Nonextensive Entropic Kernels

Andre F. T. Martins[†‡] **Mario A. T. Figueiredo**[‡]
Pedro M. Q. Aguiar[♯] **Noah A. Smith**[†]
Eric P. Xing[†]

May 2008
CMU-ML-08-106

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[†]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA,
[‡]Instituto de Telecomunicações / [♯]Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal

This work was partially supported by *Fundação para a Ciência e Tecnologia* (FCT), Portugal, grant PTDC/EEA-TEL/72572/2006. A.M. was supported by a grant from FCT through the CMU-Portugal Program and the Information and Communications Technologies Institute (ICTI) at CMU. N.S. was supported by NSF IIS-0713265 and DARPA HR00110110013. E.X. was supported by NSF DBI-0546594, DBI-0640543, and IIS-0713379.

Keywords: Tsallis entropy, positive definite kernels, Jensen-Shannon divergence, mutual information

Abstract

Recent approaches to classification of text, images, and other types of structured data, launched the quest for positive definite (pd) 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.

May 19, 2008
DRAFT

1 Introduction

In the field of kernel-based machine learning [Schölkopf and Smola, 2002], there has been recent interest in defining kernels on probability distributions, to tackle several classification problems [Moreno et al., 2003, Jebara et al., 2004, Hein and Bousquet, 2005, Lafferty and Lebanon, 2005, Cuturi et al., 2005]. By defining a parametric family S containing the distributions from which the data points (in the input space X) are generated, and defining a map from X from S (*e.g.*, through maximum likelihood estimation), a distribution in S is fitted to each datum. Therefore, a kernel that is defined on $S \times S$ automatically induces a kernel on the original input space, through map composition. In text categorization, this framework appears as an alternative to the Euclidean geometry inherent to the usual bag-of-words vector representations. In fact, approaches that map data to a statistical manifold, where well-motivated non-Euclidean metrics may be defined [Lafferty and Lebanon, 2005], often outperform SVM classifiers with linear kernels [Joachims, 1997]. Some of these kernels have a natural information theoretic interpretation, creating a bridge between kernel methods and information theory [Cuturi et al., 2005, Hein and Bousquet, 2005].

We reinforce that bridge by introducing a new class of kernels rooted in *nonextensive* information theory. The Shannon and Rényi entropies [Shannon, 1948, Rényi, 1961] share the *extensivity* property: the joint entropy of a pair of independent random variables equals the sum of the individual entropies. Abandoning this property yields the so-called nonextensive entropies [Havrda and Charvát, 1967, Tsallis, 1988], which have raised great interest among physicists in modeling certain phenomena (*e.g.*, long-range interactions and multifractals) in the construction of a nonextensive generalization of the classical Boltzmann-Gibbs statistical mechanics [Abe, 2006]. Nonextensive entropies have also been recently used in signal/image processing [Li et al., 2006] and many other areas [Gell-Mann and Tsallis, 2004]. The so-called *Tsallis entropies* form a parametric family of nonextensive entropies that includes the Shannon-Boltzmann-Gibbs entropy as a particular case.

Some attempts have been made [Furuichi, 2006] to construct a nonextensive generalization of information theory. One key concept that underlies some fundamental results in information theory is the one of *convexity*, namely via the many implications of Jensen's inequality [Jensen, 1906]; for example, the non-negativity of the *Kullback-Leibler (KL) divergence* (also called *relative entropy*) [Cover and Thomas, 1991]. The Jensen inequality underlies the concept of *Jensen-Shannon (JS) divergence*, which is a symmetrized and smoothed version of the KL divergence [Lin and Wong, 1990, Lin, 1991]. The JS divergence is widely used in areas such as statistics, machine learning, image and signal processing, and physics.

Here, we introduce new extensions of JS-type divergences by extending its two building blocks: *convexity* and the *Shannon entropy*. These divergences are then used to define new information-theoretic kernels. More specifically, our main contributions are:

- The concept of q*-convexity*, as a generalization of convexity, for which we prove a *Jensen* q*-inequality*. The related concept of *Jensen* q*-differences*, which generalize Jensen differences, is also proposed. Based on these concepts, we introduce the *Jensen-Tsallis* q*-difference*, a nonextensive generalization of the JS divergence, which is also a "mutual information" in the sense of Furuichi [2006].

- Characterization of the Jensen-Tsallis q-difference, with respect to convexity and its extrema, extending the work by Burbea and Rao [1982] and by Lin [1991] for the Jensen-Shannon divergence.

- We propose a broad family of positive definite kernels, which are interpretable as nonextensive mutual information kernels. This family ranges from the Boolean to the linear kernels, and also includes the Jensen-Shannon kernel [Hein and Bousquet, 2005].

- We extend results of Hein and Bousquet [2005] by proving positive definiteness of kernels based on the unbalanced JS divergence. A connection between these new kernels and those previously studied by Fuglede [2005] and Hein and Bousquet [2005] is also established. As a side note, we show that the parametrix approximation of the multinomial diffusion kernel introduced by [Lafferty and Lebanon, 2005] is *not* positive definite in general.

The rest of the paper is organized as follows. Section 2 reviews the concepts of nonextensive entropies, with emphasis on the Tsallis case. Section 3 introduces denormalization formulae for several entropies and divergences, to be used in later sections. Section 4 discusses Jensen differences and divergences. The concepts of q-differences and q-convexity are introduced in Section 5, where they are used to define and characterize some new divergence-type quantities. Section 6 defines the Jensen-Tsallis q-difference and derives some properties. The new family of entropic kernels is described and characterized in Section 7, after a brief review of some key results concerning positive definite kernels. Section 8 reports experiments on text categorization. Finally, Section 9 contains concluding remarks and mentions directions for future research.

2 Nonextensive entropies and Tsallis statistics

We start by a brief overview of nonextensive entropies. In what follows, $\mathbb{R}_+$ denotes the nonnegative reals, $\mathbb{R}_{++}$ denotes the strictly positive reals, $\mathbb{R}_{++} \triangleq \mathbb{R}_+ \setminus \{0\}$, and

$$\Delta^{n-1} \triangleq \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i = 1, \ \forall i \ x_i \geq 0 \right\} \tag{1}$$

denotes the $(n-1)$-dimensional simplex.

Inspired by the Shannon-Khinchin axiomatic formulation of the Shannon entropy [Khinchin, 1957, Shannon and Weaver, 1949], Suyari [2004] proposed an axiomatic framework for nonextensive entropies and a uniqueness theorem. Let $q \geq 0$ be a fixed scalar, called the *entropic index*, and let f_q be a function defined on Δ^{n-1}. Consider the following set of axioms:

(A1) *Continuity*: f_q is continuous in Δ^{n-1};

(A2) *Maximality*: For any $q \geq 0$, $n \in \mathbb{N}$, and $(p_1, \ldots, p_n) \in \Delta^{n-1}$,

$$f_q(p_1, \ldots, p_n) \leq S_q(1/n, \ldots, 1/n);$$

(A3) *Generalized additivity*: For $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, $p_{ij} \geq 0$, and $p_i = \sum_{j=1}^{m_i} p_{ij}$,

$$\begin{aligned} f_q(p_{11}, \ldots, p_{nm_i}) &= f_q(p_1, \ldots, p_n) + \\ &\quad \sum_{i=1}^{n} p_i^q f_q\left(\frac{p_{i1}}{p_i}, \ldots, \frac{p_{im_i}}{p_i}\right); \end{aligned}$$

(A4) *Expandability*: $f_q(p_1, \ldots, p_n, 0) = f_q(p_1, \ldots, p_n)$.

The Suyari axioms (A1)-(A4) uniquely determine a function $S_{q,\phi} : \Delta^{n-1} \to \mathbb{R}$ of the form

$$S_{q,\phi}(p_1, \ldots, p_n) = \begin{cases} \frac{k}{\phi(q)}\left(1 - \sum_{i=1}^{n} p_i^q\right) & \text{if } q \neq 1 \\ -k \sum_{i=1}^{n} p_i \ln p_i & \text{if } q = 1, \end{cases} \tag{2}$$

where k is a positive constant, and $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a continuous function that satisfies the following three conditions: *(i)* $\phi(q)$ has the same sign as $q-1$; *(ii)* $\phi(q)$ vanishes if and only if $q = 1$; *(iii)* ϕ is differentiable in a neighborhood of 1 and $\phi'(1) = 1$. (Note that $S_{1,\phi} = \lim_{q\to 1} S_{q,\phi}$, i.e., $S_{q,\phi}(p_1, \ldots, p_n)$, seen as a function of q, is continuous at $q = 1$.) For any ϕ satisfying these conditions, $S_{q,\phi}$ has the *pseudoadditivity* property: for any two independent random variables A and B, with probability mass functions $p_A \in \Delta^{n_A - 1}$ and $p_B \in \Delta^{n_B - 1}$, respectively, consider the new random variable $A \otimes B$ defined by the joint distibution $p_A \otimes p_B \in \Delta^{n_A n_B - 1}$; then,

$$S_{q,\phi}(A \otimes B) = S_{q,\phi}(A) + S_{q,\phi}(B) - \frac{\phi(q)}{k} S_{q,\phi}(A) S_{q,\phi}(B),$$

where we denote (as usual) $S_{q,\phi}(A) \triangleq S_{q,\phi}(p_A)$.

For $q = 1$, we recover the Shannon-Boltzmann-Gibbs (SBG) entropy,

$$S_{1,\phi}(p_1, \ldots, p_n) = H(p_1, \ldots, p_n) = -k \sum_{i=1}^{n} p_i \ln p_i, \tag{3}$$

and pseudoadditivity turns into *additivity*, i.e., $H(A \otimes B) = H(A) + H(B)$ holds.

Several proposals for ϕ have appeared in the literature[Havrda and Charvát, 1967, Daróczy, 1970, Tsallis, 1988]. In the sequel, unless stated otherwise, we set $\phi(q) = q - 1$, which yields the *Tsallis entropy*:

$$S_q(p_1, \ldots, p_n) = \frac{k}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right). \tag{4}$$

To simplify, we let $k = 1$ and write the Tsallis entropy as

$$S_q(X) \triangleq S_q(p_1, \ldots, p_n) = -\sum_{x \in X} p(x)^q \ln_q p(x), \tag{5}$$

where $\ln_q(x) \triangleq (x^{1-q} - 1)/(1-q)$ is the *q-logarithm function*, which satisfies $\ln_q(xy) = \ln_q(x) + x^{1-q} \ln_q(y)$ and $\ln_q(1/x) = -x^{q-1} \ln_q(x)$. This notation was introduced by Tsallis [1988].

Furuichi [2006] derived some information theoretic properties of Tsallis entropies. Tsallis *joint* and *conditional entropies* are defined, respectively, as

$$S_q(X,Y) \triangleq -\sum_{x,y} p(x,y)^q \ln_q p(x,y) \tag{6}$$

and

$$\begin{aligned} S_q(X|Y) &\triangleq -\sum_{x,y} p(x,y)^q \ln_q p(x|y) \\ &= \sum_y p(y)^q S_q(X|y), \end{aligned} \tag{7}$$

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

For two probability mass functions $p_X,\, p_Y \in \Delta^n$, the *Tsallis relative entropy*, generalizing the KL divergence, is defined as

$$D_q(p_X \| p_Y) \triangleq -\sum_x p_X(x) \ln_q \frac{p_Y(x)}{p_X(x)}. \tag{8}$$

Finally, the *Tsallis mutual entropy* is defined as

$$I_q(X;Y) \triangleq S_q(X) - S_q(X|Y) = S_q(Y) - S_q(Y|X), \tag{9}$$

generalizing (for $q > 1$) Shannon's mutual information [Furuichi, 2006]. In Section 6, we establish a relationship between Tsallis mutual entropy and a quantity called *Tsallis q-difference*, generalizing the one between mutual information and the JS divergence (shown for example in [Grosse et al., 2002] and recalled here, in Subsection 4.2).

Furuichi [2006] mentions also what would be an alternative generalization of Shannon's mutual information,

$$\tilde{I}_q(X;Y) \triangleq D_q(p_{X,Y} \| p_X \otimes p_Y), \tag{10}$$

where $p_{X,Y}$ is the true joint probability mass function of (X,Y) and $p_X \otimes p_Y$ denotes their joint probability if they were independent. This alternative definition has also been used as a "Tsallis mutual entropy" by Lamberti and Majtey [2003]; notice that $I_q(X;Y) \neq \tilde{I}_q(X;Y)$ in general, the case $q = 1$ being a notable exception. In Section 6, we show that this alternative definition also leads to a nonextensive analogue of the JS divergence.

3 Entropies of unnormalized measures

In this section, we consider functionals that extend the domain of the SBG and Tsallis entropies to include unnormalized measures. Although, as we are going to see, these functionals are completely characterized by their restriction to the normalized probability distributions, the denormalization expressions will be used in Section 7 to derive novel positive definite kernels inspired by mutual informations.

In order to keep generality, whenever possible we do not restrict to finite or countable sample spaces. Instead, we consider a measured space $(\mathcal{X}, \mathscr{M}, \nu)$ where $\mathcal{X}$ is Hausdorff and ν is a σ-finite Radon measure. 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 its corresponding nonnegative density; this is legitimated by the Radon-Nikodym theorem, which guarantees the existence and uniqueness (up to equivalence within measure zero) of a density function $f : \mathcal{X} \to \mathbb{R}_+$. 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 $\mathcal{A} = \mathcal{X}$. Unless otherwise stated, ν is the Lebesgue-Borel measure, if $\mathcal{X} \subseteq \mathbb{R}^n$ and int$\mathcal{X} \neq \varnothing$, or the counting measure, if $\mathcal{X}$ is countable. In the latter case integrals can be seen as finite sums or infinite series.

3.1 Denormalization of the SBG Entropy and the KL Divergence

We start by introducing functionals that extend the usual notions of SBG entropy and KLD to the broader domain of unnormalized measures.

Define $\overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{-\infty, +\infty\}$. For some functional $G : M_+(\mathcal{X}) \to \overline{\mathbb{R}}$, let the set $M_+^G(\mathcal{X}) \triangleq \{f \in M_+(\mathcal{X}) : |G(f)| < \infty\}$ be its effective domain, and $M_+^{1,G}(\mathcal{X}) \triangleq M_+^G(\mathcal{X}) \cap M_+^1(\mathcal{X})$ be its subdomain of probability measures.

The following functional [Cuturi and Vert, 2005], extends the SBG entropy from $M_+^{1,H}$ to the unnormalized measures in M_+^H:

$$H(f) = -k \int f \ln f = \int \varphi_H \circ f, \tag{11}$$

where $k > 0$ is a constant, the function $\varphi_H : \mathbb{R}_{++} \to \mathbb{R}$ is defined as

$$\varphi_H(y) = -k\, y \ln y, \tag{12}$$

and, as usual, $0 \ln 0 \triangleq 0$.

The generalized form of the KL divergence, often called *generalized I-divergence* [Csiszar, 1975], 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 ($\mu_f \ll \mu_g$). Let f and g be the densities associated with μ_f and μ_g, respectively. In terms of densities, this generalized KL divergence is

$$D(f, g) \;=\; k \int \left(g - f + f \ln \frac{f}{g} \right). \tag{13}$$

Both functionals 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 $|f| \triangleq \int f = \mu_f(\mathcal{X})$. *Consider* $f \in M_+^H(\mathcal{X})$ *and* $g \in M_+^H(\mathcal{Y})$, *and define* $f \otimes g \in M_+^H(\mathcal{X} \times \mathcal{Y})$ *as* $(f \otimes g)(x, y) \triangleq f(x)g(y)$. *Then,*

$$H(f \otimes g) = |g|\, H(f) + |f|\, H(g).$$

Naturally, if $|f| = |g| = 1$, *we recover the additivity property of the SBG entropy,* $H(f \otimes g) = H(f) + H(g)$.

Proof: Straightforward from (11) and (13). ■

3.2 Denormalization of Nonextensive Entropies

Let us now proceed similarly with the nonextensive entropies. For $q \geq 0$, let $M_+^{S_q}(\mathcal{X}) = \{f \in M_+(\mathcal{X}) : f^q \in M_+(\mathcal{X})\}$ for $q \neq 1$, and $M_+^{S_q}(\mathcal{X}) = M_+^H(\mathcal{X})$ for $q = 1$. The nonextensive counterpart of (11), defined on $M_+^{S_q}(\mathcal{X})$, is

$$S_q(f) = \int \varphi_q \circ f, \tag{14}$$

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

$$\varphi_q(y) = \begin{cases} \varphi_H(y) & \text{if } q = 1, \\ \frac{k}{\phi(q)}\,(y - y^q) & \text{if } q \neq 1, \end{cases} \tag{15}$$

and $\phi : \mathbb{R}_+ \to \mathbb{R}$ satisfies conditions *(i)-(iii)* stated following (2). The Tsallis entropy is obtained for $\phi(q) = q - 1$,

$$S_q(f) = -k \int f^q \ln_q f. \tag{16}$$

Similarly, a nonextensive generalization of the generalized KLD (13) is

$$D_q(f, g) = -\frac{k}{\phi(q)} \int \left(qf + (1 - q)g - f^q g^{1-q}\right), \tag{17}$$

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

For $|f| = |g| = 1$, several particular cases are recovered: if $\phi(q) = 1 - 2^{1-q}$, then $D_q(f, g)$ is the Havrda-Charvát or Daróczi relative entropy [Havrda and Charvát, 1967, Daróczy, 1970]; if $\phi(q) = q - 1$, then $D_q(f, g)$ is the Tsallis relative entropy (8); finally, if $\phi(q) = q(q - 1)$, then $D_q(f, g)$ is the canonical α-divergence defined by Amari and Nagaoka [2001] in the realm of information geometry (with the reparameterization $\alpha = 2q - 1$ and assuming $q > 0$ so that $\phi(q) = q(q - 1)$ conforms with the axioms).

The following proposition generalizes Proposition 1 to the nonextensive case.

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

$$S_q(cf) = c^q S_q(f) + |f|\varphi_q(c), \tag{18}$$

$$D_q(cf, cg) = cD_q(f, g), \tag{19}$$

$$D_q(cf, g) = c^q D_q(f, g) - q\,\varphi_q(c)|f| + \\ + \frac{k}{\phi(q)}(q-1)(1-c^q)|g|. \tag{20}$$

For any $f \in M_+^{S_q}(\mathcal{X})$ and $g \in M_+^{S_q}(\mathcal{Y})$,

$$S_q(f \otimes g) = |g|S_q(f) + |f|S_q(g) - \frac{\phi(q)}{k} S_q(f) S_q(g). \tag{21}$$

If $|f| = |g| = 1$, we recover the pseudo-additivity property of nonextensive entropies:

$$S_q(f \otimes g) = S_q(f) + S_q(g) - \frac{\phi(q)}{k} S_q(f) S_q(g).$$

Proof: Straightforward from (14) and (17). ■

For $\phi(q) = q - 1$, D_q is the Tsallis relative entropy and (20) reduces to

$$D_q(cf, g) = c^q D_q(f, g) - q\varphi_q(c)|f| + k(1 - c^q)|g|. \tag{22}$$

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

4 Jensen Differences and Divergences

4.1 The Jensen Difference

Jensen's inequality [Jensen, 1906] is at the heart of many important results in information theory. Let $E[.]$ denote the expectation operator. Jensen's inequality states that, if Z is an integrable random variable taking values in a set $\mathcal{Z}$, and f is a measurable convex function defined on the convex hull of $\mathcal{Z}$, then

$$f(E[Z]) \leq E[f(Z)]. \tag{23}$$

Burbea and Rao [1982] considered the scenario where $\mathcal{Z}$ is finite, and took $f \triangleq -H_\varphi$, where $H_\varphi : [a, b]^n \to \mathbb{R}$ is a concave function, called a *φ-entropy*, defined as

$$H_\varphi(z) \triangleq -\sum_i \varphi(z_i), \tag{24}$$

where $\varphi : [a, b] \to \mathbb{R}$ is convex. They studied the Jensen difference

$$J_\varphi^\pi(y_1, \ldots, y_m) \triangleq H_\varphi\left(\sum_i \pi_i\, y_i\right) - \sum_i \pi_i\, H_\varphi(y_i), \tag{25}$$

where $\pi \triangleq (\pi_1, \ldots, \pi_m) \in \Delta^{m-1}$, and each $y_1, \ldots, y_m \in [a, b]^n$.

We consider here a more general scenario, involving two measured sets $(\mathcal{X}, \mathscr{M}, \nu)$ and $(\mathcal{T}, \mathscr{T}, \tau)$, where the second is used to index the first.

Definition 3 *Let $\mu \triangleq (\mu_t)_{t\in\mathcal{T}} \in [M_+(\mathcal{X})]^{\mathcal{T}}$ be a family of measures in $M_+(\mathcal{X})$ indexed by $\mathcal{T}$, and let $\omega \in M_+(\mathcal{T})$ be a measure in $\mathcal{T}$. Define:*

$$J_\Psi^\omega(\mu) \quad \triangleq \quad \Psi\left(\int_{\mathcal{T}} \omega(t)\,\mu_t\,d\tau(t)\right) - \int_{\mathcal{T}} \omega(t)\Psi(\mu_t)\,d\tau(t) \tag{26}$$

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) *$\omega(t)\Psi(\mu_t)$ is τ-integrable.*

If $\omega \in M_+^1(\mathcal{T})$, we still call (26) *a Jensen difference.*

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

4.2 The Jensen-Shannon Divergence

Let p be a random probability distribution taking values in $\{p_t\}_{t\in\mathcal{T}}$ according to a distribution $\pi \in M_+^1(\mathcal{T})$. (In classification/estimation theory parlance, π is called the prior distribution and $p_t \triangleq p(.|t)$ the likelihood function.) Then, (26) becomes

$$J_\Psi^\pi(p) = \Psi\left(E[p]\right) - E[\Psi(p)], \tag{27}$$

where the expectations are with respect to π.

Let now $\Psi = H$, the SBG entropy. Consider the random variables T and X, taking values respectively in $\mathcal{T}$ and $\mathcal{X}$, with densities $\pi(t)$ and $p(x) \triangleq \int_{\mathcal{T}} p(x|t)\pi(t)$. Using standard notation of information theory [Cover and Thomas, 1991],

$$\begin{aligned} J^\pi(p) \triangleq J_H^\pi(p) &= H\left(\int_{\mathcal{T}} \pi(t)p_t\right) - \int_{\mathcal{T}} \pi(t)H(p_t) \\ &= H(X) - \int_{\mathcal{T}} \pi(t)H(X|T=t) \\ &= H(X) - H(X|T) \\ &= I(X;T), \end{aligned} \tag{28}$$

where $I(X;T)$ is the mutual information between X and T. (This relationship between JS divergence and mutual information was pointed out by Grosse et al. [2002].) Since $I(X;T)$ is also equal to the KL divergence between the joint distribution and the product of the marginals [Cover and Thomas, 1991], we have

$$J^{\pi}(p) = H\left(E[p]\right) - E[H(p)] = E[D(p\|E[p])]. \tag{29}$$

When $\mathcal{X}$ and $\mathcal{T}$ are finite with $|\mathcal{T}| = m$, $J^{\pi}_{H}(p_1, \ldots, p_m)$ is called the *Jensen-Shannon (JS) divergence* of $p_1, \ldots, p_m$, with weights $\pi_1, \ldots, \pi_m$ [Burbea and Rao, 1982, Lin, 1991]. Equality (29) allows two interpretations of the JS divergence: (i) the Jensen difference of the Shannon entropy of P; (ii) the expected KL divergence from p to the expectation of p.

A remarkable fact is that $J^{\pi}(p) = \min_r E[D(p\|r)]$, *i.e.*, $r^* = E[p]$ is a minimizer of $E[D(p\|r)]$ with respect to r. It has been shown that this property together with equality (29) characterize the so-called *Bregman divergences*: they hold not only for $\Psi = H$, but for any concave Ψ and the corresponding Bregman divergence, in which case J^{π}_{Ψ} is the *Bregman information*, introduced by Banerjee et al. [2005].

When $|\mathcal{T}| = 2$ and $\pi = (1/2, 1/2)$, p may be seen as a random distribution whose value on $\{p_1, p_2\}$ is chosen by tossing a fair coin. In this case, $J^{(1/2,1/2)}(p) = JS(p_1, p_2)$, where

$$\begin{aligned} JS(p_1, p_2) &\triangleq H\Big(\frac{p_1 + p_2}{2}\Big) - \frac{H(p_1) + H(p_2)}{2} \\ &= \frac{1}{2} D\Big(p_1 \Big\| \frac{p_1 + p_2}{2}\Big) + \frac{1}{2} D\Big(p_2 \Big\| \frac{p_1 + p_2}{2}\Big), \end{aligned} \tag{30}$$

as introduced by Lin [1991]. It has been shown that $\sqrt{JS}$ satisfies the triangle inequality (hence being a metric) and that, moreover, it is an Hilbertian metric [Endres and Schindelin, 2003, Topsøe, 2000], which has motivated its use in kernel-based machine learning [Cuturi et al., 2005] (see Section 7).

4.3 The Jensen-Rényi Divergence

Consider again the scenario above (Subsection 4.2), with the Rényi q-entropy

$$R_q(p) = \frac{1}{1-q} \ln \int p^q \tag{31}$$

replacing the SBG entropy. It is worth to note that the Rényi and Tsallis q-entropies are monotonically related through

$$R_q(p) = \ln\Big([1 + (1-q) S_q(p)]^{\frac{1}{1-q}} \Big), \tag{32}$$

or, using the q-logarithm function,

$$S_q(p) = \ln_q \exp R_q(p). \tag{33}$$

The Rényi q-entropy is concave for $q \in [0, 1)$ and has the SBG entropy as the limit when $q \to 1$. Letting $\Psi = R_q$, (27) becomes

$$J^{\pi}_{R_q}(p) = R_q\left(E[p]\right) - E[R_q(p)]. \tag{34}$$

Unlike in the JS divergence case, there is no counterpart of equality (29) based on the Rényi q-divergence

$$D_{R_q}(p_1\|p_2) = \frac{1}{q-1}\ln\int p_1^q\, p_2^{1-q}. \tag{35}$$

When $\mathcal{X}$ and $\mathcal{T}$ are finite, we call $J^{\pi}_{R_q}$ in (34) the *Jensen-Rényi (JR) divergence*. Furthermore, when $|\mathcal{T}| = 2$ and $\pi = (1/2, 1/2)$, we write $J^{\pi}_{R_q}(p) = JR_q(p_1, p_2)$, where

$$JR_q(p_1, p_2) = R_q\left(\frac{p_1+p_2}{2}\right) - \frac{R_q(p_1)+R_q(p_2)}{2}. \tag{36}$$

The JR divergence has been used in several signal/image processing applications, such as registration, segmentation, denoising, and classification [Hamza and Krim, 2003, He et al., 2003, Karakos et al., 2007]. In Section 7, we show that the JR divergence is (as the JS divergence) an Hilbertian metric, which is relevant for its use in kernel-based machine learning.

4.4 The Jensen-Tsallis Divergence

Burbea and Rao [1982] have defined divergences of the form (27) based on the Tsallis q-entropy S_q, defined in (16). Like the SBG entropy, but unlike the Rényi entropies, the Tsallis q-entropy, for finite $\mathcal{T}$, is an instance of a φ-entropy (see (24)). Letting $\Psi = S_q$, (27) becomes

$$J^{\pi}_{S_q}(p) = S_q\left(E[p]\right) - E[S_q(p)]. \tag{37}$$

Again, like in Subsection 4.3, if we consider the Tsallis q-divergence,

$$D_q(p_1\|p_2) = \frac{1}{1-q}\left(1 - \int {p_1}^q {p_2}^{1-q}\right), \tag{38}$$

there is no counterpart of the equality (29).

When $\mathcal{X}$ and $\mathcal{T}$ are finite, $J^{\pi}_{S_q}$ in (37) is called the *Jensen-Tsallis (JT) divergence* and it has also been applied in image processing [Hamza, 2006]. Unlike the JS divergence, the JT divergence lacks an interpretation as a mutual information. Despite this, for $q \in [1, 2]$, it exhibits joint convexity [Burbea and Rao, 1982]. In the next section, we propose an alternative to the JT divergence which, amongst other features, is interpretable as a nonextensive mutual information (in the sense of Furuichi [2006]) and is jointly convex, for $q \in [0, 1]$.

5 q-Convexity and q-Differences

5.1 Introduction

This section introduces a novel class of functions, termed *Jensen q-differences*, that generalize Jensen differences. We will later (Section 6) use them to define the *Jensen-Tsallis q-difference*, which we will propose as an alternative nonextensive generalization of the JS divergence, instead of the JT divergence discussed in Subsection 4.4.

We begin by recalling the concept of q-expectation, which is used in nonextensive thermodynamics [Tsallis, 1988].

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

$$E_q[X] \triangleq \int x\, p(x)^q. \tag{39}$$

Of course, $q = 1$ corresponds to the standard notion of expectation. For $q \neq 1$, the q-expectation does not correspond to the intuitive meaning of average/expectation (*e.g.*, $E_q[1] \neq 1$ in general). Nonetheless, it has been used in the construction of nonextensive information theoretic concepts such as the Tsallis entropy, which can be written as $S_q(X) = -E_q[\ln_q p(X)]$.

5.2 q-Convexity

We now introduce a novel concept of q-convexity and use it to derive a set of results, among which we emphasize a *q-Jensen inequality*.

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* $\lambda \in [0, 1]$*,*

$$f(\lambda x + (1-\lambda)y) \leq \lambda^q f(x) + (1-\lambda)^q f(y). \tag{40}$$

f *is said to be* q-concave *if* $-f$ *is q-convex.*

Of course, 1-convexity is the usual notion of convexity. The next proposition states the q-Jensen inequality.

Proposition 6 *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* $\pi = (\pi_1, \ldots, \pi_n) \in \Delta^{n-1}$*,*

$$f\left(\sum_{i=1}^{n} \pi_i x_i\right) \leq \sum_{i=1}^{n} \pi_i^q f(x_i). \tag{41}$$

Moreover, if f is continuous, the above still holds for countably many points $(x_i)_{i\in\mathbb{N}}$*.*

Proof: Use induction, exactly as in the proof of the standard Jensen inequality [Cover and Thomas, 1991]. If f is continuous, it interchanges with the limit sign:

$$f\left(\sum_{i=1}^{\infty} \pi_i x_i\right) = f\left(\lim \sum_{i=1}^{n} \pi_i x_i\right) = \lim f\left(\sum_{i=1}^{n} \pi_i x_i\right) \leq \lim \sum_{i=1}^{n} \pi_i^q f(x_i) = \sum_{i=1}^{\infty} \pi_i^q f(x_i). \tag{42}$$

■

Proposition 7 *Let* $f \geq 0$ *and* $q \geq r \geq 0$*; then,*

$$f \text{ is } q\text{-convex} \quad \Rightarrow \quad f \text{ is } r\text{-convex} \tag{43}$$

$$f \text{ is } r\text{-concave} \quad \Rightarrow \quad f \text{ is } q\text{-concave}. \tag{44}$$

Proof: Implication (43) results from

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &\leq \lambda^q f(x) + (1-\lambda)^q f(y) \\ &\leq \lambda^r f(x) + (1-\lambda)^r f(y), \end{aligned}$$

where the first inequality states the q-convexity of f and the second one is valid because $f(x), f(y) \geq 0$ and $t^r \geq t^q \geq 0$, for any $t \in [0, 1]$ and $q \geq r$. The proof of (44) is analogous. ■

5.3 Jensen q-Differences

We now generalize Jensen differences, formalized in Definition 3, by introducing the concept of Jensen q-differences.

Definition 8 *Let $\mu \triangleq (\mu_t)_{t\in\mathcal{T}} \in [M_+(\mathcal{X})]^{\mathcal{T}}$ be a family of measures in $M_+(\mathcal{X})$ indexed by $\mathcal{T}$, and let $\omega \in M_+(\mathcal{T})$ be a measure in $\mathcal{T}$. For $q \geq 0$, define*

$$T^{\omega}_{q,\Psi}(\mu) \;\triangleq\; \Psi\left(\int_{\mathcal{T}} \omega(t)\,\mu_t\,d\tau(t)\right) - \int_{\mathcal{T}} \omega(t)^q \Psi(\mu_t)\,d\tau(t), \tag{45}$$

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) $\omega(t)^q\Psi(\mu_t)$ *is τ-integrable.*

If $\omega \in M^1_+(\mathcal{T})$, we call the function defined in (45) *a* Jensen q-difference.

Burbea and Rao [1982] established necessary and sufficient conditions on φ for the Jensen difference of a φ-entropy to be convex. The following proposition generalizes that result, extending it to Jensen q-differences.

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 [Burbea and Rao, 1982]) function $\Psi : [0,1]^n \to \mathbb{R}$ defined by $\Psi(z) \triangleq -\sum_{i=1}^n \varphi(z_i)$. 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.*

Proof: The case $q = 1$ corresponds to the Jensen difference and was proved by Burbea and Rao [1982] (Theorem 1). Our proof extends that to $q \neq 1$.

Let $y = (y_1, ..., y_m)$, where $y_t = (y_{t1}, ..., y_{tn})$. Thus

$$\begin{aligned} T^{\pi}_{q,\Psi}(y) &= \Psi\left(\sum_{t=1}^m \pi_t y_t\right) - \sum_{t=1}^m \pi_t^q\ \Psi(y_t) \\ &= \sum_{i=1}^n \left[\sum_{t=1}^m \pi_t^q \varphi(y_{ti}) - \varphi\left(\sum_{t=1}^m \pi_t y_{ti}\right)\right], \end{aligned}$$

showing that it suffices to consider $n = 1$, where each $y_t \in [0,1]$, *i.e.*,

$$T^{\pi}_{q,\Psi}(y_1, \ldots, y_m) = \sum_{t=1}^m \pi_t^q\ \varphi(y_t) - \varphi\left(\sum_{t=1}^m \pi_t y_t\right); \tag{46}$$

this function is convex on $[0,1]^m$ if and only if, for every fixed $a_1,\ldots,a_m \in [0,1]$, and $b_1,\ldots,b_m \in \mathbb{R}$, the function

$$f(x) = T^{\pi}_{q,\Psi}(a_1 + b_1 x, \ldots, a_m + b_m x) \tag{47}$$

is convex in $\{x \in \mathbb{R} : a_t + b_t x \in [0,1],\ t = 1,\ldots,m\}$. Since f is C^2, it is convex if and only if $f''(t) \geq 0$.

We first show that convexity of f (equivalently of $T^{\pi}_{q,\Psi}$) implies convexity of φ. Letting $c_t = a_t + b_t x$,

$$f''(x) = \sum_{t=1}^{m} \pi_t^q\, b_t^2\, \varphi''(c_t) - \left(\sum_{t=1}^{m} \pi_t\, b_t\right)^2 \varphi''\left(\sum_{t=1}^{m} \pi_t\, c_t\right). \tag{48}$$

By choosing $x = 0$, $a_t = a \in [0,1]$, for $t = 1, ..., m$, and $b_1,\ldots,b_m$ satisfying $\sum_t \pi_t b_t = 0$ in (48), we get

$$f''(0) = \varphi''(a) \sum_{t=1}^{m} \pi_t^q b_t^2,$$

hence, if f is convex, $\varphi''(a) \geq 0$ thus φ is convex.

Next, we show that convexity of f also implies $(2-q)$-convexity of $-1/\varphi''$. By choosing $x = 0$ (thus $c_t = a_t$) and $b_t = \pi_t^{1-q}(\varphi''(a_t))^{-1}$, we get

$$\begin{aligned} f''(0) &= \sum_{t=1}^{m} \frac{\pi_t^{2-q}}{\varphi''(a_t)} - \left(\sum_{t=1}^{m} \frac{\pi_t^{2-q}}{\varphi''(a_t)}\right)^2 \varphi''\left(\sum_{t=1}^{m} \pi_t a_t\right) \\ &= \left(\sum_{t=1}^{m} \frac{\pi_t^{2-q}}{\varphi''(a_t)}\right) \varphi''\left(\sum_{t=1}^{m} \pi_t a_t\right) \\ &\quad \times \left[\frac{1}{\varphi''\left(\sum_{t=1}^{m} \pi_t a_t\right)} - \sum_{t=1}^{m} \frac{\pi_t^{2-q}}{\varphi''(a_t)}\right], \end{aligned}$$

where the expression inside the square brackets is the Jensen $(2-q)$-difference of $1/\varphi''$ (see Definition 8). Since $\varphi''(x) \geq 0$, the factor outside the square brackets is non-negative, thus the Jensen $(2-q)$-difference of $1/\varphi''$ is also nonnegative and $-1/\varphi''$ is $(2-q)$-convex.

Finally, we show that if φ is convex and $-1/\varphi''$ is $(2-q)$-convex, then $f'' \geq 0$, thus $T^{\pi}_{q,\Psi}$ is convex. Let $r_t = (q\pi_t^{2-q}/\varphi''(c_t))^{1/2}$ and $s_t = b_t(\pi_t^q \varphi''(c_t)/q)^{1/2}$; then, non-negativity of f'' results from the following chain of inequalities/equalities:

$$0 \leq \left(\sum_{t=1}^{m} r_t^2\right)\left(\sum_{t=1}^{m} s_t^2\right) - \left(\sum_{t=1}^{m} r_t\, s_t\right)^2 \tag{49}$$

$$= \sum_{t=1}^{m} \frac{\pi_t^{2-q}}{\varphi''(c_t)} \sum_{t=1}^{m} b_t^2 \pi_i^q \varphi''(c_t) - \left(\sum_{t=1}^{m} b_t \pi_t\right)^2 \tag{50}$$

$$\leq \frac{1}{\varphi''\left(\sum_{t=1}^{m} \pi_t c_t\right)} \sum_{t=1}^{m} b_t^2 \pi_t^q \varphi''(c_t) - \left(\sum_{t=1}^{m} b_t \pi_t\right)^2 \tag{51}$$

$$= \frac{1}{\varphi''\left(\sum_{t=1}^{m} \pi_t c_t\right)} \cdot f''(t), \tag{52}$$

where: (49) is the Cauchy-Schwarz inequality; equality (50) results from the definitions of r_t and s_t and from the fact that $r_t s_t = b_t \pi_t$; inequality (51) states the $(2-q)$-convexity of $-1/\varphi''$; equality (52) results from (48). ■

6 The Jensen-Tsallis q-Difference

6.1 Definition

As in Subsection 4.2, let p be a random probability distribution taking values in $\{p_t\}_{t\in\mathcal{T}}$ according to a distribution $\pi \in M_+^1(\mathcal{T})$. Then, we may write

$$T_{q,\Psi}^{\pi}(p) = \Psi\left(E[p]\right) - E_q[\Psi(p)], \tag{53}$$

where the expectations are with respect to π. Hence Jensen q-differences may be seen as deformations of the standard Jensen differences (27), in which the second expectation is replaced by a q-expectation.

Let now $\Psi = S_q$, the nonextensive Tsallis q-entropy. Introducing the random variables T and X, with values respectively in $\mathcal{T}$ and $\mathcal{X}$, with densities $\pi(t)$ and $p(x) \triangleq \int_{\mathcal{T}} p(x|t)\pi(t)$, we have (writing T_{q,S_q}^{π} simply as T_q^{π})

$$\begin{aligned} T_q^{\pi}(p) &= S_q\left(E[p]\right) - E_q[S_q(p)] \\ &= S_q(X) - \int_{\mathcal{T}} \pi(t)^q S_q(X|T=t) \\ &= S_q(X) - S_q(X|T) \\ &= I_q(X;T), \end{aligned} \tag{54}$$

where $S_q(X|T)$ is the Tsallis conditional entropy (7), and $I_q(X;T)$ is the Tsallis mutual information (9), as defined by Furuichi [2006]. Observe that (54) is a nonextensive analogue of (28). Since, in general, $I_q \neq \tilde{I}_q$ (see (10)), unless $q = 1$ (in that case, $I_1 = \tilde{I}_1 = I$), there is no counterpart of (29) in terms of q-differences. Nevertheless, Lamberti and Majtey [2003] have proposed a non-logarithmic version of the JS divergence, which corresponds to using $\tilde{I}_q$ for the Tsallis mutual q-entropy (although this interpretation is not explicitally mentioned by those authors).

When $\mathcal{X}$ and $\mathcal{T}$ are finite with $|\mathcal{T}| = m$, we call the quantity $T_q^{\pi}(p_1, \ldots, p_m)$ the *Jensen-Tsallis (JT) q-difference* of $p_1, \ldots, p_m$ with weights $\pi_1, \ldots, \pi_m$. Although the JT q-difference is a generalization of the JS divergence, for $q \neq 1$, the term "divergence" would be misleading in this case, since T_q^{π} may take negative values (if $q < 1$) and does not vanish in general if p is deterministic.

When $|\mathcal{T}| = 2$ and $\pi = (1/2, 1/2)$, define $T_q \triangleq T_q^{1/2,1/2}$,

$$T_q(p_1, p_2) = S_q\left(\frac{p_1 + p_2}{2}\right) - \frac{S_q(p_1) + S_q(p_2)}{2^q}. \tag{55}$$

Notable cases arise for particular values of q:

- For $q = 0$, $S_0(p) = -1 + \nu(\mathrm{supp}(p))$, where $\nu(\mathrm{supp}(p))$ denotes the measure of the support of p (recall that p is defined on the measured space $(\mathcal{X}, \mathscr{M}, \nu)$). For example, if $\mathcal{X}$ is finite and ν is the counting measure, $\nu(\mathrm{supp}(p)) = \|p\|_0$ is the so-called *0-norm* (although it's not a norm) of vector p, *i.e.*, its number of nonzero components. The Jensen-Tsallis 0-difference is thus

$$\begin{aligned} T_0(p_1, p_2) &= -1 + \nu\left(\mathrm{supp}\left(\frac{p_1 + p_2}{2}\right)\right) + 1 - \nu\left(\mathrm{supp}(p_1)\right) + 1 - \nu\left(\mathrm{supp}(p_2)\right) \\ &= 1 + \nu\left(\mathrm{supp}(p_1) \cup \mathrm{supp}(p_2)\right) - \nu\left(\mathrm{supp}(p_1)\right) - \nu\left(\mathrm{supp}(p_2)\right) \\ &= 1 - \nu\left(\mathrm{supp}(p_1) \cap \mathrm{supp}(p_2)\right); \end{aligned} \tag{56}$$

  if $\mathcal{X}$ is finite and ν is the counting measure, this becomes

$$T_0(p_1, p_2) = 1 - \|p_1 \odot p_2\|_0, \tag{57}$$

  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 JS divergence,

$$T_1(p_1, p_2) = JS(p_1, p_2). \tag{58}$$

- For $q = 2$, $S_2(p) = 1 - \langle p, p \rangle$, where $\langle a, b \rangle = \int_{\mathcal{X}} a(x) b(x) d\nu(x)$ is the inner product between a and b (which reduces to $\langle a, b \rangle = \sum_i a_i\, b_i$ if $\mathcal{X}$ is finite and ν is the counting measure). Consequently, the Tsallis 2-difference is

$$T_2(p_1, p_2) = \frac{1}{2} - \frac{1}{2}\,\langle p_1, p_2 \rangle, \tag{59}$$

  which we call the *linear difference*.

6.2 Properties of the JT q-difference

This subsection presents results regarding convexity and extrema of the JT q-difference, for several values of q, extending known properties of the JS divergence ($q = 1$).

Some properties of the JS divergence are lost in the transition to nonextensivity. For example, while the former is nonnegative and vanishes if and only if all the distributions are identical, this is not true in general with the JT q-difference. Nonnegativity of the JT q-difference is only guaranteed if $q \geq 1$, which explains why some authors (*e.g.*, Furuichi [2006]) only consider values of $q \geq 1$, when looking for nonextensive analogues of Shannon's information theory. Moreover, unless $q = 1$, it is not generally true that $T_q^\pi(p, \ldots, p) = 0$ or even that $T_q^\pi(p, \ldots, p, p') \geq T_q^\pi(p, \ldots, p, p)$. For example, the solution to the optimization problem

$$\min_{p_1 \in \Delta^n} T_q(p_1, p_2), \tag{60}$$

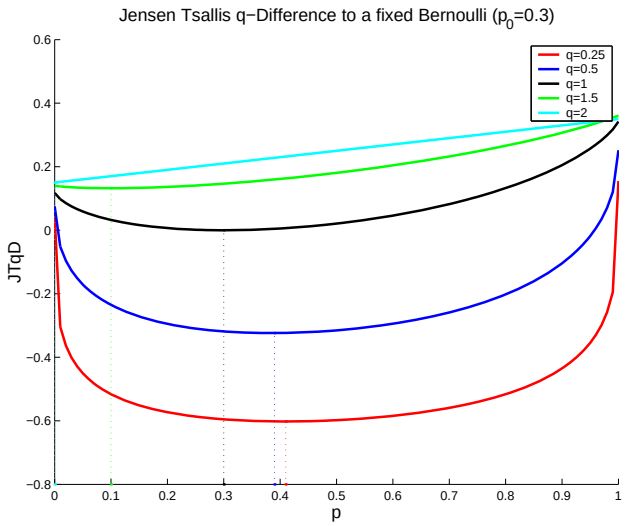

Figure 1: Jensen-Tsallis q-difference between two Bernoulli distributions, $p_1 = (p_0, 1-p_0)$ and $p_2 = (p, 1-p)$, for several values of the entropic index q. The first Bernoulli was kept fixed ($p_0 = 0.3$) and the second was varied in the range $p \in [0, 1]$. We may observe that, for $q \in [0, 1)$ the minimizant of the JT q-difference is closer to the uniform distribution, $p_2^* \to (0.5, 0.5)$ as $q \to 0$, while for $q \in (1, 2]$ it is closer to the degenerate distribution, $p_2^* \to (0, 1)$ as $q \to 2$. In the extensive case ($q = 1$), the minimizant is $p_2^* = p_1$.

is, in general, different from p_2, unless if $q = 1$. Instead, this minimizer is closer to the uniform distribution, if $q \in [0, 1)$, and closer to a degenerate distribution, for $q \in (1, 2]$. This is not so surprising: recall that $T_2(p_1, p_2) = \frac{1}{2} - \frac{1}{2}\langle p_1, p_2 \rangle$; in this case, (60) becomes a linear program, and the solution is not p_2, but $p_1^* = \delta_j$, where $j = \arg\max_i p_{2i}$ (see Fig.1).

We start by recalling a basic result, which essentially confirms that Tsallis entropies satisfy one of the Suyari axioms (see Axiom A2 in Section 1), which states that entropies should be maximized by uniform distributions.

Proposition 10 *Let $\mathcal{X}$ be a finite set. The uniform distribution maximizes the Tsallis entropy for any $q \geq 0$.*

Proof: Consider the problem

$$\max_p S_q(p), \quad \text{subject to } \textstyle\sum_i p_i = 1 \text{ and } p_i \geq 0.$$

Equating the gradient of the Lagrangian to zero, yields

$$\tfrac{\partial}{\partial p_i}\left(S_q(p) + \lambda(\textstyle\sum_i p_i - 1)\right) = -q(q-1)^{-1} p_i^{q-1} + \lambda = 0,$$

for all i. Since all these equations are identical, the solution is the uniform distribution, which is a maximum, due to the concavity of S_q. ■

The following corollary of Proposition 9 establishes the joint convexity of the JT q-difference, for $q \in [0,1]$. (Interestingly, this "complements" the joint convexity of the JT divergence (37), for $q \in [1,2]$, which was proved by Burbea and Rao [1982].)

Corollary 11 *Let $\mathcal{T}$ and $\mathcal{X}$ be finite sets with cardinalities m and n, respectively. For $q \in [0,1]$, the JT q-difference is a jointly convex function on $M_+^{1,S_q}(\mathcal{X})$. Formally, let $\{p_t^{(i)}\}_{t\in\mathcal{T}}$, and $i = 1,\ldots,l$, be a collection of l sets of probability distributions on $\mathcal{X}$; then, for any $(\lambda_1,\ldots,\lambda_l) \in \Delta^{l-1}$,*

$$T_q^\pi\left(\sum_{i=1}^l \lambda_i p_1^{(i)},\ldots,\sum_{i=1}^l \lambda_i p_m^{(i)}\right) \le \sum_{i=1}^l \lambda_i T_q^\pi(p_1^{(i)},\ldots,p_m^{(i)}).$$

Proof: Observe that the Tsallis entropy (5) of a probability distribution $p_t = \{p_{t1},...,p_{tn}\}$ can be written as

$$S_q(p_t) = -\sum_{i=1}^n \varphi(p_{ti}), \quad \text{where} \quad \varphi_q(x) = \frac{x - x^q}{1-q};$$

thus, from Proposition 9, T_q^π is convex if and only if φ_q is convex and $-1/\varphi_q''$ is $(2-q)$-convex. Since $\varphi_q''(x) = q\,x^{q-2}$, φ_q is convex for $x \ge 0$ and $q \ge 0$. To show the $(2-q)$-convexity of $-1/\varphi_q''(x) = -(1/q)x^{2-q}$, for $x_t \ge 0$, and $q \in [0,1]$, we use a version of the power mean inequality [Steele, 2006],

$$-\left(\sum_{i=1}^l \lambda_i\, x_i\right)^{2-q} \le -\sum_{i=1}^l (\lambda_i\, x_i)^{2-q} = -\sum_{i=1}^l \lambda_i^{2-q}\, x_i^{2-q},$$

thus concluding that $-1/\varphi_q''$ is in fact $(2-q)$-convex. ■

The next corollary, which results from the previous one, provides an upper bound for the JT q-difference, for $q \in [0,1]$. Although this result is weaker than that of Proposition 13 below, we include it since it provides insight about the upper extrema of the JT q-difference.

Corollary 12 *Let $\mathcal{X}$, $\mathcal{T}$ and q be as in Corollary 11. Then, $T_q^\pi(p_1,\ldots,p_m) \le S_q(\pi)$.*

Proof: From Corollary 11, for $q \in [0,1]$, $T_q^\pi(p_1,\ldots,p_m)$ is convex. Since its domain is a convex polytope (the cartesian product of m simplices), its maximum occurs on a vertex, *i.e.*, when each argument p_t is a degenerate distribution at x_t, denoted δ_{x_t}. In particular, if $|\mathcal{X}| \ge |\mathcal{T}|$, this maximum occurs at the vertex corresponding to disjoint degenerate distributions, *i.e.*, such that $x_i \ne x_j$ if $i \ne j$. At this maximum,

$$\begin{aligned} T_q^\pi(\delta_{x_1},\ldots,\delta_{x_m}) &= S_q\left(\sum_{t=1}^m \pi_t \delta_{x_t}\right) - \sum_{t=1}^m \pi_t S_q(\delta_{x_t}) \\ &= S_q\left(\sum_{t=1}^m \pi_t \delta_{x_t}\right) \qquad (61) \\ &= S_q(\pi) \qquad (62) \end{aligned}$$

where the equality in (61) results from $S_q(\delta_{x_t}) = 0$. Notice that this maximum may not be achieved if $|\mathcal{X}| < |\mathcal{T}|$. ■

The next proposition establishes (upper and lower) bounds for the JT q-difference, extending Corollary 12 to any non-negative q and to countable $\mathcal{X}$ and $\mathcal{T}$.

Proposition 13 *Let $\mathcal{T}$ and $\mathcal{X}$ be countable sets. For $q \geq 0$,*

$$T_q^\pi(p_1, \ldots, p_m) \leq S_q(\pi), \tag{63}$$

and, if $|\mathcal{X}| \geq |\mathcal{T}|$, the maximum is reached for a set of disjoint degenerate distributions. As in Corollary 12, this maximum may not be attained if $|\mathcal{X}| < |\mathcal{T}|$.

For $q \geq 1$,

$$T_q^\pi(p_1, \ldots, p_m) \geq 0, \tag{64}$$

and the minimum is attained in the pure deterministic case, i.e., when all distributions are equal to same degenerate distribution.

For $q \in [0, 1]$ and $\mathcal{X}$ a finite set with $|\mathcal{X}| = n$,

$$T_q^\pi(p_1, \ldots, p_m) \geq S_q(\pi)[1 - n^{1-q}]. \tag{65}$$

This lower bound (which is zero or negative) is attained when all distributions are uniform.

Proof: The proof of (63), for $q \geq 0$, results from

$$\begin{aligned} T_q^\pi(p_1, \ldots, p_m) &= \frac{1}{q-1}\left[1 - \sum_{j=1}^{n}\left(\sum_{t=1}^{m}\pi_t p_{tj}\right)^q - \sum_{t=1}^{m}\pi_t^q\left(1 - \sum_{j=1}^{n} p_{tj}^q\right)\right] \\ &= S_q(\pi) + \frac{1}{q-1}\sum_{j=1}^{n}\left[\sum_{t=1}^{m}(\pi_t p_{tj})^q - \left(\sum_{t=1}^{m}\pi_t p_{tj}\right)^q\right] \\ &\leq S_q(\pi), \end{aligned} \tag{66}$$

where the inequality holds since, for $y_i \geq 0$: if $q \geq 1$, then $\sum_i y_i^q \leq (\sum_i y_i)^q$; if $q \in [0, 1]$, then $\sum_i y_i^q \geq (\sum_i y_i)^q$.

The proof that $T_q^\pi \geq 0$ for $q \geq 1$, uses the notion of q-convexity. Since $\mathcal{X}$ is countable, the Tsallis entropy is as in (4), thus $S_q \geq 0$. Since $-S_q$ is 1-convex, then, by Proposition 7, it is also q-convex for $q \geq 1$. Consequently, from the q-Jensen inequality (Proposition 6), for finite $\mathcal{T}$, with $|\mathcal{T}| = m$,

$$T_q^\pi\,(p_1, \ldots, p_m) = S_q\left(\sum_{t=1}^{m}\pi_t p_t\right) - \sum_{t=1}^{m}\pi_t^q S_q(p_t) \geq 0.$$

Since S_q is continuous, so is T_q^π, thus the inequality is valid in the limit as $m \to \infty$, which proves the assertion for $\mathcal{T}$ countable. Finally, $T_q^\pi(\delta_1, \ldots, \delta_1, \ldots) = 0$, where δ_1 is some degenerate distribution.

Finally, to prove (65), for $q \in [0,1]$ and $\mathcal{X}$ finite,

$$\begin{aligned} T_q^\pi(p_1,\ldots,p_m) &= S_q\left(\sum_{t=1}^m \pi_t p_t\right) - \sum_{t=1}^m \pi_t^q S_q(p_t) \\ &\geq \sum_{t=1}^m \pi_t S_q(p_t) - \sum_{t=1}^m \pi_t^q S_q(p_t) \qquad (67) \\ &= \sum_{t=1}^m (\pi_t - \pi_t^q) S_q(p_t) \\ &\geq S_q(U) \sum_{t=1}^m (\pi_t - \pi_t^q) \qquad (68) \\ &= S_q(\pi)[1 - n^{1-q}]. \qquad (69) \end{aligned}$$

where the inequality (67) results from S_q being concave, and the inequality 68 holds since $\pi_t - \pi_t^q \leq 0$, for $q \in [0,1]$, and the uniform distribution U maximizes S_q (Proposition 10), with $S_q(U) = (1-n^{1-q})/(q-1)$. ■

Finally, the next proposition characterizes the convexity/concavity of the JT q-difference.

Proposition 14 *Let $\mathcal{T}$ and $\mathcal{X}$ be countable sets. The JT q-difference is convex in each argument, for $q \in [0,2]$, and concave in each argument, for $q \geq 2$.*

Proof: Notice that the JT q-difference can be written as $T_q^\pi(p_1,\ldots,p_m) = \sum_j \psi(p_{1j},\ldots,p_{mj})$, with

$$\psi(y_1,\ldots,y_m) = \frac{1}{q-1}\left[\sum_i (\pi_i - \pi_i^q) y_i + \sum_i \pi_i^q y_i^q - \left(\sum_i \pi_i y_i\right)^q\right].$$

It suffices to consider the second derivative of ψ with respect to y_1. Introducing $z = \sum_{i=2}^m \pi_i\, y_i$,

$$\begin{aligned} \frac{\partial^2 \psi}{\partial y_1^2} &= q\left[\pi_1^q\, y_1^{q-2} - \pi_1^2\,(\pi_1\, y_1 + z)^{q-2}\right] \\ &= q\,\pi_1^2\left[(\pi_1\, y_1)^{q-2} - (\pi_1\, y_1 + z)^{q-2}\,\right]. \qquad (70) \end{aligned}$$

Since $\pi_1\, y_1 \leq (\pi_1\, y_1 + z) \leq 1$, the quantity in (70) is nonnegative for $q \in [0,2]$ and non-positive for $q \geq 2$. ■

7 Nonextensive mutual information kernels

7.1 Introduction

In this section we consider the application of extensive and nonextensive entropies to define kernels on measures; since kernels involve pairs of measures, throughout this section $|\mathcal{T}| = 2$. Based

on the denormalization formulae presented in Section 3, we devise novel kernels related to the JS divergence and the JT q-difference, which allow setting a weight for each argument of the kernel. We will call them *weighted Jensen-Tsallis kernels*. We also introduce kernels related to the JR divergence (Subsection 4.3) and the JT divergence (Subsection 4.4). Finally, we establish a connection between the Tsallis kernels and a family of kernels investigated by Hein et al. [2004a] and Fuglede [2005], giving those kernels a new information-theoretic interpretation.

7.2 Positive and negative definite kernels

We start by recalling basic concepts from kernel theory [Schölkopf and Smola, 2002]; in the following, $\mathcal{X}$ denotes a nonempty set.

Definition 15 *Let $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function, i.e., a function satisfying $\varphi(y, x) = \varphi(x, y)$, for all $x, y \in \mathcal{X}$. φ is called a* positive definite *(pd) kernel if and only if*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \varphi(x_i, x_j) \geq 0 \tag{71}$$

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

Definition 16 *Let $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric. ψ is called a* negative definite *(nd) kernel if and only if*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \psi(x_i, x_j) \leq 0 \tag{72}$$

for all $n \in \mathbb{N}$, $x_i, \ldots, x_n \in \mathcal{X}$ and $c_i, \ldots, c_n \in \mathbb{R}$, satisfying the additional constraint $c_1 + \ldots + c_n = 0$. In this case, $-\psi$ is called conditionally *pd; obviously, positive definiteness implies conditional positive definiteness.*

Both the sets of pd and nd kernels are closed under pointwise sums/integrations, the former being also closed under pointwise products; moreover, both sets are closed under pointwise convergence. While pd kernels correspond to inner products via embedding in a Hilbert space, nd kernels that vanish on the diagonal and are positive anywhere else, correspond to squared Hilbertian distances. These facts, and the following ones, are shown by Berg et al. [1984].

Proposition 17 *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*

$$\varphi(x, y) = \psi(x, x_0) + \psi(y, x_0) - \psi(x, y) - \psi(x_0, x_0). \tag{73}$$

Then, φ is pd if and only if ψ is nd.

Proposition 18 *The function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a nd kernel if and only if $\exp(-t\psi)$ is pd for all $t > 0$.*

Proposition 19 *The function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a nd kernel if and only if $(t + \psi)^{-1}$ is pd for all $t > 0$.*

Lemma 20 *If ψ is nd 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 $\ln(1 + \psi)$.*

Lemma 21 *If $f : \mathcal{X} \to \mathbb{R}$ satisfies $f \geq 0$, then, for $\alpha \in [1, 2]$, the function $\psi_\alpha(x, y) = -(f(x) + f(y))^\alpha$ is a nd kernel.*

The following definition has been used in a machine learning context by Cuturi and Vert [2005], following Berg et al. [1984].

Definition 22 *Let $(\mathcal{X}, +)$ be a semigroup.*[1] *A function $\varphi : \mathcal{X} \to \mathbb{R}$ is called pd (in the semigroup sense) if $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined as $k(x, y) = \varphi(x + y)$, is a pd kernel. Likewise, φ is called nd if k is a nd kernel. Accordingly, these are called* semigroup kernels.

7.3 Jensen-Shannon and Tsallis kernels

The basic result that allows deriving pd kernels based on the JS divergence and, more generally, on the JT q-difference, is the fact that the denormalized Tsallis q-entropies (14) are nd functions on $M_+^{S_q}(\mathcal{X})$, for $q \in [0, 2]$. Of course, this includes the denormalized SBG entropy (11) as a particular case, corresponding to $q = 1$. Although part of the proof was given by Berg et al. [1984] (and by Topsøe [2000] and Cuturi and Vert [2005] for the Shannon entropy case), we present a complete proof here.

Proposition 23 *For $q \in [0, 2]$, the denormalized Tsallis q-entropy S_q is a nd function on $M_+^{S_q}(\mathcal{X})$.*

Proof: Since nd kernels are closed under pointwise integration, it suffices to prove that φ_q (see (15)) is nd on $\mathbb{R}_+$. For $q \neq 1$, $\varphi_q(y) = (q-1)^{-1}(y - y^q)$. Let's consider two cases separately: if $q \in [0, 1)$, $\varphi_q(y)$ equals a positive constant times $-\iota + \iota^q$, where $\iota(y) = y$ is the identity map defined on $\mathbb{R}_+$. Since the set of nd functions is closed under sums, we only need to show that both $-\iota$ and ι^q are nd. Both ι and $-\iota$ are nd, as can easily be seen from the definition; besides, since ι is nd and nonnegative, Lemma 20 guarantees that ι^q is also nd. For the second case, where $q \in (1, 2]$, $\varphi_q(y)$ equals a positive constant times $\iota - \iota^q$. It only remains to show that $-\iota^q$ is nd for $q \in (1, 2]$: Lemma 21 guarantees that the kernel $k(x, y) = -(x + y)^q$ is nd; therefore $-\iota^q$ is a nd function.

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

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

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

[1] Recall that $(\mathcal{X}, +)$ is a *semigroup* if $+$ is a binary operation in $\mathcal{X}$ that is associative and has an identity element.

The following lemma [Berg et al., 1984] will also be necessary below.

Lemma 24 *The function $\zeta_q : \mathbb{R}_{++} \to \mathbb{R}$, defined as $\zeta_q(y) = y^{-q}$ is pd, for $q \in [0, 1]$.*

Proof: We need to show that $k_q(x, y) : \mathbb{R}_{++} \times \mathbb{R}_{++} \to \mathbb{R}$, defined as $k_q(x, y) = \zeta_q(x + y)$, is pd, for $q \in [0, 1]$. The proof results from observing that

$$k_q(x, y) = (x + y)^{-q} = \lim_{t \to 0^+} \left[t + (x + y)^q\right]^{-1}, \tag{74}$$

which is always well defined because $x + y > 0$, combined with the following facts: from Lemma 20, since $(x, y) \mapsto x + y$ is nd and nonnegative, $(x, y) \mapsto (x + y)^q$ is nd; from Proposition 19, $(x, y) \mapsto \left[t + (x + y)^q\right]^{-1}$ is pd for any $t > 0$; the set of pd kernels is closed under limits. ■

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 JS-based (and other) kernels in two ways:

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

Definition 25 (weighted Jensen-Tsallis kernels) *The kernel $\widetilde{k}_q : M_+^{S_q}(\mathcal{X}) \times M_+^{S_q}(\mathcal{X}) \to \mathbb{R}$ is defined as*

$$\begin{aligned} \widetilde{k}_q(\mu_1, \mu_2) &\triangleq \widetilde{k}_q(\omega_1 p_1, \omega_2 p_2) \\ &= \Big(S_q(\pi) - T_q^{\pi}(p_1, p_2)\Big)\,(\omega_1 + \omega_2)^q, \end{aligned}$$

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 $k_q : \Big(M_+^{S_q}(\mathcal{X}) \setminus \{0\}\Big)^2 \to \mathbb{R}$ is defined as

$$k_q(\mu_1, \mu_2) \triangleq k_q(\omega_1 p_1, \omega_2 p_2) = S_q(\pi) - T_q^{\pi}(p_1, p_2).$$

Recalling (54), notice that $S_q(\pi) - T_q^{\pi}(p_1, p_2) = S_q(T) - I_q(X; T) = S_q(T|X)$ can be interpreted as the *Tsallis posterior conditional entropy*. Hence, k_q can be seen (in Bayesian classification terms) as a nonextensive expected measure of uncertainty in correctly identifying the class, given the prior $\pi = (\pi_1, \pi_2)$, and a random sample from the mixture distribution $\pi_1 p_1 + \pi_2 p_2$. The more similar the two distributions are, the greater this uncertainty.

Proposition 26 *The kernel $\widetilde{k}_q$ is pd, for $q \in [0, 2]$. The kernel k_q is pd, for $q \in [0, 1]$.*

Proof: With $\mu_1 = \omega_1 p_1$ and $\mu_2 = \omega_2 p_2$ and using the denormalization formula of Proposition 2, we obtain $\widetilde{k}_q(\mu_1, \mu_2) = -S_q(\mu_1 + \mu_2) + S_q(\mu_1) + S_q(\mu_2)$. Now invoke Prop. 17 with $\psi = S_q$ (which is nd by Prop. 23), $x = \mu_1$, $y = \mu_2$, and $x_0 = 0$ (the null measure). Observe now that $k_q(\mu_1, \mu_2) = \widetilde{k}_q(\mu_1, \mu_2)(\omega_1 + \omega_2)^{-q}$. Since the product of two pd kernels is a pd kernel and (Prop. 24) $(\omega_1 + \omega_2)^{-q}$ is a pd kernel, for $q \in [0, 1]$, we conclude that k_q is pd. ■

As we can see, the weighted Jensen-Tsallis kernels have two inherent properties: they are parameterized by the entropic index q and they allow their arguments to be unbalanced, *i.e.*, to have different weights ω_i. We now mention some instances of kernels where each of these degrees of freedom is suppressed.

We start by the following subfamily of kernels, obtained by setting $q = 1$.

Definition 27 (weighted Jensen-Shannon kernels) *The kernel $\widetilde{k}_{WJS} : (M_+^H(\mathcal{X}))^2 \to \mathbb{R}$ is defined as $\widetilde{k}_{WJS} \triangleq \widetilde{k}_1$, i.e.,*

$$\begin{aligned} \widetilde{k}_{WJS}(\mu_1, \mu_2) &= \widetilde{k}_{WJS}(\omega_1 p_1, \omega_2 p_2) \\ &= (H(\pi) - J^\pi(p_1, p_2))\,(\omega_1 + \omega_2), \end{aligned}$$

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 $k_{WJS} : \left(M_+^H(\mathcal{X}) \setminus \{0\}\right)^2 \to \mathbb{R}$ is simply $k_{WJS} \triangleq k_1$, i.e.,

$$k_{WJS}(\mu_1, \mu_2) = k_{WJS}(\omega_1 p_1, \omega_2 p_2) = H(\pi) - J^\pi(p_1, p_2).$$

Corollary 28 *The weighted Jensen-Shannon kernels $\widetilde{k}_{WJS}$ and k_{WJS} are pd.*

Proof: Invoke Proposition 26 with $q = 1$. ■

The following family of *weighted exponentiated JS kernels*, generalize the so-called *exponentiated JS* kernel, that has been used, and shown to be pd, by Cuturi and Vert [2005].

Definition 29 (Exponentiated JS kernel) *The kernel $k_{EJS} : M_+^1(\mathcal{X}) \times M_+^1(\mathcal{X}) \to \mathbb{R}$ is defined, for $t > 0$, as*

$$k_{EJS}(p_1, p_2) = \exp\left[-t\, JS\,(p_1, p_2)\right], \tag{75}$$

Definition 30 (Weighted exponentiated JS kernels) *The kernel $k_{WEJS} : M_+^H(\mathcal{X}) \times M_+^H(\mathcal{X}) \to \mathbb{R}$ is defined, for $t > 0$, as*

$$\begin{aligned} k_{WEJS}(\mu_1, \mu_2) &= \exp[t\, k_{WJS}(\mu_1, \mu_2)] \\ &= \exp(t\, H(\pi)) \exp\left[-t J^\pi(p_1, p_2)\right]. \end{aligned} \tag{76}$$

Corollary 31 *The kernels k_{WEJS} are pd. In particular, k_{EJS} is pd.*

Proof: Results from Proposition 18 and Corollary 28. Notice that although k_{WEJS} is pd, none of its two exponential factors in (76) is pd. ■

We now keep $q \in [0, 2]$ but consider the weighted JT kernel family restricted to normalized measures, $k_q|_{(M_+^1(\mathcal{X}))^2}$. This corresponds to setting uniform weights ($\omega_1 = \omega_2 = 1/2$); note that in this case $\widetilde{k}_q$ and k_q collapse into the same kernel,

$$\widetilde{k}_q(p_1, p_2) = k_q(p_1, p_2) = \ln_q(2) - T_q(p_1, p_2). \tag{77}$$

Prop. 26 tells us that these kernels are pd for $q \in [0, 2]$. Remarkably, we recover three well-known particular cases for $q \in \{0, 1, 2\}$. We start by the Jensen-Shannon kernel, introduced and shown to be pd by Hein et al. [2004a]; it is a particular case of a weighted Jensen-Shannon kernel in Definition 27.

Definition 32 (Jensen-Shannon kernel) *The kernel $k_{JS} : M_+^1(\mathcal{X}) \times M_+^1(\mathcal{X}) \to \mathbb{R}$ is defined as*

$$k_{JS}(p_1, p_2) = \ln 2 - JS(p_1, p_2).$$

Corollary 33 *The kernel k_{JS} is pd.*

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

Finally, we study two other particular cases of the family of Tsallis kernels: the Boolean and linear kernels.

Definition 34 (Boolean kernel) *Let the kernel $k_{Bool} : M_+^{S_0,1}(\mathcal{X}) \times M_+^{S_0,1}(\mathcal{X}) \to \mathbb{R}$ be defined as $k_{Bool} = k_0$, i.e.,*

$$k_{Bool}(p_1, p_2) = \nu\left(\operatorname{supp}(p_1) \cap \operatorname{supp}(p_2)\right), \tag{78}$$

i.e., $k_{Bool}(p_1, p_2)$ equals the measure of the intersection of the supports (cf. the result (56)*). In particular, if $\mathcal{X}$ is finite and ν is the counting measure, the above may be written as*

$$k_{Bool}(p_1, p_2) = \|p_1 \odot p_2\|_0. \tag{79}$$

Definition 35 (Linear kernel) *Let the kernel $k_{lin} : M_+^{S_2,1}(\mathcal{X}) \times M_+^{S_2,1}(\mathcal{X}) \to \mathbb{R}$ be defined as*

$$k_{lin}(p_1, p_2) = \frac{1}{2}\,\langle p_1, p_2 \rangle. \tag{80}$$

Corollary 36 *The kernels k_{Bool} and k_{lin} are pd.*

Proof: Invoke Proposition 26 with $q = 0$ and $q = 2$. Notice that, for $q = 2$, we just recover the well-known property of the inner product kernel [Schölkopf and Smola, 2002], which is equal to k_{lin} up to a scalar. ■

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

One of the key features of our generalization is that the kernels are defined on unnormalized measures, with arbitrary mass. This is relevant, for example, in applications of kernels on empirical measures (*e.g.*, word counts, pixel intensity histograms); instead of the usual step of normalization [Hein et al., 2004a], we may leave these empirical measures unnormalized, thus allowing objects of different size (*e.g.*, total number of words in a document, total number of image pixels) to be weighted differently. Another possibility opened by our generalization 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.

7.4 Other kernels based on Jensen differences and q-differences

It is worth to note that the Jensen-Rényi and the Jensen-Tsallis divergences also yield positive definite kernels, albeit there are not any obvious "weighted generalizations" like the ones presented above for the Tsallis kernels.

Proposition 37 (Jensen-Rényi and Jensen-Tsallis kernels) *For any $q \in [0, 2]$, the kernel*

$$(p_1, p_2) \mapsto S_q\left(\frac{p_1 + p_2}{2}\right)$$

and the (unweighted) Jensen-Tsallis divergence J_{S_q} (37) *are nd kernels on $M_+^1(\mathcal{X}) \times M_+^1(\mathcal{X})$.*

Also, for any $q \in [0, 1]$, the kernel

$$(p_1, p_2) \mapsto R_q\left(\frac{p_1 + p_2}{2}\right)$$

and the (unweighted) Jensen-Rényi divergence J_{R_q} (34) *are nd kernels on $M_+^1(\mathcal{X}) \times M_+^1(\mathcal{X})$.*

Proof: The fact that $(p_1, p_2) \mapsto S_q(\frac{p_1+p_2}{2})$ is nd results from the embedding $x \mapsto x/2$ and Prop. 23. Since $(p_1, p_2) \mapsto \frac{S_q(p_1)+S_q(p_2)}{2}$ is trivially nd, we have that J_{S_q} is a sum of nd functions, which turns it nd. To prove the negative definiteness of the kernel $(p_1, p_2) \mapsto R_q\left(\frac{p_1+p_2}{2}\right)$, notice first that the kernel $(x, y) \mapsto (x + y)/2$ is clearly nd. From Lemma 20 and integrating, we have that $(p_1, p_2) \mapsto \int \left(\frac{p_1+p_2}{2}\right)^q$ is nd for $q \in [0, 1]$. From the same lemma we have that $(p_1, p_2) \mapsto \ln\left(t + \int \left(\frac{p_1+p_2}{2}\right)^q\right)$ is nd for any $t > 0$. Since $\int \left(\frac{p_1+p_2}{2}\right)^q > 0$, the nonnegativity of $(p_1, p_2) \mapsto R_q\left(\frac{p_1+p_2}{2}\right)$ follows by taking the limit $t \to 0$. By the same argument as above, we conclude that J_{R_q} is nd. ■

As a consequence, we have from Lemma 18 that the following kernels are pd for any $t > 0$:

$$\begin{aligned} \tilde{k}_{\text{EJR}}(p_1, p_2) &= \exp\left(-tR_q\left(\frac{p_1+p_2}{2}\right)\right) = \\ &= \left(\int\left(\frac{p_1+p_2}{2}\right)^q\right)^{-\frac{t}{1-q}}, \end{aligned} \tag{81}$$

and its "normalized" counterpart,

$$\begin{aligned} k_{\text{EJR}}(p_1, p_2) &= \exp(-tJ_{R_q}(p_1, p_2)) = \\ &= \left(\frac{\int\left(\frac{p_1+p_2}{2}\right)^q}{\sqrt{\int p_1^q \int p_2^q}}\right)^{-\frac{t}{1-q}}. \end{aligned} \tag{82}$$

Although we could have derived its positive definiteness without ever referring the Rényi entropy, the latter has in fact a suggestive interpretation: it corresponds to an exponentiation of the Jensen-Rényi divergence; it generalizes the case $q = 1$ which corresponds to the exponentiated Jensen-Shannon kernel.

Finally, we point a relationship between the Jensen-Tsallis divergences and a family of difference kernels introduced by Fuglede [2005],

$$\psi_{\alpha,\beta}(x, y) = \left(\frac{x^\alpha + y^\alpha}{2}\right)^{1/\alpha} - \left(\frac{x^\beta + y^\beta}{2}\right)^{1/\beta}, \tag{83}$$

Fuglede [2005] derived the negative definiteness of the above family of kernels provided $1 \leq \alpha \leq \infty$ and $1/2 \leq \beta \leq \alpha$; he went further by providing representations for these kernels. Hein et al. [2004a] used the fact that the integration $\int \psi_{\alpha,\beta}(x(t), y(t))d\tau(t)$ is also nd to derive a family of pd kernels for probability measures that included the Jensen-Shannon kernel.

We start by noting the following property of the extended Tsallis entropy, that is very easy to establish:

$$S_q(\mu) = q^{-1}S_{1/q}(\mu^q) \tag{84}$$

As a consequence, we have that

$$J_{S_q}(y_1, y_2) = S_q\left(\frac{y_1+y_2}{2}\right) - \left(\frac{S_q(y_1)+S_q(y_2)}{2}\right) = \tag{85}$$

$$= r\left[S_r\left(\left(\frac{x_1^r+x_2^r}{2}\right)^{1/r}\right) - \frac{S_r(x_1)+S_r(x_2)}{2}\right] \triangleq \tag{86}$$

$$\triangleq r\tilde{J}_{S_r}(x_1, x_2) \tag{87}$$

where we made the substitutions $r \triangleq q^{-1}$, $x_1 \triangleq y_1^q$ and $x_2 \triangleq y_2^q$, and introduced

$$\begin{aligned} \tilde{J}_{S_r}(x_1, x_2) &= S_r\left(\left(\frac{x_1^r+x_2^r}{2}\right)^{1/r}\right) - \frac{S_r(x_1)+S_r(x_2)}{2} = \\ &= (r-1)^{-1}\int\left[\left(\frac{x_1^r+x_2^r}{2}\right)^{1/r} - \frac{x_1+x_2}{2}\right]. \end{aligned} \tag{88}$$

Since J_{S_q} is nd for $q \in [0, 2]$, we have that $\tilde{J}_{S_r}$ is nd for $r \in [1/2, \infty]$.

Notice that while J_{S_q} may be interpreted as "the difference between the Tsallis q-entropy of the mean and the mean of the Tsallis q-entropies", $\tilde{J}_{S_q}$ may be interpreted as "the difference between the Tsallis q-entropy of the q-power mean and the mean of the Tsallis q-entropies".

From (88) we have that

$$\int \psi_{\alpha,\beta}(x, y) = (\alpha - 1)\tilde{J}_{S_\alpha}(x, y) - (\beta - 1)\tilde{J}_{S_\beta}(x, y), \tag{89}$$

so the family of probabilistic kernels studied in Hein et al. [2004a] can be written in terms of Jensen-Tsallis divergences.

7.5 The heat kernel approximation

The diffusion kernel for statistical manifolds, recently proposed by Lafferty and Lebanon [2005], is grounded in information geometry [Amari and Nagaoka, 2001]. It models the diffusion of "information" over a statistical manifold according to the heat equation. Since in the case of the multinomial manifold (the relative interior of Δ^n), the diffusion kernel has no closed form, the authors adopt the so-called "first-order parametrix expansion," which resembles the Gaussian kernel replacing the Euclidean distance by the geodesic distance that is induced when the manifold is endowed with a Riemannian structure given by the Fisher information (we refer to Lafferty and Lebanon [2005] for further details). The resulting heat kernel approximation is

$$k_{\text{heat}}(p_1, p_2) = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{4t}\, d_g^2(p_1, p_2)\right), \tag{90}$$

where $t > 0$ and $d_g(p_1, p_2) = 2 \arccos\left(\sum_i \sqrt{p_{1i} p_{2i}}\right)$. Whether k_{heat} is pd has been an open problem [Hein et al., 2004b, Zhang et al., 2005]. Let $\mathbb{S}^n_+$ be the positive orthant of the n-dimensional sphere, *i.e.*,

$$\mathbb{S}^n_+ = \left\{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid \sum_{i=1}^{n+1} x_i^2 = 1,\ \forall i\ x_i \geq 0\right\}.$$

The problem can be restated as follows: is there an isometric embedding from $\mathbb{S}^n_+$ to some Hilbert space? In this section we answer that question in the negative.

Proposition 38 *Let $n \geq 2$. For sufficiently large t, the kernel k_{heat} is* not *pd.*

Proof: From Prop. 18, k_{heat} is pd, for all $t > 0$, if and only if d_g^2 is nd. We provide a counterexample, using the following four points in Δ^2: $p_1 = (1, 0, 0)$, $p_2 = (0, 1, 0)$, $p_3 = (0, 0, 1)$ and $p_4 = (1/2, 1/2, 0)$. The squared distance matrix $[D_{ij}] = [d_g^2(p_i, p_j)]$ is

$$D = \frac{\pi^2}{4} \cdot \begin{bmatrix} 0 & 4 & 4 & 1 \\ 4 & 0 & 4 & 1 \\ 4 & 4 & 0 & 4 \\ 1 & 1 & 4 & 0 \end{bmatrix}. \tag{91}$$

Taking $c = (-4, -4, 1, 7)$ we have $c^T D c = 2\pi^2 > 0$, showing that D is not nd. Although p_1, p_2, p_3, p_4 lie on the boundary of Δ^2, continuity of d_g^2 implies that it is not nd on the relative interior of Δ^2. The case $n > 2$ follows easily, by appending zeros to the four vectors above. ■

8 Experiments

We illustrate the performance of the proposed nonextensive kernels, in comparison with common kernels, for SVM text classification. We performed experiments in two standard datasets: *Reuters-21578*[2] and *WebKB*.[3] Since our objective was to evaluate the kernels, we considered a simple binary classification task that tries to discriminate among the two largest categories of each dataset; this led us to the *earn-vs-acq* classification task for the first dataset, and *stud-vs-fac* (students' *vs.* faculty webpages) in the second dataset.

After the usual preprocessing steps of stemming and stop-word removal, we mapped text documents into probability distributions over words using the bag-of-words model and maximum likelihood estimation (which corresponds to normalizing term frequency using the ℓ_1-norm), which we denote by *tf*. We also used the *tf-idf* measure, which penalizes terms that occur in many documents. To weight the documents for the Tsallis kernels, we tried four strategies: uniform weighting, word counts, square root of the word counts, and one plus the logarithm of the word counts; however, for both tasks, uniform weighting revealed the best strategy, which may be due to the fact that documents in both collections are usually short and do not differ much in size.

As baselines, we used the linear kernel with ℓ_2 normalization, commonly used for this task, and the heat kernel approximation (90) [Lafferty and Lebanon, 2005], which is known to outperform the former, albeit not being guaranteed to be pd for an arbitrary choice of τ (see (90)), as shown above. This parameter and the SVM C parameter were tuned with cross-validation over the training set. The SVM-Light package (`http://svmlight.joachims.org/`) was used to solve the SVM quadratic optimization problem.

Figs. 2–3 summarize the results. We report the performance of the Tsallis kernels as a function of the entropic index. For comparison, we also plot the performance of an instance of a Tsallis kernel with q tuned through cross-validation. For the first task, this kernel and the two baselines exhibit similar performance for both the *tf* and the *tf-idf* representations; differences are not statiscally significant. In the second task, the Tsallis kernel outperformed the ℓ_2-normalized linear kernel for both representations, and the heat kernel for *tf-idf*; the differences are statistically significant (using the unpaired t test at the 0.05 level). Regarding the influence of the entropic index, we observe that in both tasks, the optimum value of q is usually higher for *tf-idf* than for *tf*.

The results on these two problems are representative of the typical relative performance of the kernels considered: in almost all tested cases, both the heat kernel and the Tsallis kernels (for a suitable value of q) outperform the ℓ_2-normalized linear kernel; the Tsallis kernels are competitive with the heat kernel.

9 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

[2] `www.daviddlewis.com/resources/testcollections`.

[3] `www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data`.

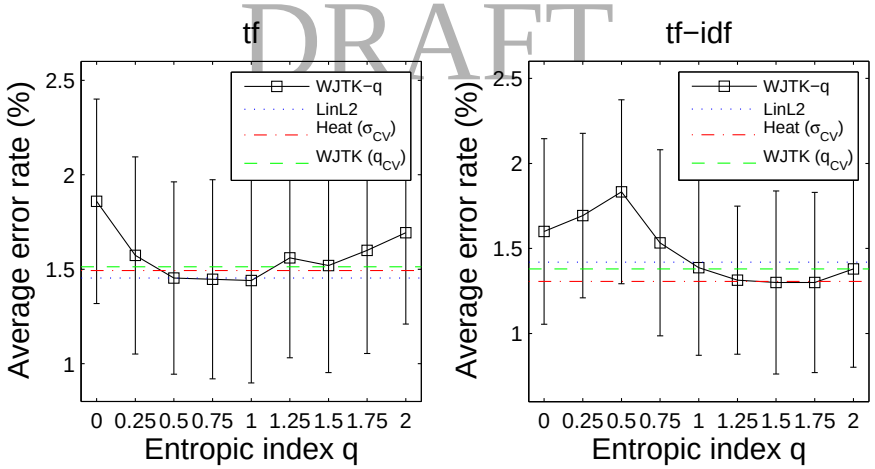

Figure 2: Results for *earn-vs-acq* using *tf* and *tf-idf* representations. The error bars represent ± 1 standard deviation on 30 runs. Training (resp. testing) with 200 (resp. 250) samples per class.

the key features of the new kernels is that they are defined on unnormalized 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 [Hein et al., 2004a], 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 Jensen-Shannon divergence; denormalization formulae for several entropies and divergences.

References

S. Abe. Foundations of nonextensive statistical mechanics. In *Chaos, Nonlinearity, Complexity*. Springer, 2006.

Shun-Ichi Amari and H. Nagaoka. *Methods of Information Geometry (Translations of Mathematical Monographs)*. Oxford University Press, 2001.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005. URL `http://www.jmlr.org/papers/v6/banerjee05b.html`.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, Berlin, 1984.

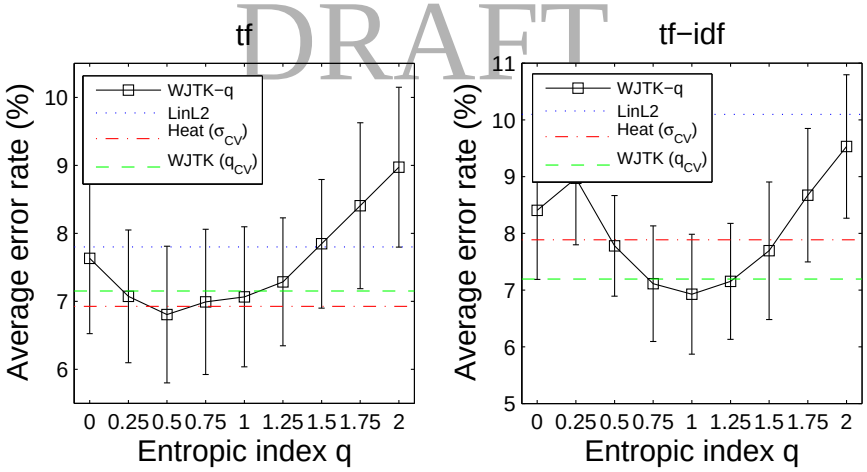

Figure 3: Results for *stud-vs-fac*.

Jacob Burbea and C. Radhakrishna Rao. On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, 28(3):489–495, 1982.

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

I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

Marco Cuturi and Jean-Philippe Vert. Semigroup kernels on finite sets. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pages 329–336. MIT Press, Cambridge, MA, 2005.

Marco Cuturi, Kenji Fukumizu, and Jean-Philippe Vert. Semigroup kernels on measures. *J. Mach. Learn. Res.*, 6:1169–1198, 2005. ISSN 1533-7928.

Zoltán Daróczy. Generalized information functions. *Information and Control*, 16(1):36–51, March 1970.

Dominik M. Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

Bent Fuglede. Spirals in Hilbert space. with an application in information theory. *Expositiones Mathematicae*, 25(1):23–46, 2005.

Shigeru Furuichi. Information theoretical properties of Tsallis entropies. *Journal of Mathematical Physics*, 47(2):023302, 2006. doi: 10.1063/1.2165744. URL `http://link.aip.org/link/?JMP/47/023302/1`.

M. Gell-Mann and C. Tsallis. *Nonextensive entropy: interdisciplinary applications*. Oxford University Press, 2004.

I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan J. Oliver, and H. E. Stanley. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E*, 65, 2002.

A. Hamza and H. Krim. Image registration and segmentation by maximizing the jensen-rényi divergence. In *Proc. of International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 147–163, Lisbon, Portugal, 2003. Springer.

A. Ben Hamza. A nonextensive information-theoretic measure for image edge detection. *Journal of Electronic Imaging*, 15-1:13011.1–13011.8, 2006.

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

Y. He, A. Hamza, and H. Krim. A generalized divergence measure for robust image registration. *IEEE Trans. on Signal Processing*, 51(5):1211–1220, 2003.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In Z. Ghahramani and R. Cowell, editors, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*. 2005.

M. Hein, T.N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in svm's. In *Proceedings of th 26th DAGM Symposium 3175, 270-277*, pages 270–277. 2004a.

Matthias Hein, T.N. Lal, and Olivier Bousquet. Hilbertian metrics on probability measures and their application in svm's. September 28 2004b.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

J. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30:175–193, 1906.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. Technical Report LS VIII-Report, Universität Dortmund, Dortmund, Germany, 1997.

D. Karakos, S. Khudanpur, J. Eisner, and C. Priebe. Iterative denoising using Jensen-Rényi divergences with an application to unsupervised document categorization. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 509–512, Baltimore, MD, 2007.

A. Ya. Khinchin. *Mathematical Foundations of Information Theory*. Dover, New York, 1957.

John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163, 2005.

P. W. Lamberti and A. P. Majtey. Non-logarithmic Jensen-Shannon divergence. *Physica A Statistical Mechanics and its Applications*, 329:81–90, November 2003. doi: 10.1016/S0378-4371(03)00566-1.

Yan Li, Xiaoping Fan, and Gang Li. Image segmentation based on Tsallis-entropy and Renyi-entropy and their comparison. In *IEEE International Conference on Industrial Informatics*, pages 943–948, 2006.

J. Lin and S. Wong. A new directed divergence measure and its characterization. *International Journal of General Systems*, 17:73–81, 1990.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Pedro J. Moreno, Purdy Ho, and Nuno Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *NIPS*. MIT Press, 2003. ISBN 0-262-20152-6. URL `http://books.nips.cc/papers/files/nips16/NIPS2003_SP02.pdf`.

A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, volume 1, pages 547–561, Berkely, 1961. Univ. Calif. Press.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002. URL `http://www.learning-with-kernels.org`.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948.

C.E. Shannon and W.W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Ill., 1949.

J. Steele. *The Cauchy-Schwarz Master Class*. Cambridge University Press, Cambridge, 2006.

Hiroki Suyari. Generalization of shannon-khinchin axioms to nonextensive systems and the uniqueness theorem for the nonextensive entropy. *IEEE Transactions on Information Theory*, 50(8):1783–1787, 2004.

Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

Dell Zhang, Xi Chen, and Wee Sun Lee. Text classification with kernels on the multinomial manifold. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 266–273, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-034-5. doi: http://doi.acm.org/10.1145/1076034.1076081.