The *concave conjugate* of a function f (concave or not) is the function

$$f_\star(\xi) = \inf\{\langle x, \xi \rangle - f(x) | x \in X\}, \quad \xi \in \Lambda.$$

A remarkable fact is that convex conjugacy acts as an involution on certain classes of functions. For example, if $f: \mathbb{R}^d \to [-\infty, \infty]$ is a lower-semicontinuous proper convex function, then

$$f^{\star\star} := (f^\star)^\star = f.$$

Given a convex subset $C \in \mathbb{R}^d$, we call *relative interior* of C the interior of C with respect to its *affine hull* aff C. Recall that aff C is the smallest affine subspace that contains C. The relative interior of C is denoted by ri C. For example, if C is a closed segment in $\mathbb{R}^2$, its interior is empty while its relative interior is the segment without its ends. It can be shown that the relative interior of a nonempty convex set is necessarily nonempty.

Theorem 1. *(**Fenchel**) Let* f *and* g *be functions on* $\mathbb{R}^d$ *respectively* proper convex *and* proper concave *such that*

$$\text{ri dom}\, f \cap \text{ri dom}\, g \neq \varnothing. \tag{16}$$

Then

$$\eta := \inf_{\mathbf{x} \in \mathbb{R}^d} \{f(\mathbf{x}) - g(\mathbf{x})\} = \sup_{\boldsymbol{\xi} \in \mathbb{R}^d} \{g_\star(\boldsymbol{\xi}) - f^\star(\boldsymbol{\xi})\}$$

and the supremum *is attained.*

The above theorem asserts equality between the optimal values of two problems, together with attainment in the second one. Is customary to call these underlying optimization problems the *primal problem* and the *dual problem*, respectively.

In the above theorem, both the primal and dual are finite dimensional. Howover, problems such as ($\mathscr{P}$) have constraints involving some linear mapping. The following theorem will make it possible to *dualize* infinite dimensional problems with finitely many linear constraints.

Theorem 2. *Let be given:*

(i) L *and* Λ*, real vector spaces;*

(ii) $\langle \cdot, \cdot \rangle$*, a bilinear form on* $L \times \Lambda$;

(iii) $\mathbb{A}: L \to \mathbb{R}^d$*, a linear mapping;*

(iv) $F: L \to (-\infty, \infty]$*, a proper convex function;*

(v) $g: \mathbb{R}^d \to [-\infty, \infty)$*, a proper concave function.*

Assume that $\mathbb{A}$ admits a *formal adjoint mapping* $\mathbb{A}^\star$, that is, a linear mapping $\mathbb{A}^\star\colon \mathbb{R}^d \to \Lambda$ such that $\langle \mathbb{A}x,\, \mathbf{y}\rangle = \langle x,\, \mathbb{A}^\star\mathbf{y}\rangle$ for every $x \in L$ and every $\mathbf{y} \in \mathbb{R}^d$. Then, under the qualification condition

$$(QC) \quad \text{ri}\,(\mathbb{A}\,\text{dom}\,F) \cap \text{ri}\,(\text{dom}\,g) \neq \varnothing,$$

one has

$$\eta := \inf_{x\in X}\{F(x) - g(\mathbb{A}x)\} = \max_{\boldsymbol{\lambda}\in\mathbb{R}^d}\{g_\star(\boldsymbol{\lambda}) - F^\star(\mathbb{A}^\star\boldsymbol{\lambda})\}.$$

This theorem is the corner stone of what is is referred to as *partially finite convex programming*. Various forms appeared in the literature (see in particular [36, 37]). The selected form is as in [38], where no topological structure on the infinite dimensional side is requested. The optimization problems

$$\text{Minimize}\ (F - g\circ\mathbb{A}) \quad \text{and} \quad \text{Maximize}\ (g_\star - F^\star\circ\mathbb{A}^\star)$$

are respectively referred to as the *primal* and *dual* problems. The function $D := g_\star - F^\star\circ\mathbb{A}^\star$ appearing in the dual problem is referred to as the *dual function*. Again, the theorem asserts the equality between the optimal values of the primal and dual problems, together with *dual attainment*. The next result will provide conditions that will guarantee *primal attainment* as well.

Theorem 3. *(**Primal attainment**) With the notation and assumptions of the previous theorem, assume in addition that*

$$(QC^\star) \quad \text{ri dom}\, g_\star \cap \text{ri dom}(F^\star\circ\mathbb{A}^\star) \neq \varnothing.$$

Suppose further that

(a) $F^{\star\star} = F$ *and* $g_{\star\star} = g$;

(b) there exists $\bar{\boldsymbol{\lambda}}$ *dual optimal and* $\bar{x} \in \partial F^\star(\mathbb{A}^\star\bar{\boldsymbol{\lambda}})$ *such that* $F^\star\circ\mathbb{A}^\star$ *has gradient* $\mathbb{A}\bar{x}$ *at* $\bar{\boldsymbol{\lambda}}$.

Then $\bar{x}$ is primal optimal.

The latter result provides not only a condition for primal attainment, but it also makes appear as a watermark the possibility of a link between primal and dual solutions. The bi-conjugate relationships in Assumption (a) are central in the theorem, and the difficulty in our problem is to prove that the entropy satisfies this property. It turns out that, in our context, it is possible to compute the conjugate of the Boltzmann-Shannon entropy by *conjugating through the integral sign*.

An integral functional is a functional of the form

$$\mathscr{H}(p) = \int_V h(p(\mathbf{x}),\, \mathbf{x})\, \mathrm{d}\mu(\mathbf{x}), \qquad u \in L. \tag{17}$$

Here, V is assumed to be endowed with a σ-algebra of measurable sets and with a measure denoted by μ; the function h is called the integrand, and the argument p is assumed to pertain to some space of measurable functions L. In our context, it is enough to consider such functional on the familiar space $L = L^1(V)$, implicitly endowed with the Borel σ-algebra and the Lebesgue measure. Moreover, dependence of h in its second argument is not vital here, and we are only interested here in the case where $h(p(\mathbf{x}), \mathbf{x}) = h_\circ(p(\mathbf{x}))$.

Clearly, $h_\circ$ is a lower semi-continuous convex proper function so that it satisfies $h_\circ^{**} = h_\circ$. Conjugating $\mathscr{H}$ is elegantly performed by conjugating the integrand, as we shall see now. Following Rockafellar, we say that '*a space L of measurable functions is decomposable if it is stable under bounded alteration on sets of finite measure*.' Otherwise expressed, L is decomposable if and only if it contains all functions of the form

$$\mathbb{1}_T p_\circ + \mathbb{1}_{T^c} p,$$

in which T has finite measure, $p_\circ$ is a measurable function such that the set $p_\circ(T)$ is bounded, and p is any member of L. Here, 1_T denotes the characteristic function of T: $1_T(\mathbf{x})$ equals 1 if $\mathbf{x} \in T$ and $1_T(\mathbf{x})$ equals zero otherwise; and T^c denotes the complement of T. One can easily see that the familiar L^1-spaces are decomposable, which includes our workspace $L^1(V)$.

Theorem 4. *(**Rockafellar**) Let* L *and* Λ *be spaces of measurable functions on* Ω *paired by means of the standard integral bilinear form*

$$\langle f,\, \varphi\rangle = \int_V f(\mathbf{x})\varphi(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Let $\mathscr{H}$ be the functional of integrand $h_\circ$, that is,

$$\mathscr{H}(p) = \int_V h_\circ(p(\mathbf{x}))\,\mathrm{d}\mathbf{x},$$

with $h_\circ$ proper convex and lower semi-continuous. Assume that L *is decomposable and that $\mathscr{H}$ has nonempty effective domain. Then*

$$\mathscr{H}^\star(\varphi) = \int h_\circ^\star(\varphi(\mathbf{x}))\,\mathrm{d}\mathbf{x}$$

for every $\varphi \in \Lambda$, and $\mathscr{H}^\star$ is convex on Λ.

Applying the latter theorem with $L = L^1(V)$, $\Lambda = L^\infty(V)$ and $h_\circ$ the above defined integrand of the Boltzmann-Shannon neg-entropy we see that, in our case,

$$\mathscr{H}^\star(\varphi) = \int_V h_\circ^\star(\varphi(\mathbf{x}))\,\mathrm{d}\mathbf{x}.$$

By means of an easy computation, we can see that the function $h_\circ^\star$ is given by

$$h_\circ^\star(\tau) = \exp(\tau - 1), \quad \tau \in \mathbb{R}.$$

Finally, since $\Lambda = L^\infty(V)$ is also decomposable, we obtain that

$$\mathscr{H}^{\star\star}(p) = \int_V h_\circ^{\star\star}(p(\mathbf{x}))\,\mathrm{d}\mathbf{x} = \int_V h_\circ(p(\mathbf{x}))\,\mathrm{d}\mathbf{x} = \mathscr{H}(p).$$

We conclude that our entropy satisfies the bi-conjugacy relationship requested in theorem 3.

Before returning to our specific problem, let us state one more result, in which an explicit relationship between primal and dual solutions is obtained.

Theorem 5. *(**Primal-dual relationship**) With the notation and assumptions of theorem 2 assume in addition that* dom *D has nonempty interior, that $\mathscr{H}$ is an integral functional of integrand* h *such that conjugacy through the integral sign is permitted. Assume that, as in theorem 3, $\mathscr{H}^{\star\star} = \mathscr{H}$ and $g_{\star\star} = g$. Assume finally that the conjugate integrand $h^\star$ is differentiable over $\mathbb{R}$, and that there exists some dual-optimal vector $\bar{\boldsymbol{\lambda}}$ in int* dom *D. If*

$$\bar{p}(\mathbf{x}) \coloneqq h^{\star\prime}([\mathbb{A}^\star\bar{\boldsymbol{\lambda}}](\mathbf{x}), \mathbf{x}) \in L,$$

then $\bar{p}$ is a primal solution.

We are now ready to get back to our specific entropy problem.

4. Maximum entropy densities

4.1. The case of Boltzmann-Shannon entropy

Problem $(\mathscr{P})$ can be written as

$$\text{Minimize} \quad \mathscr{H}(p) - g_\circ(\mathbb{A}_\circ p)$$

over the space $L^1(V)$, in which

$$\mathbb{A}_\circ p = (\mathbb{I}p, \mathbb{A}p) \in \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$$

and $g_\circ$ is the function on $\mathbb{R}^2$ given by

$$g_\circ(\eta_\circ, \eta) = -\delta_{\{E_\circ\}}(\eta) - \delta_{\{1\}}(\eta_\circ).$$

The use of the indicator functions $\delta_{\{1\}}(\cdot)$ and $\delta_{\{E_\circ\}}(\cdot)$ enables to encode the constraints in $(\mathscr{P})$. Straightforward computations show that the adjoint mapping $\mathbb{A}_\circ^\star\colon \mathbb{R}^2 \to L^\infty(V)$ is given by

$$\mathbb{A}_\circ^\star(\lambda_\circ, \lambda)(\mathbf{x}) = \lambda_\circ + \lambda H(\mathbf{x}).$$

and that

$$(g_\circ)_\star(\lambda_\circ, \lambda) = \lambda_\circ + \lambda E_\circ.$$

Accounting for the fact that, as we have seen above, the entropy can be conjugated by conjugating through the integral sign, the dual problem reads:

$$\text{Maximize} \quad D(\lambda_\circ, \lambda) \coloneqq \lambda_\circ + \lambda E_\circ - \exp(\lambda_\circ - 1)\int_V \exp(\lambda H(\mathbf{x}))\,\mathrm{d}\mathbf{x}.$$

The function D is obviously concave and differentiable on $\mathbb{R}^2$. Its stationary points must satisfy the system

$$\begin{cases} 0 &= 1 - \exp(\bar{\lambda}_\circ - 1)\int_V \exp\left(\bar{\lambda}H(\mathbf{x})\right) \mathrm{d}\mathbf{x}, \\ 0 &= E_\circ - \exp(\bar{\lambda}_\circ - 1)\int_V H(\mathbf{x})\exp\left(\bar{\lambda}H(\mathbf{x})\right) \mathrm{d}\mathbf{x}, \end{cases}$$

which reduces to

$$0 = E_\circ - \frac{\int_V H(\mathbf{x})\exp\left(\bar{\lambda}H(\mathbf{x})\right) \mathrm{d}\mathbf{x}}{\int_V \exp\left(\bar{\lambda}H(\mathbf{x})\right) \mathrm{d}\mathbf{x}}. \tag{18}$$

Notice that the equality in (18) is also the first order optimality condition of the problem,

$$(\tilde{\mathscr{D}}) \quad \left| \begin{array}{ll} \text{Maximize} & \lambda E_\circ - \ln \int_V \exp(\lambda H(\mathbf{x})) \mathrm{d}\mathbf{x} \\ \text{s.t.} & \lambda \in \mathbb{R}. \end{array} \right.$$

Proposition 6. *The function*

$$\tilde{D}(\lambda) \coloneqq \lambda E_\circ - \ln \int_V \exp(\lambda H(\mathbf{x})) \mathrm{d}\mathbf{x} \tag{19}$$

to be maximized in Problem $(\tilde{\mathscr{D}})$ *is concave, smooth and everywhere finite.*

The function $h_\circ^\star(\tau) = \exp(\tau - 1)$ obviously meets the requirements of theorem 5. Provided we can obtain a dual solution $(\bar{\lambda}_\circ, \bar{\lambda})$, the optimal density is then given by

$$\bar{p}(\mathbf{x}) = \exp\left[\bar{\lambda}_\circ - 1 + \bar{\lambda}H(\mathbf{x})\right] = \frac{\exp(\bar{\lambda}H(\mathbf{x}))}{\int_V \exp(\bar{\lambda}H(\mathbf{x})) \mathrm{d}\mathbf{x}}, \tag{20}$$

where $\bar{\lambda}$ maximizes the function $\tilde{D}$.

Note that $\bar{p}(\mathbf{x})$ corresponds to the canonical ensemble $p(\mathbf{\Gamma}; \beta)$ under the identification $\lambda = -\beta$, and also $\tilde{\mathscr{D}}(\lambda)$ corresponds to the negative of the maximized entropy $\mathcal{S}$ in units of k_B.

4.2. The case of Tsallis entropy

Tsallis entropy of a probability density p is define as the integral

$$\mathscr{S}_q(p) = \frac{1}{q-1}\left(1 - \int p(\mathbf{x})^q \mathrm{d}\mathbf{x}\right) \tag{21}$$

provided $p \in L^q(V)$. The computations differ depending on whether $q > 1$ or $q \in (0, 1)$.

4.2.1. Case q > 1

Maximizing the Tsallis entropy is equivalent to minimizing the integral functional

$$\mathscr{T}_q(p) = \int h_q(p(\mathbf{x})) \mathrm{d}\mathbf{x}, \tag{22}$$

in which

$$h_q(t) = \begin{cases} \dfrac{1}{q}t^q & \text{if } t \geqslant 0, \\ \infty & \text{if } t < 0. \end{cases} \tag{23}$$

Notice first that $L^q(V) \subset L^1(V)$, since the Lebesgue is a finite measure on V. An immediate consequence is that $L^1(V) \cap L^q(V) = L^q(V)$, and our optimization problem takes place in the decomposable space $L^q(\mathbb{R}^n)$. As in the case of the Boltzmann-Shannon entropy, the integrand h_q is proper convex and lower semi-continuous. We now proceed to compute its convex conjugate. We have:

$$h_q^\star(\tau) = \sup_{t\in\mathbb{R}} (t\tau - h_q(t)) = \sup_{t\in\mathbb{R}_+} (t\tau - q^{-1}t^q).$$

If $\tau \leqslant 0$, the above supremum is attained at $t = 0$, so that $h_q^\star(\tau) = 0$. Suppose now that $\tau > 0$. The function $t \mapsto t\tau - q^{-1}t^q$ is differentiable on $\mathbb{R}_+^*$. Its derivative, the function $t \mapsto \tau - t^{q-1}$, vanishes at

$$t = \tau^{\frac{1}{q-1}}.$$

Therefore, on denoting q' the conjugate exponent of q (defined by the relationship $q^{-1} + q'^{-1} = 1$), we obtain:

$$h_q^\star(\tau) = \tau^{\frac{1}{q-1}}\tau - q^{-1}\tau^{\frac{q}{q-1}} = \left(1 - \frac{1}{q}\right)\tau^{\frac{q}{q-1}} = q'^{-1}\tau^{q'}.$$

In summary, the conjugate is given by

$$h_q^\star(\tau) = \begin{cases} \dfrac{1}{q'}\tau^{q'} & \text{if } \tau > 0, \\ 0 & \text{if } \tau \leqslant 0. \end{cases}$$

As in the case of the Boltzmann-Shannon entropy, the conjugate integrand is everywhere finite and differentiable. Everything is then similar to the Boltzmann-Shannon entropy case, except for the trick consisting in considering Problem $\tilde{\mathscr{D}}$; The dual function is now given by

$$D(\lambda_\circ, \lambda) = \lambda_\circ + \lambda E_\circ - \int h_q^\star(\lambda_\circ + \lambda H(\mathbf{x}))\,\mathrm{d}\mathbf{x} = \lambda_\circ + \lambda E_\circ - \frac{1}{q'}\int_{\lambda_\circ + \lambda H(\mathbf{x}) > 0} (\lambda_\circ + \lambda H(\mathbf{x}))^{q'}\,\mathrm{d}\mathbf{x}.$$

Its effective domain (as a concave function) is the set of $(\lambda_\circ, \lambda) \in \mathbb{R}^2$ such that the function $\mathbf{x} \mapsto h_q^\star(\lambda_\circ + \lambda H(\mathbf{x}))$ is integrable on V. If H is bounded on V, then the domain of D is $\mathbb{R}^2$. In this case, the optimality system is, as usual, provided by Fermat's principle, which reads here:

$$\begin{cases} 1 = \displaystyle\int_{\lambda_\circ + \lambda H(\mathbf{x}) > 0} (\lambda_\circ + \lambda H(\mathbf{x}))^{q'-1}\,\mathrm{d}\mathbf{x}, \\ E_\circ = \displaystyle\int_{\lambda_\circ + \lambda H(\mathbf{x}) > 0} H(\mathbf{x})(\lambda_\circ + \lambda H(\mathbf{x}))^{q'-1}\,\mathrm{d}\mathbf{x}. \end{cases}$$

If $(\bar{\lambda}_\circ, \bar{\lambda})$ denotes a solution to the above system, more likely to be obtained via the maximization of $D(\lambda_\circ, \lambda)$, then the optimal probability is given by

$$\bar{p}(\mathbf{x}) = h_p^{\star\prime}(\bar{\lambda}_\circ + \bar{\lambda}H(\mathbf{x})) = \begin{cases} 0 & \text{if } \bar{\lambda}_\circ + \bar{\lambda}H(\mathbf{x}) \leqslant 0, \\ (\bar{\lambda}_\circ + \bar{\lambda}H(\mathbf{x}))^{\frac{1}{q-1}} & \text{if } \bar{\lambda}_\circ + \bar{\lambda}H(\mathbf{x}) > 0. \end{cases} \tag{24}$$

4.2.2. Case q ∈ *(0, 1)*

It is readily seen that maximizing the Tsallis entropy in this case is equivalent to minimizing the integral functional

$$\mathscr{T}_q(p) = \int h_q(p(\mathbf{x}))\,\mathrm{d}\mathbf{x}, \quad \text{with} \quad h_q(t) = \begin{cases} -\dfrac{1}{q}t^q & \text{if } t \geqslant 0, \\ \infty & \text{if } t < 0. \end{cases}$$

The above functional is well-defined on the vector space $L^1(V) \cap L^q(V)$. Note that the mapping

$$f \mapsto \left(\int |f|^q\right)^{1/q}$$

fails to be a norm, as is the case when $q \geqslant 1$, since it does not satisfy the triangle inequality. However, the following holds:

- the functional $N_q(f) \coloneqq \int |f|^q$ satisfies the triangle inequality $\int |f_1 + f2|^q \leqslant |f_1|^q + |f_2|^q$;
- $L^q(V)$ is complete metric space with the distance

$$d(f_1, f_2) \coloneqq \int |f_1 - f_2|^q.$$

An immediate consequence of the first point is that $L^q(V)$ is decomposable. Let $f \in L^q(V)$, let $T \subset V$ be a measurable set of finite measure, and let $f_\circ$ be a (measurable and) bounded on T. Then,

$$\int |\mathbb{1}_T f_\circ + \mathbb{1}_{T^c} f|^q \leqslant \int_T |f_\circ|^q + \int_{T^c} |f|^q;$$

In the right hand term, the first integral is finite since $|f_\circ|^q$ is bounded on T and T has finite measure, and the second integral is also finite since it is bounded above by $N_q(f)$ and $N_q(f)$ is finite.

The decomposability of both $L^q(V)$ and $L^1(V)$ implies, of course, that of their intersection. This will enable us to conjugate $\mathscr{T}_q$ through the integral sign.

Let us now compute the conjugate if the function h_q. As in the previous case, we have:

$$h_q^\star(\tau) = \sup_{t\in\mathbb{R}} (t\tau - h_q(t)) = \sup_{t\in\mathbb{R}_+} (t\tau + q^{-1}t^q).$$

It is easy to see that, if $\tau \geqslant 0$, then $h_q^\star(\tau) = \infty$. Suppose now that $\tau < 0$, and let us search for a maximizer of the function $t \mapsto t\tau + q^{-1}t^q$ on $\mathbb{R}_-^*$. The latter function is clearly differentiable on $\mathbb{R}_+^*$. Its derivative, the function $t \mapsto \tau + t^{q-1}$, vanishes at

$$t = (-\tau)^{\frac{1}{q-1}}.$$

We therefore have

$$h_q^\star(\tau) = (-\tau)^{\frac{1}{q-1}}\tau + \frac{1}{q}(-\tau)^{\frac{q}{q-1}} = \frac{1-q}{q}(-\tau)^{\frac{q}{q-1}}.$$

In terms of the conjugate exponent q' (which is now negative), the conjugate function is given by

$$h_q^\star(\tau) = \begin{cases} \infty & \text{if } \tau \geqslant 0, \\ -\dfrac{1}{q'}(-\tau)^{q'} & \text{if } \tau < 0. \end{cases}$$

The dual function is given by

$$D(\lambda_\circ, \lambda) = \lambda_\circ + \lambda E_\circ - \int h_q^\star(\lambda_\circ + \lambda H(\mathbf{x}))\, \mathrm{d}\mathbf{x}. \tag{25}$$

We observe right away that, unlike in the cases of Tsallis entropy with $q > 1$ or Boltzmann-Shannon entropy, the conjugate integrand takes infinite values on nonnegative arguments. However, $h_q^\star$ remains differentiable on its domain $\mathbb{R}_-^*$, with derivative

$$(h_q^\star)'(\tau) = (-\tau)^{q'-1} = (-\tau)^{\frac{1}{q-1}}.$$

Provided that we can find a maximizer $(\bar\lambda_\circ, \bar\lambda)$ of the dual function, the optimal density is then given by

$$\bar p(\mathbf{x}) = h_p^{\star\prime}(\bar\lambda_\circ + \bar\lambda H(\mathbf{x})) = (-(\bar\lambda_\circ + \bar\lambda H(\mathbf{x})))^{\frac{1}{q-1}}. \tag{26}$$

4.3. The case of Tsallis entropy with escort distribution

This case involves the generalized form of entropy $\mathscr{S}_q$ but including the use of so-called escort probabilities, giving in principle a more flexible treatment of non-extensive systems [39]. The escort distributions in the context of Tsallis entropy in fact weights the probability densities of states differently, which could be useful in scenarios where certain states are more relevant or significant. The interpretation of these escort distributions is an interesting issue, from their connection to fractality [18, 21] to the view regarding them as interpolation between distributions [40]. Nevertheless, consistently applying escort distributions results in undefined behavior, and only achieves the desired outcomes when employed in an *ad-hoc* manner. Indeed, as a matter of example, it has been shown that 'any deformed entropy expression, maximized with the escort averaged constraints, yields that the Shannon entropy is equal to the logarithm of the ordinary canonical partition function i.e. $S = \ln(Z_S)$ instead of the correct thermodynamic relation' [41].

In this section we depict this formulation rigorously as an optimization problem.

Remark 7. The Tsallis entropy is usually optimized under constraint on the so-called *escort distribution*, which is defined as

$$P(\mathbf{x}) = \frac{p(\mathbf{x})^q}{\int p(\mathbf{x})^q\, \mathrm{d}\mathbf{x}}. \tag{27}$$

It is readily seen that P is a probability distribution whenever p is a probability distribution. The maximum Tsallis entropy problem would then read:

$$(\mathscr{P}_{\mathrm{ESC}}) \quad \left| \begin{array}{ll} \text{Maximize} & \mathscr{S}_q(p) \\ \text{subject to} & p \in L^1(V) \cap L^q(V), \\ & \displaystyle\int_V p(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1, \\ & \displaystyle\int_V p(\mathbf{x})^q H(\mathbf{x})\, \mathrm{d}\mathbf{x} = \mathscr{E}_\circ, \end{array} \right.$$

in which, again,

$$\mathscr{S}_q(p) = \begin{cases} \dfrac{1}{q-1}\left(1 - \displaystyle\int p(\mathbf{x})^q \, \mathrm{d}\mathbf{x}\right) & \text{if } p \geqslant 0 (\text{almost everywhere}), \\ \infty & \text{otherwise.} \end{cases}$$

Here the replacement of $E_\circ$ by $\mathscr{E}_\circ$ is justified by the normalizing denominator in (27). See e.g. [17], pages 88-89. In this case, we see that, unless $q = 1$, the last constraint in ($\mathscr{P}_{\mathrm{ESC}}$) fails to be linear, yielding a nonconvex optimization problem. In this case, the optimization problem is more difficult. In addition to the difficulties we observed when dealing with variational calculus, formal derivatives and so on, the nonconvexity entails the possibility for minima to be local and not global. Moreover, it seems questionable to impose some moment constraints on the original density p and some other moment constraints on the corresponding escort distribution.

5. Conclusion

In this paper, we have clarified the derivation of Maximum Entropy distributions in statistical physics, both in the case of Shannon entropy and that of Tsallis entropy. We believe this clarification was necessary, since the usual variational calculus approach is not sufficient to guarantee optimality in the corresponding infinite-dimensional optimization problems.

The standard problem is efficiently addressed in both the Shannon and Tsallis cases. However, the use of escort distributions can be solved without resorting to Fenchel's duality, since the resolution is trivial in this case. This resolution reveals that little can be done with such formalism, since the problem has either no solution or infinitely many solutions. On the other hand, when the escort distribution is used for the expectation constraint on the Hamiltonian but not for normalization, we show that the usual q-exponential family solution is recovered. Thus we are forced with the choice of either use the escort distributions inconsistently (since the argument of the entropy is not the same as the distributions used in the constraints) or otherwise deal with an undetermined solution.

Finally, our results imply that escort distributions can be used just in very particular cases of non-linearity in an ad-hoc manner.

Acknowledgments

We are grateful for the feedback from three anonymous reviewers. SD and YN thankfully acknowledge funding from ANID FONDECYT 1220651 grant.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Pierre Maréchal https://orcid.org/0000-0001-7121-5699
Yasmín Navarrete https://orcid.org/0000-0002-4115-8213
Sergio Davis https://orcid.org/0000-0003-2757-332X

References

[1] Richardson W H 1972 Bayesian-based iterative method of image restoration∗ *J. Opt. Soc. Am.* **62** 55–9
[2] Gull S F and Skilling J 1984 Maximum entropy method in image processing *IEE Proceedings F (Communications, Radar and Signal Processing)* 131 646–59
[3] Phillips S J and Dudik M 2008 Modeling of species distributions with maxent: new extensions and a comprehensive evaluation *Ecography* **31** 161–75
[4] Elith J *et al* 2006 Novel methods improve prediction of species' distributions from occurrence data *Ecography* **29** 129–51
[5] Narayan R and Nityananda R 1986 Maximum entropy image restoration in astronomy *Annu. Rev. Astron. Astrophys.* **24** 127–70
[6] Davis S, Navarrete Y and Gutiérrez G 2014 A maximum entropy model for opinions in social groups *Eur. Phys. J.*B **87** 78
[7] Xu B and Wang Z 2012 Test maxent in social strategy transitions with experimental two-person constant sum 2 × 2 games *Results in Physics* **2** 127–34
[8] Rioux G, Choksi R, Hoheisel T, Maréchal P and Scarvelis C 2020 The maximum entropy on the mean method for image deblurring *Inverse Prob.* **37** 015011

[9] Gibbs J W 1902 *Elementary Principles in Statistical Mechanics* (Charles Scribner's Sons)
[10] Jaynes E T 1957 Information theory and statistical mechanics *Phys. Rev.* **106** 620–30
[11] Shore J and Johnson R W 1980 Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy *IEEE Trans. Inf. Theory* **26** 26–37
[12] Lima J, Silva R and Santos J 2000 Plasma oscillations and nonextensive statistics *Phys. Rev.* E **61** 3260–3
[13] Pavlos G P, Iliopoulos A C, Zastenker G N, Zelenyi L M, Karakatsanis L P, Riazantseva M O, Xenakis M N and Pavlos E G 2015 Tsallis non-extensive statistics and solar wind plasma complexity *Phys.* A **422** 113–35
[14] Arimitsu T and Arimitsu N 2000 Analysis of fully developed turbulence in terms of Tsallis statistics *Phys. Rev.* E **61** 3237
[15] Komatsu N, Kiwata T and Kimura S 2012 Transition of velocity distributions in collapsing self-gravitating N-body systems *Phys. Rev.* E **85** 021132
[16] Peterson J, Dixit P D and Dill K A 2013 A maximum entropy framework for nonexponential distributions *Proc. Nat. Acad. Sci.* 110 20380–5
[17] Tsallis C 2009 *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*
[18] Naudts J 2011 *Generalised Thermostatistics* (Springer)
[19] Abe S 2003 Geometry of escort distributions *Phys. Rev.* E **68** 031101
[20] Abe S 2006 Why q-expectation values must be used in nonextensive statistical mechanics *Astrophys. Space Sci.* **305** 241–5
[21] Beck C and Schögl F 1995 *Thermodynamics of chaotic systems: An introduction* (Cambridge University Press)
[22] Nauenberg M 2003 Critique of *q*-entropy for thermal statistics *Phys. Rev.* E **67** 036114
[23] Pressé S, Ghosh K, Lee J and Dill K A 2013 Principles of maximum entropy and maximum caliber in statistical physics *Rev. Mod. Phys.* **85** 1115–41
[24] Bagci G B and Oikonomou T 2013 Tsallis power laws and finite baths with negative heat capacity *Phys. Rev.* E **88** 042126
[25] Pressé S 2014 Nonadditive entropy maximization is inconsistent with Bayesian updating *Phys. Rev.* E **90** 052149
[26] Pessoa P and Costa B A 2020 Comment on Tsallis, C. Black hole entropy: a closer look *Entropy* **22** 17
Pessoa P and Costa B A 2020 Comment on Tsallis, C. Black hole entropy: a closer look *Entropy* **22** 1110
[27] Pessoa P, Costa F X and Caticha A 2021 Entropic dynamics on Gibbs statistical manifolds *Entropy* **23** 494
[28] Caticha A 2021 Entropy, information, and the updating of probabilities *Entropy* **23** 895
[29] Bercher J-F 2008 Tsallis distribution as a standard maximum entropy solution with 'tail' constraint *Phys. Lett.* A **372** 5657–9
[30] Hernando A, Plastino A and Plastino A R 2012 MaxEnt and dynamical information *Eur. Phys. J.* B **85** 1–8
[31] Visser M 2013 Zipf's law, power laws and maximum entropy *New J. Phys.* **15** 043021
[32] Oikonomou T, Kaloudis K and Bagci G B 2021 The q-exponentials do not maximize the Rényi entropy *Phys.* A **578** 126126
[33] Ramshaw J D 2022 Maximum entropy and constraints in composite systems *Phys. Rev.* E **105** 024138
[34] Beck C and Cohen E G D 2003 Superstatistics *Physica* A **322** 267–75
[35] Jizba P and Korbel J 2019 Maximum entropy principle in statistical inference: case for non-Shannonian entropies *Phys. Rev. Lett.* **122** 120601
[36] Borwein J M and Lewis A S 1992 Partially finite convex programming, part i: Quasi relative interiors and duality theory *Math. Program.* **57** 15–48
[37] Borwein J M and Lewis A S 1992 Partially finite convex programming, part ii: explicit lattice models *Math. Program.* **57** 49–83
[38] Maréchal P 1998 *On the principle of maximum entropy on the mean as a methodology for the regularization of inverse problems* ed B Grigelionis (Walter de Gruyter GmbH) pp 481–92
[39] Kalogeropoulos N 2012 *Escort Distributions and Tsallis Entropy* arXiv:1206.5127
[40] Bercher J-F 2011 On escort distributions, q-Gaussians and Fisher information *AIP Conf. Proc* 1305 208–15
[41] Oikonomou T and Bagci G B 2017 *Impossible Mission: Entropy Maximization with Escort Averages* arXiv:1704.04721