Entropy **2011**, *13*, 1170-1185; doi:10.3390/e13061170

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Geometry of q-Exponential Family of Probability Distributions

Shun-ichi Amari [1,⋆] **and Atsumi Ohara** [2,⋆]

[1] Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Hirosawa 2-1, Wako-shi, Saitama 351-0198, Japan

[2] Department of Electrical and Electronics Engineering, Graduate School of Engineering, University of Fukui, Bunkyo 3-9-1, Fukui-shi, Fukui 910-8507, Japan

⋆ Authors to whom correspondence should be addressed; E-Mails: amari@brain.riken.jp (S.-i.A.); ohara@fuee.u-fukui.ac.jp (A.O.).

Received: 11 February 2011; in revised form: 1 June 2011 / Accepted: 2 June 2011 / Published: 14 June 2011

Abstract: The Gibbs distribution of statistical physics is an exponential family of probability distributions, which has a mathematical basis of duality in the form of the Legendre transformation. Recent studies of complex systems have found lots of distributions obeying the power law rather than the standard Gibbs type distributions. The Tsallis q-entropy is a typical example capturing such phenomena. We treat the q-Gibbs distribution or the q-exponential family by generalizing the exponential function to the q-family of power functions, which is useful for studying various complex or non-standard physical phenomena. We give a new mathematical structure to the q-exponential family different from those previously given. It has a dually flat geometrical structure derived from the Legendre transformation and the conformal geometry is useful for understanding it. The q-version of the maximum entropy theorem is naturally induced from the q-Pythagorean theorem. We also show that the maximizer of the q-escort distribution is a Bayesian MAP (Maximum A posteriori Probability) estimator.

Keywords: q-exponential family; q-entropy; information geometry; q-Pythagorean theorem; q-Max-Ent theorem; conformal transformation

---

1. Introduction

Statistical physics is founded on the Gibbs distribution for microstates, which forms an exponential family of probability distributions known in statistics. Important macro-quantities such as energy, entropy, free energy, *etc.* are connected with it. However, recent studies show that there are non-standard complex systems which are subject to the power law instead of the exponential law of the Gibbs type distributions. See [1,2] as well as extensive literatures cited in them.

Tsallis [3] defined the q-entropy to elucidate various physical phenomena of this type, followed by many related research works on this subject (see, [1]). The concept of the q-Gibbs distribution or q-exponential family of probability distributions is naturally induced from this framework (see also [4]). However, its mathematical structure has not yet been explored enough [2,5,6], while the Gibbs type distribution has been studied well as the exponential family of distributions [7]. We need a mathematical (geometrical) foundation to study the properties of the q-exponential family. This paper presents a geometrical foundation for the q-exponential family based on information geometry [8], giving geometrical definitions of the q-potential function, q-entropy and q-divergence in a unified way.

We define the q-geometrical structure consisting of a Riemannian metric and a pair of dual affine connections. By using this framework, we prove that a family of q-exponential distributions is dually flat, in which the q-Pythagorean theorem holds. This naturally induces the corresponding q-maximum entropy theorem similarly to the case of the Tsallis q-entropy [1,9,10]. The q-structure is ubiquitous since the family S_n of all discrete probability distributions can always be endowed with the structure of the q-exponential family for arbitrary q. It is possible to generalize the q-structure to any family of probability distributions. Further, it has a close relation with the α-geometry [8], which is one of information geometric structure of constant curvature. This new dually flat structure, different from the old one given rise to from the invariancy in information geometry, can be also obtained by conformal flattening of the α-geometry [11,12], using a technique in the conformal and projective geometry [13–15].

The present framework prepares mathematical tools for analyzing physical phenomena subject to the power law. The Legendre transformation again plays a fundamental role for deriving the geometrical dual structure. There exist lots of applications of q-geometry to information theory ([16] and others) and statistics, including Bayes q-statistics.

It is possible to generalize our framework to a more general non-linear family of distributions by using a positive convex function instead of q-exponential function (See [2,17]). A good example is the κ-exponential family [18–20], but we do not state it here.

2. q-Gibbs or q-Exponential Family of Distributions

2.1. q-Logarithm and q-Exponential Function

It is the first step to generalize the logarithm and exponential functions to include a family of power functions, where the logarithm and exponential functions are included as the limiting case [1,5,21]. This

was also used for defining the α-family of distributions in information geometry [8]. We define the q-logarithm by

$$\log_q(u) = \frac{1}{1-q}\left(u^{1-q} - 1\right), \quad u > 0 \tag{1}$$

and its inverse function, the q-exponential, by

$$\exp_q(u) = \{1 + (1-q)u\}^{\frac{1}{1-q}}, \quad u > -1/(1-q) \tag{2}$$

for a positive q with $q \neq 1$. The limiting case $q \to 1$ reduces to

$$\log_1(u) = \log u \tag{3}$$

$$\exp_1(u) = \exp u \tag{4}$$

so that $\log_q$ and $\exp_q$ are defined for $q > 0$.

2.2. *q-Exponential Family*

The standard form of an exponential family of distributions is written as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\sum \theta^i x_i - \psi(\boldsymbol{\theta})\right\} \tag{5}$$

with respect to an adequate measure $\mu(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \cdots, x_n)$ is a set of random variables and $\boldsymbol{\theta} = (\theta^1, \cdots, \theta^n)$ are the canonical parameters to describe the underlying system. The Gibbs distribution is of this type. Here, $\psi(\boldsymbol{\theta})$ is called the free energy, which is the cumulant generating function.

The power version of the Gibbs distribution is written as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp_q\left\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_q(\boldsymbol{\theta})\right\} \tag{6}$$

$$\log_q\left\{p(\boldsymbol{x}, \boldsymbol{\theta})\right\} = \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_q(\boldsymbol{\theta}) \tag{7}$$

where $\boldsymbol{\theta} \cdot \boldsymbol{x} = \sum \theta^i x_i$. This is the q-Gibbs distribution or q-exponential family [4], which we denote by S, where the domain of $\boldsymbol{x}$ is restricted such that $p(\boldsymbol{x}, \boldsymbol{\theta}) > 0$ holds. The function $\psi_q(\boldsymbol{\theta})$, called the q-free energy or q-potential function, is determined from the normalization condition:

$$\int \exp_q\left\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_q(\boldsymbol{\theta})\right\} d\boldsymbol{x} = 1 \tag{8}$$

where we replaced $d\mu(\boldsymbol{x})$ by $d\boldsymbol{x}$ for brevity's sake. The function ψ_q depends on q, but we hereafter neglect suffix q in most cases. Research on the q-exponential family can be found, for example, in [2,4,19]. The q-Gaussian distribution is given by

$$p(x, \mu, \sigma) = \exp_q\left\{-\frac{(x-\mu)^2}{2\sigma^2} - \psi(\mu, \sigma)\right\} \tag{9}$$

and is studied in [22–25] in detail. Here, we need to introduce a vector random variable $\boldsymbol{x} = (x, x^2)$ and a new parameter $\boldsymbol{\theta}$, which is a vector-valued function of μ and σ, to represent it in the standard form (7). It is an interesting observation that the domain of $\boldsymbol{x}$ in the q-Gaussian case depends on q if $0 < q < 1$. Hence, that q- and q'-Gaussian are in general not absolutely continuous when $q \neq q'$.

It should be remarked that the q-exponential family itself is the same as the α-family of distributions in information geometry [8]. Here, we introduce a different geometrical structure, generalizing the result of [24].

We mainly use the family S_n of discrete distributions over $(n+1)$ elements $X = \{x_0, x_1, \cdots, x_n\}$, although we can easily extend the results to the case of continuous random variables. Here, random variable x takes values over X. We also treat the case of $0 < q < 1$, and the limiting cases of $q = 0$ or 1 give the well-known ones.

Let us put $p_i = \text{Prob}\{x = x_i\}$ and denote the probability distribution by vector $\boldsymbol{p} = (p_0, p_1, \cdots, p_n)$, where

$$\sum_{i=1}^{n} p_i = 1 \tag{10}$$

The probability of x is also written as

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x) \tag{11}$$

where

$$\delta_i(x) = \begin{cases} 1, & x = x_i, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Theorem 1 The family S_n of discrete probability distributions has the structure of a q-exponential family for any q.

Proof We take $\log_q$ of distribution $p(x)$ of (11). For any function $f(u)$, we have

$$f\left\{\sum_{i=1}^{n} p_i \delta_i(x)\right\} = \sum_{i=0}^{n} f\left(p_i\right) \delta_i(x) \tag{13}$$

By taking

$$\delta_0(x) = 1 - \sum_{i=1}^{n} \delta_i(x) \tag{14}$$

into account, discrete distribution (11) can be rewritten in the form (8) as

$$\log_q p(x) = \frac{1}{1-q}\left\{\sum_{i=1}^{n}\left(p_i^{1-q} - p_0^{1-q}\right)\delta_i(x) + p_0^{1-q} - 1\right\} \tag{15}$$

where

$$p_0 = 1 - \sum_{i=1}^{n} p_i \tag{16}$$

is treated as a function of $(p_1, \cdots, p_n)$. Hence, S_n is q-exponential family (6) for any q, with the following q-canonical parameters, random variables and q-potential function:

$$\theta^i = \frac{1}{1-q}\left(p_i^{1-q} - p_0^{1-q}\right), \quad i = 1, \cdots, n \tag{17}$$

$$x_i = \delta_i(x) \tag{18}$$

$$\psi(\boldsymbol{\theta}) = -\log_q p_0 \tag{19}$$

This completes the proof. □

Note that the q-potential $\psi(\boldsymbol{\theta})$ and the canonical parameter $\boldsymbol{\theta}$ depend on q as is seen in (17) and (19). It should also be remarked that Theorem 1 does not contradict to the theorem 1 in [19] stating that a parametrized family of probability distributions can belong to at most one q-exponential family. The author considers an m-dimensional parametrized submanifold in S_n with $m < n$ where the canonical parameter depending on q is given via the variational principle. Therefore, by denoting the q-canonical parameter by $\boldsymbol{\theta}_q \in \mathbf{R}^m$, we can restate his theorem in terms of geometry that a linear submanifold parametrized by $\boldsymbol{\theta}_q \in \mathbf{R}^m$ is not a linear submanifold parametrized by $\boldsymbol{\theta}_{q'} \in \mathbf{R}^m$ when $q' \neq q$. On the other hand, the present theorem states that there exists the q-canonical parameter $\boldsymbol{\theta}_q \in \mathbf{R}^n$ on whole S_n for any q and the manifold has linear structure with respect to any $\boldsymbol{\theta}_q$. This is a surprising new finding.

2.3. q-Potential Function

We study the q-geometrical structure of S. The q-log-likelihood is a linear form defined by

$$l_q(\boldsymbol{x}, \boldsymbol{\theta}) = \log_q p(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \theta^i x_i - \psi(\boldsymbol{\theta}) \tag{20}$$

By differentiating it with respect to θ^i, with the abbreviated notation $\partial_i = \frac{\partial}{\partial \theta^i}$, we have

$$\begin{aligned} \partial_i l_q(\boldsymbol{x}, \boldsymbol{\theta}) &= x_i - \partial_i \psi(\boldsymbol{\theta}) & (21) \\ \partial_i \partial_j l_q(\boldsymbol{x}, \boldsymbol{\theta}) &= -\partial_i \partial_j \psi(\boldsymbol{\theta}) & (22) \end{aligned}$$

From this we have the following important theorem.

Theorem 2 The q-free energy or q-potential $\psi_q(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}_q$.

Proof We omit the suffix q for simplicity's sake. We have

$$\begin{aligned} \partial_i p(\boldsymbol{x}, \boldsymbol{\theta}) &= p(\boldsymbol{x}, \boldsymbol{\theta})^q \left(x_i - \partial_i \psi\right) & (23) \\ \partial_i \partial_j p(\boldsymbol{x}, \boldsymbol{\theta}) &= q p(\boldsymbol{x}, \boldsymbol{\theta})^{2q-1} \left(x_i - \partial_i \psi\right)\left(x_j - \partial_j \psi\right) - p(\boldsymbol{x}, \boldsymbol{\theta})^q \partial_i \partial_j \psi & (24) \end{aligned}$$

The following identities hold:

$$\begin{aligned} \int \partial_i p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} &= \partial_i \int p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = 0 & (25) \\ \int \partial_i \partial_j p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} &= \partial_i \partial_j \int p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = 0 & (26) \end{aligned}$$

Here, we define an important functional

$$h_q(\boldsymbol{\theta}) = h_q[p(\boldsymbol{x}, \boldsymbol{\theta})] = \int p(\boldsymbol{x}, \boldsymbol{\theta})^q d\boldsymbol{x} \tag{27}$$

in particular for discrete S_n,

$$h_q(\boldsymbol{p}) = \sum_{i=0}^{n} p_i^q \tag{28}$$

for $0 < q < 1$. This function plays a key role in the following. From (25) and (26), by using (23) and (24), we have

$$\partial_i \psi(\boldsymbol{\theta}) = \frac{1}{h_q(\boldsymbol{\theta})} \int x_i p(\boldsymbol{x}, \boldsymbol{\theta})^q d\boldsymbol{x} \tag{29}$$

$$\partial_i \partial_j \psi(\boldsymbol{\theta}) = \frac{q}{h_q(\boldsymbol{\theta})} \int (x_i - \partial_i \psi)(x_j - \partial_j \psi) p(\boldsymbol{x}, \boldsymbol{\theta})^{2q-1} d\boldsymbol{x} \tag{30}$$

The latter shows that $\partial_i \partial_j \psi(\boldsymbol{\theta})$ is positive-definite, and hence ψ is convex. □

2.4. *q-Divergence*

A convex function $\psi(\boldsymbol{\theta})$ makes it possible to define a divergence of the Bregman-type between two probability distributions $p(\boldsymbol{x}, \boldsymbol{\theta}_1)$ and $p(\boldsymbol{x}, \boldsymbol{\theta}_2)$ [8,26,27]. It is given by using the gradient $\nabla = \partial/\partial\boldsymbol{\theta}$,

$$D_q[p(\boldsymbol{x}, \boldsymbol{\theta}_1) : p(\boldsymbol{x}, \boldsymbol{\theta}_2)] = \psi(\boldsymbol{\theta}_2) - \psi(\boldsymbol{\theta}_1) - \nabla\psi(\boldsymbol{\theta}_1) \cdot (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \tag{31}$$

satisfying the non-negativity condition

$$D_q[p(\boldsymbol{x}, \boldsymbol{\theta}_1) : p(\boldsymbol{x}, \boldsymbol{\theta}_2)] \geq 0 \tag{32}$$

with equality when and only when $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. This gives a q-divergence in S_n different from the invariant divergence of S_n [28]. The divergence is canonical in the sense that it is uniquely determined in accordance with dually flat structure of q-exponential family in Sections 3 and 4. The canonical divergence is different from the α-divergence or conventional Tsallis relative entropy used in information geometry (See the discussion in the end of this subsection). Note that it is used in [16].

Theorem 3 For two discrete distributions $p(x) = \boldsymbol{p}$ and $r(x) = \boldsymbol{r}$, the q-divergence is given by

$$D_q[\boldsymbol{p} : \boldsymbol{r}] = \frac{1}{(1-q)h_q(\boldsymbol{p})} \left(1 - \sum_{i=0}^{n} p_i^q r_i^{1-q}\right) \tag{33}$$

Proof The potentials are, from (19),

$$\psi(\boldsymbol{p}) = -\log_q p_0, \quad \psi(\boldsymbol{r}) = -\log_q r_0 \tag{34}$$

for $\boldsymbol{p}$ and $\boldsymbol{r}$. We need to calculate $\nabla\psi(\boldsymbol{\theta})$ given in (29). In our case, $x_i = \delta_i(x)$ and hence

$$\partial_i \psi = \frac{p_i^q}{h_q(\boldsymbol{p})} \tag{35}$$

By using this and (17), we obtain (33). □

It is useful to consider a related probability distribution,

$$\hat{p}_q(\boldsymbol{x}) = \frac{1}{h_q[p(\boldsymbol{x})]} p(\boldsymbol{x})^q \tag{36}$$

for defining the q-expectation. This is called the q-escort probability distribution [1,4,29]. Introducing the q-expectation of random variable $f(\boldsymbol{x})$ by

$$E_{\hat{p}}[f(\boldsymbol{x})] = \frac{1}{h_q[p(\boldsymbol{x})]} \int p(\boldsymbol{x})^q f(\boldsymbol{x}) d\boldsymbol{x} \tag{37}$$

we can rewrite the q-divergence (31) for $p(\boldsymbol{x}), r(\boldsymbol{x}) \in S$ as

$$D_q\left[p(\boldsymbol{x}) : r(\boldsymbol{x})\right] = E_{\hat{p}}\left[\log_q p(\boldsymbol{x}) - \log_q r(\boldsymbol{x})\right] \tag{38}$$

because of the relations (20) and (29). The expression (38) is also valid on the exterior of $S \times S$ when it is integrable. This is different from the definition of the Tsallis relative entropy [30,31]

$$\tilde{D}_q[p(\boldsymbol{x}) : r(\boldsymbol{x})] = \frac{1}{1-q}\left(1 - \int p(\boldsymbol{x})^q r(\boldsymbol{x})^{1-q} d\boldsymbol{x}\right) \tag{39}$$

which is equal to the well-known α-divergence up to a constant factor where $\alpha = 1 - 2q$ (see [8,28]), satisfying the invariance criterion. We have

$$D_q[p(\boldsymbol{x}) : r(\boldsymbol{x})] = \frac{1}{h_q[p(\boldsymbol{x})]} \tilde{D}_q[p(\boldsymbol{x}) : r(\boldsymbol{x})] \tag{40}$$

This is a conformal transformation of divergence, as we see in the following. See also the derivation based on affine differential geometry [12].

2.5. *q-Riemannian Metric*

When $\boldsymbol{\theta}_2$ is infinitesimally close to $\boldsymbol{\theta}_1$, by putting $\boldsymbol{\theta}_1 = \boldsymbol{\theta}$, $\boldsymbol{\theta}_2 = \boldsymbol{\theta} + d\boldsymbol{\theta}$ and using the Taylor expansion, we have

$$D_q\left[p(\boldsymbol{x}, \boldsymbol{\theta}) : p(\boldsymbol{x}, \boldsymbol{\theta} + d\boldsymbol{\theta})\right] = \sum g_{ij}^q(\boldsymbol{\theta}) d\theta^i d\theta^j \tag{41}$$

where

$$g_{ij}^{(q)} = \partial_i \partial_j \psi(\boldsymbol{\theta}) \tag{42}$$

is a positive-definite matrix. We call $\left[g_{ij}^{(q)}(\boldsymbol{\theta})\right]$ the q-Fisher information matrix. When $q = 1$, this reduces to the ordinary Fisher information matrix given by

$$g_{ij}^{(1)}(\boldsymbol{\theta}) = g_{ij}^F(\boldsymbol{\theta}) = E\left[\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_j \log p(\boldsymbol{x}, \boldsymbol{\theta})\right] \tag{43}$$

The positive-definite matrix $g_{ij}^{(q)}(\boldsymbol{\theta})$ defines a Riemannian metric on S_n, giving it the q-Riemannian structure.

When a metric tensor $g_{ij}(\boldsymbol{\theta})$ is transformed to

$$\tilde{g}_{ij}(\boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}) g_{ij}(\boldsymbol{\theta}) \tag{44}$$

by a positive function $\sigma(\boldsymbol{\theta})$, we call it a conformal transformation. See, e.g., [13–15,32]. The conformal transformation of divergence induces that of the Riemannian metric.

Theorem 4 The q-Fisher information metric is given by a conformal transformation of the Fisher information metric g^F_{ij} as

$$g^{(q)}_{ij}(\boldsymbol{\theta}) = \frac{q}{h_q(\boldsymbol{\theta})} g^F_{ij}(\boldsymbol{\theta}) \tag{45}$$

Proof The q-metric is derived from the Taylor expansion of $D_q\left[\boldsymbol{p} : \boldsymbol{p} + d\boldsymbol{p}\right]$. We have

$$\begin{aligned} D_q\left[p(\boldsymbol{x},\boldsymbol{\theta}) : p(\boldsymbol{x},\boldsymbol{\theta}+d\boldsymbol{\theta})\right] &= \frac{1}{(1-q)h_q(\boldsymbol{\theta})}\left\{1 - \int p(\boldsymbol{x},\boldsymbol{\theta})^q p(\boldsymbol{x},\boldsymbol{\theta}+d\boldsymbol{\theta})^{1-q} d\boldsymbol{x}\right\} \\ &= \frac{q}{h_q(\boldsymbol{\theta})}\left\{\int \frac{1}{p(\boldsymbol{x},\boldsymbol{\theta})} \partial_i p(\boldsymbol{x},\boldsymbol{\theta}) \partial_j p(\boldsymbol{x},\boldsymbol{\theta}) d\boldsymbol{x}\right\} d\theta^i d\theta^j \end{aligned} \tag{46}$$

using the identities (25) and (26). When $q = 1$, this is the Fisher information given by (43). Hence, the q-Fisher information is given by (45). □

A Riemannian metric defines the length of a tangent vector $\boldsymbol{X} = (X^1, \cdots, X^n)$ at $\boldsymbol{\theta}$ by

$$\|\boldsymbol{X}\|^2 = \sum g_{ij}(\boldsymbol{\theta}) X^i X^j \tag{47}$$

Similarly, for two tangent vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, their inner product is defined by

$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \sum g_{ij} X^i Y^j \tag{48}$$

When $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle$ vanishes, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be orthogonal. The orthogonality, or more generally the angle, of two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ does not change by a conformal transformation, although their magnitudes change.

3. Dually Flat Structure of q-Exponential Family

3.1. Legendre Transformation and q-Entropy

Given a convex function $\psi(\boldsymbol{\theta})$, the Legendre transformation is defined by

$$\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta}) \tag{49}$$

where $\nabla = (\partial/\partial\theta^i)$ is the gradient. Since the correspondence between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is one-to-one, we may consider $\boldsymbol{\eta}$ as another coordinate system of S.

The dual potential function is defined by

$$\varphi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}} \{\boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta})\} \tag{50}$$

which is convex with respect to $\boldsymbol{\eta}$. The original coordinates are recovered from the inverse transformation given by

$$\boldsymbol{\theta} = \nabla \varphi(\boldsymbol{\eta}) \tag{51}$$

where $\nabla = (\partial/\partial\eta_i)$, so that $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are in dual correspondence.

The following theorem gives explicit relations among these quantities.

Theorem 5 The dual coordinates $\boldsymbol{\eta}$ are given by

$$\boldsymbol{\eta} = E_{\hat{p}}[\boldsymbol{x}] \tag{52}$$

and the dual potential is given by

$$\varphi(\boldsymbol{\eta}) = \frac{1}{1-q}\left\{\frac{1}{h_q(\boldsymbol{p})} - 1\right\} \tag{53}$$

Proof The relation (52) is immediate from (29). From the Legendre duality, the dual potential satisfies

$$\varphi(\boldsymbol{\eta}) + \psi(\boldsymbol{\theta}) - \boldsymbol{\theta}\cdot\boldsymbol{\eta} = 0 \tag{54}$$

when $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ correspond to each other by $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$. Therefore,

$$\begin{aligned}
\varphi(\boldsymbol{\eta}) &= \sum_{i=1}^{n}\theta^i\eta_i - \psi(\boldsymbol{\theta}) && (55)\\
&= E_{\hat{p}}\left[\log_q p(\boldsymbol{x},\boldsymbol{\theta})\right] && (56)\\
&= \frac{1}{(1-q)h_q(\boldsymbol{\theta})}\left(1 - \int p^q(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}\right) && (57)\\
&= \frac{1}{1-q}\left(\frac{1}{h_q(\boldsymbol{\theta})} - 1\right) && (58)
\end{aligned}$$

This is a convex function of $\boldsymbol{\eta}$. □

We call the q-dual potential

$$\varphi(\boldsymbol{\eta}) = E\left[\log_q p(\boldsymbol{x},\boldsymbol{\theta})\right] = \frac{1}{1-q}\left\{\frac{1}{h_q} - 1\right\} \tag{59}$$

the negative q-entropy, because it is the Legendre-dual of the q-free energy $\psi(\boldsymbol{\theta})$. There are various definitions of q-entropy. The Tsallis q-entropy [3] is originally defined by

$$H_{\text{Tsallis}} = \frac{1}{1-q}\left(h_q - 1\right) \tag{60}$$

while the Rényi q-entropy [33] is

$$H_{\text{Rényi}} = \frac{1}{1-q}\log h_q \tag{61}$$

They are mutually related by monotone functions. When $q \to 1$, all of them reduce to the Shannon entropy.

Our definition of

$$H_q = \frac{1}{1-q}\left(1 - \frac{1}{h_q}\right) = \frac{H_{\text{Tsallis}}}{h_q} \tag{62}$$

is also monotonically connected with the previous ones, but is more natural from the point of view of q-geometry. The entropy H_q has been known as the normalized q-entropy, which was studied in [16,34–37].

3.2. q-Dually Flat Structure

There are two dually coupled coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ in q-exponential family S with two potential functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ for each q. Two affine structures are introduced by the two convex functions ψ and φ. See information geometry of dually flat space [8]. Although S is a Riemannian manifold given by the q-Fisher information matrix (45), we may nevertheless regard S as an affine manifold where $\boldsymbol{\theta}$ is an affine coordinate system. They represent intensive quantities of a physical system. Dually, we introduce a dual affine structure to S, where $\boldsymbol{\eta}$ is another affine coordinate system. They represent extensive quantities. We can define two types of straight lines or geodesics in S due to the q-affine structures.

For two distributions $p(\boldsymbol{x}, \boldsymbol{\theta}_1)$ and $p(\boldsymbol{x}, \boldsymbol{\theta}_2)$ in S, a curve $p(\boldsymbol{x}, \boldsymbol{\theta}(t))$ is said to be a q-geodesic connecting them, when

$$\boldsymbol{\theta}(t) = t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2 \tag{63}$$

where t is the parameter of the curve. Dually, in terms of dual coordinates $\boldsymbol{\eta}$, when

$$\boldsymbol{\eta}(t) = t\boldsymbol{\eta}_1 + (1-t)\boldsymbol{\eta}_2 \tag{64}$$

holds, the curve is said to be a dual q-geodesic.

More generally, the q-geodesic connecting two distribution $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ is given by

$$\log_q p(\boldsymbol{x}, t) = t \log_q p_1(\boldsymbol{x}) + (1-t) \log_q p_2(\boldsymbol{x}) - c(t) \tag{65}$$

where $c(t)$ is a normalizing term. This is rewritten as

$$p(\boldsymbol{x}, t)^{1-q} = tp_1(\boldsymbol{x})^{1-q} + (1-t)p_2(\boldsymbol{x})^{1-q} - c(t) \tag{66}$$

Dually, the dual q-geodesic connecting $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ is given by using the escort distributions as

$$\hat{p}(\boldsymbol{x}, t) = t\hat{p}_1(\boldsymbol{x}) + (1-t)\hat{p}_2(\boldsymbol{x}) \tag{67}$$

Since the manifold S has a q-Riemannian structure, the orthogonality of two tangent vectors is defined by the Riemannian metric. We rewrite the orthogonality of two geodesics in terms of the affine coordinates. Let us consider two small deviations $d_1p(\boldsymbol{x})$ and $d_2p(\boldsymbol{x})$ of $p(\boldsymbol{x})$, that is, from $p(\boldsymbol{x})$ to $p(\boldsymbol{x}) + d_1p(\boldsymbol{x})$ and $p(\boldsymbol{x}) + d_2p(\boldsymbol{x})$, which are regarded as two (infinitesimal) tangent vectors of S at $p(\boldsymbol{x})$.

Lemma 1 The inner product of two deviations d_1p and d_2p is given by

$$\langle d_1p(\boldsymbol{x}), d_2p(\boldsymbol{x}) \rangle = \int d_1\hat{p}(\boldsymbol{x}) d_2 \log_q p(\boldsymbol{x}) d\boldsymbol{x} \tag{68}$$

Proof By simple calculations, we have

$$\int d_1\hat{p}(\boldsymbol{x}) d_2 \log_q p(\boldsymbol{x}) d\boldsymbol{x} = \frac{q}{h_q} \int \frac{d_1p(\boldsymbol{x}) d_2p(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x} \tag{69}$$

of which the right-hand side is the Riemannian inner product in the form of (46). □

Corollary. Two curves $\boldsymbol{\theta}_1(t)$ and $\boldsymbol{\eta}_2(t)$, intersecting at $t = 0$, are orthogonal when $\langle \dot{\boldsymbol{\theta}}_1(0), \dot{\boldsymbol{\eta}}_2(0) \rangle = 0$. Here, $\dot{\boldsymbol{\theta}}_1(t)$ and $\dot{\boldsymbol{\eta}}_2(t)$ denote derivatives of $\boldsymbol{\theta}_1(t)$ and $\boldsymbol{\eta}_2(t)$ by t, respectively.

The two geodesics and the orthogonality play a fundamental role in S as will be seen in the following.

4. q-Pythagorean and q-Max-Ent Theorems

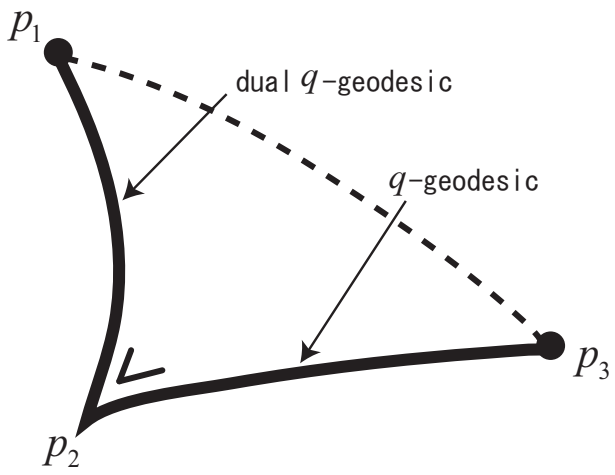
A dually flat Riemannian manifold admits the generalized Pythagorean theorem and the related projection theorem [8]. We state them in our case.

q-Pythagorean Theorem. For three distributions $p_1(\boldsymbol{x}), p_2(\boldsymbol{x})$ and $p_3(\boldsymbol{x})$ in S, it holds that

$$D_q\left[p_1 : p_2\right] + D_q\left[p_2 : p_3\right] = D_q\left[p_1 : p_3\right] \tag{70}$$

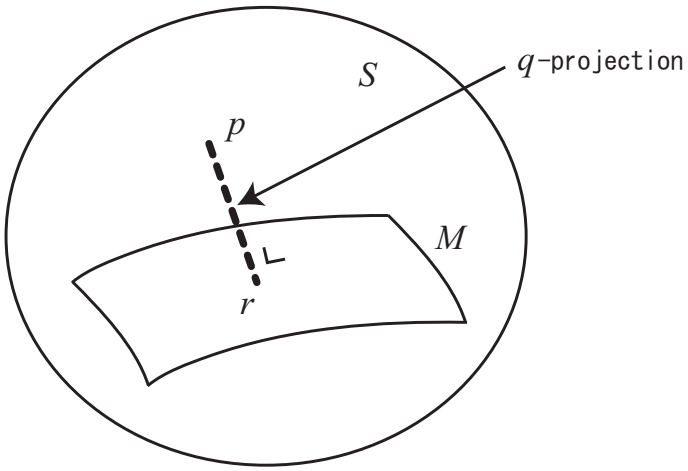
when the dual geodesic connecting $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ is orthogonal at $p_2(\boldsymbol{x})$ to the geodesic connecting $p_2(\boldsymbol{x})$ and $p_3(\boldsymbol{x})$ (see Figure 1).

Figure 1. q-Pythagorean theorem.



Given a distribution $p(\boldsymbol{x}) \in S$ and a submanifold $M \subset S$, a distribution $r(\boldsymbol{x}) \in M$ is said to be the q-projection (dual q-projection) of $p(\boldsymbol{x})$ to M, when the q-geodesic (dual q-geodesic) connecting $p(\boldsymbol{x})$ and $r(\boldsymbol{x})$ is orthogonal to M at $r(\boldsymbol{x})$ (Figure 2).

Figure 2. q-projection of p to M.



q-Projection Theorem. Let M be a submanifold of S. Given $p(\boldsymbol{x}) \in S$, the point $r(\boldsymbol{x}) \in M$ that minimizes $D_q[p(\boldsymbol{x}) : r(\boldsymbol{x})]$ is given by the dual q-projection of $p(\boldsymbol{x})$ to M. The point $r(\boldsymbol{x}) \in M$ that minimizes $D_q[r(\boldsymbol{x}) : p(\boldsymbol{x})]$ is given by the q-projection of $p(\boldsymbol{x})$ to M.

We show that the well-known q-max-ent theorem in the case of Tsallis q-entropy [1,4,9,11] is a direct consequence of the above q-Pythagorean and q-projection theorems.

q-Max-Ent Theorem. Probability distributions maximizing the q-entropies $H_{\text{Tsallis}}, H_{\text{Rényi}}$ and H_q under q-linear constraints for m random variables $c_k(\boldsymbol{x})$ and various values of a_k

$$E_{\hat{p}}\left[c_k(\boldsymbol{x})\right] = a_k, \quad k = 1, \cdots, m \tag{71}$$

form a q-exponential family

$$\log_q p(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \theta^i c_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) \tag{72}$$

The proof is easily obtained by the standard analytical method. Here, we give a geometrical proof. Let us consider the subspace $M^* \subset S$ whose member $p(\boldsymbol{x})$ satisfies the m constraints

$$E_{\hat{p}}\left[c_k(\boldsymbol{x})\right] = \int \hat{p}(\boldsymbol{x}) c_k(\boldsymbol{x}) d\boldsymbol{x} = a_k, \quad k = 1, \cdots, m. \tag{73}$$

Since the constraints are linear in the dual affine coordinates $\boldsymbol{\eta}$ or $\hat{p}(\boldsymbol{x})$, M^* is a linear subspace of S with respect to the dual affine connection. Let $p_0(\boldsymbol{x}, \boldsymbol{\theta}_0)$ be the uniform distribution defined by $\boldsymbol{\theta}_0 = 0$, which implies $p_0(\boldsymbol{x}, \boldsymbol{\theta}_0) = \text{const}$ from (6). Let $\bar{p}(\boldsymbol{x}) \in M^*$ be the q-projection of $p_0(\boldsymbol{x})$ to M^* (Figure 3). Then, the divergence $D_q\left[p : p_0\right]$ from $p(\boldsymbol{x}) \in M^*$ to $p_0(\boldsymbol{x})$ is decomposed as

$$D_q\left[p : p_0\right] = D_q\left[p : \bar{p}\right] + D_q\left[\bar{p} : p_0\right] \tag{74}$$

Let $\boldsymbol{\eta}_p$ be the dual coordinates of $p(\boldsymbol{x})$. Since the divergence is written as

$$D_q\left[p : p_0\right] = \psi\left(\boldsymbol{\theta}_0\right) + \varphi\left(\boldsymbol{\eta}_p\right) - \boldsymbol{\theta}_0 \cdot \boldsymbol{\eta}_p \tag{75}$$

the minimizer of $D_q\left[p : p_0\right]$ among $p(\boldsymbol{x}) \in M^*$ is just $\bar{p}(\boldsymbol{x})$, which is also the maximizer of the entropy $-\varphi\left(\boldsymbol{\eta}_p\right)$.

The trajectories of $\bar{p}(\boldsymbol{x})$ for various values of a_k form a flat subspace orthogonal to M^*, implying that they form a q-exponential family of the form (6) (see Figure 3). The tangent directions $d\hat{p}(\boldsymbol{x})$ of M^* satisfies

$$\int d\hat{p}(\boldsymbol{x}) c_k(\boldsymbol{x}) d\boldsymbol{x} = 0, \quad k = 1, \cdots, m. \tag{76}$$

Hence, a q-exponential family of the form

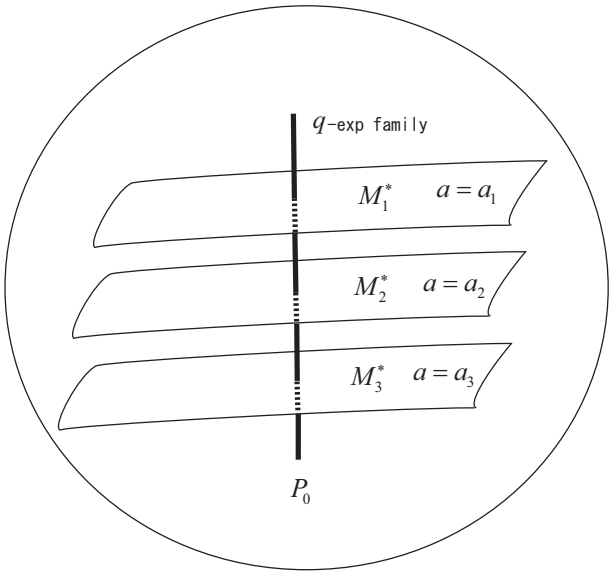
$$\log_q p(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{i=1}^{m} \xi_i d_i(\boldsymbol{x}) - \psi(\boldsymbol{\xi}) \tag{77}$$

is orthogonal to M^*, when

$$\int d\hat{p}(\boldsymbol{x}) d \log_q p(\boldsymbol{x}, \boldsymbol{\xi}) d\boldsymbol{x} = 0 \tag{78}$$

This implies that $d_i(\boldsymbol{x}) = c_i(\boldsymbol{x})$. Hence, we have the q-exponential family (72) that maximizes the q-entropies.

Figure 3. q-Max-Ent theorem.



5. q-Bayesian MAP Estimator

Given N iid observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ from a statistical model $M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\}$, we have

$$p\left(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\xi}\right) = \prod_{i=1}^{N} p\left(\boldsymbol{x}_i, \boldsymbol{\xi}\right) \tag{79}$$

Since $\log_q u$ is a monotonically increasing function, the maximizer of the q-likelihood

$$l_q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\xi}) = \log_q p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\xi}) \tag{80}$$

is the same as the ordinary maximum likelihood estimator (mle). However, the maximizer of the q-escort distribution that maximizes the q-escort log-likelihood,

$$\frac{1}{q}\hat{l}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\xi}) = \log p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\xi}) - \frac{1}{q}\log h_q(\boldsymbol{\xi}) \tag{81}$$

is different from this. We show that the q-mle is a Bayesian MAP (maximum a posteriori probability) estimator. This clarifies the meaning of the q-escort mle.

The q-escort mle is the maximizer of the q-escort distribution,

$$\hat{\boldsymbol{\xi}}_q = \arg\max \hat{p}\left(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\xi}\right) \tag{82}$$

Theorem 6 The q-escort mle $\hat{\boldsymbol{\xi}}_q$ is the Bayesian MAP estimator with the prior distribution

$$\pi(\boldsymbol{\xi}) = h_q(\boldsymbol{\xi})^{-N/q} \tag{83}$$

Proof The Bayesian MAP is the maximizer of the posterior distribution with prior $\pi(\boldsymbol{\xi})$

$$p\left(\boldsymbol{\xi}|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\right) = \frac{\pi(\boldsymbol{\xi})p\left(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\xi}\right)}{p\left(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\right)} \tag{84}$$

which also maximizes

$$(\pi(\boldsymbol{\xi})p\left(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N,\boldsymbol{\xi}\right))^q\,,\quad \text{for } q>0 \tag{85}$$

On the other hand, the q-escort mle is the maximizer of

$$\hat{p}\left(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N,\boldsymbol{\xi}\right)=\prod_{i=1}^{N}\hat{p}(\boldsymbol{x}_i,\boldsymbol{\xi})=\prod_{i=1}^{N}\frac{p\left(\boldsymbol{x}_i,\boldsymbol{\xi}\right)^q}{h_q(\boldsymbol{\xi})} \tag{86}$$

Hence, when

$$\pi(\boldsymbol{\xi})=h_q(\boldsymbol{\xi})^{-N/q} \tag{87}$$

the two estimators are identical. □

The theorem shows that the Bayesian prior has a peak at the maximizer of our q-entropy H_q.

6. Conclusions

Much attention has been recently paid to the probability distributions subject to the power law, instead of the exponential law, since Tsallis proposed the q-entropy and related theories. The power law is also found in various communication networks. It is now a hot topic of research.

However, we do not have a geometrical foundation while that for the ordinary family of probability distributions is given by information geometry [8]. The present paper tried to give a geometrical foundation to the q-family of probability distributions. We introduced a new notion of the q-geometry. The q-structure is ubiquitous in the sense that the family of all the discrete probability distributions (and the family of all the continuous probability distributions, if we neglect delicate problems involved in the infinite dimensionality) belongs to the q-exponential family of distributions for any q. That is, we can introduce the q-geometrical structure to an arbitrary family of probability distributions, because any parametrized family of probability distributions forms a submanifold embedded in the entire manifold.

The q-structure consists of a Riemannian metric together with a pair of dually coupled affine connections, which sits in the framework of the standard information geometry. However, the q-structure is essentially different from the standard one derived by the invariance criterion of the manifold of probability distributions. We have a novel look on the theory related to the q-entropy from a viewpoint of conformal transformation. This leads us to unified definitions of various quantities such as the q-entropy, q-divergence, q-potential function and their duals, as well as new interpretations of known quantities.

This is a geometrical foundation and we expect that the paper contributes to provide further developments in this field.

References

1. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics*; Springer: New York, NY, USA, 2009.
2. Naudts, J. *Generalised Thermostatistics*; Springer: London, UK, 2011.
3. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
4. Naudts, J. The q-exponential family in statistical Physics. *Cent. Eur. J. Phys.* **2009**, *7*, 405–413.

5. Suyari, H. Mathematical structures derived from the q-multinomial coefficient in Tsallis statistics. *Physica A* **2006**, *368*, 63–82.
6. Suyari, H.; Wada, T. Multiplicative duality, q-triplet and μ, ν, q-relation derived from the one-to-one correspondence between the (μ, ν)-multinomial coefficient and Tsallis entropy S_q. *Physica A* **2008**, *387*, 71–83.
7. Barndorff-Nielsen, O.E. *Information and Exponential Families in Statistical Theory*. Wiley: New York, NY, USA, 1978.
8. Amari, S.; Nagaoka, H. *Methods of Information Geometry (Translations of Mathematical Monographs)*; Oxford University Press: Oxford, UK, 2000.
9. Ohara, A. Geometry of distributions associated with Tsallis statistics and properties of relative entropy minimization. *Phys. Lett. A* **2007**, *370*, 184–193.
10. Furuichi, S. On the maximum entropy principle and the minimization of the Fisher information in Tsallis statistics. *J. Math. Phys.* **2009** *50*, 013303.
11. Ohara, A. Geometric study for the Legendre duality of generalized entropies and its application to the porous medium equation. *Eur. Phys. J. B* **2009**, *70*, 15–28.
12. Ohara, A.; Matsuzoe, H.; Amari, S. A dually flat structure with escort probability and its application to alpha-Voronoi diagrams. *arXiv* **2010**, arXiv:cond-mat/1010.4965.
13. Kurose, T. On the Divergence of 1-conformally Flat Statistical Manifolds. *Tôhoku Math. J.* **1994**, *46*, 427–433.
14. Matsuzoe, H. Geometry of contrast functions and conformal geometry. *Hiroshima Math. J.* **1999**, *29*, 175–191.
15. Kurose, T. Conformal-projective geometry of statistical manifolds. *Interdisciplinary Information Sciences* **2002**, *8*, 89–100.
16. Yamano, T. Information theory based on non-additive information content. *Phys. Rev. E* **2001**, *63*, 046105.
17. Naudts, J. Estimators, escort probabilities, and phi-exponential families in statistical physics. *J. Ineq. Pure Appl. Math.* **2004**, *5*, 102.
18. Pistone, G. kappa-exponential models from the geometrical viewpoint. *Eur. Phys. J. B* **2009**, *70*, 29–37.
19. Naudts, J. Generalized exponential families and associated entropy functions. *Entropy* **2008**, *10*, 131–149.
20. Kaniadakis, G.; Lissia, M.; Scarfone, A.M. Deformed logarithms and entropies. *Physica A* **2004**, *340*, 41–49.
21. Yamano, T. Some properties of q-logarithmic and q-exponential functions in Tsallis statistics. *Physica A* **2002**, *305*, 486–496.
22. Tsallis, C.; Levy, S.V.F.; Souza, A.M.C.; Maynard, R. Statistical-mechanical foundation of the ubiquity of Levy distributions in nature. *Phys. Rev. Lett.* **1995**, *75*, 3589–3593, Erratum *Phys. Rev. Lett.* **1996**, *77*, 5442.
23. Tanaka, M. A consideration on the family of q-Gaussian distributions. *IEICE (Japan)* **2002**, *J85-D2*, 161–173 (in Japanese).

24. Zhang, Z.; Zhong, F.; Sun, H. Information geometry of the power inverse Gaussian distribution. *Appl. Sci.* **2007**, *9*, 194–203.
25. Ohara, A.; Wada, T. Information geometry of q-Gaussian densities and behaviors of solutions to related diffusion equations. *J. Phys. A: Math. Theor.* **2010**, *43*, 035002.
26. Wada, T. Generalized log-likelihood functions and Bregman divergences. *J. Math. Phys.* **2009**, *50*, 113301.
27. Cichocki, A.; Cruces, S.; Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
28. Amari, S. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Trans. Inform. Theor.* **2009**, *55*, 4925–4931.
29. Beck, C.; Schlögl, F. *Thermodynamics of Chaotic Systems*; Cambridge University Press: Cambridge, UK, 1993.
30. Borland, L.; Plastino, A.R.; Tsallis C. Information gain within nonextensive thermostatistics. *J. Math. Phys.* **1998**, *39*, 6490–6501.
31. Furuichi, S. Fundamental properties of Tsallis relative entropy. *J. Math. Phys.* **2004**, *45*, 4868–4877.
32. Okamoto, I.; Amari, S.; Takeuchi, K. Asymptotic theory of sequential estimation procedures for curved exponential families. *Ann. Stat.* **1991**, *19*, 961–981.
33. Rényi, A. On measures of entropy and information. In Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1960; pp. 547–561.
34. Landsberg, P.T.; Vedral, V. Distributions and channel capacities in generalized statistical mechanics. *Phys. Lett. A* **1998**, *247*, 211–217.
35. Rajagopal, A.K.; Abe, S. Implications of form invariance to the structure of nonextensive entropies. *Phys. Rev. Lett.* **1999**, *83*, 1711–1714.
36. Yamano, T. Source coding theorem based on a nonadditive information content. *Physica A* **2002**, *305*, 190–195.
37. Wada, T.; Scarfone, A.M. Connections between Tsallis' formalisms employing the standard linear average energy and ones employing the normalized q-average energy. *Phys. Lett. A* **2005**, *335*, 351–362.