Nonextensive Entropic Kernels

Andre F. T. Martins, Mario A. T. Figueiredo
Pedro M. Q. Aguiar, Noah Smith, Eric P. Xing

August 2008
CMU-ML-08-106

Carnegie Mellon

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Nonextensive Entropic Kernels	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Machine Learning Department,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Positive definite kernels on probability measures have been recently applied in classification problems involving text, images, and other types of structured data. Some of these kernels are related to classic information theoretic quantities, such as (Shannon?s) mutual information and the Jensen-Shannon (JS) divergence. Meanwhile, there have been recent advances in nonextensive generalizations of Shannon?s information theory. This paper bridges these two trends by introducing nonextensive information theoretic kernels on probability measures, based on new JS-type divergences. These new divergences result from extending the the two building blocks of the classical JS divergence: convexity and Shannon?s entropy. The classical notion of convexity is extended to the wider concept of q-convexity, for which we prove a Jensen q-inequality. Based on this inequality, we introduce Jensen-Tsallis (JT) q-differences, a nonextensive generalization of the JS divergence, and define a k-th order JT q-difference between stochastic processes. We then define a new family of nonextensive mutual information kernels, which allow weights to be assigned to their arguments, and which includes the Boolean, JS, and linear kernels as particular cases. Nonextensive string kernels are also defined that subsume the p-spectrum kernel. We illustrate the performance of these kernels on text categorization tasks, in which documents are modeled both as bags-of-words and as sequences of characters.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**50**	

Nonextensive Entropic Kernels

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

August 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 and by the European Commission under project SIMBAD. 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: Positive definite kernels, nonextensive information theory, Tsallis entropy, Jensen-Shannon divergence, string kernels.

Abstract

Positive definite kernels on probability measures have been recently applied in classification problems involving text, images, and other types of structured data. Some of these kernels are related to classic information theoretic quantities, such as (Shannon's) mutual information and the Jensen-Shannon (JS) divergence. Meanwhile, there have been recent advances in nonextensive generalizations of Shannon's information theory. This paper bridges these two trends by introducing nonextensive information theoretic kernels on probability measures, based on new JS-type divergences. These new divergences result from extending the the two building blocks of the classical JS divergence: convexity and Shannon's entropy. The classical notion of convexity is extended to the wider concept of q-convexity, for which we prove a Jensen q-inequality. Based on this inequality, we introduce Jensen-Tsallis (JT) q-differences, a nonextensive generalization of the JS divergence, and define a k-th order JT q-difference between stochastic processes. We then define a new family of nonextensive mutual information kernels, which allow weights to be assigned to their arguments, and which includes the Boolean, JS, and linear kernels as particular cases. Nonextensive string kernels are also defined that subsume the p-spectrum kernel. We illustrate the performance of these kernels on text categorization tasks, in which documents are modeled both as bags-of-words and as sequences of characters.

1 Introduction

In kernel-based machine learning [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004], there has been recent interest in defining kernels on probability distributions, to tackle several problems involving structured data [Desobry et al., 2007, Moreno et al., 2004, 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 assumed to have been generated, and defining a map from X from S (*e.g.*, through maximum likelihood estimation), a distribution in S may be 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 statistical manifolds, equipped with well-motivated non-Euclidean metrics [Lafferty and Lebanon, 2005], often outperform support vector machine (SVM) classifiers with linear kernels [Joachims, 2002]. Some of these kernels have a natural information theoretic interpretation, establishing a bridge between kernel methods and information theory [Cuturi et al., 2005, Hein and Bousquet, 2005].

The main goal of this paper is to widen that bridge; we do that by introducing a new wide class of kernels rooted in *nonextensive* information theory, which contains previous information theoretic kernels as particular elements. 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, Lindhard, 1974, Lindhard and Nielsen, 1971, Tsallis, 1988], which have raised great interest among physicists in modeling certain phenomena (*e.g.*, long-range interactions and multifractals) and 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* [Havrda and Charvát, 1967, Tsallis, 1988] form a parametric family of nonextensive entropies that includes the Shannon-Boltzmann-Gibbs entropy as a particular case. Some attempts have been made to construct a nonextensive generalization of information theory [Furuichi, 2006].

Convexity is a key concept underlying several fundamental results in information theory, *e.g.*, the non-negativity of the *Kullback-Leibler (KL) divergence* (also called *relative entropy*), namely via the many implications of Jensen's inequality [Cover and Thomas, 1991, Jensen, 1906]. Jensen's inequality also 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.

In this paper, we introduce new extensions of JS-type divergences by generalizing its two pillars: *convexity* and *Shannon's entropy*. These divergences are then used to define new information-theoretic kernels between probability distributions. 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 extrema, extending the work by Burbea and Rao [1982] and by Lin [1991] for the JS divergence.

- Definition of k-th order joint and conditional Jensen-Tsallis q-differences for families of stochastic processes, and derivation of a chain rule.

- We propose a broad family of (nonextensive information theoretic) 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 JS kernel proposed by Hein and Bousquet [2005].

- We define a family of (nonextensive information theoretic) positive definite kernels between stochastic processes, which subsume well-known string kernels like the p-spectrum kernel [Leslie et al., 2002].

- 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 by 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. In Section 6, we define the Jensen-Tsallis q-difference and derive some of its properties; in that section, we also define k-th order Jensen-Tsallis q-differences for families of stochastic processes. The new family of entropic kernels is introduced and characterized in Section 7, after a brief review of some key results concerning positive definite kernels; that section also presents a brief review of string kernels, and introduces nonextensive kernels between stochastic processes. Section 7 ends by proving that the parametrix approximation of the multinomial diffusion kernel is not positive definite. Section 8 reports experiments on text categorization using both a bag-of-words and a sequential representation of documents. Finally, Section 9 contains concluding remarks and discusses directions for future research.

Earlier and shorter versions of this work have appeared in Martins et al. [2008a] and Martins et al. [2008b].

2 Nonextensive entropies and Tsallis statistics

We start with a brief overview of nonextensive entropies. In what follows, $\mathbb{R}_+$ denotes the nonnegative reals, $\mathbb{R}_{++}$ denotes the strictly positive reals, 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 Shannon's 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 $(p_1, \ldots, p_n) \in \Delta^{n-1}$,

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

(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 $S_{1,\phi} = \lim_{q \to 1} S_{q,\phi}$, thus $S_{q,\phi}(p_1, \ldots, p_n)$, 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 $p_B \in$

Δ^{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 $q = 1$, Suyari's axioms 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

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

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 *Jensen-Tsallis q-difference*,

generalizing the one between mutual information and the JS divergence (shown, *e.g.*, by Grosse et al. [2002], and recalled below, in Subsection 4.2).

Furuichi [2006] also mentions an alternative generalization of Shannon's mutual information, defined as

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

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 of a "Tsallis mutual entropy" has also been used 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 Shannon-Boltzmann-Gibbs and Tsallis entropies to include unnormalized measures. Although, as shown below, these functionals are completely characterized by their restriction to the normalized probability distributions, the denormalization expressions will play an important role 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 Shannon-Boltzmann-Gibbs Entropy and the KL Divergence

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 Shannon-Boltzmann-Gibbs 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 (denoted $\mu_f \ll \mu_g$). 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 Shannon-Boltzmann-Gibbs 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 equation (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 KL divergence (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 $\mu_f \ll \mu_g$:*

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

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

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

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

$$H_\varphi(z) \triangleq -\sum_{i=1}^{n} \varphi(z_i), \tag{24}$$

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

$$J_\varphi^\pi(y_1, \ldots, y_m) \triangleq H_\varphi\left(\sum_{t=1}^{m} \pi_t\, y_t\right) - \sum_{t=1}^{m} \pi_t\, H_\varphi(y_t), \tag{25}$$

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:*

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

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 Shannon-Boltzmann-Gibbs 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 $|\mathcal{T}| = m$, $J_H^\pi(p_1, \ldots, p_m)$ 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:

- the Jensen difference of the Shannon entropy of p;
- 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_Ψ^π is the *Bregman information* [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

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

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[1] [Endres and Schindelin, 2003, Topsøe, 2000], which has motivated its use in kernel-based machine learning [Cuturi et al., 2005, Hein and Bousquet, 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 Shannon-Boltzmann-Gibbs entropy. It is worth noting 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 Shannon-Boltzmann-Gibbs 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 [Ben-Hamza and Krim, 2003, He et al., 2003, Karakos et al., 2007]. In Section 7, we show that the JR divergence is (like the JS divergence) an Hilbertian metric, which is relevant for its use in kernel-based machine learning.

[1] A metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is Hilbertian if there is some Hilbert space $\mathcal{H}$ and an isometry $f : \mathcal{X} \to \mathcal{H}$ such that $d^2(x, y) = \langle f(x) - f(y), f(x) - f(y) \rangle_{\mathcal{H}}$ holds for any $x, y \in \mathcal{X}$ [Hein and Bousquet, 2005].

4.4 The Jensen-Tsallis Divergence

Burbea and Rao [1982] have defined Jensen-type divergences of the form (27) based on the Tsallis q-entropy S_q, defined in (16). Like the Shannon-Boltzmann-Gibbs 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 [Ben-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*, which generalize Jensen differences. Later (in Section 6), use will these functions 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, used by Tsallis [1988] in nonextensive thermodynamics.

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 match the intuitive meaning of average/expectation (*e.g.*, $E_q[1] \neq 1$, in general). The q-expectation is a convenient concept in nonextensive information theory; *e.g.*, it yields a very compact form for the Tsallis entropy: $S_q(X) = -E_q[\ln_q p(X)]$.

5.2 q-Convexity

We now introduce the novel concept of q-convexity and use it to derive a set of results, namely the *Jensen q-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 $\lambda \in [0, 1]$,*

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

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

Of course, 1-convexity is the usual notion of convexity. The next proposition states the Jensen q-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: In the finite case, the proof can be carried out trivially, by induction, exactly as in the proof of the standard Jensen inequality [Cover and Thomas, 1991]. If f is continuous, it commutes with taking limits, thus

$$f\left(\sum_{i=1}^{\infty} \pi_i \, x_i\right) = f\left(\lim_{n\to\infty} \sum_{i=1}^{n} \pi_i \, x_i\right) = \lim_{n\to\infty} f\left(\sum_{i=1}^{n} \pi_i \, x_i\right) \leq \lim_{n\to\infty} \sum_{i=1}^{n} \pi_i^q f(x_i) = \sum_{i=1}^{\infty} \pi_i^q \, f(x_i).$$

■

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

$$f \text{ is } q\text{-convex} \quad \Rightarrow \quad f \text{ is } r\text{-convex} \tag{42}$$

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

Proof: Implication (42) results from

$$f(\lambda x + (1-\lambda)y) \;\leq\; \lambda^q f(x) + (1-\lambda)^q f(y) \;\leq\; \lambda^r f(x) + (1-\lambda)^r f(y),$$

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 (43) is similar. ■

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*

$$T^{\omega}_{q,\Psi}(\mu) \;\triangleq\; \Psi\left(\int_{\mathcal{T}} \omega(t)\;\mu_t\, d\tau(t)\right) - \int_{\mathcal{T}} \omega(t)^q\;\Psi(\mu_t)\, d\tau(t), \tag{44}$$

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 (44) *a* Jensen q-difference.

Burbea and Rao [1982] established necessary and sufficient conditions on φ for the Jensen difference of a φ-entropy (see (24)) 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 as $\Psi(z) \triangleq -\sum_{i=1}^n \varphi(z_i)$. 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.*

The proof is rather long, thus it is relegated to Appendix A.

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

$$T^{\pi}_{q,\Psi}(p) = \Psi\left(E[p]\right) - E_q[\Psi(p)], \tag{45}$$

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^{π}_{q,S_q} simply as T^{π}_q)

$$\begin{aligned} T^{\pi}_q(p) &= S_q\left(E[p]\right) - E_q[S_q(p)] \\ &= S_q(X) - \int_{\mathcal{T}} \pi(t)^q S_q(X|T=t) \\ &= S_q(X) - S_q(X|T) \\ &= I_q(X;T), \end{aligned} \tag{46}$$

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 (46) 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 $T^{\pi}_q(p_1, \ldots, p_m)$ 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^{π}_q 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 $T_q \triangleq T^{1/2,1/2}_q$,

$$T_q(p_1, p_2) = S_q\left(\frac{p_1 + p_2}{2}\right) - \frac{S_q(p_1) + S_q(p_2)}{2^q}. \tag{47}$$

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 is not a norm) of vector p, *i.e.*, its number of nonzero components. The Jensen-Tsallis 0-difference is thus

$$\begin{aligned} T_0(p_1, p_2) &= -1 + \nu\left(\mathrm{supp}\left(\frac{p_1 + p_2}{2}\right)\right) + 1 - \nu\left(\mathrm{supp}(p_1)\right) + 1 - \nu\left(\mathrm{supp}(p_2)\right) \\ &= 1 + \nu\left(\mathrm{supp}(p_1) \cup \mathrm{supp}(p_2)\right) - \nu\left(\mathrm{supp}(p_1)\right) - \nu\left(\mathrm{supp}(p_2)\right) \\ &= 1 - \nu\left(\mathrm{supp}(p_1) \cap \mathrm{supp}(p_2)\right); \end{aligned} \tag{48}$$

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

$$T_0(p_1, p_2) = 1 - \|p_1 \odot p_2\|_0, \tag{49}$$

  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,

$$T_1(p_1, p_2) = JS(p_1, p_2). \tag{50}$$

- 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

$$T_2(p_1, p_2) = \frac{1}{2} - \frac{1}{2}\, \langle p_1, p_2\rangle, \tag{51}$$

  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; *e.g.*, 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 of the optimization problem

$$\min_{p_1 \in \Delta^n} T_q(p_1, p_2), \tag{52}$$

is, in general, different from p_2, unless $q = 1$. Instead, this minimizer is closer to the uniform distribution if $q \in [0, 1)$, and closer to a degenerate distribution, for $q \in (1, 2]$ (see Fig. 1). 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, (52) becomes a linear program, and the solution is not p_2, but $p_1^* = \delta_j$, where $j = \arg\max_i p_{2i}$.

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. ■

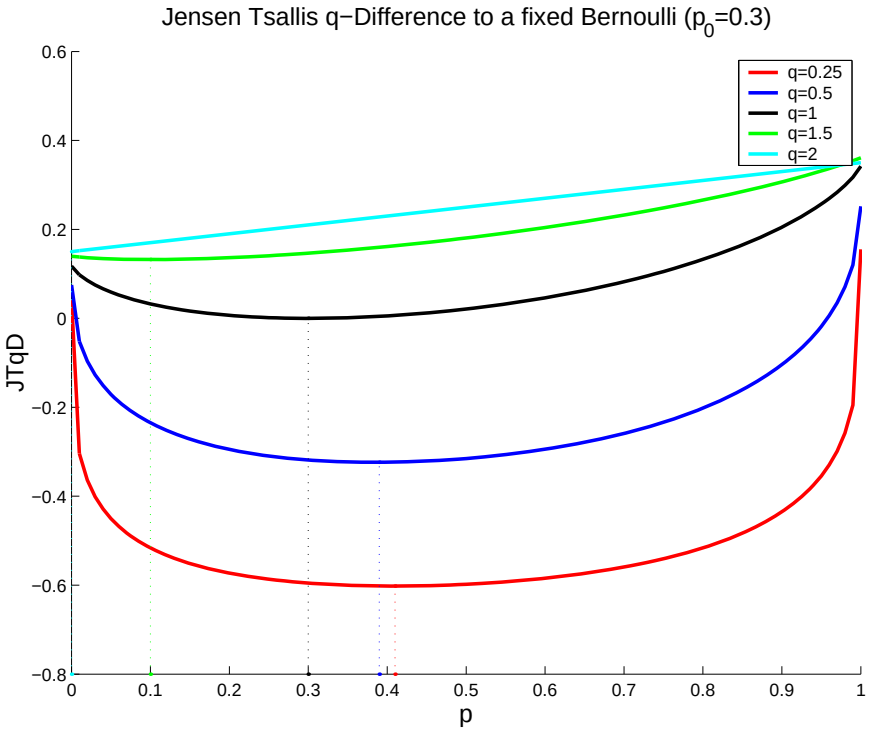

Figure 1: Jensen-Tsallis q-difference between two Bernoulli distributions, $p_1 = (0.3, 0.7)$ and $p_2 = (p, 1-p)$, for several values of the entropic index q. Observe that, for $q \in [0, 1)$, the minimizer of the JT q-difference approaches the uniform distribution $(0.5, 0.5)$ as q approaches 0; for $q \in (1, 2]$, this minimizer approaches the degenerate distribution, as $q \to 2$.

The next 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 $(\lambda_1, \ldots, \lambda_l) \in \Delta^{l-1}$,*

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

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 $x \geq 0$ and $q \geq 0$. To show the $(2-q)$-convexity

of $-1/\varphi_q''(x) = -(1/q)x^{2-q}$, for $x_t \geq 0$, and $q \in [0,1]$, we use a version of the power mean inequality [Steele, 2006],

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

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 $q \in [0,1]$. (Notice that this result is weaker than that of Proposition 13 below.)

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

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 $|\mathcal{X}| \geq |\mathcal{T}|$, this maximum occurs at the vertex corresponding to disjoint degenerate distributions, *i.e.*, such that $x_i \neq x_j$ if $i \neq j$. At this maximum,

$$\begin{aligned} T_q^\pi(\delta_{x_1}, \ldots, \delta_{x_m}) &= S_q\left(\sum_{t=1}^{m} \pi_t \delta_{x_t}\right) - \sum_{t=1}^{m} \pi_t S_q(\delta_{x_t}) \\ &= S_q\left(\sum_{t=1}^{m} \pi_t \delta_{x_t}\right) \qquad (53) \\ &= S_q(\pi) \qquad (54) \end{aligned}$$

where the equality in (53) results from $S_q(\delta_{x_t}) = 0$. Notice that this maximum may not be achieved if $|\mathcal{X}| < |\mathcal{T}|$. ■

The next proposition (proved in Appendix B) 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 $q \geq 0$,*

$$T_q^\pi(p_1, \ldots, p_m) \leq S_q(\pi), \qquad (55)$$

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 $q \geq 1$,

$$T_q^\pi(p_1, \ldots, p_m) \geq 0, \qquad (56)$$

and the minimum is attained in the purely 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 $|\mathcal{X}| = n$,

$$T_q^\pi(p_1, \ldots, p_m) \geq S_q(\pi)[1 - n^{1-q}]. \qquad (57)$$

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

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 $z = \sum_{i=2}^m \pi_i\, y_i$,

$$\begin{aligned}\frac{\partial^2 \psi}{\partial y_1^2} &= q\left[\pi_1^q\, y_1^{q-2} - \pi_1^2\,(\pi_1\, y_1 + z)^{q-2}\right] \\ &= q\,\pi_1^2\left[(\pi_1\, y_1)^{q-2} - (\pi_1\, y_1 + z)^{q-2}\right]. \end{aligned} \tag{58}$$

Since $\pi_1\, y_1 \leq (\pi_1\, y_1 + z) \leq 1$, the quantity in (58) is nonnegative for $q \in [0, 2]$ and non-positive for $q \geq 2$. ■

6.3 Joint and conditional JT q-differences and a chain rule

This subsection introduces joint and conditional JT q-differences, which will later be used as a contrast measure between stochastic processes. A chain rule is derived that relates conditional and joint JT q-differences.

Definition 15 *Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{T}$ be measured sets. Let $(p_t)_{t\in\mathcal{T}} \in [M_+^1(\mathcal{X} \times \mathcal{Y})]^{\mathcal{T}}$ be a family of measures in $M_+^1(\mathcal{X} \times \mathcal{Y})$ indexed by $\mathcal{T}$, and let p be a random probability distribution taking values in $\{p_t\}_{t\in\mathcal{T}}$ according to a distribution $\pi \in M_+^1(\mathcal{T})$. Consider also:*

- *for each $t \in \mathcal{T}$, the marginals $p_t(Y) \in M_+^1(\mathcal{Y})$,*
- *for each $t \in \mathcal{T}$ and $y \in \mathcal{Y}$, the conditionals $p_t(X|Y = y) \in M_+^1(\mathcal{X})$,*
- *the mixture $r(X, Y) \triangleq \int_{\mathcal{T}} \pi(t)\, p_t(X, Y) \in M_+^1(\mathcal{X} \times \mathcal{Y})$,*
- *the marginal $r(Y) \in M_+^1(\mathcal{Y})$,*
- *for each $y \in \mathcal{Y}$, the conditionals $r(X|Y = y) \in M_+^1(\mathcal{X})$.*

For notational convenience, we also append a subscript to p to emphasize its joint or conditional dependency of the random variables X and Y, i.e., *$p_{XY} \triangleq p$, and $p_{X|Y}$ denotes a random conditional probability distribution taking values in $\{p_t(.|Y)\}_{t\in\mathcal{T}}$ according to the distribution π.*

For $q \geq 0$, we call joint JT q-difference *of p_{XY} to*

$$T_q^\pi(p_{XY}) \triangleq T_q^\pi(p) = S_q(r) - E_{q,T\sim\pi(T)}[S_q(p_t)] \tag{59}$$

and conditional JT q-difference *of* $p_{X|Y}$ *to*

$$T_q^\pi(p_{X|Y}) \triangleq E_{q,Y\sim r(Y)}\left[S_q(r(.|Y=y))\right] - E_{q,T\sim\pi(T)}\left[E_{q,Y\sim p_t(Y)}\left[S_q(p_t(.|Y=y))\right]\right], \quad (60)$$

where we appended the random variables being used in each q*-expectation, for the sake of clarity.*

Note that the joint JT q-difference is just the usual JT q-difference of the joint random variable $X \times Y$, which equals (cf. (46))

$$T_q^\pi(p_{XY}) \;=\; S_q(X,Y) - S_q(X,Y|T) \;=\; I_q(X\times Y;T), \quad (61)$$

and the conditional JT q-difference is nothing but the usual JT q-difference with all entropies replaced by conditional entropies (conditioned on Y). Indeed, expression (60) can be rewritten as:

$$T_q^\pi(p_{X|Y}) \;=\; S_q(X|Y) - S_q(X|T,Y) \;=\; I_q(X;T|Y), \quad (62)$$

i.e., the conditional JT q-difference may also interpreted as a Tsallis mutual information, as in (46), but now *conditioned* on the random variable Y.

Note also that, for $q = 1$ (the extensive case), (60) may also be rewritten in terms of the conditional KL divergences,

$$\begin{aligned} J^\pi(p_{X|Y}) \triangleq T_1^\pi(p_{X|Y}) &= E_{Y\sim r(Y)}\left[H(r(.|Y=y))\right] - E_{T\sim\pi(T)}\left[E_{Y\sim p_t(Y)}\left[H(p_t(.|Y=y))\right]\right] \\ &= E_{T\sim\pi(T)}\left[E_{Y\sim r(Y)}\left[D(p_t(.|Y=y)\|r(.|Y=y))\right]\right]. \end{aligned} \quad (63)$$

Proposition 16 *The following chain rule holds:*

$$T_q^\pi(p_{XY}) = T_q^\pi(p_{X|Y}) + T_q^\pi(p_Y) \quad (64)$$

Proof: Writing the joint/conditional JT q-differences as joint/conditional mutual informations (61)-(62) and invoking the chain rule provided by (7), we have that

$$\begin{aligned} I(X;T|Y) + I(Y;T) &= H(X|T,Y) - H(X|Y) + H(Y|T) - H(Y) \\ &= H(X,Y|T) - H(X,Y), \end{aligned} \quad (65)$$

which is the joint JT q-difference associated with the random variable $X \times Y$. ■

Let us now turn our attention to the case where $Y = X^k$ for some $k \in \mathbb{N}$. In the following, the notation $(A_n)_{n\in\mathbb{N}}$ denotes a stationary ergodic process with values on some finite alphabet $\mathcal{A}$.

Definition 17 *Let* $\mathcal{X}$ *and* $\mathcal{T}$ *be measured sets, with* $\mathcal{X}$ *finite, and let* $\mathscr{F} = [(X_n)_{n\in\mathbb{N}}]^{\mathcal{T}}$ *be a family of stochastic processes (taking values on the alphabet* $\mathcal{X}$*) indexed by* $\mathcal{T}$*. The* k-th order JT q-difference *of* $\mathscr{F}$ *is defined, for* $k = 1, \ldots, n$*, as*

$$T_{q,k}^{joint,\pi}(\mathscr{F}) \triangleq T_q^\pi(p_{X^k}) \quad (66)$$

and the k-th order conditional JT q-difference *of* $\mathscr{F}$ *is defined, for* $k = 1, \ldots, n$*, as*

$$T_{q,k}^{cond,\pi}(\mathscr{F}) \triangleq T_q^\pi(p_{X|X^k}), \quad (67)$$

and, for $k = 0$*, as* $T_{q,0}^{cond,\pi}(\mathscr{F}) \triangleq T_{q,1}^{joint,\pi}(\mathscr{F}) = T_q^\pi(p_X)$.

Proposition 18 *The joint and conditional k-th order JT q-differences are related through:*

$$T_{q,k}^{joint,\pi}(\mathscr{F}) = \sum_{i=0}^{k-1} T_{q,i}^{cond,\pi}(\mathscr{F}) \tag{68}$$

Proof: Use Proposition 16 and induction. ■

6.4 Asymptotic Analysis in the Extensive Case

We now focus on the extensive case ($q = 1$) for a brief asymptotic analysis of the k-th order joint and conditional JT 1-differences (or *conditional Jensen-Shannon divergences*) when k goes to infinity.

The conditional Jensen-Shannon divergence was introduced by El-Yaniv et al. [1998] to address the *two-sample problem* for strings emitted by Markovian sources. Given two strings s and t, the goal is to decide whether they were emitted by the same source or by different sources. Under some fair assumptions, the most likely k-th order Markovian joint source of s and t is governed by a distribution $\hat{r}$ given by

$$\hat{r} = \arg\min_{r} \lambda D(\hat{p}_s \| r) + (1-\lambda) D(\hat{p}_t \| r). \tag{69}$$

where $D(.\|.)$ are conditional KL divergences, $\hat{p}_s$ and $\hat{p}_t$ are the empirical $(k-1)$-th order conditionals associated with s and t, respectively, and $\lambda = |s|/(|s|+|t|)$ is the length ratio. The solution of the optimization problem is

$$\hat{r}(a|c) = \frac{\lambda\, \hat{p}_s(c)}{\lambda\, \hat{p}_s(c) + (1-\lambda)\, \hat{p}_t(c)}\, \hat{p}_s(a|c) + \frac{(1-\lambda)\, \hat{p}_t(c)}{\lambda\, \hat{p}_s(c) + (1-\lambda)\, \hat{p}_t(c)}\, \hat{p}_t(a|c), \tag{70}$$

where $a \in \mathcal{A}$ is a symbol and $c \in \mathcal{A}^{k-1}$ is a context; this can be rewritten as $\hat{r}(a,c) = \lambda \hat{p}_s(a,c) + (1-\lambda)\hat{p}_t(a,c)$; *i.e.*, the optimum in (69) is a mixture of $\hat{p}_s$ and $\hat{p}_t$ weighted by the string lengths. Notice that, at the minimum, we have

$$D(\hat{p}_s \| \hat{r}) + (1-\lambda) D(\hat{p}_t \| \hat{r}) = JS_k^{\text{cond},(\lambda,1-\lambda)}(\hat{p}_s, \hat{p}_t). \tag{71}$$

It is tempting to investigate the asymptotic behavior of the conditional and joint JS divergences, when $k \to \infty$; however, unlike other asymptotic information theoretic quantities, like the entropy rate or the cross entropy rate, this behavior fails to characterize the sources s and t. Intuitively, this is justified by the fact that observing more and more symbols drawn from the mixture of the two sources rapidly decreases the uncertainty about which source generated the sample. Indeed, from the asymptotic equipartition property of stationary ergodic sources [Cover and Thomas, 1991], we have that $\lim_{k\to\infty} \frac{1}{k} H(p_{X_k}) = \lim_{k\to\infty} H(p_{X|X_k})$, which implies

$$\lim_{k\to\infty} JS_k^{\text{cond},\pi} \;=\; \lim_{k\to\infty} \frac{1}{k} JS_k^{\text{joint},\pi} \;\leq\; \lim_{k\to\infty} \frac{1}{k} H(\pi) \;=\; 0, \tag{72}$$

where we used the fact that the JS divergence is upper-bounded by the entropy of the mixture $H(\boldsymbol{\pi})$ (see Proposition 13). Since the conditional JS divergence must be non-negative, we therefore conclude that $\lim_{k\to\infty} JS_k^{\text{cond},\pi} = 0$, pointwise.

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; these kernels allow setting a weight for each argument, thus will be called *weighted Jensen-Tsallis kernels*. We also introduce kernels related to the JR divergence (Subsection 4.3) and the JT divergence (Subsection 4.4), and establish a connection between the Tsallis kernels and a family of kernels investigated by Hein et al. [2004] and Fuglede [2005], placing those kernels under a new information-theoretic light. After that, we give a brief overview of string kernels, and using the results of Subsection 6.3, we devise k-th order Jensen-Tsallis kernels between stochastic processes that subsume the well-known p-spectrum kernel of Leslie et al. [2002]. Finally, we show that the parametrix approximation of the multinomial diffusion kernel, proposed by Lafferty and Lebanon [2005], is not positive definite in general.

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 19 *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*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \varphi(x_i, x_j) \geq 0 \tag{73}$$

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

Definition 20 *Let $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric. ψ is called a* negative definite *(nd) kernel if and only if*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \psi(x_i, x_j) \leq 0 \tag{74}$$

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.*

The sets of pd and nd kernels are both 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 propositions and lemmas, are shown in Berg et al. [1984].

Proposition 21 *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*

$$\varphi(x, y) = \psi(x, x_0) + \psi(y, x_0) - \psi(x, y) - \psi(x_0, x_0). \tag{75}$$

Then, φ is pd if and only if ψ is nd.

Proposition 22 *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 23 *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 24 *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 25 *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 [Berg et al., 1984] has been used in a machine learning context by Cuturi and Vert [2005].

Definition 26 *Let $(\mathcal{X}, +)$ be a semigroup.*[2] *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 Shannon-Boltzmann-Gibbs 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 27 *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 24 guarantees that ι^q is also nd. For the second case, where $q \in (1, 2]$,

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

$\varphi_q(y)$ equals a positive constant times $\iota - \iota^q$. It only remains to show that $-\iota^q$ is nd for $q \in (1, 2]$: Lemma 25 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. ■

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

Lemma 28 *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

$$k_q(x, y) = (x + y)^{-q} = \lim_{t \to 0^+} \left[t + (x + y)^q\right]^{-1}, \tag{76}$$

which is always well defined because $x + y > 0$, combined with the following facts: from Lemma 24, since $(x, y) \mapsto x + y$ is nd and nonnegative, $(x, y) \mapsto (x + y)^q$ is nd; from Proposition 23, $(x, y) \mapsto \left[t + (x + y)^q\right]^{-1}$ 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 29 (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*

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

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 $k_q : \left(M_+^{S_q}(\mathcal{X}) \setminus \{0\}\right)^2 \to \mathbb{R}$ 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 (46), 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 30 *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 Proposition 21 with $\psi = S_q$ (which is nd by Proposition 27), $x = \mu_1$, $y = \mu_2$, 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 (Proposition 28) $(\omega_1 + \omega_2)^{-q}$ 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 31 (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 32 *The weighted Jensen-Shannon kernels $\widetilde{k}_{WJS}$ and k_{WJS} are pd.*

Proof: Invoke Proposition 30 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 33 (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*

$$k_{EJS}(p_1, p_2) = \exp\left[-t\, JS\,(p_1, p_2)\right]. \tag{77}$$

Definition 34 (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*

$$\begin{aligned} k_{WEJS}(\mu_1, \mu_2) &= \exp[t\, k_{WJS}(\mu_1, \mu_2)] \\ &= \exp(t\, H(\pi)) \exp\left[-tJ^\pi(p_1, p_2)\right]. \end{aligned} \tag{78}$$

Corollary 35 *The kernels k_{WEJS} are pd. In particular, k_{EJS} is pd.*

Proof: Results from Proposition 22 and Corollary 32. Notice that although k_{WEJS} is pd, none of its two exponential factors in (78) 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,

$$\widetilde{k}_q(p_1, p_2) = k_q(p_1, p_2) = \ln_q(2) - T_q(p_1, p_2). \tag{79}$$

Proposition 30 guarantees 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. [2004]; it is a particular case of a weighted Jensen-Shannon kernel in Definition 31.

Definition 36 (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 37 *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 38 (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.,

$$k_{Bool}(p_1, p_2) = \nu\left(\operatorname{supp}(p_1) \cap \operatorname{supp}(p_2)\right), \tag{80}$$

i.e., *$k_{Bool}(p_1, p_2)$ equals the measure of the intersection of the supports (cf. the result* (48)*). In particular, if $\mathcal{X}$ is finite and ν is the counting measure, the above may be written as*

$$k_{Bool}(p_1, p_2) = \|p_1 \odot p_2\|_0. \tag{81}$$

Definition 39 (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*

$$k_{lin}(p_1, p_2) = \frac{1}{2} \langle p_1, p_2 \rangle. \tag{82}$$

Corollary 40 *The kernels* k_{Bool} *and* k_{lin} *are pd.*

Proof: Invoke Proposition 30 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., 2004], 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 noting 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 41 (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 Proposition 27. 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 24 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 22 that the following kernels are pd for any $t > 0$:

$$\tilde{k}_{\text{EJR}}(p_1, p_2) = \exp\left(-tR_q\left(\frac{p_1 + p_2}{2}\right)\right) = \left(\int \left(\frac{p_1 + p_2}{2}\right)^q\right)^{-\frac{t}{1-q}}, \tag{83}$$

and its "normalized" counterpart,

$$k_{\text{EJR}}(p_1, p_2) = \exp(-tJ_{R_q}(p_1, p_2)) = \left(\frac{\int \left(\frac{p_1+p_2}{2}\right)^q}{\sqrt{\int p_1^q \int p_2^q}}\right)^{-\frac{t}{1-q}}. \tag{84}$$

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 out a relationship between the Jensen-Tsallis divergences (Subsection 4.4) and a family of difference kernels introduced by Fuglede [2005],

$$\psi_{\alpha,\beta}(x, y) = \left(\frac{x^\alpha + y^\alpha}{2}\right)^{1/\alpha} - \left(\frac{x^\beta + y^\beta}{2}\right)^{1/\beta}. \tag{85}$$

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. [2004] 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:

$$S_q(\mu) = q^{-1}S_{1/q}(\mu^q) \tag{86}$$

As a consequence, we have that

$$\begin{aligned} J_{S_q}(y_1, y_2) &= S_q\left(\frac{y_1 + y_2}{2}\right) - \left(\frac{S_q(y_1) + S_q(y_2)}{2}\right) && (87)\\ &= r\left[S_r\left(\left(\frac{x_1^r + x_2^r}{2}\right)^{1/r}\right) - \frac{S_r(x_1) + S_r(x_2)}{2}\right] && (88)\\ &\triangleq r\tilde{J}_{S_r}(x_1, x_2) && (89) \end{aligned}$$

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

$$\begin{aligned}\tilde{J}_{S_r}(x_1, x_2) &= S_r\left(\left(\frac{x_1^r + x_2^r}{2}\right)^{1/r}\right) - \frac{S_r(x_1) + S_r(x_2)}{2} \\ &= (r-1)^{-1}\int\left[\left(\frac{x_1^r + x_2^r}{2}\right)^{1/r} - \frac{x_1 + x_2}{2}\right]. \end{aligned} \tag{90}$$

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 (90) we have that

$$\int \psi_{\alpha,\beta}(x, y) = (\alpha - 1)\tilde{J}_{S_\alpha}(x, y) - (\beta - 1)\tilde{J}_{S_\beta}(x, y), \tag{91}$$

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

7.5 k-th order Jensen-Tsallis string kernels

This subsection introduces a new class of string kernels inspired by the k-th order JT q-difference introduced in Subsection 6.3. Although we refer to them as "string kernels," they are more generally kernels between stochastic processes.

Several string kernels (*i.e.*, kernels operating on the space of strings) have been proposed in the literature [Haussler, 1999, Lodhi et al., 2002, Leslie et al., 2002, Vishwanathan and Smola, 2003, Shawe-Taylor and Cristianini, 2004]. These are kernels defined on $\mathcal{A}^* \times \mathcal{A}^*$, where $\mathcal{A}^*$ is the Kleene closure of a finite alphabet $\mathcal{A}$ (*i.e.*, the set of all finite strings formed by characters in $\mathcal{A}$ together with the empty string ϵ.) The *p-spectrum kernel* [Leslie et al., 2002] is associated with a feature space indexed by $\mathcal{A}^p$ (the set of length-p strings). The feature representation of a string s, $\Phi^p(s) \triangleq (\phi_u^p(s))_{u\in\mathcal{A}^p}$, counts the number of times each $u \in \mathcal{A}^p$ occurs as a substring of s,

$$\phi_u^p(s) = |\{(v_1, v_2) : s = v_1 u v_2\}|. \tag{92}$$

The p-spectrum kernel is then defined as the standard inner product in $\mathbb{R}^{|\mathcal{A}|^p}$

$$k_{\text{SK}}^p(s, t) = \left\langle \Phi^p(s), \Phi^p(t) \right\rangle. \tag{93}$$

A more general kernel is the *weighted all-substrings kernel* [Vishwanathan and Smola, 2003], which takes into account the contribution of all the substrings weighted by their length. This kernel can be viewed as a conic combination of p-spectrum kernels and can be written as

$$k_{\text{WASK}}(s, t) = \sum_{p=1}^{\infty} \alpha_p \, k_{\text{SK}}^p(s, t), \tag{94}$$

where α_p is often chosen to decay exponentially with p and truncated; for example, $\alpha_p = \lambda^p$, if $p_{\min} \leq p \leq p_{\max}$, and $\alpha_p = 0$, otherwise, where $0 < \lambda < 1$ is the decaying factor.

Both k^p_{SK} and k_{WASK} are trivially positive definite, the former by construction and the latter because it is a conic combination of positive definite kernels. A remarkable fact is that both kernels may be computed in $O(|s| + |t|)$ time (*i.e.*, with cost that is linear in the length of the strings), as shown by Vishwanathan and Smola [2003], by using data structures such as suffix trees or suffix arrays [Gusfield, 1997]. Moreover, with s fixed, any kernel $k(s,t)$ may be computed in time $O(|t|)$, which is particularly useful for classification applications.

We will now see how Jensen-Tsallis kernels may be used as string kernels. In Subsection 6.3, we have introduced the concept of *joint* and *conditional* JT q-differences. We have seen that joint JT q-differences are just JT q-differences in a product space of the form $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$; for k-th order joint JT q-differences this product space is of the form $\mathcal{A}^k = \mathcal{A} \times \mathcal{A}^{k-1}$. Therefore, they still yield positive definite kernels as those introduced in Definition 29, where $\mathcal{X} = \mathcal{A}^k$. The next definition and proposition summarize these statements.

Definition 42 (k-th order weighted JT kernels) *Let $\mathscr{S}(\mathcal{A})$ be the set of stationary and ergodic stochastic processes that take values on the alphabet $\mathcal{A}$. For $k \in \mathbb{N}$ and $q \in [0,2]$, let the kernel $\widetilde{k}_{q,k} : (\mathbb{R}_+ \times \mathscr{S}(\mathcal{A}))^2 \to \mathbb{R}$ be defined as*

$$\begin{aligned} \widetilde{k}_{q,k}((\omega_1, s_1), (\omega_2, s_2)) &\triangleq \widetilde{k}_q(\omega_1 p_{s_1,k}, \omega_2 p_{s_2,k}) \\ &= \left(S_q(\pi) - T^{\text{joint},\pi}_{q,k}(s_1, s_2)\right)(\omega_1 + \omega_2)^q, \end{aligned} \tag{95}$$

where $p_{s_1,k}$ and $p_{s_2,k}$ are the k-th order joint probability functions associated with the stochastic sources s_1 and s_2, and $\pi = (\omega_1/(\omega_1+\omega_2), \omega_2/(\omega_1+\omega_2))$.

Let the kernel $k_{q,k} : (\mathbb{R}_{++} \times \mathscr{S}(\mathcal{A}))^2 \to \mathbb{R}$ be defined as

$$\begin{aligned} k_{q,k}((\omega_1, s_1), (\omega_2, s_2)) &\triangleq k_q(\omega_1 p_{s_1,k}, \omega_2 p_{s_2,k}) \\ &= \left(S_q(\pi) - T^{\text{joint},\pi}_{q,k}(s_1, s_2)\right), \end{aligned} \tag{96}$$

Proposition 43 *The kernel $\widetilde{k}_{q,k}$ is pd, for $q \in [0,2]$. The kernel $k_{q,k}$ is pd, for $q \in [0,1]$.*

Proof: Define the map $g : \mathbb{R}_+ \times \mathscr{S}(\mathcal{A}) \to \mathbb{R}_+ \times M_+^{1,S_q}(\mathcal{A}^k)$ as $(\omega, s) \mapsto g(\omega, s) = (\omega, p_{s,k})$. From Proposition 30, the kernel $\widetilde{k}_q(g(\omega_1, s_1), g(\omega_2, s_2))$ is pd and therefore so is $\widetilde{k}_{q,k}((\omega_1, s_1), (\omega_2, s_2))$; proceed analogously for $k_{q,k}$. ■

At this point, one might wonder whether the "k-th order conditional JT kernel" $\widetilde{k}^{\text{cond}}_{q,k}$ that would be obtained by replacing $T^{\text{joint},\pi}_{q,k}$ with $T^{\text{cond},\pi}_{q,k}$ in (95)-(96) is also pd. Formula (68) shows that such "conditional JT kernel" is a difference between two joint JT kernels, which is inconclusive. The following proposition shows that $\widetilde{k}^{\text{cond}}_{q,k}$ and $k^{\text{cond}}_{q,k}$ are not pd in general. The proof, which is in Appendix C, proceeds by building a counterexample.

Proposition 44 *Let $\widetilde{k}^{cond}_{q,k}$ be defined as $\widetilde{k}^{cond}_{q,k}(s_1, s_2) \triangleq \left(S_q(\pi) - T^{cond,\pi}_{q,k}(s_1, s_2)\right)(\omega_1 + \omega_2)^q$; and $k^{cond}_{q,k}$ be defined as $k^{cond}_{q,k}(s_1, s_2) \triangleq \left(S_q(\pi) - T^{cond,\pi}_{q,k}(s_1, s_2)\right)$. It holds that $\widetilde{k}^{cond}_{q,k}$ and $k^{cond}_{q,k}$ are not pd in general.*

Despite the negative result in Proposition 44, the chain rule in Proposition 18 still allows us to define pd kernels by combining conditional JT q-differences.

Proposition 45 *Let $(\beta_k)_{k\in\mathbb{N}}$ be a non-increasing infinitesimal sequence,* i.e. *satisfying*

$$\beta_0 \geq \beta_1 \geq \ldots \geq \beta_n \to 0 \tag{97}$$

Any kernel of the form

$$\sum_{k=0}^{\infty} \beta_k \, \widetilde{k}_{q,k}^{cond} \tag{98}$$

is pd for $q \in [0, 2]$; and any kernel of the form

$$\sum_{k=0}^{\infty} \beta_k \, k_{q,k}^{cond} \tag{99}$$

is pd for $q \in [0, 1]$, provided both series above converge pointwise.

Proof: From the chain rule, we have that (defining the 0-th order joint JT q-difference as $\widetilde{k}_{q,0} \triangleq 0$)

$$\sum_{k=0}^{\infty} \beta_k \, \widetilde{k}_{q,k}^{\text{cond}} \;=\; \sum_{k=0}^{\infty} \beta_k \, (\widetilde{k}_{q,k+1} - \widetilde{k}_{q,k}) \;=\; \lim_{n\to\infty} \sum_{k=1}^{n} \alpha_k \, \widetilde{k}_{q,k} + \beta_n \, \widetilde{k}_{q,n+1} \;=\; \sum_{k=1}^{\infty} \alpha_k \, \widetilde{k}_{q,k} \tag{100}$$

with $\alpha_k = \beta_{k-1} - \beta_k$ (the term $\lim \beta_n \widetilde{k}_{q,n+1}$ was dropped because $\beta_n \to 0$ and $\widetilde{k}_{q,n+1}$ is bounded). Since $(\beta_k)_{k\in\mathbb{N}}$ is non-increasing, we have that $(\alpha_k)_{k\in\mathbb{N}\setminus\{0\}}$ is non-negative, which makes (100) the pointwise limit of a conic combination of pd kernels, and therefore a pd kernel. The proof for $\sum_{k=0}^{\infty} \beta_k k_{q,k}^{\text{cond}}$ is analogous. ■

Notice that if we set $\beta_0 = \ldots = \beta_{k-1} = 1$ and $\beta_j = 0, \; \forall j \geq k$, in the above proposition, we recover the k-th order joint JT q-difference.

Finally, notice that, in the same way that the linear kernel is a special case of a JT kernel when $q = 2$ (see Cor. 40), the p-spectrum kernel (93) is a particular case of a p-th order joint JT kernel, and the weighted all substrings kernel (94) is a particular case of a combination of joint JT kernels in the form (98), both obtained when we set $q = 2$ and the weights ω_1 and ω_2 equal to the length of the strings. Therefore, we conclude that the JT string kernels introduced in this section subsume these two well-known string kernels.

7.6 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

$$k_{\text{heat}}(p_1, p_2) = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{4t}\, d_g^2(p_1, p_2)\right), \tag{101}$$

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., 2004, 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 46 *Let $n \geq 2$. For sufficiently large t, the kernel k_{heat} is* not *pd.*

Proof: From Proposition 22, 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

$$D = \frac{\pi^2}{4} \cdot \begin{bmatrix} 0 & 4 & 4 & 1 \\ 4 & 0 & 4 & 1 \\ 4 & 4 & 0 & 4 \\ 1 & 1 & 4 & 0 \end{bmatrix}. \tag{102}$$

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 information theoretic kernels, in comparison with common kernels, for SVM-based text classification. We performed experiments with two standard datasets: *Reuters-21578*[3] and *WebKB*.[4] 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. Two different frameworks were considered: modeling documents as bags-of-words, and modeling them as strings of characters. Therefore, both bags-of-words kernels and string kernels were employed for each task.

[3] Available at `www.daviddlewis.com/resources/testcollections`.

[4] Available at `www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data`.

8.1 Documents as bags-of-words

For the bags-of-words framework, 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; this corresponds to normalizing the *term frequencies* (tf) using the ℓ_1-norm, and is referred to as *tf* [Joachims, 2002, Manning and Schütze, 1999]. We also used the *tf-idf* (term frequencyï¿½inverse document frequency) representation, which penalizes terms that occur in many documents [Joachims, 2002, Manning and Schütze, 1999]. 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 [Joachims, 2002], and the heat kernel approximation (101) [Lafferty and Lebanon, 2005], which is known to outperform the former, albeit not being guaranteed to be pd for an arbitrary choice of t (see (101)), as shown above. This parameter and the SVM C parameter were tuned by cross-validation over the training set. The SVM-Light package (available at `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 q. For comparison, we also plot the performance of an instance of a Tsallis kernel with q tuned by 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 statistically 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.

8.2 Documents as strings

In the second set of experiments, each document is mapped into a probability distribution over character p-grams, using maximum likelihood estimation; we did experiments for $p = 3, 4, 5$. To weight the documents for the p-th order joint Jensen-Tsallis kernels, four strategies were attempted: uniform weighting, document lengths (in characters), square root of the document lengths, and one plus the logarithm of the document lengths. For the *earn-vs-acq* task, all strategies performed similarly, with a slight advantage for the square root and logarithm of the document lengths; for the *stud-vs-fac* task, uniform weighting revealed the best strategy. For simplicity, all experiments reported here use uniform weighting.

As baselines, we used the p-spectrum kernel (PSK, see (93)) for the values of p referred above, and the weighted all substrings kernel (WASK, see (94)) with decaying factor tuned to $\lambda = 0.75$

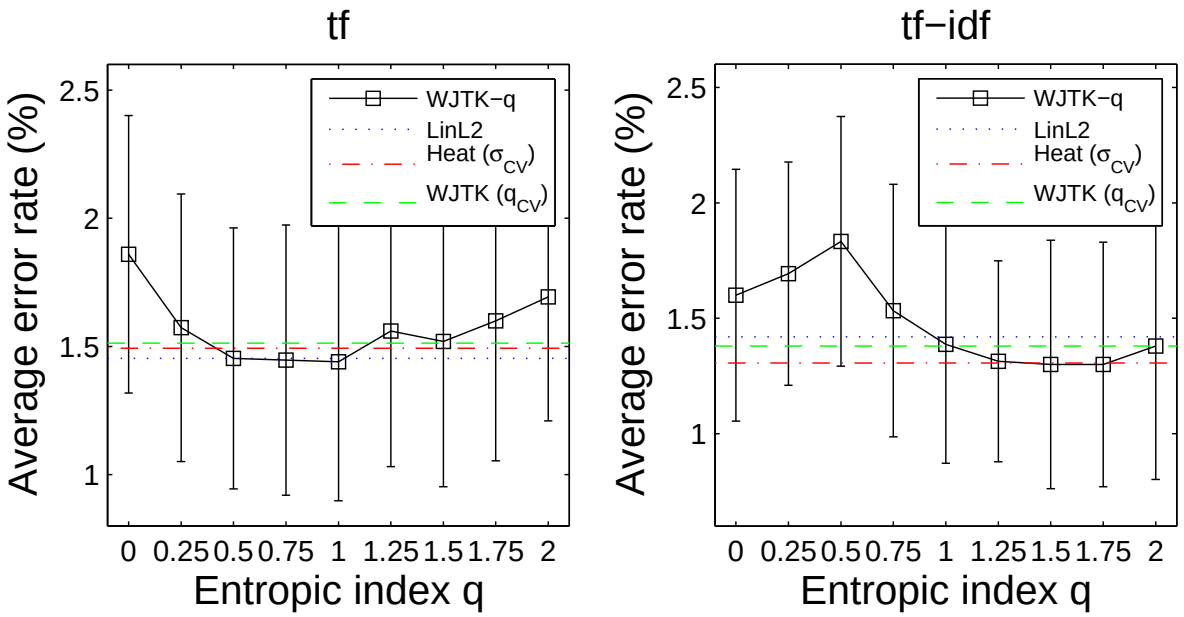

Figure 2: Results for *earn-vs-acq* using *tf* and *tf-idf* representations. The error bars represent ± 1 standard deviation on 30 runs. Training (resp. testing) with 200 (resp. 250) samples per class.

(which yielded the best results), with $p_{\min} = p$ set to the values above, and $p_{\max} = \infty$. The SVM C parameter was tuned by cross-validation over the training set.

Figs. 4–5 summarize the results. For the first task, the JT string kernel and the WASK outperformed the PSK (with statistical significance for $p = 3$), all kernels performed similarly for $p = 4$, and the JT string kernel outperformed the WASK for $p = 5$; all other differences are not statiscally significant. In the second task, the JT string kernel outperformed both the WASK and the PSK (and the WASK outperformed the PSK), with statistical significance for $p = 3, 4, 5$. Furthermore, by comparing Fig. 3 and Fig. 5, we also observe that the 5-th order JT string kernel remarkably outperforms all bags-of-words kernels for the *stud-vs-fac* task, even though it does not use or build any sort of language model at the word level.

9 Conclusions

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 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., 2004], 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. In addition,

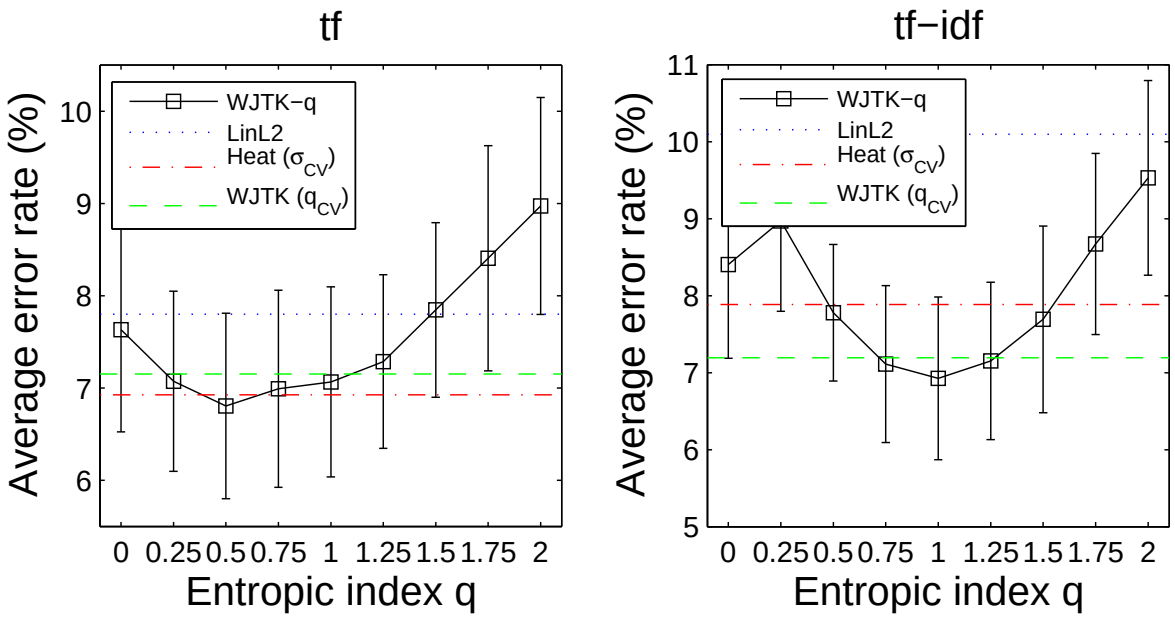

Figure 3: Results for *stud-vs-fac*.

we define positive definite kernels between stochastic processes that subsume well-known string kernels.

The new kernels, and the proofs of positive definiteness, rely on other main 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.

We have reported experiments in which these new kernels were used in support vector machines for text classification tasks. Although the reported experiments do not allow drawing strong conclusions, they show that the new kernels are competitive with the state-of-the-art, in some cases yielding a significant performance improvement.

A Proof of Proposition 9

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 $y = (y_1, \ldots, y_m)$, where $y_t = (y_{t1}, \ldots, y_{tn})$. 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.*,

$$T^{\pi}_{q,\Psi}(y_1, \ldots, y_m) = \sum_{t=1}^{m} \pi_t^q \, \varphi(y_t) - \varphi\left(\sum_{t=1}^{m} \pi_t y_t\right); \tag{103}$$

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

$$f(x) = T^{\pi}_{q,\Psi}(a_1 + b_1 x, \ldots, a_m + b_m x) \tag{104}$$

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 $c_t = a_t + b_t x$,

$$f''(x) = \sum_{t=1}^{m} \pi_t^q \, b_t^2 \, \varphi''(c_t) - \left(\sum_{t=1}^{m} \pi_t \, b_t\right)^2 \varphi''\left(\sum_{t=1}^{m} \pi_t \, c_t\right). \tag{105}$$

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 (105), 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

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

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:

$$0 \leq \left(\sum_{t=1}^{m} r_t^2\right)\left(\sum_{t=1}^{m} s_t^2\right) - \left(\sum_{t=1}^{m} r_t \, s_t\right)^2 \tag{106}$$

$$= \sum_{t=1}^{m} \frac{\pi_t^{2-q}}{\varphi''(c_t)} \sum_{t=1}^{m} b_t^2 \pi_i^q \varphi''(c_t) - \left(\sum_{t=1}^{m} b_t \pi_t\right)^2 \tag{107}$$

$$\leq \frac{1}{\varphi''\left(\sum_{t=1}^{m} \pi_t c_t\right)} \sum_{t=1}^{m} b_t^2 \pi_t^q \varphi''(c_t) - \left(\sum_{t=1}^{m} b_t \pi_t\right)^2 \tag{108}$$

$$= \frac{1}{\varphi''\left(\sum_{t=1}^{m} \pi_t c_t\right)} \cdot f''(t), \tag{109}$$

where: (106) is the Cauchy-Schwarz inequality; equality (107) results from the definitions of r_t and s_t and from the fact that $r_t s_t = b_t \pi_t$; inequality (108) states the $(2-q)$-convexity of $-1/\varphi''$; equality (109) results from (105). ■

B Proof of Proposition 13

Proof: The proof of (55), for $q \geq 0$, results from

$$\begin{aligned} T_q^\pi(p_1, \ldots, p_m) &= \frac{1}{q-1}\left[1 - \sum_{j=1}^n \left(\sum_{t=1}^m \pi_t p_{tj}\right)^q - \sum_{t=1}^m \pi_t^q \left(1 - \sum_{j=1}^n p_{tj}^q\right)\right] \\ &= S_q(\pi) + \frac{1}{q-1}\sum_{j=1}^n \left[\sum_{t=1}^m (\pi_t p_{tj})^q - \left(\sum_{t=1}^m \pi_t p_{tj}\right)^q\right] \\ &\leq S_q(\pi), \end{aligned} \tag{110}$$

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 (57), for $q \in [0,1]$ and $\mathcal{X}$ finite,

$$\begin{aligned} T_q^\pi(p_1, \ldots, p_m) &= S_q\left(\sum_{t=1}^m \pi_t p_t\right) - \sum_{t=1}^m \pi_t^q S_q(p_t) \\ &\geq \sum_{t=1}^m \pi_t S_q(p_t) - \sum_{t=1}^m \pi_t^q S_q(p_t) \qquad (111) \\ &= \sum_{t=1}^m (\pi_t - \pi_t^q) S_q(p_t) \\ &\geq S_q(U) \sum_{t=1}^m (\pi_t - \pi_t^q) \qquad (112) \\ &= S_q(\pi)[1 - n^{1-q}]. \qquad (113) \end{aligned}$$

where the inequality (111) results from S_q being concave, and the inequality 112 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 $S_q(U) = (1 - n^{1-q})/(q-1)$. ■

C Proof of Proposition 44

Proof: We show a counterexample with $q = 1$ (the extensive case), $\pi = (1/2, 1/2)$ and $k = 1$, that discards both cases. It suffices to show that $\sqrt{JS_1^{\text{cond}}} \triangleq \sqrt{T_{1,1}^{\text{cond},(1/2,1/2)}}$ violates the triangle inequality for some choice of stochastic processes s_1, s_2, s_3 and therefore is a not a squared distance; this in turn implies that $\sqrt{JS_1^{\text{cond}}}$ is not nd and, from Proposition 21, that the above two kernels are not pd. We define s_1, s_2, s_3 to be stationary first order Markov processes in a binary alphabet $\mathcal{A} = \{0, 1\}$ defined by the following transition matrices, respectively:

$$S_1 = \lim_{\epsilon \to 0} \begin{bmatrix} 1-\epsilon & \epsilon \\ 1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/4 & 3/4 \end{bmatrix}, \tag{114}$$

$$S_2 = \lim_{\epsilon \to 0} \begin{bmatrix} 3/4 & 1/4 \\ \epsilon & 1-\epsilon \end{bmatrix} = \begin{bmatrix} 3/4 & 1/4 \\ 0 & 1 \end{bmatrix}, \tag{115}$$

and

$$S_3 = \lim_{\epsilon \to 0} \begin{bmatrix} \epsilon & 1-\epsilon \\ 1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1/4 & 3/4 \end{bmatrix}, \tag{116}$$

whose stationary distributions are

$$\sigma_1 = \lim_{\epsilon \to 0} \frac{1}{1+4\epsilon} \begin{bmatrix} 1 \\ 4\epsilon \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{117}$$

$$\sigma_2 = \lim_{\epsilon \to 0} \frac{1}{1+4\epsilon} \begin{bmatrix} 4\epsilon \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{118}$$

and

$$\sigma_3 = \lim_{\epsilon \to 0} \frac{1}{5-4\epsilon} \begin{bmatrix} 1 \\ 4-4\epsilon \end{bmatrix} = \begin{bmatrix} 1/5 \\ 4/5 \end{bmatrix}. \tag{119}$$

The matrix of first order conditional JT 1-differences (or first order conditional Jensen-Shannon divergences) is

$$\begin{bmatrix} 0 & 0 & \frac{3}{5}H(\frac{5}{6}) \\ * & 0 & \frac{9}{10}H(\frac{8}{9}) - \frac{2}{5}H(\frac{1}{4}) \\ * & * & 0 \end{bmatrix} \approx \begin{bmatrix} 0 & 0 & 0.390 \\ * & 0 & 0.128 \\ * & * & 0 \end{bmatrix}, \tag{120}$$

which fails to be negative definite, since

$$\sqrt{JS_1^{\text{cond}}(s_1, s_2)} + \sqrt{JS_1^{\text{cond}}(s_2, s_3)} < \sqrt{JS_1^{\text{cond}}(s_1, s_3)}, \tag{121}$$

which violates the triangle inequality required for $\sqrt{JS_1^{\text{cond}}}$ to be a metric.

Interestingly, the 0-th order conditional Jensen-Shannon divergence matrix (this one ensured to be negative definite because it equals a standard Jensen-Shannon divergence matrix) is

$$\begin{bmatrix} 0 & 1 & H(\frac{2}{5}) - \frac{1}{2}H(\frac{1}{5}) \\ * & 0 & H(\frac{1}{10}) - \frac{1}{2}H(\frac{1}{5}) \\ * & * & 0 \end{bmatrix} \approx \begin{bmatrix} 0 & 1 & 0.610 \\ * & 0 & 0.108 \\ * & * & 0 \end{bmatrix}. \tag{122}$$

From the chain rule (68), we have that the sum of the matrices (120) and (122) is the second order joint Jensen-Shannon divergence, and therefore is also guaranteed to be negative definite. ■

References

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

S. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2001.

A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

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

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

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

J. Burbea and C. 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.

M. Cuturi and J.-P. Vert. Semigroup kernels on finite sets. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 329–336. MIT Press, Cambridge, MA, 2005.

M. Cuturi, K. Fukumizu, and J.-P. Vert. Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:1169–1198, 2005.

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

F. Desobry, M. Davy, and W. Fitzgerald. Density kernels on unordered sets for kernel-based signal processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP'2007*, 2007.

R. El-Yaniv, S. Fine, and N. Tishby. Agnostic classification of markovian sequences. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 465–471. MIT Press, Cambridge, MA, 1998.

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

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

S. Furuichi. Information theoretical properties of Tsallis entropies. *Journal of Mathematical Physics*, 47(2), 2006.

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.

D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1997.

D. Haussler. Convolution kernels on discrete structures, 1999. URL `citeseer.ist.psu.edu/haussler99convolution.html`.

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

Y. He, A. Ben-Hamza, and H. Krim. A generalized divergence measure for robust image registration. *IEEE Transactions 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. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in SVMs. In *Proceedings of the 26th DAGM Symposium*, pages 270–277. Springer, 2004.

T. Jebara, R. Kondor, and A. 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.

T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, 2002.

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

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

P. Lamberti and A. Majtey. Non-logarithmic Jensen-Shannon divergence. *Physica A Statistical Mechanics and its Applications*, 329:81–90, 2003.

C. Leslie, E. Eskin, and W. Noble. The spectrum kernel: A string kernel for svm protein classification. In *Proceedings of the Pacific Symposium on Biocomputing 2002*, pages 564–575, 2002.

Y. Li, X. Fan, and G. 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. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

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

J. Lindhard. *On the theory of measurement and its consequences in statistical dynamics*. Munksgaard, Copenhagen, 1974.

J. Lindhard and V. Nielsen. *Studies in statistical dynamics*. Munksgaard, Copenhagen, 1971.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

A. Martins, P. Aguiar, and M. Figueiredo. Tsallis kernels on measures. In *Proceedings of the IEEE Information Theory Workshop – ITW'08*, Porto, Portugal, 2008a.

A. Martins, M. Figueiredo, P. Aguiar, N. Smith, and E. Xing. Nonextensive entropic kernels. In *Proceedings of the International Conference on Machine Learning – ICML'08*, Helsinki, Finland, 2008b.

P. Moreno, P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

A. Rényi. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics, and Probability*, volume 1, pages 547–561, Berkeley, 1961. University of California Press.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002.

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

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

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

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

H. 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.

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

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

S. Vishwanathan and A. Smola. Fast kernels for string and tree matching. In K. Tsuda, B. Schölkopf, and J.P. Vert, editors, *Kernels and Bioinformatics*, Cambridge, MA, 2003. MIT Press.

D. Zhang, X. Chen, and W. Lee. Text classification with kernels on the multinomial manifold. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 266–273, New York, NY, 2005. ACM Press.

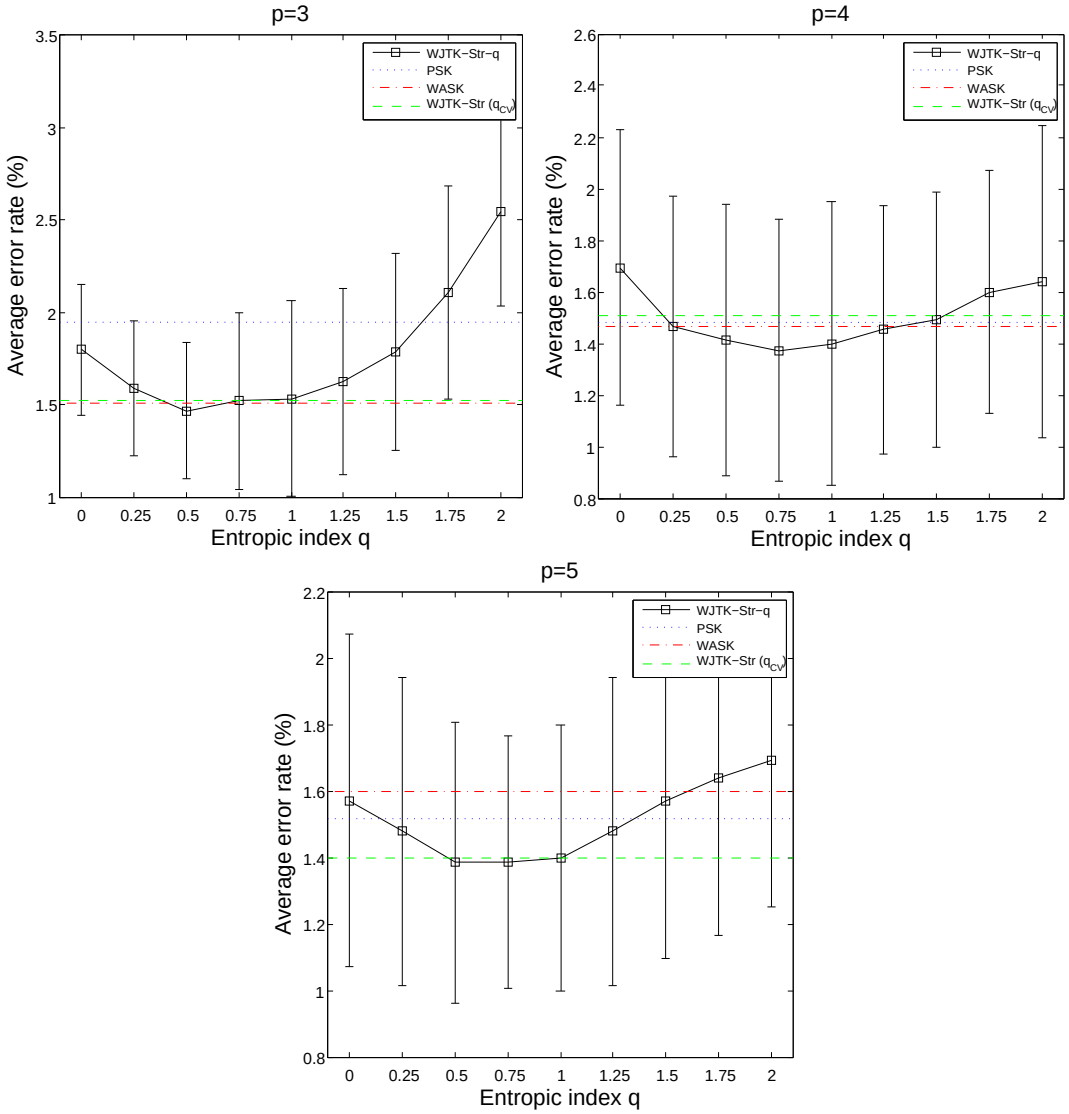

Figure 4: Results for *earn-vs-acq* using string kernels and $p = 3, 4, 5$. The error bars represent ± 1 standard deviation on 15 runs. Training (resp. testing) with 200 (resp. 250) samples per class.

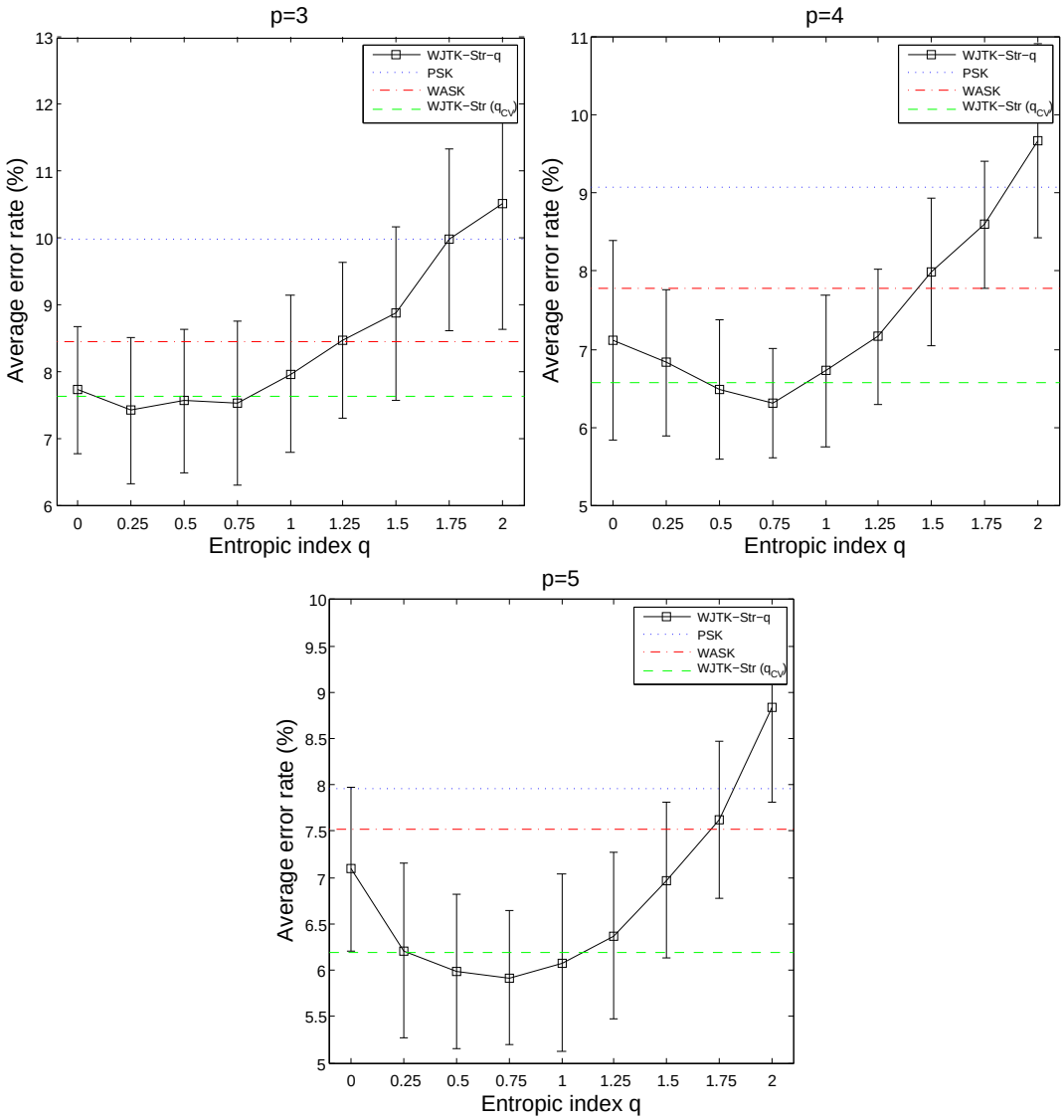

Figure 5: Results for *stud-vs-fac* using string kernels.

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon