Tsallis Kernels on Measures

André F. T. Martins*†
*Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Pedro M. Q. Aguiar‡
‡Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisboa, Portugal

Mário A. T. Figueiredo†
†Instituto de Telecomunicações
Instituto Superior Técnico
Lisboa, Portugal

***Abstract*—Recent approaches to classification of text, images, and other types of structured data, launched the quest for positive definite (p.d.) kernels on probability measures. In particular, kernels based on the Jensen-Shannon (JS) divergence and other information-theoretic quantities have been proposed. We introduce new JS-type divergences, by extending its two building blocks: convexity and Shannon's entropy. These divergences are then used to define new information-theoretic kernels on measures. In particular, we introduce a new concept of q-convexity, for which a Jensen q-inequality is proved. Based on this inequality, we introduce the Jensen-Tsallis q-difference, a nonextensive generalization of the Jensen-Shannon divergence. Furthermore, we provide denormalization formulae for entropies and divergences, which we use to define a family of nonextensive information-theoretic kernels on measures. This family, grounded in nonextensive entropies, extends Jensen-Shannon divergence kernels, and allows assigning weights to its arguments.**

***Index Terms*—Positive definite kernels, nonextensive entropy, Tsallis entropy, Jensen-Shannon divergence, convexity.**

I. Introduction

In the field of kernel-based machine learning [1], there has been recent interest in defining positive definite kernels on probability measures, with applications in classification of text, images, and other types of structured data [2], [3], [4]. In particular, kernels based on the Jensen-Shannon divergence (JSD [5]) and other (Shannon) information-theoretic quantities have been considered by several authors [2], [3].

Over the years, several generalizations of the Shannon entropy have been proposed [6], [7], [8]. Rényi entropies are arguably the best known of these, with several applications (e.g., [9], [10]). The Rényi and Shannon entropies are both *additive*: the joint entropy of independent variables is the sum of the individual entropies. In the so-called *nonextensive* (namely Tsallis) entropies [7], [8], [11], the additivity property is abandoned. Tsallis entropies have been used to formulate nonextensive statistical mechanics [12], [13] and, recently, in signal/image processing [14], [15], [16].

Convexity is a key concept in information theory, namely via the ubiquitous *Jensen's inequality* (JI) [17], [18]. The JI underlies the concept of JSD, which has been used in statistics, machine learning, image and signal processing, and physics.

In this paper, we introduce new JSD-type divergences, by extending its two building blocks: convexity and Shannon's entropy. These divergences are then used to define new (and recover previous) information-theoretic kernels on measures. More specifically, our main contributions are:

- A new concept of q-convexity, for which a *Jensen q-inequality* (JqI) is proved. Based on the JqI, we introduce a the *Jensen-Tsallis q-difference*, (JTqD) a nonextensive generalization of the JSD.
- Characterization of the JTqD, with respect to convexity and extrema, extending the work in [19], [5] for the JSD.
- Denormalization formulae for entropies and divergences, which we use to define a family of nonextensive information-theoretic kernels on measures. This family (which contains JSD kernels [2] as particulars cases) is novel in two ways: it is grounded in nonextensive entropies; it allows assigning weights to its arguments.

All the proofs omitted in this paper can be found in [20].

II. Tsallis Entropies

Let Δ^{n-1} be the simplex in $\mathbb{R}^n$. The Tsallis entropy $S_q : \Delta^{n-1} \to \mathbb{R}$, defined as

$$S_q(p_1, \ldots, p_n) = \frac{(1 - \sum_{i=1}^n p_i^q)}{q-1} = -\sum_{x \in X} p(x)^q \, \ln_q p(x) \quad (1)$$

where $\ln_q(x) := (x^{1-q} - 1)/(1-q)$ is the q-logarithm, satisfies the axioms for nonextensive entropies introduced in [21].

Tsallis joint and conditional entropies are defined as

$$\begin{aligned} S_q(X,Y) &:= -\sum_{x,y} p(x,y)^q \ln_q p(x,y) \\ S_q(X|Y) &:= -\sum_{x,y} p(x,y)^q \ln_q p(x|y) \end{aligned}$$

and the chain rule $S_q(X,Y) = S_q(X) + S_q(Y|X)$ holds [22].

For two pmfs p_X, $p_Y \in \Delta^n$, the *Tsallis relative entropy*, generalizing the KLD, is defined as

$$D_q(p_X \| p_Y) := -\sum_x p_X(x) \ln_q(p_Y(x)/p_X(x)). \quad (2)$$

III. Entropies of unnormalized measures

We consider functionals that extend the domain of the Shannon and Tsallis entropies to unnormalized measures. Although they are completely characterized by their restriction to the normalized counterparts, these denormalizations will be used in Section VI to derive novel positive definite kernels.

Partially supported by the Portuguese *Fundação para a Ciência e Tecnologia* (grant PTDC/EEA-TEL/72572/2006 and the CMU-Portugal program) and by the EU FP7 project SIMBAD.

Let $(\mathcal{X}, \mathscr{M}, \nu)$ be a measured space, where $\mathcal{X}$ is Hausdorff and ν a σ-finite Radon measure (usually the Lebesgue-Borel measure, if $\mathcal{X} \subseteq \mathbb{R}^n$ and $\text{int}\mathcal{X} \neq \varnothing$, or the counting measure, if $\mathcal{X}$ is countable). We denote by $M_+(\mathcal{X})$ the set of *finite* Radon ν-absolutely continuous measures on $\mathcal{X}$, and by $M_+^1(\mathcal{X})$ the subset of those which are probability measures. For simplicity, we often identify each measure in $M_+(\mathcal{X})$ or $M_+^1(\mathcal{X})$ with the corresponding nonnegative density. In the sequel, Lebesgue-Stieltjes integrals of the form $\int_{\mathcal{A}} f(x) d\nu(x)$ are often written as $\int_{\mathcal{A}} f$, or simply $\int f$, if $\mathcal{A} = \mathcal{X}$.

For some functional $G : M_+(\mathcal{X}) \to \overline{\mathbb{R}}$, let $M_+^G(\mathcal{X}) := \{f \in M_+(\mathcal{X}) : |G(f)| < \infty\}$ and $M_+^{1,G}(\mathcal{X}) := M_+^G(\mathcal{X}) \cap M_+^1(\mathcal{X})$. The following functional [23], extends the SBG entropy from $M_+^{1,H}$ to unnormalized measures in M_+^H (with $0 \log 0 := 0$)

$$H(f) = -k \int f \log f = \int \varphi_H \circ f, \tag{3}$$

where $k \in \mathbb{R}_{++}$, and $\varphi_H : \mathbb{R}_{++} \to \mathbb{R}$ is defined as

$$\varphi_H(y) = -k\, y \log y. \tag{4}$$

The generalized KLD is a directed divergence between two measures $\mu_f, \mu_g \in M_+^H(\mathcal{X})$, such that μ_f is μ_g-absolutely continuous ($\mu_f \ll \mu_g$). In terms of densities,

$$D(f, g) = k \int \left(g - f + f \log \frac{f}{g} \right). \tag{5}$$

Both H and D are completely determined by their restriction to the normalized measures, as the next proposition shows.

Proposition 1: The following equalities hold for any $c \in \mathbb{R}_{++}$ and $f, g \in M_+^H(\mathcal{X})$, with $\mu_f \ll \mu_g$:

$$\begin{aligned} H(cf) &= c\, H(f) + |f|\, \varphi_H(c), \\ D(cf, cg) &= c\, D(f, g), \\ D(cf, g) &= c\, D(f, g) - |f|\, \varphi_H(c) + k\,(1 - c)\,|g|, \end{aligned}$$

where $|f| := \int f = \mu_f(\mathcal{X})$.

Proof: Straightforward from (3) and (5). ∎

For $q \geq 0$, let $M_+^{S_q}(\mathcal{X}) := \{f \in M_+(\mathcal{X}) : f^q \in M_+(\mathcal{X})\}$. The Tsallis counterpart of (3), defined on $M_+^{S_q}(\mathcal{X})$, is

$$S_q(f) = \int \varphi_q \circ f, \tag{6}$$

where $\varphi_q : \mathbb{R}_{++} \to \mathbb{R}$ is given by

$$\varphi_q(y) = \begin{cases} \varphi_H(y) & \text{if } q = 1, \\ \frac{k}{q-1}\,(y - y^q) & \text{if } q \neq 1. \end{cases} \tag{7}$$

Similarly, a nonextensive version of (5) is

$$D_q(f, g) = -\frac{k}{q-1} \int \left(qf + (1-q)g - f^q g^{1-q} \right), \tag{8}$$

for $q \neq 1$, and $D_1(f, g) := \lim_{q \to 1} D_q(f, g) = D(f, g)$.

Proposition 2: The following equalities hold for any $c \in \mathbb{R}_{++}$ and $f, g \in M_+^{S_q}(\mathcal{X})$, with $\mu_f \ll \mu_g$:

$$S_q(cf) = c^q S_q(f) + |f| \varphi_q(c), \tag{9}$$

$$D_q(cf, cg) = c D_q(f, g), \tag{10}$$

$$D_q(cf, g) = c^q D_q(f, g) - q \varphi_q(c) |f| + k(1 - c^q)|g|. \tag{11}$$

Proof: Straightforward from (6) and (8). ∎

Naturally, all the equalities in Prop. 1 are obtained by taking the limit $q \to 1$ in those of Prop. 2.

IV. Jensen Differences and Divergences

Definition 3 (Jensen difference (JD)): Consider two measured sets $(\mathcal{X}, \mathscr{M}, \nu)$ and $(\mathcal{T}, \mathscr{T}, \tau)$. Let $\mu := \{\mu_t\}_{t \in \mathcal{T}} \in [M_+(\mathcal{X})]^{\mathcal{T}}$ be a set of measures in $M_+(\mathcal{X})$ indexed by $\mathcal{T}$, and let $\omega \in M_+(\mathcal{T})$ be a measure in $\mathcal{T}$. The JD is defined as

$$J_\Psi^\omega(\mu) := \Psi \left(\int_{\mathcal{T}} \omega(t)\, \mu_t \, d\tau(t) \right) - \int_{\mathcal{T}} \omega(t) \Psi(\mu_t)\, d\tau(t) \tag{12}$$

where: *(i)* Ψ is a concave functional such that $\text{dom}\, \Psi \subseteq M_+(\mathcal{X})$; *(ii)* $\omega(t)\mu_t(x)$ is τ-integrable, for all $x \in \mathcal{X}$; *(iii)* $\int_{\mathcal{T}} \omega(t)\mu_t d\tau(t) \in \text{dom}\, \Psi$; *(iv)* $\mu_t \in \text{dom}\, \Psi$, for all $t \in \mathcal{T}$; *(v)* $\omega(t)\Psi(\mu_t)$ is τ-integrable.

In the following subsections, we consider several instances of Definition 3, leading to several Jensen-type divergences.

A. The Jensen-Shannon Divergence

Let P be a random probability distribution with values in $\{p_t\}_{t \in \mathcal{T}}$ following a distribution $\pi \in M_+^1(\mathcal{T})$. Then,

$$J_\Psi^\pi(\{p_t\}_{t \in \mathcal{T}}) = \Psi\left(E[P]\right) - E[\Psi(P)], \tag{13}$$

where the expectations are with respect to π. Letting $\Psi = H$, the Shannon entropy, we have $J^\pi := J_H^\pi$.

If $\mathcal{X}$ and $\mathcal{T}$ are finite with $|\mathcal{T}| = m$, $J_H^\pi(p_1, \ldots, p_m)$ is the JSD of $p_1, \ldots, p_m$, with weights $\pi_1, \ldots, \pi_m$ [19], [5]. For $|\mathcal{T}| = 2$ and $\pi = (\frac{1}{2}, \frac{1}{2})$, we have $J^{(\frac{1}{2}, \frac{1}{2})}(P) = JS(p_1, p_2)$,

$$JS(p_1, p_2) = H((p_1 + p_2)/2) - (H(p_1) + H(p_2))/2 \tag{14}$$

as introduced in [5]. It has been shown that $\sqrt{JS}$ satisfies the triangle inequality and that it is an Hilbertian metric [24], [25].

B. The Jensen-Tsallis Divergence

Divergences of the form (13), based on the Tsallis entropy have been studied in [19]. Letting $\Psi = S_q$, (13) becomes

$$J_{S_q}^\pi(\{p_t\}_{t \in \mathcal{T}}) = S_q\left(E[P]\right) - E[S_q(P)]. \tag{15}$$

For finite $\mathcal{X}$ and $\mathcal{T}$, $J_{S_q}^\pi$ is called the *Jensen-Tsallis divergence* (JTD) and it has been applied in image processing [26].

V. q-Convexity and Jensen q-Differences

A. Introduction and Definitions

Definition 4: The unnormalized *q-expectation* of a random variable X, with probability density p, is

$$E_q[X] := \int x\, p^q(x)\, dx. \tag{16}$$

For $q \neq 1$, the q-expectation does not correspond to the intuitive meaning of expectation. Nonetheless, it has been used in the construction of nonextensive information theory; *e.g.*, the Tsallis entropy can be written as $S_q(X) = -E_q[\ln_q p(X)]$.

We now introduce q-convexity and derive several related results, namely the *Jensen q-inequality* (JqI).

Definition 5: Let $q \in \mathbb{R}$ and $\mathcal{X}$ be a convex set. A function $f : \mathcal{X} \to \mathbb{R}$ is q-convex if for any $x, y \in \mathcal{X}$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda^q f(x) + (1-\lambda)^q f(y). \tag{17}$$

Naturally, f is q-concave if $-f$ is q-convex, and 1-convexity is simply standard convexity.

Proposition 6 (The Jensen q-Inequality): If $f : \mathcal{X} \to \mathbb{R}$ is q-convex, then for any $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ and $\pi = (\pi_1, \ldots, \pi_n) \in \Delta^{n-1}$,

$$f\left(\sum \pi_i x_i\right) \leq \sum \pi_i^q f(x_i). \tag{18}$$

Proof: By induction, as in the proof of the JI [18]. ∎

Proposition 7: Let $f \geq 0$ and $q \geq q' \geq 0$; then, q-convexity implies q'-convexity.

Definition 8 (Jensen q-Differences (JqD)): Let $\mu := \{\mu_t\}_{t\in\mathcal{T}} \in [M_+(\mathcal{X})]^{\mathcal{T}}$ be a class of measures in $\mathcal{X}$ indexed by $\mathcal{T}$, and let $\omega \in M_+(\mathcal{T})$ be a measure in $\mathcal{T}$. For $q \geq 0$, define

$$T^{\omega}_{q,\Psi}(\mu) := \Psi\left(\int_{\mathcal{T}} \omega(t)\,\mu_t\,d\tau(t)\right) - \int_{\mathcal{T}} \omega^q(t)\Psi(\mu_t)\,d\tau(t) \tag{19}$$

where: *(i)* Ψ is a concave functional such that $\operatorname{dom}\Psi \subseteq M_+(\mathcal{X})$; *(ii)* $\omega(t)\mu_t(x)$ is τ-integrable for all $x \in \mathcal{X}$; *(iii)* $\int_{\mathcal{T}} \omega(t)\mu_t d\tau(t) \in \operatorname{dom}\Psi$; *(iv)* $\mu_t \in \operatorname{dom}\Psi$, for all $t \in \mathcal{T}$; *(v)* $\omega^q(t)\Psi(\mu_t)$ is τ-integrable.

Conditions for the Jensen difference to be convex were given in [19]. The following proposition generalizes that result, extending it to JqD.

Proposition 9: Let $\mathcal{T}$ and $\mathcal{X}$ be finite sets, with $|\mathcal{T}| = m$ and $|\mathcal{X}| = n$, and let $\pi \in M^1_+(\mathcal{T})$. Let $\varphi : [0, 1] \to \mathbb{R}$ be a function of class C^2 and consider the (φ-entropy [19]) function $\Psi : [0, 1]^n \to \mathbb{R}$ defined by $\Psi(z) := -\sum_{i=1}^n \varphi(z_i)$. Then, the q-difference $T^{\pi}_{q,\Psi} : [0, 1]^{nm} \to \mathbb{R}$ is convex if and only if φ is convex and $-1/\varphi''$ is $(2-q)$-convex.

B. The Jensen-Tsallis q-Difference

Definition 10 (Jensen-Tsallis q-Difference (JTqD)): In the conditions of Definition 8, the JTqD, denoted T^{π}_q, is defined as $T^{\pi}_q := T^{\pi}_{q,S_q}$.

When $|\mathcal{T}| = 2$ and $\pi = (1/2, 1/2)$, define $T_q := T_q^{1/2,1/2}$. Notable cases arise for particular values of q:

- For $q = 0$, $S_0(p) = -1 + \|p\|_0$, where $\|p\|_0$ denotes the so-called *0-norm* (although it's not a norm) of vector p, *i.e.*, its number of nonzero components. The JT0D is thus

$$T_0(p_1, p_2) = 1 - \|p_1 \odot p_2\|_0, \tag{20}$$

  where $\odot$ denotes the Hadamard-Schur (*i.e.*, elementwise) product. We call T_0 the *Boolean difference*.
- For $q = 1$, since $S_1(p) = H(p)$, T_1 is the standard JSD.
- For $q = 2$, $S_2(p) = 1 - \langle p, p\rangle$, where $\langle x, y\rangle = \sum_i x_i\, y_i$ denotes inner product. Consequently, the JT2D is $T_2(p_1, p_2) = (1-\langle p_1, p_2\rangle)/2$, we call *linear difference*.

We now present results regarding convexity and extrema of the JTqD, extending known properties of the JSD ($q = 1$), some of which are lost in the transition to nonextensivity. For example, while the JSD is nonnegative and vanishes iff all the distributions are identical, this is not true in general for the JTqD. Nonnegativity of the JTqD is only guaranteed if $q \geq 1$, explaining why some authors (*e.g.*, [22]) only consider $q \geq 1$, when developing nonextensive information theories.

The following propositions establish convexity properties of the JTqD (complementing the joint convexity of the JTD, for $q \in [1, 2]$, proved in [19]) and provide upper and lower bounds for the JTqD.

Proposition 11: Let $\mathcal{T}$ and $\mathcal{X}$ be finite sets with cardinalities m and n, respectively. For $q \in [0, 1]$, the JTqD is a jointly convex function on $(M_+^{1,S_q}(\mathcal{X}))^{\mathcal{T}}$.

Proposition 12: Let $\mathcal{T}$ and $\mathcal{X}$ be countable sets. The JTqD is convex in each argument, for $q \in [0, 2]$, and concave in each argument, for $q \geq 2$.

Proposition 13: Let $\mathcal{T}$ and $\mathcal{X}$ be countable sets. For $q \geq 0$, $T^{\pi}_q(p_1, \ldots, p_m) \leq S_q(\pi)$, with the bound reached for a set of disjoint degenerate distributions. For $q \geq 1$, $T^{\pi}_q(p_1, \ldots, p_m) \geq 0$, with the minimum attained in the pure deterministic case, *i.e.*, when all distributions are equal to the same degenerate one. For $q \in [0, 1]$ and $\mathcal{X}$ a finite set with $|\mathcal{X}| = n$, $T^{\pi}_q(p_1, \ldots, p_m) \geq S_q(\pi)[1 - n^{1-q}]$. This lower bound (which can be negative) is attained when all distributions are uniform.

VI. Tsallis Kernels

A. Positive and negative definite kernels

We start by recalling basic concepts from kernel theory [1]; in the following, $\mathcal{X}$ denotes a nonempty set.

Definition 14: Let $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function, *i.e.*, satisfying $\varphi(y, x) = \varphi(x, y)$, for all $x, y \in \mathcal{X}$. φ is called a *positive definite* (p.d.) *kernel* if and only if

$$\sum_{i=1}^n \sum_{j=1}^n c_i\, c_j\, \varphi(x_i, x_j) \geq 0 \tag{21}$$

for all $n \in \mathbb{N}$, $x_i, \ldots, x_n \in \mathcal{X}$ and $c_i, \ldots, c_n \in \mathbb{R}$.

Definition 15: Let $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric. ψ is called a *negative definite* (n.d.) *kernel* if and only if

$$\sum_{i=1}^n \sum_{j=1}^n c_i\, c_j\, \psi(x_i, x_j) \leq 0 \tag{22}$$

for all $n \in \mathbb{N}$, $x_i, \ldots, x_n \in \mathcal{X}$ and $c_i, \ldots, c_n \in \mathbb{R}$, such that $c_1 + \ldots + c_n = 0$. In this case, $-\psi$ is called *conditionally* p.d.

Both the sets of p.d. and n.d. kernels are convex cones (closed under pointwise sums and integrations), the former being closed under pointwise products; moreover, both sets are closed under pointwise convergence. Proofs of these facts and of the following propositions can be found in [27].

Proposition 16: Let $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function, and $x_0 \in \mathcal{X}$. Let $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be given by

$\varphi(x,y) = \psi(x,x_0) + \psi(y,x_0) - \psi(x,y) - \psi(x_0,x_0)$. Then, φ is p.d. if and only if ψ is n.d.

Proposition 17: The function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a n.d. kernel if and only if $\exp(-t\psi)$ is p.d. for all $t > 0$.

Proposition 18: The function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a n.d. kernel if and only if $(t+\psi)^{-1}$ is p.d. for all $t > 0$.

Proposition 19: If ψ is n.d. and nonnegative on the diagonal, *i.e.*, $\psi(x,x) \geq 0$ for all $x \in \mathcal{X}$, then so are ψ^α, for $\alpha \in [0,\, 1]$, and $\log(1+\psi)$.

Proposition 20: Let $f : \mathcal{X} \to \mathbb{R}$ with $f \geq 0$; then, for $\alpha \in [1,2]$, $\psi_\alpha(x,y) = -(f(x)+f(y))^\alpha$ is a n.d. kernel.

The following definition has been used in a machine learning context [23], following [27].

Definition 21 (Semigroup Kernels): Let $(\mathcal{X}, +)$ be a semigroup. A function $\varphi : \mathcal{X} \to \mathbb{R}$ is called p.d. (in the semigroup sense) if $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined as $k(x,y) = \varphi(x+y)$, is a p.d. kernel. Likewise, φ is called n.d. if k is a n.d. kernel.

B. Jensen-Shannon and Tsallis kernels

The basic result underlying JSD- and JTqD-based p.d. kernels is the fact, shown in the following proposition, that the denormalized Tsallis q-entropies (6) are n.d. functions on $M_+^{S_q}(\mathcal{X})$, for $q \in [0,2]$. Of course, this includes the denormalized Shannon entropy (3) as a particular case (for $q = 1$). Although, for the Shannon entropy case, part of the proof is in [27], [25], [23], we present a general proof here.

Proposition 22: For $q \in [0,2]$, the denormalized Tsallis q-entropy S_q is a n.d. function on $M_+^{S_q}(\mathcal{X})$.

Proof: Since n.d. kernels are closed under pointwise integration, it suffices to prove that φ_q (see (7)) is n.d. on $\mathbb{R}_+$. For $q \neq 1$, $\varphi_q(y) = (q-1)^{-1}(y - y^q)$, thus $\varphi_q = |q-1|^{-1}(\xi_q + \gamma_q)$, where $\xi_q(y) = y\,\mathrm{sign}(q-1)$ and $\gamma_q(y) = y^q\,\mathrm{sign}(1-q)$, both defined on $\mathbb{R}_+$. Since the set of n.d. functions is closed under sums and multiplications by non-negative scalars, this reduces to showing that both ξ_q and γ_q are n.d. Function ξ_q is both n.d. and p.d. for any q. For $q \in [0,1[$, $\gamma_q = \xi_a^q$, for any $a > 1$; since ξ_a is n.d. and nonnegative, Prop. 19 guarantees that γ_q is also n.d. For $q \in]1,2]$, Prop. 20 guarantees that $k(x,y) = -(x+y)^q$ is n.d., thus so is γ_q.

For $q = 1$, we use the fact that,

$$\varphi_1(x) = \varphi_H(x) = -x\log x = \lim_{q\to 1} \frac{x - x^q}{q-1} = \lim_{q\to 1} \varphi_q(x),$$

where the limit is obtained by L'Hôpital's rule; since the set of n.d. functions is closed under limits, $\varphi_1(x)$ is n.d. ■

We are now in a position to present the main contribution of this section, which is a family of *weighted Jensen-Tsallis kernels*, generalizing the JSD-based (and other) kernels in three ways

- they allow using unnormalized measures;
- they allow using different weights for each of the two arguments;
- they extend the mutual information feature of the JSD kernel to the nonextensive scenario.

Definition 23 (weighted Jensen-Tsallis kernels (WJSK)): The kernel $\widetilde{\varphi}_q : M_+^{S_q}(\mathcal{X}) \times M_+^{S_q}(\mathcal{X}) \to \mathbb{R}$ is defined as

$$\widetilde{\varphi}_q(\mu_1,\mu_2) = \left(S_q(\pi) - T_q^\pi(p_1,p_2)\right)(\omega_1+\omega_2)^q,$$

where $p_1 = \mu_1/\omega_1$ and $p_2 = \mu_2/\omega_2$ are the normalized counterparts of μ_1 and μ_2, with corresponding masses $\omega_1, \omega_2 \in \mathbb{R}_+$, and $\pi = (\omega_1/(\omega_1+\omega_2), \omega_2/(\omega_1+\omega_2))$.

The kernel $\varphi_q : \left(M_+^{S_q}(\mathcal{X}) \setminus \{0\}\right)^2 \to \mathbb{R}$ is defined as

$$\varphi_q(\mu_1,\mu_2) = S_q(\pi) - T_q^\pi(p_1,p_2).$$

Proposition 24: The kernel $\widetilde{\varphi}_q$ is p.d., for $q \in [0,2]$.

Proof: Writing $\mu_1 = \omega_1 p_1$ and $\mu_2 = \omega_2 p_2$ and using the denormalization formulae of Prop. 2, we obtain, after algebra, $\widetilde{\varphi}_q(\mu_1,\mu_2) = -T_{q,S_q}^{(1,1)}(\mu_1,\mu_2)$. Since $-T_{q,S_q}^{(1,1)}(\mu_1,\mu_2) = -S_q(\mu_1+\mu_2) + S_q(\mu_1) + S_q(\mu_2) = -S_q(\mu_1+\mu_2) + S_q(\mu_1+\mu_0) + S_q(\mu_2+\mu_0) - S_q(\mu_0+\mu_0)$, with $\mu_0 = 0$, and S_q is n.d. (Prop. 22), Prop. 16 guarantees that $-T_{q,S_q}^{(1,1)}$ is p.d. ■

Proposition 25: The kernel φ_q is p.d., for $q \in [0,1]$.

Proof: Observe that $\varphi_q(\mu_1,\mu_2) = \widetilde{\varphi}_q(\mu_1,\mu_2)(\omega_1+\omega_2)^{-q}$. The result follows from the fact that the product of two p.d. kernels is a p.d. kernel and $(\omega_1+\omega_2)^{-q}$ is a p.d. kernel, for $q \in [0,1]$ (see [27]). ■

The following are particular cases of WJTK, for $q = 1$.

Definition 26 (weighted JS kernel (WJSK)): The kernel $\widetilde{\varphi} : (M_+^H(\mathcal{X}))^2 \to \mathbb{R}$ is defined as $\widetilde{\varphi} = \widetilde{\varphi}_1$, *i.e.*,

$$\widetilde{\varphi}(\mu_1,\mu_2) = \left(H(\pi) - J^\pi(p_1,p_2)\right)(\omega_1+\omega_2),$$

where $p_1 = \mu_1/\omega_1$ and $p_2 = \mu_2/\omega_2$ are the normalized counterpart of μ_1 and μ_2, and $\pi = (\omega_1/(\omega_1+\omega_2), \omega_2/(\omega_1+\omega_2))$.

Analogously, the kernel $\varphi : \left(M_+^H(\mathcal{X}) \setminus \{0\}\right)^2 \to \mathbb{R}$ is simply $\varphi = \varphi_1$, *i.e.*,

$$\varphi(\mu_1,\mu_2) = H(\pi) - J^\pi(p_1,p_2).$$

Corollary 27: The WJSK $\widetilde{\varphi}$ and φ are p.d.

Proof: Invoke Props. 24 and 25 with $q = 1$. ■

The JS kernel (JSK), introduced and shown to be p.d. in [2], is now simply a particular case of the WJSK in Def. 26.

Definition 28 (JSK): The kernel $k_{\mathrm{JS}} : (M_+^1(\mathcal{X}))^2 \to \mathbb{R}$ is defined as $k_{\mathrm{JS}}(p_1,p_2) = 1 - JS(p_1,p_2)$.

Corollary 29: The kernel k_{JS} is p.d.

Proof: k_{JS} is the restriction of φ to $(M_+^1(\mathcal{X}))^2$. ■

The so-called *exponentiated JSK* (EJSK), next defined, has been used (and shown to be p.d.) by several authors [23].

Definition 30 (EJSK): Let the kernel $k_{\mathrm{EJS}} : (M_+^1(\mathcal{X}))^2 \to \mathbb{R}$ be defined (for $t > 0$) as $k_{\mathrm{EJS}}(p_1,p_2) = \exp\left[-t\, JS\,(p_1,p_2)\right]$.

Corollary 31: The EJSK is p.d.

Proof: Invoke Prop. 17 and the fact that k_{JS} is n.d. ■

Next, we introduce a weighted generalization of the EJS kernel, which allows unnormalized measures as its arguments.

Definition 32 (Weighted EJSK (WEJSK): Define the kernel $k_{\text{WEJS}} : M_+^H(\mathcal{X}) \times M_+^H(\mathcal{X}) \to \mathbb{R}$, for $t > 0$, as

$$k_{\text{WEJS}}(\mu_1, \mu_2) = \exp(t\, H(\pi)) \exp\left[-tJ^{\pi}(p_1, p_2)\right]. \quad (23)$$

Corollary 33: The kernel k_{WEJS} is p.d.

Proof: From Prop. 17 and Cor. 27. Notice that although k_{WEJS} is p.d., none of its exponential factors in (23) is p.d. ■

Finally, we study two particular (nonextensive) members cases of the family of Tsallis kernels.

Definition 34 (Boolean kernel): Let the kernel $k_{\text{Boole}} : M_+^{S_0}(\mathcal{X}) \times M_+^{S_0}(\mathcal{X}) \to \mathbb{R}$ be defined as $k_{\text{Boole}} = \widetilde{\varphi}_0$, *i.e.*,

$$k_{\text{Boole}}(\mu_1, \mu_2) = (\text{card}(\pi) - 1)\, \text{card}(\mu_1 \odot \mu_2). \quad (24)$$

Definition 35 (Linear kernel): Let the kernel $k_{\text{lin}} : M_+^1(\mathcal{X}) \times M_+^1(\mathcal{X}) \to \mathbb{R}$ be defined as

$$k_{\text{lin}}(p_1, p_2) = \frac{1}{2}\, \langle p_1, p_2 \rangle. \quad (25)$$

Corollary 36: The kernel k_{Boole} is p.d.

Proof: Invoke Prop. 24, with $q = 0$. ■

Corollary 37: The kernel k_{lin} is p.d.

Proof: This well-known property of the inner product kernel [1], also results from Prop. 24, since $k_{\text{lin}}(p_1, p_2) = \varphi_2(p_1, p_2) = \widetilde{\varphi}_2(p_1, p_2)/4$. ■

In conclusion, the Boolean kernel, the JSK, and the linear kernel, are simply particular elements of the much wider family of Tsallis kernels, continuously parameterized by $q \in [0, 2]$. Furthermore, the Tsallis kernels are a particular subfamily of the even wider set of weighted Tsallis kernels.

VII. Conclusion

In this paper we have introduced a new family of positive definite kernels between measures, which contain previous information-theoretic kernels on probability measures as particular cases. One of the key features of the new kernels is that they are defined on non-normalized measures (not necessarily normalized probabilities). This is relevant, *e.g.*, for kernels on empirical measures (such as word counts, pixel intensity histograms); instead of the usual step of normalization [2], we may leave these empirical measures unnormalized, thus allowing objects of different size (*e.g.*, documents of different lengths, images with different sizes) to be weighted differently. Another possibility is the explicit inclusion of weights: given two normalized measures, they can be multiplied by arbitrary (positive) weights before being fed to the kernel function.

Technically, the new kernels, and the proofs of positive definiteness, are supported on other contributions of this paper: the new concept of q-convexity, for which we proved a *Jensen q-inequality*; the concept of *Jensen-Tsallis q-difference*, a nonextensive generalization of the JSD; denormalization formulae for several entropies and divergences.

We are currently experimentally assessing the performance of these new kernels, namely on text classification problems.

References

[1] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002.

[2] M. Hein and O. Bousquet, "Hilbertian metrics and positive definite kernels on probability measures," in *Proc. 10th Intern. Workshop on Artificial Intelligence and Stats.*, 2005.

[3] M. Cuturi, K. Fukumizu, and J.-P. Vert, "Semigroup kernels on measures," *J. Mach. Learn. Res.*, vol. 6, pp. 1169–1198, 2005.

[4] J. Lafferty and G. Lebanon, "Diffusion kernels on statistical manifolds," *J. Mach. Learn. Res.*, vol. 6, pp. 129–163, 2005.

[5] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Th.*, vol. 37, pp. 145–151, 1991.

[6] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, vol. 1. Berkely: Univ. Calif. Press, 1961, pp. 547–561.

[7] M. E. Havrda and F. Charvát, "Quantification method of classification processes: concept of structural α-entropy," *Kybernetika*, vol. 3, pp. 30–35, 1967.

[8] Z. Daróczy, "Generalized information functions," *Information and Control*, vol. 16, pp. 36–51, Mar. 1970.

[9] J. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. Signal Proc.*, vol. 52, pp. 2210–2221, 2004.

[10] Neemuchwala, A. O. Hero, and P. Carson, "Image matching using α-entropy measures and entropic graphs," *Signal Processing*, vol. 85, pp. 277–296, 2005.

[11] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *J. Stat. Physics*, vol. 52, pp. 479–487, 1988.

[12] S. Abe, "Foundations of nonextensive statistical mechanics," in *Chaos, Nonlinearity, Complexity*, Springer, 2006.

[13] S. Abe and Y. Okamoto, *Nonextensive Statistical Mechanics and Its Applications*, Springer, 2001.

[14] J. Hu, W. Tung, and J. Gao, "Modeling sea clutter as a nonstationary and nonextensive random process," in *IEEE Int. Conf. Radar*, Gainesville, FL, 2006.

[15] Y. Li, X. Fan, and G. Li, "Image segmentation based on Tsallis-entropy and Renyi-entropy and their comparison," in *IEEE Int. Conf. Indust. Informatics*, 2006.

[16] S. Martin, G. Morison, W. Nailon, and T. Durrani, "Fast and accurate image registration using Tsallis entropy and simultaneous perturbation stochastic approximation," *Electronics Letters*, vol. 40, pp. 595–597, 2004.

[17] J. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, pp. 175–193, 1906.

[18] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 1991.

[19] J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Trans. Inf. Th.*, vol. 28, pp. 489–495, 1982.

[20] A. Martins, M. Figueiredo, and P. Aguiar, "On nonextensive entropies and divergences and their application to kernels on measures," 2008, to be submitted.

[21] H. Suyari, "Generalization of Shannon-Khinchin axioms to nonextensive systems and the uniqueness theorem for the nonextensive entropy," *IEEE Trans. Inf. Th.*, vol. 50, pp. 1783–1787, 2004.

[22] S. Furuichi, "Information theoretical properties of Tsallis entropies," *J. Math. Phy.*, vol. 47, 2006.

[23] M. Cuturi and J.-P. Vert, "Semigroup kernels on finite sets," in *Advances in Neural Information Processing Systems 17*, pp. 329–336, MIT Press, 2005.

[24] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Th.*, vol. 49, pp. 1858–1860, 2003.

[25] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inf. Th.*, vol. 46, pp. 1602–1609, 2000.

[26] A. B. Hamza, "A nonextensive information-theoretic measure for image edge detection," *Journal of Electronic Imaging*, vol. 15-1, pp. 13 011.1–13 011.8, 2006.

[27] C. Berg, J. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups*, Springer, 1984.