Nonextensive Entropic Kernels

André F. T. Martins[†‡] AFM@CS.CMU.EDU
Mário A. T. Figueiredo[‡] MARIO.FIGUEIREDO@LX.IT.PT
Pedro M. Q. Aguiar[♯] AGUIAR@ISR.IST.UTL.PT
Noah A. Smith[†] NASMITH@CS.CMU.EDU
Eric P. Xing[†] EPXING@CS.CMU.EDU

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

Abstract

Positive definite kernels on probability measures have been recently applied in structured data classification problems. Some of these kernels are related to classic information theoretic quantities, such as mutual information and the Jensen-Shannon divergence. Meanwhile, driven by recent advances in Tsallis statistics, nonextensive generalizations of Shannon's information theory have been proposed. This paper bridges these two trends. We introduce the *Jensen-Tsallis q-difference*, a generalization of the Jensen-Shannon divergence. 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, Jensen-Shannon, and linear kernels as particular cases. We illustrate the performance of these kernels on text categorization tasks.

1. Introduction

There has been recent interest in kernels on probability distributions, to tackle several classification problems (Moreno *et al.*, 2003; Jebara *et al.*, 2004; Hein & Bousquet, 2005; Lafferty & Lebanon, 2005; Cuturi *et al.*, 2005). By mapping data points to fitted distributions in a parametric family where a kernel is defined, a kernel is automatically induced on the original input space. In text categorization, this appears as an alternative to the Euclidean geometry inherent to the usual bag-of-words vector representations. In fact, approaches that map data to a statistical manifold, where well-motivated non-Euclidean metrics may be defined (Lafferty & Lebanon, 2005), outperform SVM classifiers with linear kernels (Joachims, 1997). Some of these kernels have a natural information theoretic interpretation, creating a bridge between kernel methods and information theory (Cuturi *et al.*, 2005; Hein & Bousquet, 2005).

We reinforce that bridge by introducing a new class of kernels rooted in *nonextensive* (NE) information theory. The Shannon and Rényi entropies (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 NE entropies (Havrda & Charvát, 1967; Tsallis, 1988), which have raised great interest among physicists in modeling certain phenomena (*e.g.*, long-range interactions and multifractals) and as generalizations of Boltzmann-Gibbs statistical mechanics (Abe, 2006). NE entropies have also been recently used in signal/image processing (Li *et al.*, 2006) and other areas (Gell-Mann & Tsallis, 2004).

The main contributions of this paper are:

- Based on the new concept of q-convexity and a related q-Jensen inequality, we introduce the *Jensen-Tsallis q-difference*, a NE generalization of the Jensen-Shannon (JS) divergence.
- We propose a broad family of positive definite (pd) kernels, which are interpretable as NE mutual information (MI) kernels. This family ranges from the Boolean to the linear kernels, and also includes the JS kernel (Hein & Bousquet, 2005).
- We extend results of Hein and Bousquet (2005) by proving positive definiteness of kernels based on the unbalanced JS divergence. As a side note, we

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

show that the parametrix approximation of the multinomial diffusion kernel introduced by Lafferty and Lebanon (2005) is *not* pd in general.

Our main purpose is to present new theoretical insights about kernels on measures by unifying some well-known instances into a common parametrized family. This family allows reinterpreting these kernels in light of NE information theory, a connection that to our knowledge had not been presented before. The fact that some members of this family are novel pd kernels leads us to include a set of text categorization experiments that illustrates their effectiveness.

The paper is organized as follows. Sec. 2 reviews NE entropies, while Jensen differences and divergences are discussed in Sec. 3. In Sec. 4, the concepts of q-differences and q-convexity are introduced and used to define the Jensen-Tsallis q-difference. Sec. 5 presents the new family of entropic kernels. Sec. 6 reports experiments on text categorization and Sec. 7 presents concluding remarks and future research directions.

Although, for simplicity, we focus on discrete distributions on finite sets, most results are valid in arbitrary measured spaces, as shown by Martins *et al.* (2008).

2. Nonextensive Information Theory

Let X denote a random variable (rv) taking values in a finite set $\mathcal{X} = \{x_1, \ldots, x_n\}$ according to a probability distribution P_X. An entropy function is said to be *extensive* if it is additive over independent variables. For example, the Shannon entropy (Cover & Thomas, 1991), $H(X) \triangleq -\mathbb{E}[\ln P_X]$, is extensive: if X and Y are independent, then $H(X,Y) = H(X)+H(Y)$. Another example is the family of Rényi entropies (Rényi, 1961), parameterized by $q \geq 0$,

$$R_q(X) \triangleq \frac{1}{1-q} \ln \sum_{i=1}^{n} P_X(x_i)^q, \tag{1}$$

which includes Shannon's entropy as a special case when $q \to 1$.

In classic information theory, extensivity is considered desirable, and is enforced axiomatically (Khinchin, 1957), to express the idea borrowed from thermodynamics that "independent systems add their entropies." In contrast, the *Tsallis entropies* abandon the extensivity requirement (Tsallis, 1988). These NE entropies, denoted $S_q(X)$, are defined as follows:

$$S_q(X) \triangleq -\mathbb{E}_q(\ln_q P_X) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} P_X(x_i)^q\right), \tag{2}$$

where $\mathbb{E}_q(f) \triangleq \sum_{i=1}^{n} P(x_i)^q f(x_i)$ is the unnormalized *q-expectation*, and $\ln_q(y) \triangleq (y^{1-q}-1)/(1-q)$ is the so-called *q-logarithm*. It is noteworthy that when $q \to 1$, we get $\mathbb{E}_q \to \mathbb{E}$, $\ln_q \to \ln$, and $S_q \to H$; *i.e.*, the family of Tsallis entropies also includes Shannon's entropy. For the Tsallis family, when X and Y are independent, extensivity no longer holds; instead, we have

$$S_q(X,Y) = S_q(X) + S_q(Y) - (q-1)S_q(X)S_q(Y), \tag{3}$$

where the parameter $q \geq 0$ is called *entropic index*.

While statistical physics has been the main application of Tsallis entropies, some attempts have been made to produce NE generalizations of classic information theory results (Furuichi, 2006). As for the Shannon entropy, the Tsallis *joint* and *conditional* entropies are defined as $S_q(X,Y) \triangleq -\mathbb{E}_q[\ln_q P_{XY}]$ and $S_q(X|Y) \triangleq -\mathbb{E}_q[\ln_q P_{X|Y}]$, respectively, and follow a chain rule $S_q(X,Y) = S_q(X) + S_q(Y|X)$. Similarly, Furuichi (2006) defines the Tsallis MI as

$$I_q(X;Y) \triangleq S_q(X) - S_q(X|Y) = I_q(Y;X), \tag{4}$$

generalizing (for $q > 1$) Shannon's MI. This NE version of the MI underlies one of the central contributions of this paper: the Jensen-Tsallis q-difference (Sec. 4).

For reasons that will become clear in Sec. 5, it is convenient to extend the domain of Tsallis entropies to *unnormalized* measures, *i.e.*, in $\mathbb{R}^n_+ \triangleq \{\boldsymbol{\mu} \in \mathbb{R}^n \mid \forall i\ \mu_i \geq 0\}$, but not necessarily in the probability simplex $\mathbb{P}^{n-1} \triangleq \{\mathbf{p} \in \mathbb{R}^n \mid \sum_{i=1}^{n} p_i = 1,\ \forall i\ p_i \geq 0\}$. The Tsallis entropy of a measure $\boldsymbol{\mu}$ in $\mathbb{R}^n_+$ is[1]

$$S_q(\boldsymbol{\mu}) \triangleq -\sum_{i=1}^{n} \mu_i^q \ln_q \mu_i = \sum_{i=1}^{n} \varphi_q(\mu_i), \tag{5}$$

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

$$\varphi_q(y) = -y^q \ln_q y = \begin{cases} -y \ln y, & \text{if } q = 1, \\ (y - y^q)/(q-1), & \text{if } q \neq 1. \end{cases} \tag{6}$$

This extension does not add expressive power, since function (5) is completely determined by its values on $\mathbb{P}^{n-1}$, as shown by the following proposition (the proof is straightforward).

Proposition 1 *The following denormalization formula holds for any $c \geq 0$ and $\boldsymbol{\mu} \in \mathbb{R}^n_+$:*

$$S_q(c\boldsymbol{\mu}) = c^q S_q(\boldsymbol{\mu}) + \varphi_q(c)\|\boldsymbol{\mu}\|_1, \tag{7}$$

where $\|\boldsymbol{\mu}\|_1 \triangleq \sum_{i=1}^{n} \mu_i$ is the ℓ_1-norm of $\boldsymbol{\mu}$.

[1] In the following, we represent normalized and unnormalized measures as vectors in $\mathbb{R}^n$, and we use those as arguments of entropy functions, *e.g.*, we write $H(\boldsymbol{\pi})$ to denote $H(X)$ where $X \sim P(X)$, with $\pi_i = P(x_i)$.

This fact will be used in a constructive way in Sec. 5 to devise a family of pd NE entropic kernels.

3. Jensen Differences and Divergences

Jensen's inequality is at the heart of many important results in information theory. Let the rv Z take values on a finite set $\mathcal{Z}$. Jensen's inequality states that if f is a convex function defined on the convex hull of $\mathcal{Z}$, then $f(\mathbb{E}[Z]) \leq \mathbb{E}[f(Z)]$. The nonnegative quantity $\mathbb{E}[f(Z)] - f(\mathbb{E}[Z])$ is known as *Jensen difference* and has been studied by Burbea and Rao (1982) when $-f$ is some form of generalized entropy. Here, we are interested in the case where $Z \in \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m\}$ is a *random measure*, where each $\boldsymbol{\mu}_j \in \mathbb{R}^n_+$, with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m) \in \mathbb{P}^{m-1}$. The Jensen difference induced by a (concave) generalized entropy Ψ is

$$\begin{aligned} J^{\boldsymbol{\pi}}_{\Psi}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m) &\triangleq \Psi\left(\sum_{j=1}^m \pi_j \boldsymbol{\mu}_j\right) - \sum_{j=1}^m \pi_j \Psi(\boldsymbol{\mu}_j) \\ &= \Psi\left(\mathbb{E}[Z]\right) - \mathbb{E}[\Psi(Z)], \end{aligned} \tag{8}$$

Below, we show examples of Jensen differences that have been applied in machine learning. In Sec. 4, we provide a NE generalization of the Jensen difference.

Jensen-Shannon (JS) Divergence Consider a classification problem with m classes, $Y \in \mathcal{Y} = \{1, \ldots, m\}$, with *a priori* probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m) \in \mathbb{P}^{m-1}$. Let $\mathbf{p}_j = (p_{j1}, \ldots, p_{jn}) \in \mathbb{P}^n$ for $j = 1, \ldots, m$, where $p_{ji} \triangleq P(X = x_i | Y = j)$, be the corresponding class-conditional distributions.

Letting Ψ in (8) be H, the Shannon entropy, the resulting Jensen difference $J^{\boldsymbol{\pi}}_H(\mathbf{p}_1, \ldots, \mathbf{p}_m)$ is known as the JS divergence of $\mathbf{p}_1, \ldots, \mathbf{p}_m$, with weights $\pi_1, \ldots, \pi_m$ (Burbea & Rao, 1982; Lin, 1991). In this instance of the Jensen difference,

$$J^{\boldsymbol{\pi}}_H(\mathbf{p}_1, \ldots, \mathbf{p}_m) = I(X; Y), \tag{9}$$

where $I(X;Y) = H(X) - H(X|Y)$ is the MI between X and Y (Banerjee *et al.*, 2005).

For $m = 2$ and $\boldsymbol{\pi} = (\frac{1}{2}, \frac{1}{2})$, we denote the ensuing $J_H^{(\frac{1}{2}, \frac{1}{2})}(\mathbf{p}_1, \mathbf{p}_2)$ as $JS(\mathbf{p}_1, \mathbf{p}_2)$:

$$JS(\mathbf{p}_1, \mathbf{p}_2) = H((\mathbf{p}_1 + \mathbf{p}_2)/2) - (H(\mathbf{p}_1) + H(\mathbf{p}_2))/2.$$

It can be shown that that $\sqrt{JS}$ satisfies the triangle inequality and is a *Hilbertian metric*[2] (Endres & Schindelin, 2003; Topsøe, 2000), which has motivated its use in kernel-based machine learning.

[2] 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 & Bousquet, 2005).

Jensen-Rényi (JR) Divergence Let $\Psi = R_q$, which is concave for $q \in [0, 1)$; then, (8) becomes

$$J^{\boldsymbol{\pi}}_{R_q}(\mathbf{p}_1, \ldots, \mathbf{p}_m) = R_q\left(\mathbb{E}[\mathbf{p}]\right) - \mathbb{E}[R_q(\mathbf{p})]. \tag{10}$$

We call $J^{\boldsymbol{\pi}}_{R_q}$ the JR divergence. When $m = 2$ and $\boldsymbol{\pi} = (1/2, 1/2)$, we write $J^{\boldsymbol{\pi}}_{R_q}(\mathbf{p}) = JR_q(\mathbf{p}_1, \mathbf{p}_2)$, where

$$JR_q(\mathbf{p}_1, \mathbf{p}_2) = R_q\left(\frac{\mathbf{p}_1 + \mathbf{p}_2}{2}\right) - \frac{R_q(\mathbf{p}_1) + R_q(\mathbf{p}_2)}{2}.$$

The JR divergence has been used in signal processing applications (Karakos *et al.*, 2007). We show in Sect. 5.3 that $\sqrt{JR_q}$ is also an Hilbertian metric.

Jensen-Tsallis (JT) Divergence Divergences of the form (8), with $\Psi = S_q$, are known as JT divergences (Burbea & Rao, 1982) and were recently used in image processing (Hamza, 2006). Unlike the JS divergence, the JT divergence lacks a MI interpretation; in Sec. 4, we introduce an alternative to the JT divergence, which is interpretable as a NE MI in the sense of Furuichi (2006).

4. Jensen q-Differences

We now introduce *Jensen q-differences*, a generalization of Jensen differences. As described shortly, a special case of the Jensen q-difference is the *Jensen-Tsallis q-difference*, which is an NE generalization of the JS divergence, and provides the building block for the NE entropic kernels to be introduced in Sec. 5. We begin by introducing the concept of "q-convexity", which satisfies a Jensen-type inequality.

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

$$f(\lambda x + (1 - \lambda) y) \leq \lambda^q f(x) + (1 - \lambda)^q f(y). \tag{11}$$

f is q-concave *if $-f$ is q-convex.*

Naturally, 1-convexity is the usual convexity. The next proposition states the q-Jensen inequality and is easily proved by induction, like the standard Jensen inequality (Cover & Thomas, 1991). It also states that the property of q-convexity gets stronger as q increases.

Proposition 2 *If $f : \mathcal{X} \to \mathbb{R}$ is q-convex and $f \geq 0$, then, for any $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ and $\boldsymbol{\pi} \in \mathbb{P}^{n-1}$:*

$$f\left(\sum_{i=1}^n \pi_i x_i\right) \leq \sum_{i=1}^n \pi_i^q f(x_i). \tag{12}$$

Moreover, if $q \geq r \geq 0$, we have:

$$f \text{ is } q\text{-convex} \quad \Rightarrow \quad f \text{ is } r\text{-convex} \tag{13}$$

$$f \text{ is } r\text{-concave} \quad \Rightarrow \quad f \text{ is } q\text{-concave.} \tag{14}$$

Based on the q-Jensen inequality, we can now consider *Jensen q-differences* of the form $\mathbb{E}_q[f(Z)] - f(\mathbb{E}[Z])$, which are nonnegative if f is q-convex. As in Sec. 3, we focus on the scenario where Z is a random measure and $-f = \Psi$ is an entropy function, yielding

$$\begin{aligned} T^{\boldsymbol{\pi}}_{q,\Psi}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m) &\triangleq \Psi\left(\sum_{t=1}^m \pi_t\, \boldsymbol{\mu}_t\right) - \sum_{t=1}^m \pi_t^q \Psi(\boldsymbol{\mu}_t) \\ &= \Psi\left(\mathbb{E}[Z]\right) - \mathbb{E}_q[\Psi(Z)]. \end{aligned} \quad (15)$$

The Jensen q-difference is a deformation of the Jensen 1-difference (8), in which the second expectation is replaced by a q-expectation. We are now ready to introduce the class of Jensen-Tsallis q-differences.

Jensen-Tsallis q-Differences Consider again the classification problem used in the description of the JS divergence, but replacing the Jensen difference with the Jensen q-difference and the Shannon entropy with the Tsallis q-entropy, *i.e.*, letting $\Psi = S_q$ in (15). We obtain (writing $T^{\boldsymbol{\pi}}_{q,S_q}$ simply as $T^{\boldsymbol{\pi}}_q$):

$$T^{\boldsymbol{\pi}}_q(\mathbf{p}_1, \ldots, \mathbf{p}_m) = S_q(X) - S_q(X|Y) = I_q(X;Y), \quad (16)$$

where $S_q(X|Y)$ is the Tsallis conditional q-entropy, and $I_q(X;Y)$ is the Tsallis MI (cf. (4)). Note that (16) is an NE analogue of (9), *i.e.* the Jensen-Tsallis q-differences are NE mutual informations.

We call $T^{\boldsymbol{\pi}}_q(\mathbf{p}_1, \ldots, \mathbf{p}_m)$ the Jensen-Tsallis q-difference of $\mathbf{p}_1, \ldots, \mathbf{p}_m$ with weights $\pi_1, \ldots, \pi_m$.

When $m = 2$ and $\boldsymbol{\pi} = (1/2, 1/2)$, define $T_q \triangleq T^{(1/2,1/2)}_q$,

$$T_q(\mathbf{p}_1, \mathbf{p}_2) = S_q\left(\frac{\mathbf{p}_1 + \mathbf{p}_2}{2}\right) - \frac{S_q(\mathbf{p}_1) + S_q(\mathbf{p}_2)}{2^q}. \quad (17)$$

Three special cases are obtained for $q \in \{0, 1, 2\}$:

$$\begin{aligned} S_0(\mathbf{p}) &= -1 + \|\mathbf{p}\|_0; & T_0(\mathbf{p}_1, \mathbf{p}_2) &= 1 - \|\mathbf{p}_1 \odot \mathbf{p}_2\|_0 \\ S_1(\mathbf{p}) &= H(\mathbf{p}); & T_1(\mathbf{p}_1, \mathbf{p}_2) &= JS(\mathbf{p}_1, \mathbf{p}_2) \\ S_2(\mathbf{p}) &= 1 - \langle \mathbf{p}, \mathbf{p} \rangle; & T_2(\mathbf{p}_1, \mathbf{p}_2) &= \frac{1}{2} - \frac{1}{2} \langle \mathbf{p}_1, \mathbf{p}_2 \rangle \end{aligned}$$

where $\|\mathbf{x}\|_0$ is the number of nonzeros in $\mathbf{x}$, $\odot$ denotes the Hadamard-Schur (elementwise) product, and $\langle \cdot, \cdot \rangle$ is the inner product.

The JT q-difference is an NE generalization of the JS divergence, and some of the latter's properties are lost in general. Since Tsallis entropies are 1-concave, Prop. 2 guarantees q-concaveness only for $q \geq 1$. Therefore, nonnegativity is only guaranteed for JT q-differences when $q \geq 1$; for this reason some authors only consider this range of values (Furuichi, 2006). Moreover, unless $q = 1$ (the JS divergence), it is not generally true that $T^{\boldsymbol{\pi}}_q(\mathbf{p}, \ldots, \mathbf{p}) = 0$ or even that $T^{\boldsymbol{\pi}}_q(\mathbf{p}, \ldots, \mathbf{p}, \mathbf{p}') \geq T^{\boldsymbol{\pi}}_q(\mathbf{p}, \ldots, \mathbf{p}, \mathbf{p})$. For example,

$$\operatorname{argmin}_{\mathbf{p}_1 \in \mathbb{P}^{n-1}} T_q(\mathbf{p}_1, \mathbf{p}_2) \quad (18)$$

can be different from $\mathbf{p}_2$, unless $q = 1$. In general, the minimizer is closer to either the uniform distribution (if $q \in [0, 1)$) or a degenerate distribution[3] (for $q \in (1, 2]$). For these reasons, the term "divergence" is misleading and we use the term "difference." Other properties of JT q-differences (convexity, lower/upper bounds) are studied by Martins *et al.* (2008).

5. Nonextensive Entropic Kernels

Using the denormalization formula (7), we now introduce kernels based on the JS divergence and the JT q-difference, which allow weighting their arguments. In this section, $m = 2$ (kernels involve pairs of measures).

5.1. Background on Kernels

We begin with some basic results on kernels (Schölkopf & Smola, 2002). Below, $\mathcal{X}$ denotes a nonempty set; $\mathbb{R}_+$ denote the nonnegative reals, and $\mathbb{R}_{++} \triangleq \mathbb{R}_+ \setminus \{0\}$.

Definition 2 *Let $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function, i.e., $\varphi(y, x) = \varphi(x, y)$, for all $x, y \in \mathcal{X}$. φ is called a pd kernel if and only if*

$$\sum_{i=1}^n \sum_{j=1}^n c_i\, c_j\, \varphi(x_i, x_j) \geq 0, \quad (19)$$

for any integer n, $x_i, \ldots, x_n \in \mathcal{X}$ and $c_i, \ldots, c_n \in \mathbb{R}$. A symmetric function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a negative definite *(nd) kernel if and only if*

$$\sum_{i=1}^n \sum_{j=1}^n c_i\, c_j\, \psi(x_i, x_j) \leq 0, \quad (20)$$

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

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

[3] Notice that $T_2(\mathbf{p}_1, \mathbf{p}_2) = \frac{1}{2} - \frac{1}{2}\langle \mathbf{p}_1, \mathbf{p}_2 \rangle$; in this case, (18) becomes a linear program, and the solution is $\mathbf{p}_1^* = \mathbf{e}_j$, where $j = \operatorname{argmax}_i p_{2i}$.

Proposition 3 *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*

$$\varphi(x, y) = \psi(x, x_0) + \psi(y, x_0) - \psi(x, y) - \psi(x_0, x_0).$$

Then, φ is pd if and only if ψ is nd.

Proposition 4 *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 5 *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 $t > 0$.*

Proposition 6 *If ψ is nd and $\psi(x, x) \geq 0$, for all $x \in \mathcal{X}$, then so are ψ^α, for $\alpha \in [0,\, 1]$, and $\ln(1 + \psi)$.*

Proposition 7 *If $f : \mathcal{X} \to \mathbb{R}$ satisfies $f \geq 0$, then, for $\alpha \in [1, 2]$, the function $-(f(x) + f(y))^\alpha$ is a nd kernel.*

5.2. 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 are nd functions[4] on $\mathbb{R}^n_+$, for $q \in [0, 2]$. Of course, this includes the denormalized Shannon entropy as a particular case, corresponding to $q = 1$. Partial proofs are given by Berg *et al.* (1984), Topsøe (2000), and Cuturi *et al.* (2005); we present here a complete proof.

Proposition 8 *For $q \in [0, 2]$, the denormalized Tsallis q-entropy S_q is an nd function on $\mathbb{R}^n_+$.*

Proof: Since nd kernels are closed under pointwise summation, it suffices to prove that φ_q (see (6)) is nd on $\mathbb{R}_+$. For $q \neq 1$, $\varphi_q(y) = (q - 1)^{-1}(y - y^q)$. If $q \in [0, 1)$, φ_q equals $-\iota + \iota^q$ times a positive constant, where ι is the identity ($\iota(y) = y$) 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, which is easily seen from the definition; besides, since ι is nd and nonnegative, Prop. 6 implies that ι^q is also nd. For $q \in (1, 2]$, φ_q equals $\iota - \iota^q$ times a positive constant. It remains to show that $-\iota^q$ is nd for $q \in (1, 2]$; since $k(x, y) = -(x+y)^q$ is nd (Prop. 7), so is ι^q. For $q = 1$, since the set of nd functions is closed under limits,

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

it follows that φ_1 is nd. □

The following proposition, proved by Berg *et al.* (1984), will also be used below.

[4] A function $f : \mathcal{X} \to \mathbb{R}$ is called pd (resp. nd) if $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined as $k(x, y) = f(x + y)$, is a pd (resp. nd) kernel (Berg *et al.*, 1984).

Proposition 9 *The function $\zeta_q : \mathbb{R}_{++} \to \mathbb{R}$, defined as $\zeta_q(y) = y^{-q}$ is pd, for $q \in [0, 1]$.*

We now present the main contribution of this section, the family of *weighted JT kernels*, generalizing the JS divergence kernels in two ways: **(i)** they apply to unnormalized measures (equivalently, they allow weighting the arguments differently); **(ii)** they extend the MI nature of the JS divergence kernel to the NE case.

Definition 3 (weighted Jensen-Tsallis kernels) *The kernel $\widetilde{k}_q : (\mathbb{R}^n_+)^2 \to \mathbb{R}$ is defined as*

$$\begin{aligned} \widetilde{k}_q(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) &= \widetilde{k}_q(\omega_1 \mathbf{p}_1, \omega_2 \mathbf{p}_2) \\ &\triangleq \left[S_q(\boldsymbol{\pi}) - T_q^{\boldsymbol{\pi}}(\mathbf{p}_1, \mathbf{p}_2)\right](\omega_1 + \omega_2)^q, \end{aligned}$$

where $\mathbf{p}_1 = \boldsymbol{\mu}_1/\omega_1$ and $\mathbf{p}_2 = \boldsymbol{\mu}_2/\omega_2$ are the normalized counterparts of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, with corresponding weights $\omega_1, \omega_2 \in \mathbb{R}_+$, and $\boldsymbol{\pi} = (\omega_1/(\omega_1 + \omega_2), \omega_2/(\omega_1 + \omega_2))$.

The kernel $k_q : (\mathbb{R}^n_{++})^2 \to \mathbb{R}$ is defined as

$$k_q(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = k_q(\omega_1 \mathbf{p}_1, \omega_2 \mathbf{p}_2) \triangleq S_q(\boldsymbol{\pi}) - T_q^{\boldsymbol{\pi}}(\mathbf{p}_1, \mathbf{p}_2).$$

Recalling (16), notice $S_q(Y) - I_q(X; Y) = S_q(Y|X)$ can be interpreted as the *Tsallis posterior conditional entropy*. Hence, k_q can be seen (in Bayesian classification terms) as a NE expected measure of uncertainty in correctly identifying the class given the prior $\boldsymbol{\pi} = (\pi_1, \pi_2)$ and a random sample from the mixture distribution $\pi_1 \mathbf{p}_1 + \pi_2 \mathbf{p}_2$. The more similar the two distributions are, the greater this uncertainty.

Proposition 10 *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 $\boldsymbol{\mu}_1 = \omega_1 \mathbf{p}_1$ and $\boldsymbol{\mu}_2 = \omega_2 \mathbf{p}_2$ and using the denormalization formula (7), we obtain $\widetilde{k}_q(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = -S_q(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + S_q(\boldsymbol{\mu}_1) + S_q(\boldsymbol{\mu}_2)$. Now invoke Prop. 3 with $\psi = S_q$ (which is nd by Prop. 8), $x = \boldsymbol{\mu}_1$, $y = \boldsymbol{\mu}_2$, and $x_0 = \mathbf{0}$ (the null measure). Observe now that $k_q(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \widetilde{k}_q(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)(\omega_1 + \omega_2)^{-q}$. Since the product of two pd kernels is a pd kernel and (Prop. 9) $(\omega_1 + \omega_2)^{-q}$ is a pd kernel, for $q \in [0, 1]$, k_q is pd. □

As we can see, the weighted JT 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.

Weighted JS Kernel Setting $q = 1$, we obtain an extensive subfamily that contains unbalanced versions of the JS kernel (Hein & Bousquet, 2005). Namely, we

get the pd kernels:

$$\begin{aligned}\widetilde{k}_1(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) &= [H(\boldsymbol{\pi}) - J^{\boldsymbol{\pi}}(\mathbf{p}_1, \mathbf{p}_2)]\,(\omega_1 + \omega_2),\\ k_1(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) &= H(\boldsymbol{\pi}) - J^{\boldsymbol{\pi}}(\mathbf{p}_1, \mathbf{p}_2). \end{aligned} \quad (21)$$

Exponentiated Weighted JS Kernel Using Prop. 4, we have that exponentiated weighted JS kernel $k_{EWJS} : \mathbb{R}^n_+ \to \mathbb{R}$,

$$\begin{aligned} k_{EWJS}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) &\triangleq \exp[t\, k_1(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)] \\ &= \exp(t\, H(\boldsymbol{\pi}))\exp\left[-tJ^{\boldsymbol{\pi}}(\mathbf{p}_1, \mathbf{p}_2)\right] \end{aligned} \quad (22)$$

is also pd for any $t > 0$. This generalizes the exponentiated JS kernel $k_{EJS}(\mathbf{p}_1, \mathbf{p}_2) \triangleq \exp\left[-t\, JS\,(\mathbf{p}_1, \mathbf{p}_2)\right]$ (Cuturi *et al.*, 2005).

We now keep $q \in [0, 2]$ but consider the weighted JT kernel family restricted to normalized measures, $k_q|_{(\mathbb{P}^{n-1})^2}$. 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,

$$\widetilde{k}_q(\mathbf{p}_1, \mathbf{p}_2) = k_q(\mathbf{p}_1, \mathbf{p}_2) = \ln_q(2) - T_q(\mathbf{p}_1, \mathbf{p}_2). \quad (23)$$

Prop. 10 tells us that these kernels are pd for $q \in [0, 2]$. Remarkably, we recover three well-known particular cases for $q \in \{0, 1, 2\}$.

Jensen-Shannon kernel (JSK) For $q = 1$, we obtain the JS kernel, $k_{JS} : (\mathbb{P}^{n-1})^2 \to \mathbb{R}$,

$$k_{JS}(\mathbf{p}_1, \mathbf{p}_2) = \ln(2) - JS(\mathbf{p}_1, \mathbf{p}_2), \quad (24)$$

introduced and shown pd by Hein and Bousquet (2005).

Boolean kernel For $q = 0$, we obtain the kernel $k_0 = k_{Bool} : (\mathbb{P}^{n-1})^2 \to \mathbb{R}$,

$$k_{Bool}(\mathbf{p}_1, \mathbf{p}_2) = \|\mathbf{p}_1 \odot \mathbf{p}_2\|_0. \quad (25)$$

Linear kernel For $q = 2$, we obtain the kernel $k_2 = k_{lin} : (\mathbb{P}^{n-1})^2 \to \mathbb{R}$,

$$k_{lin}(\mathbf{p}_1, \mathbf{p}_2) = \frac{1}{2}\,\langle \mathbf{p}_1, \mathbf{p}_2 \rangle. \quad (26)$$

Summarizing, Boolean, JS, and linear kernels, are members of the much wider family of Tsallis kernels, continuously parameterized by $q \in [0, 2]$. Furthermore, Tsallis kernels are a particular subfamily of the even wider set of weighted Tsallis kernels.

A key feature of our generalization is that the kernels are defined on unnormalized measures. This is relevant for empirical measures (*e.g.*, term counts, image histograms); instead of the usual normalization (Hein & Bousquet, 2005), these empirical measures may be left unnormalized, allowing objects of different sizes to have different weights. Another possibility is the explicit inclusion of weights (ω_i): given an input set of normalized measures, each can be multiplied by an arbitrary (positive) weight before computing the kernel.

5.3. Other Kernels Based on Jensen Differences

Other pd kernels may be devised inspired by Jensen-Rényi and Jensen-Tsallis divergences (Section 3). For example, it is a direct consequence of Prop. 6 that, for $q \in [0, 1]$, $(\mathbf{p}_1, \mathbf{p}_2) \mapsto R_q\big(\frac{\mathbf{p}_1 + \mathbf{p}_2}{2}\big)$, and therefore JR_q, are nd kernels on $(\mathbb{P}^{n-1})^2$. We can then make use of Prop. 4 to derive pd kernels via exponentiation; for example, the *exponentiated Jensen-Rényi kernel* (pd for $q \in [0, 1]$ and $t \geq 0$):

$$\begin{aligned} k_{EJR}(\mathbf{p}_1, \mathbf{p}_2) &\triangleq \exp(-t\, J_{R_q}(\mathbf{p}_1, \mathbf{p}_2)) \\ &= \left(\frac{\sum_i \left(\frac{p_{1i} + p_{2i}}{2}\right)^q}{\sqrt{\sum_i p_{1i}^q \sum_i p_{2i}^q}}\right)^{-\frac{t}{1-q}}. \end{aligned} \quad (27)$$

However, these kernels are no longer interpretable as MIs, and arbitrary weights are not allowed. Martins *et al.* (2008) also show that a related family of pd kernels for probability measures introduced by Hein and Bousquet (2005) can be written as differences between JT-type divergences.

5.4. The Heat Kernel Approximation

The diffusion kernel for statistical manifolds, recently proposed by Lafferty and Lebanon (2005), is grounded in information geometry. It models the diffusion of "information" over the manifold through the heat equation. Since in the case of the multinomial manifold 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 induced by the Fisher information metric. The resulting heat kernel approximation is

$$k_{\,heat}(\mathbf{p}_1, \mathbf{p}_2) = (4\pi\tau)^{-\frac{n}{2}} \exp\left(-\frac{1}{4t}\, d_g^2(\mathbf{p}_1, \mathbf{p}_2)\right), \quad (28)$$

where $\tau > 0$ and $d_g(\mathbf{p}_1, \mathbf{p}_2) = 2 \arccos\left(\sum_i \sqrt{p_{1i} p_{2i}}\right)$. Whether $k_{\,heat}$ is pd has been an open problem (Hein *et al.*, 2004; Zhang *et al.*, 2005).

Proposition 11 *Let $n \geq 2$. For sufficiently large τ, the kernel k_{heat} is* not *pd.*

Proof: From Prop. 4, k_{heat} is pd, for all $\tau > 0$, if and only if d_g^2 is nd. We provide a counterexample, using the following four points in $\mathbb{P}^2$: $\mathbf{p}_1 = (1, 0, 0)$, $\mathbf{p}_2 = (0, 1, 0)$, $\mathbf{p}_3 = (0, 0, 1)$ and $\mathbf{p}_4 = (1/2, 1/2, 0)$. The squared distance matrix $[D_{ij}] = [d_g^2(\mathbf{p}_i, \mathbf{p}_j)]$ is

$$D = \frac{\pi^2}{4} \cdot \begin{bmatrix} 0 & 4 & 4 & 1 \\ 4 & 0 & 4 & 1 \\ 4 & 4 & 0 & 4 \\ 1 & 1 & 4 & 0 \end{bmatrix}. \tag{29}$$

Taking $\mathbf{c} = (-4, -4, 1, 7)$ we have $\mathbf{c}^T D \mathbf{c} = 2\pi^2 > 0$, showing that D is not nd. Although $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4$ lie on the boundary of $\mathbb{P}^2$, continuity of d_g^2 implies that it is not nd. The case $n > 2$ follows easily, by appending zeros to the four vectors above. □

6. Experiments

We illustrate the performance of the proposed NE kernels, in comparison with common kernels, for SVM text classification. We performed experiments in two standard datasets: *Reuters-21578* and *WebKB*.[5] 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* (student *vs.* faculty webpages) in the second dataset.

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

As baselines, we used the linear kernel with ℓ_2 normalization, commonly used for this task, and the heat kernel approximation (28) (Lafferty & Lebanon, 2005), which is known to outperform the former, albeit not being guaranteed to be pd for an arbitrary choice of τ (see 28), as shown above. This parameter and the SVM C parameter were tuned with cross-validation over the training set. The SVM-Light package (`http://svmlight.joachims.org/`) was used to solve the SVM quadratic optimization problem.

Figs. 1–2 summarize the results. We report the performance of the Tsallis kernels as a function of the entropic index. For comparison, we also plot the performance of an instance of a Tsallis kernel with q tuned through cross-validation. For the first task, this kernel and the two baselines exhibit similar performance for both the *tf* and the *tf-idf* representations; differences are not statiscally significant. In the second task, the Tsallis kernel outperformed the ℓ_2-normalized linear kernel for both representations, and the heat kernel for *tf-idf*; the differences are statistically significant (using the unpaired t test at the 0.05 level). Regarding the influence of the entropic index, we observe that in both tasks, the optimum value of q is usually higher for *tf-idf* than for *tf*.

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

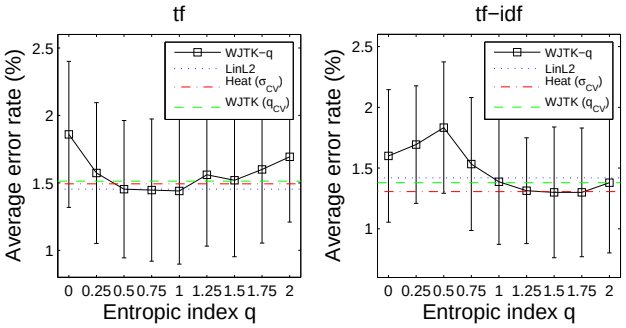

Figure 1. Results for *earn-vs-acq* using *tf* and *tf-idf* representations. The error bars represent ±1 standard deviation on 30 runs. Training (resp. testing) with 200 (resp. 250) samples per class.

7. Conclusion

We have introduced a new family of positive definite kernels between measures, which contains some well-known kernels as particular cases. These kernels are defined on unnormalized measures, which makes them suitable for use on empirical measures (*e.g.*, word counts or pixel intensity histograms), allowing objects of different sizes to be weighted differently. The family is parameterized by the entropic index, a key concept in Tsallis statistics, and includes as extreme cases the Boolean and the linear kernels. The new kernels, and

[5] Available at `http://www.daviddlewis.com/resources/testcollections` and `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data`, respectively.

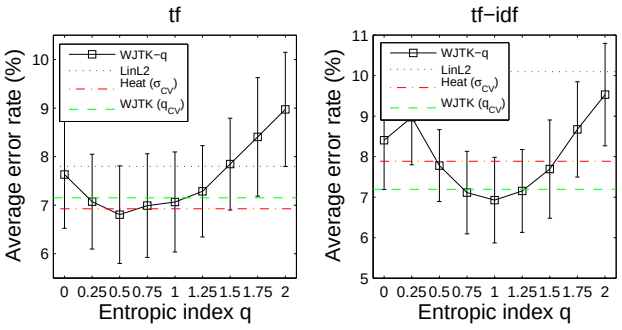

Figure 2. Results for *stud-vs-fac*.

the proofs of positive definiteness, are supported by the other contributions of this paper: the new concept of q-convexity, the underlying Jensen q-inequality, and the concept of *Jensen-Tsallis q-difference*, a nonextensive generalization of the Jensen-Shannon divergence. Experimentally, kernels in this family outperformed the linear kernel in the task of text classification and achieved similar results to the first-order approximation of the multinomial diffusion kernel. They have the advantage, however, of being pd, which fails to happen with the latter kernel, as also shown in this paper.

Future research will concern applying Jensen-Tsallis q-differences to other learning problems, like clustering, possibly exploiting the fact that they accept more than two arguments.

Acknowledgments

The authors thank the reviewers for helpful comments and Guy Lebanon for fruitful discussions on the heat kernel. This work was partially supported by *Fundação para a Ciência e Tecnologia* (FCT), Portugal, grant PTDC/EEA-TEL/72572/2006. A.M. was supported by a grant from FCT through the CMU-Portugal Program and the Information and Communications Technologies Institute (ICTI) at CMU. N.S. was supported by NSF IIS-0713265 and DARPA HR00110110013. E.X. was supported by NSF DBI-0546594, DBI-0640543, and IIS-0713379.

References

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

Banerjee, A., Merugu., S., Dhillon, I. S. & Ghosh, J. (2005). Clustering with Bregman divergences. *JMLR*, *6*, 1705–1749.

Berg, C., Christensen, J., & Ressel, P. (1984). *Harmonic analysis on semigroups*. Springer.

Burbea, J., & Rao, C. (1982). On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Inf. Theory*, *28*, 489–495.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. Wiley.

Cuturi, M., Fukumizu, K., & Vert, J.-P. (2005). Semigroup kernels on measures. *JMLR*, *6*, 1169–1198.

Endres, D., & Schindelin, J. (2003). A new metric for probability distributions. *IEEE Trans. Inf. Theory*, *49*, 1858–1860.

Fuglede, B. (2005). Spirals in Hilbert space. *Expositiones Mathematicae*, *25*, 23–46.

Furuichi, S. (2006). Information theoretical properties of Tsallis entropies. *J. Math. Physics*, *47*, no. 2.

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

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

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

Hein, M., & Bousquet, O. (2005). Hilbertian metrics and positive definite kernels on probability measures. *Proc. 10th AISTATS*.

Hein, M., Lal, T., & Bousquet, O. (2004). Hilbertian metrics on probability measures and their application in SVMs. *DAGM-Symposium*, 270-277.

Jebara, T., Kondor, R., & Howard, A. (2004). Probability product kernels. *JMLR*, *5*, 819–844.

Joachims, T. (1997). Text categorization with support vector machines: Learning with many relevant features (T.R.). Universität Dortmund.

Karakos, D., Khudanpur, S., Eisner, J., & Priebe, C. (2007). Iterative denoising using Jensen-Rényi divergences with an application to unsupervised document categorization. *Proc. IEEE ICASSP*.

Khinchin, A. (1957). *Mathematical foundations of information theory*. Dover.

Lafferty, J., & Lebanon, G. (2005). Diffusion kernels on statistical manifolds. *JMLR*, *6*, 129–163.

Li, Y., Fan, X., & Li, G. (2006). Image segmentation based on Tsallis-entropy and Rényi-entropy and their comparison. *IEEE Intern. Conf. on Industrial Informatics* (pp. 943–948).

Lin, J. (1991). Divergence measures based on Shannon entropy. *IEEE Trans. Inf. Theory*, *37*, 145–151.

Martins, A.F.T., Figueiredo, M.A.T., Aguiar, P.M.Q., Smith, N.A., Xing, E.P. (2008). Nonextensive entropic kernels. T.R. CMU-ML-08-106.

Moreno, P., Ho, P., & Vasconcelos, N. (2003). A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. *NIPS*.

Rényi, A. (1961). On measures of entropy and information. *Proc. 4th Berkeley Symp. Math. Statist. and Prob.* (pp. 547–561). Univ. Calif. Press.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. MIT Press.

Topsøe, F. (2000). Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory*, *46*, 1602–1609.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *J. Stats. Physics*, *52*, 479–487.

Zhang, D., Chen, X., & Lee, W. (2005). Text classification with kernels on the multinomial manifold. *Proc. ACM SIGIR*.