Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

On Power-Law Kernels, Corresponding Reproducing Kernel Hilbert Space and Applications

Debarghya Ghoshdastidar and **Ambedkar Dukkipati**
Department of Computer Science and Automation
Indian Institute of Science, Bangalore - 560012.
email: {debarghya.g,ad}@csa.iisc.ernet.in

Abstract

The role of kernels is central to machine learning. Motivated by the importance of power-law distributions in statistical modeling, in this paper, we propose the notion of power-law kernels to investigate power-laws in learning problem. We propose two power-law kernels by generalizing Gaussian and Laplacian kernels. This generalization is based on distributions, arising out of maximization of a generalized information measure known as nonextensive entropy that is very well studied in statistical mechanics. We prove that the proposed kernels are positive definite, and provide some insights regarding the corresponding Reproducing Kernel Hilbert Space (RKHS). We also study practical significance of both kernels in classification and regression, and present some simulation results.

1 Introduction

The notion of 'power-law' distributions is not recent, and they first arose in economics in the studies of Pareto (1906) hundred years ago. Later, power-law behavior was observed in various fields such as physics, biology, computer science etc. (Gutenberg and Richter 1954; Barabási and Albert 1999), and hence the phrase "ubiquitous power-laws". Though the term was first coined for distributions with a negative constant exponent, *i.e.*, $f(x) \propto x^{-\alpha}$, the meaning of the term has expanded in due course of time to include various fat-tailed distributions, *i.e.*, distributions decaying at a slower rate than Gaussian distribution. This class is also referred to as generalized Pareto distributions.

On the other hand, though the generalizations of information measures were proposed in the beginning of the birth of information theory, only (relatively) recently their connections with power-law distributions have been established. While maximization of Shannon entropy gives rise to exponential distributions, these generalized measures give power-law distributions. This actually led to a dramatic increase in interest in generalized information measures and their application to statistics.

Indeed, the starting point of the theory of generalized measures of information is due to Alfred Rényi (1960). Another generalization was introduced by Havrda and Charvát (1967), and then studied by Tsallis (1988) in statistical mechanics that is known as Tsallis entropy or nonextensive entropy. Tsallis entropy involves a parameter q, and it retrieves Shannon entropy as $q \to 1$. The Shannon-Khinchin axioms of Shannon entropy have been generalized to this case (Suyari 2004), and this entropy functional has been studied in information theory, statistics and many other fields. Tsallis entropy has been used to study power-law behavior in different cases like finance, earthquakes and network traffic (Sato 2010; Abe and Suzuki 2003; 2005).

In kernel based machine learning, positive definite kernels are considered as a measure of similarity between points (Scholköpf and Smola 2002). The choice of kernel is critical to the performance of the learning algorithms, and hence, many kernels have been studied in literature (Cristianini and Shawe-Taylor 2004). Kernels based on information theoretic quantities are also commonly used in text mining and image processing (He, Hamza, and Krim 2003). However, such kernels are defined on probability measures. Probability kernels based on Tsallis entropy have also been studied in (Martins et al. 2009).

In this work, we are interested in kernels based on maximum entropy distributions. It turns out that Gaussian, Laplacian, Cauchy kernels, which have been extensively studied in machine learning, have corresponding distributions, which are maximum entropy distributions. This motivates us to look into kernels that correspond to maximum Tsallis entropy distributions, also termed as Tsallis distributions. These distributions have inherent advantages as they are generalizations of exponential distributions, and they exhibit power-law nature (Sato 2010; Ghoshdastidar, Dukkipati, and Bhatnagar 2012). In fact, the value of q controls the nature of the power-law tails.

In this paper, we propose a new kernel based on q-Gaussian distribution, which is a generalization of Gaussian, obtained by maximizing Tsallis entropy under certain moment constraints. Further, we introduce a generalization of the Laplace distribution following the same lines, and propose a similar q-Laplacian kernel. We give some insights into reproducing kernel Hilbert spaces (RKHS) of these kernels. We prove that the proposed kernels are positive definite over a range of values of q. We demonstrate the effect of these kernel by applying them to machine learning tasks:

classification and regression by SVMs. We provide results indicating that in some cases, the proposed kernels perform better than their counterparts (Gaussian and Laplacian kernels) for certain values of q.

2 Tsallis distributions

Tsallis entropy can be obtained by generalizing the information of a single event in the definition of Shannon entropy as shown by Tsallis (1988), where natural logarithm is replaced with q-logarithm defined as $\ln_q x = \frac{x^{1-q}-1}{1-q}$, $q \in \mathbb{R}$, $q > 0$, $q \neq 1$. Tsallis entropy in a continuous case is defined as (Dukkipati, Bhatnagar, and Murty 2007)

$$H_q(p) = \frac{1 - \int_{\mathbb{R}} \big(p(x)\big)^q \mathrm{d}x}{q-1}, \quad q \in \mathbb{R}, q > 0, q \neq 1. \quad (1)$$

This function retrieves the differential Shannon entropy functional as $q \to 1$. It is called nonextensive because of its pseudo-additive nature (Tsallis 1988).

Kullback's minimum discrimination theorem (Kullback 1959) establishes connections between statistics and information theory. A special case is Jaynes' maximum entropy principle (Jaynes 1957), by which exponential distributions can be obtained by maximizing Shannon entropy functional, subject to some moment constraints. Using the same principle, maximizing Tsallis entropy under the following constraint

$$q\text{-mean } \langle x \rangle_q := \frac{\int_{\mathbb{R}} x \big(p(x)\big)^q \mathrm{d}x}{\int_{\mathbb{R}} \big(p(x)\big)^q \mathrm{d}x} = \mu, \quad (2)$$

results in a distribution known as q-exponential distribution (Tsallis, Mendes, and Plastino 1998), which is of the form

$$p(x) = \frac{1}{\mu} \exp_q \left(-\frac{x}{(2-q)\mu} \right), \quad (3)$$

where the q-exponential, $\exp_q(z)$, is expressed as

$$\exp_q(z) = \big(1 + (1-q)z\big)_+^{\frac{1}{1-q}}. \quad (4)$$

The condition $y_+ = \max(y, 0)$ in (4) is called the Tsallis cut-off condition, which ensures existence of q-exponential. If a constraint based on the second moment,

$$q\text{-variance } \langle x^2 \rangle_q := \frac{\int_{\mathbb{R}} (x-\mu)^2 \big(p(x)\big)^q \mathrm{d}x}{\int_{\mathbb{R}} \big(p(x)\big)^q \mathrm{d}x} = \sigma^2, \quad (5)$$

is considered along with (2), one obtains the q-Gaussian distribution (Prato and Tsallis 1999) defined as

$$p(x) = \frac{\Lambda_q}{\sigma\sqrt{3-q}} \exp_q \left(-\frac{(x-\mu)^2}{(3-q)\sigma^2} \right), \quad (6)$$

where Λ_q is the normalizing constant (Prato and Tsallis 1999). However, instead of (2), if the constraint

$$\langle |x| \rangle_q := \frac{\int_{\mathbb{R}} |x| \big(p(x)\big)^q \mathrm{d}x}{\int_{\mathbb{R}} \big(p(x)\big)^q \mathrm{d}x} = \beta, \quad (7)$$

is considered, then maximization of Tsallis entropy with only this constraint leads to a q-variant of the doubly exponential or Laplace distribution centered at zero. A translated version of the distribution can be written as

$$p(x) = \frac{1}{2\beta} \exp_q \left(-\frac{|x-\mu|}{(2-q)\beta} \right). \quad (8)$$

As $q \to 1$, we retrieve the exponential, Gaussian and Laplace distributions as special cases of (3), (6) and (8), respectively. The above distributions can be extended to a multi-dimensional setting in a way similar to Gaussian and Laplacian distributions, by incorporating 2-norm and 1-norm in (6) and (8), respectively.

3 Proposed Kernels

Based on the above discussion, we define the q-Gaussian kernel $G_q : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, for a given $q \in \mathbb{R}$, as

$$G_q(x, y) = \exp_q \left(-\frac{\|x-y\|_2^2}{(3-q)\sigma^2} \right) \text{ for all } x, y \in \mathcal{X}, \quad (9)$$

where $\mathcal{X} \subset \mathbb{R}^N$ is the input space, and $q, \sigma \in \mathbb{R}$ are two parameters controlling the behavior of the kernel, satisfying the conditions $q \neq 1$, $q \neq 3$ and $\sigma \neq 0$. For $1 < q < 3$, the term inside the bracket is non-negative and hence, the kernel is of the form

$$G_q(x, y) = \left(1 + \frac{(q-1)}{(3-q)\sigma^2} \|x-y\|_2^2 \right)^{\frac{1}{1-q}}, \quad (10)$$

where $\|.\|_2$ is the Euclidean norm. On similar lines, we use (8) to define the q-Laplacian kernel $L_q : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$$L_q(x, y) = \exp_q \left(-\frac{\|x-y\|_1}{(2-q)\beta} \right) \text{ for all } x, y \in \mathcal{X}, \quad (11)$$

where $\|.\|_1$ is the 1-norm, and $q, \beta \in \mathbb{R}$ satisfy the conditions $q \neq 1$, $q \neq 2$ and $\beta > 0$. As before, for $1 < q < 2$, the kernel can be written as

$$L_q(x, y) = \left(1 + \frac{(q-1)}{(2-q)\beta} \|x-y\|_1 \right)^{\frac{1}{1-q}}. \quad (12)$$

Due to the power-law tail of the Tsallis distributions for $q > 1$, in case of the above kernels, similarity decreases at a slower rate than the Gaussian and Laplacian kernels with increasing distance. The rate of decrease in similarity

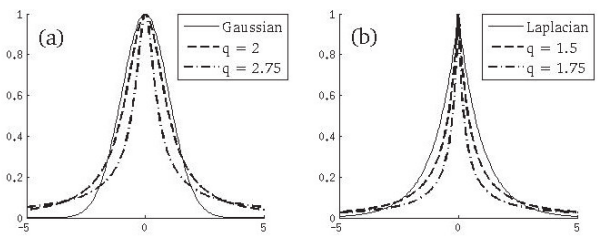

Figure 1: Example plots for (a) q-Gaussian and (b) q-Laplacian kernels with $\sigma = \beta = 1$.

is controlled by the parameter q, and leads to better performance in some machine learning tasks, as shown later. Figure 1 shows how the similarity decays for both q-Gaussian and q-Laplacian kernels in the one-dimensional case. It can be seen that as q increases, the initial decay becomes more rapid, while towards the tails, the decay becomes slower.

We now show that for certain values of q, the proposed kernels satisfy the property of positive definiteness, which is essential for them to be useful in learning theory. Berg, Christensen, and Ressel (1984) have shown that for any symmetric kernel function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, there exists a mapping $\Phi : \mathcal{X} \mapsto \mathcal{H}$, $\mathcal{H}$ being a higher dimensional space, such that $K(x, y) = \Phi(x)^T \Phi(y)$, for all $x, y \in \mathcal{X}$ if and only if K is positive definite (p.d.), *i.e.*, given any set of points $\{x_1, x_2, \ldots, x_n\} \subset \mathcal{X}$, the $n \times n$ matrix $\mathbb{K}$, such that $\mathbb{K}_{ij} = K(x_i, x_j)$, is positive semi-definite.

We first state some of the results presented in (Berg, Christensen, and Ressel 1984), which are required to prove positive definiteness of the proposed kernel.

Lemma 1. *For a p.d. kernel $\varphi : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, $\varphi \geqslant 0$, the following conditions are equivalent:*

1. *$-\log \varphi$ is negative definite (n.d.), and*
2. *φ^t is p.d. for all $t > 0$.*

Lemma 2. *Let $\varphi : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a n.d. kernel, which is strictly positive, then $\frac{1}{\varphi}$ is p.d.*

We state the following proposition, which is a general result providing a method to generate p.d. power-law kernels, given that their exponential counterpart is p.d.

Proposition 3. *Given a p.d. kernel $\varphi : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ of the form $\varphi(x, y) = \exp\big(-f(x, y)\big)$, where $f(x, y) \geqslant 0$ for all $x, y \in \mathcal{X}$, the kernel $\phi : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ given by*

$$\phi(x, y) = \big(1 + cf(x, y)\big)^k, \quad \text{for all } x, y \in \mathcal{X}, \tag{13}$$

is p.d., provided the constants c and k satisfy the conditions $c > 0$ and $k < 0$.

Proof. Since, φ is p.d., it follows from Lemma 1 that the kernel $f = -\log \varphi$ is n.d. Thus, for any $c > 0$, $(1 + cf)$ is n.d., and as $f \geqslant 0$, we can say $(1 + cf)$ is strictly positive. So, application of Lemma 2 leads to the fact that $\frac{1}{(1+cf)}$ is p.d. Finally, using Lemma 1, we can claim $(1 + cf)^k$ is p.d. for all $c > 0$ and $k < 0$. □

From Proposition 3 and positive definiteness of Gaussian and Laplacian kernels, we can show that the proposed q-Gaussian and q-Laplacian kernels are p.d. for certain ranges of q. However, strikingly, it turns out that over this range, the kernels exhibit power-law behavior.

Corollary 4. *For $1 < q < 3$, the q-Gaussian kernel, as defined in* (10)*, is positive definite.*

Corollary 5. *For $1 < q < 2$, the q-Laplacian kernel, as defined in* (12)*, is positive definite for all $\beta > 0$.*

Now, we show that some of the popular kernels can be obtained as special cases of the proposed kernels. The Gaussian kernel is defined as

$$\psi_1(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \tag{14}$$

where $\sigma \in \mathbb{R}$, $\sigma > 0$. We can retrieve the Gaussian kernel (14) when $q \to 1$ in the q-Gaussian kernel (10). The Rational Quadratic kernel is of the form

$$\psi_2(x, y) = \left(1 - \frac{\|x - y\|_2^2}{\|x - y\|_2^2 + c}\right), \tag{15}$$

where $c \in \mathbb{R}$, $c > 0$. Substituting $q = 2$ in (10), we obtain (15) with $c = \sigma^2$. The Laplacian kernel is defined as

$$\psi_3(x, y) = \exp\left(-\frac{\|x - y\|_1}{\sigma}\right), \tag{16}$$

where $\sigma \in \mathbb{R}$, $\sigma > 0$. We can retrieve (16) as $q \to 1$ in the q-Laplacian kernel (12).

4 Reproducing Kernel Hilbert Space

As discussed earlier, kernels map the data points to a higher dimensional feature space, also called the Reproducing Kernel Hilbert Space (RKHS) that is unique for each positive definite kernel (Aronszajn 1950). The significance of RKHS for support vector kernels using Bochner's theorem (Bochner 1959), which provides a RKHS in Fourier space for translation invariant kernels, is stated in (Smola, Schölkopf, and Müller 1998). Other approaches also exist that lead to explicit description of the Gaussian kernel (Steinwart, Hush, and Scovel 2006), but this approach does not work for the proposed kernels as Taylor series expansion of the q-exponential function (4) does not converge for $q > 1$. So, we follow the Bochner's approach.

We state Bochner's theorem, and then use the method presented in (Hofmann, Schölkopf, and Smola 2008) to show how it can be used to construct the RKHS for a p.d. kernel.

Theorem 6 (Bochner). *A continuous kernel $\varphi(x, y) = \varphi(x - y)$ on $\mathbb{R}^d$ is positive definite if and only if $\varphi(t)$ is the Fourier transform of a non-negative measure, i.e., there exists $\rho \geqslant 0$ such that $\rho(\omega)$ is the inverse Fourier transform of $\varphi(t)$.*

Then, the RKHS of the kernel φ is given by

$$\mathcal{H}_\varphi = \left\{ f \in L^2(\mathbb{R}) \middle| \int_{-\infty}^{\infty} \frac{|\hat{f}(\omega)|^2}{\rho(\omega)} \mathrm{d}\omega < \infty \right\} \tag{17}$$

with the inner product defined as

$$\langle f, g \rangle_\varphi = \int_{-\infty}^{\infty} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\rho(\omega)} \mathrm{d}\omega, \tag{18}$$

where $\hat{f}(\omega)$ is the Fourier transform of $f(t)$ and $L^2(\mathbb{R})$ is set of all functions on $\mathbb{R}$, square integrable with respect to the Lebesgue measure.

It must be noted here that in our case, the existence and non-negativity of the inverse Fourier transform ρ is obvious due to the positive definiteness of the proposed kernels (Corollaries 4 and 5). Hence, to describe the RKHS it is enough to determine an expression for ρ for both the kernels.

We define the functions corresponding to the q-Gaussian and q-Laplacian kernels, respectively, as

$$\varphi_G(t) = \left(1 + \frac{(q-1)}{(3-q)\sigma^2}\sum_{j=1}^{N} t_j^2\right)^{\frac{1}{1-q}}, \quad q \in (1,3), \tag{19}$$

and

$$\varphi_L(t) = \left(1 + \frac{(q-1)}{(2-q)\beta}\sum_{j=1}^{N} |t_j|\right)^{\frac{1}{1-q}}, \quad q \in (1,2), \tag{20}$$

where $\beta, \sigma \in \mathbb{R}$, $\beta > 0$ and $t = (t_1, \ldots, t_N) \in \mathbb{R}^N$. We derive expressions for their inverse Fourier transforms $\rho_G(\omega)$ and $\rho_L(\omega)$, respectively. For this, we require a technical result (Gradshteyn and Ryzhik 1994, Eq. 4.638(3)), which is stated in the following lemma.

Lemma 7. *Let $s \in (0,\infty)$ and $p_i, q_i, r_i \in (0,\infty)$ for $i = 1, 2, \ldots, N$ be constants, then the N-dimensional integral*

$$\int_0^\infty \int_0^\infty \cdots \int_0^\infty \frac{\prod_{i=1}^N x_i^{p_i-1}}{\left(1 + \sum_{i=1}^N (r_i x_i)^{q_i}\right)^s} \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_N$$
$$= \frac{\Gamma\left(s - \sum_{i=1}^N \frac{p_i}{q_i}\right)}{\Gamma(s)} \prod_{i=1}^N \left(\frac{\Gamma\left(\frac{p_i}{q_i}\right)}{q_i r_i^{p_i q_i}}\right) .$$

We now derive the inverse Fourier transforms. We prove the result for Proposition 8. The proof of Proposition 9 proceeds similarly.

Proposition 8. *The inverse Fourier transform for $\varphi_G(t)$ is given by*

$$\rho_G(\omega) = \frac{1}{\left(\frac{\sqrt{2}(q-1)}{(3-q)\sigma^2}\right)^N \Gamma\left(\frac{1}{q-1}\right)} \times$$
$$\sum_{b=0}^{\infty} \frac{(-1)^b}{b!} \Gamma\left(\frac{1}{q-1} - \frac{N}{2} - b\right) \left(\frac{(3-q)\sigma^2 \|\omega\|_2}{2(q-1)}\right)^{2b}. \tag{21}$$

Proof. By definition,

$$\rho_G(\omega) = (2\pi)^{-N/2} \int_{\mathbb{R}^N} \exp(\mathrm{i}t \cdot \omega) \varphi_G(t) \mathrm{d}t . \tag{22}$$

Expanding the exponential term, we have

$$\exp(\mathrm{i}t \cdot \omega) = \prod_{j=1}^N \left(\cos(t_j\omega_j) + \mathrm{i}\sin(t_j\omega_j)\right) .$$

Since, both $\cos(t_j\omega_j)$ are $\varphi_G(t)$ are even functions for every t_j, while $\sin(t_j\omega_j)$ is an odd function, hence integrating over $\mathbb{R}^N$, all terms with a sin component become zero. Further, the remaining term is odd, and hence, the integral is same in every orthant. So the expression reduces to

$$\rho_G(\omega) =$$
$$\left(\frac{2}{\pi}\right)^{\frac{N}{2}} \int_0^\infty \ldots \int_0^\infty \left(1 + c\sum_{j=1}^N t_j^2\right)^{\frac{1}{1-q}} \prod_{j=1}^N \cos(t_j\omega_j) \mathrm{d}t_1 \ldots \mathrm{d}t_N \tag{23}$$

where $c = \frac{(q-1)}{(3-q)\sigma^2}$. Each of the cosine term can be expanded in form of an infinite series as

$$\cos(t_j\omega_j) = \sum_{m_j=0}^{\infty} (-1)^{m_j} \frac{\omega_j^{2m_j} t_j^{2m_j}}{(2m_j)!} .$$

Substituting in (23) and using Lemma 7, we obtain

$$\rho_G(\omega) = \frac{1}{\left(\sqrt{2\pi}c\right)^N \Gamma\left(\frac{1}{q-1}\right)} \times$$
$$\sum_{m_1,..,m_N=0}^{\infty} \left(-\frac{1}{c^2}\right)^{\sum_{j=1}^N m_j} \Gamma\left(\frac{1}{q-1} - \frac{N}{2} - \sum_{j=1}^N m_j\right) g(\omega) \tag{24}$$

where

$$g(\omega) = \prod_{j=1}^N \frac{\omega_j^{2m_j} \Gamma\left(m_j + \frac{1}{2}\right)}{(2m_j)!} . \tag{25}$$

Using expansion of gamma function for half integers, we can write (25) as

$$g(\omega) = \pi^{N/2} \prod_{j=1}^N \frac{\omega_j^{2m_j}}{4^{m_j} m_j!} . \tag{26}$$

Substituting in (24) and using $b = \sum_{j=1}^N m_j$, we have

$$\rho_G(\omega) = \frac{1}{\left(\sqrt{2}c\right)^N \Gamma\left(\frac{1}{q-1}\right)} \times$$
$$\sum_{b=0}^{\infty} \left(\frac{-1}{4c^2}\right)^b \Gamma\left(\frac{1}{q-1} - \frac{N}{2} - b\right) \sum_{\substack{m_1,\ldots,m_N \\ \sum_{k=1}^N m_k = b}} \frac{\omega_1^{2m_1} \ldots \omega_N^{2m_N}}{m_1! \ldots m_N!} \tag{27}$$

We arrive at the claim by observing that the terms in the inner summation in (27) are similar to terms of multinomial expansion of $\frac{1}{b!}\left(\omega_1^2 + \ldots + \omega_N^2\right)^b$. □

It can be observed that the above result agrees with the fact that inverse Fourier transform of radial functions are radial in nature. We present corresponding result for q-Laplacian kernel.

Proposition 9. *The inverse Fourier transform for $\varphi_L(t)$ is given by*

$$\rho_L(\omega) = \frac{1}{\left(\frac{\sqrt{\pi}(q-1)}{(2-q)\beta\sqrt{2}}\right)^N \Gamma\left(\frac{1}{q-1}\right)} \times$$
$$\sum_{b=0}^{\infty} (-1)^b \Gamma\left(\frac{1}{q-1} - N - 2b\right) \left(\frac{(2-q)\beta}{(q-1)}\right)^{2b} g_b(\omega), \tag{28}$$

where

$$g_b(\omega) = \sum_{\substack{m_1,\ldots,m_N \in \mathbb{N} \\ \sum_{k=1}^N m_k = b}} \omega_1^{2m_1} \omega_2^{2m_2} \ldots \omega_N^{2m_N} \tag{29}$$

with $\omega_1, \ldots, \omega_N$ being the components of ω.

5 Performance Comparison

In this section, we apply the q-Gaussian and q-Laplacian kernels in classification and regression. We provide insights into the behavior of these kernels through examples. We also compare the performance of the kernels for different values of q, and also with the Gaussian, Laplacian (*i.e.*, when $q \to 1$), and polynomial kernels using various data sets from UCI repository (Frank and Asuncion 2010). The simulations have been performed using LIBSVM (Chang and Lin 2011). Table 1 provides a description of the data sets used. The last few data sets have been used for regression.

	Data Set	Class	Attribute	Instance
1	Acute Inflammations	2	6	120
2	Australian Credit*	2	14	690
3	Blood Transfusion	2	4	748
4	Breast Cancer*	2	9	699
5	Iris	3	4	150
6	Mammographic Mass	2	5	830
7	Statlog (Heart)*	2	13	270
8	Tic-Tac-Toe	2	9	958
9	Vertebral Column	3	6	310
10	Wine*	3	13	178
11	Auto MPG	–	8	398
12	Servo	–	4	167
13	Wine Quality (red)	–	12	1599

Table 1: Data Sets (sets marked * have been normalized).

5.1 Kernel SVM

Support Vector Machines (SVMs) are one of the most important class of kernel machines. While linear SVMs, using inner product as similarity measure, are quite common, other variants using various kernel functions, mostly Gaussian, are also used in practice. Use of kernels leads to non-linear separating hyperplanes, which sometimes provide better classification. Now, we formulate a SVM based on the proposed kernels. For the q-Gaussian kernel (10), it leads to an optimization problem with the following dual form:

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \exp_q \left(-\frac{\|x_i - x_j\|_2^2}{(3-q)\sigma^2} \right)$$

subject to $\alpha_i \geqslant 0$, $i = 1, \ldots, n$, and $\sum_{i=1}^{n} \alpha_i y_i = 0$, where, $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ are the training data points and $\{y_1, \ldots, y_n\} \subset \{-1, 1\}$ are the true classes. The optimization problem for the q-Laplacian kernel (12) can be formulated by using $\exp_q \left(-\frac{\|x_i - x_j\|_1}{(2-q)\beta} \right)$ in the above expression.

The two-dimensional example in Figure 2 illustrates the nature of hyperplanes that can be obtained using various kernels. The decision boundaries are more flexible for q-Laplacian and q-Gaussian kernels. Further, viewing the Laplacian and Gaussian kernels as special cases ($q \to 1$), it can be said that increase in the value of q leads to more flexibility of the decision boundaries.

We compare the performance of the proposed kernels with Gaussian and Laplacian kernel SVMs for various values of q. The results of 5-fold cross validation using multiclass

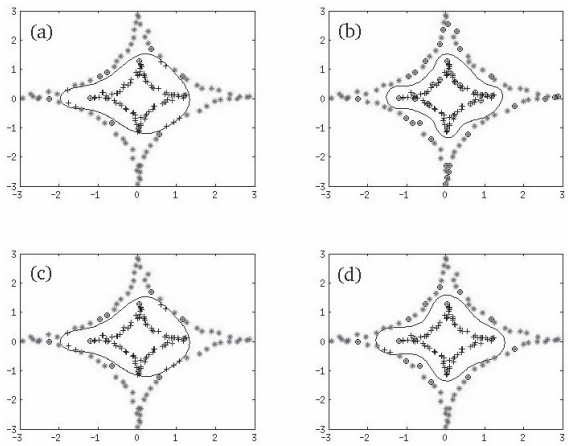

Figure 2: Decision boundaries using (a) Gaussian, (b) q-Gaussian ($q = 2.95$), (c) Laplacian, and (d) q-Laplacian ($q = 1.95$) kernel SVMs.

SVM are shown in Table 2. Further, the power-law nature reminds practitioners of the popular polynomial kernels

$$P_d(x, y) = \left(x^T y + c\right)^d, \qquad \text{for } x, y \in \mathbb{R}^N,$$

where the parameters $c \in (0, \infty)$ and $d \in \mathbb{N}$. Hence, we also provide the accuracies obtained using these kernels. We have fixed particular σ for each data set, and consider β is fixed at $\beta = \sigma^2$. For polynomial kernels, we consider $c = 0$, while d is varied. The best values of q among all Gaussian type and Laplacian type kernels are marked for each data set. In case of the polynomial kernels, we only mark those cases where the best results among these kernels is better or comparable with the best cases of Gaussian or Laplacian types. We note here that in the simulations, the polynomial kernels required normalization of few other data sets as well.

The results indicate the significance of tuning the parameter q. For most cases, the q-Gaussian and q-Laplacian kernels tend to perform better than their exponential counterparts. This can be justified by the flexibility of the separating hyperplane achieved. However, it has been observed (not demonstrated here) that for very high or very low values of σ (or β), the kernels give similar results, which happens because the power-law and the exponential natures cannot be distinguished in such cases. The polynomial kernels, though sometimes comparable to the proposed kernels, rarely improves upon the performance of the power-law kernels.

5.2 Kernel Regression

In linear basis function models for regression, given a set of data points, the output function is approximated as a linear combination of fixed non-linear functions as $f(x) = w_0 + \sum_{j=1}^{M} w_j \phi_j(x)$, where $\{\phi_1(.), \ldots, \phi_M(.)\}$ are the basis functions, usually chosen as $\phi_j(x) = \psi(x, x_j)$, $x_1, \ldots, x_M$ being the given data points, and ψ a p.d. kernel. The constants $\{w_0, w_1, \ldots, w_M\}$ are obtained by minimizing least squared error. Such an optimization can be formulated as an ϵ-Support Vector type problem (Smola and Schölkopf 2004).

Data Sets		1	2	3	4	5	6	7	8	9	10
Parameter ($\sigma = \sqrt{\beta}$)		10	15	5	5	2	25	5	1.5	50	1
Gaussian ($q \to 1$)		86.67	76.96	**77.27**	96.63	97.33	79.28	82.96	**89.46**	**87.10**	97.19
q-Gaussian	$q = 1.25$	86.67	82.46	77.14	96.63	97.33	79.40	82.96	89.25	**87.10**	97.19
	$q = 1.50$	86.67	83.19	76.87	96.93	97.33	79.52	83.33	89.25	**87.10**	97.75
	$q = 1.75$	88.33	**86.38**	76.60	96.93	**98.00**	79.76	83.33	88.94	86.45	97.75
	$q = 2.00$	88.33	85.80	76.74	96.93	**98.00**	79.88	83.33	88.62	86.13	97.75
	$q = 2.25$	89.17	85.51	76.60	96.93	**98.00**	79.40	83.33	87.68	85.48	**98.31**
	$q = 2.50$	91.67	85.51	76.34	96.93	97.33	80.00	**84.07**	85.49	85.48	**98.31**
	$q = 2.75$	98.33	85.51	76.47	**97.22**	96.67	**80.48**	**84.07**	84.34	85.16	**98.31**
	$q = 2.95$	**100**	85.51	75.53	**97.22**	96.67	80.12	82.22	75.99	85.16	97.75
Laplacian ($q \to 1$)		93.33	**86.23**	77.81	97.07	**96.67**	81.69	**83.70**	**94.89**	76.45	**98.88**
q-Laplacian	$q = 1.25$	95.83	85.51	**77.94**	97.07	**96.67**	81.57	**83.70**	92.80	77.42	**98.88**
	$q = 1.50$	97.50	85.51	77.27	97.07	**96.67**	81.81	**83.70**	89.67	77.10	**98.88**
	$q = 1.75$	**100**	85.51	77.14	97.51	**96.67**	82.29	83.33	84.55	78.39	**98.88**
	$q = 1.95$	**100**	85.51	75.67	**97.80**	96.00	**83.73**	82.96	71.09	**86.77**	95.51
Polynomial	$d = 1$ (linear)	**100**	85.51	72.86	97.07	**98.00**	82.17	**83.70**	65.34	85.16	97.19
	$d = 2$	**100**	85.22	76.47	96.19	96.67	**83.86**	80.37	86.53	76.77	96.63
	$d = 5$	**100**	80.72	76.47	95.61	95.33	83.61	74.81	**94.15**	64.84	94.94
	$d = 10$	**100**	76.23	76.47	94.00	94.67	81.69	74.81	88.73	59.03	93.26

Table 2: Percentage of correct classification in kernel SVM using 5-fold cross validation.

The kernels defined in (10) and (12) can also be used in this case as shown in the example in Figure 3, where ϵ-SV regression is used to reconstruct a sine wave from 20 uniformly spaced sampled data points in $[0, \pi]$.

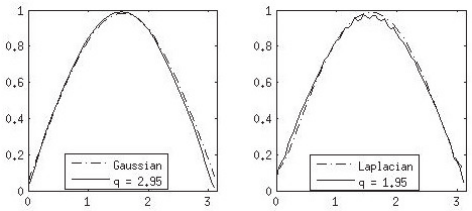

Figure 3: Sine curve obtained by ϵ-SVR using Gaussian, q-Gaussian ($q = 2.95$), Laplacian and q-Laplacian ($q = 1.95$) kernels with $\sigma = \sqrt{\beta} = 2$ and $\epsilon = 0.01$.

The performance of the proposed kernels have been compared with polynomial, Gaussian and Laplacian kernels for various values of q using data sets 11, 12 and 13. The results of 5-fold cross validation using ϵ-SVR ($\epsilon = 0.1$) are shown in Table 3. We fixed particular $\beta = \sigma^2$ for each data set. Though Laplacian kernel seems to outperform its power-law variants, the q-Gaussians dominate the performance of Gaussian kernel. The results further indicate that the error is a relatively smooth function of q, and does not have a fluctuating behavior, though its trend seems to depend on the data. The relative performance of the polynomial kernels is poor.

6 Conclusion

In this paper, we proposed a power-law generalization of Gaussian and Laplacian kernels based on Tsallis distributions. They retain their properties in the classical case as $q \to 1$. Further, due to their power-law nature, the tails of the proposed kernels decay at a slower rate than their exponential counterparts, which in turn broadens the use of these kernels in learning tasks.

Data Sets		11	12	13
Parameter ($\sigma = \sqrt{\beta}$)		1	2	10
Gaussian ($q \to 1$)		11.1630	0.9655	0.4916
q-Gaussian	$q = 1.25$	11.0694	0.9218	0.4883
	$q = 1.50$	10.9674	0.9035	0.4853
	$q = 1.75$	10.8826	**0.8986**	0.4823
	$q = 2.00$	10.7406	0.9005	0.4781
	$q = 2.25$	10.5661	0.9072	0.4734
	$q = 2.50$	**10.4428**	0.9424	0.4661
	$q = 2.75$	10.4796	1.0698	0.4595
	$q = 2.95$	12.2427	1.5439	**0.4419**
Laplacian ($q \to 1$)		**9.7681**	**0.5398**	0.4298
q-Laplacian	$q = 1.25$	10.2052	0.5532	0.4223
	$q = 1.50$	10.9578	0.6055	0.4123
	$q = 1.75$	13.2213	0.7910	0.3961
	$q = 1.95$	17.7303	1.6934	**0.3784**
Polynomial	$d = 1$ (linear)	13.3765	1.9047	0.4357
	$d = 2$	10.5835	2.2740	**0.4268**
	$d = 5$	16.8173	2.3305	0.5485
	$d = 10$	52.4609	2.7358	10.5518

Table 3: Mean Squared Error in kernel SVR.

We showed that the proposed kernels are positive definite for certain range of q, and presented results pertaining to the RKHS of the proposed kernels using Bochner's theorem. We also demonstrated the performance of the proposed kernels in support vector classification and regression.

The power-law behavior was recognized long time back in many problems in the context of statistical analysis. Recently power-law distributions have been studied in machine learning communities. As far as our knowledge, this is the first paper that introduces and studies power-law kernels, leading to the notion of a *"fat-tailed kernel machine"*.

References

Abe, S., and Suzuki, N. 2003. Itineration of the internet over nonequilibrium stationary states in Tsallis statistics. *Physical Review E* 67(016106).

Abe, S., and Suzuki, N. 2005. Scale-free statistics of time interval between successive earthquakes. *Physica A: Statistical Mechanics and its Applications* 350:588–596.

Aronszajn, N. 1950. Theory of reproducing kernels. *Transactions of American Mathematical Society* 68(3):337–404.

Barabási, A. L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509–512.

Berg, C.; Christensen, J. P. R.; and Ressel, P. 1984. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer.

Bochner, S. 1959. *Lectures on Fourier Integral*. Princeton N.J.: Princeton University Press.

Chang, C. C., and Lin, C. J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):27:1–27:27.

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

Dukkipati, A.; Bhatnagar, S.; and Murty, M. N. 2007. On measure-theoretic aspects of nonextensive entropy functionals and corresponding maximum entropy prescriptions. *Physica A: Statistical Mechanics and its Applications* 384(2):758–774.

Frank, A., and Asuncion, A. 2010. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences: http://archive.ics.uci.edu/ml.

Ghoshdastidar, D.; Dukkipati, A.; and Bhatnagar, S. 2012. q-Gaussian based smoothed functional algorithms for stochastic optimization. In *International Symposium on Information Theory*. IEEE.

Gradshteyn, I. S., and Ryzhik, I. M. 1994. *Table of Integrals, Series and Products (5th ed.)*. Elsevier.

Gutenberg, B., and Richter, C. F. 1954. *Seismicity of the Earth and Associated Phenomena (ed. 2)*. Princeton, NJ: Princeton University Press.

Havrda, J., and Charvát, F. 1967. Quantification method of classification processes: Concept of structural a-entropy. *Kybernetika* 3(1):30–35.

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

Hofmann, T.; Schölkopf, B.; and Smola, A. J. 2008. Kernel methods in machine learning. *Annals of Statistics* 36(3):1171–1220.

Jaynes, E. T. 1957. Information theory and statistical mechanics. *The Physical Review* 106(4):620–630.

Kullback, S. 1959. *Information theory and statistics*. N.Y.: John Wiley and Sons.

Martins, A. F. T.; Smith, N. A.; Xing, E. P.; Aguiar, P. M. Q.; and Figueiredo, M. A. T. 2009. Nonextensive information theoretic kernels on measures. *Journal of Machine Learning Research* 10:935–975.

Pareto, V. 1906. Manuale di economica politica. *Societa Editrice Libraria*.

Prato, D., and Tsallis, C. 1999. Nonextensive foundation of Lévy distributions. *Physical Review E.* 60(2):2398–2401.

Rényi, A. 1960. Some fundamental questions of information theory. *MTA III. Oszt. Közl.* 10:251–282. (reprinted in (Turán 1976), pp. 526-552).

Sato, A. H. 2010. q-Gaussian distributions and multiplicative stochastic processes for analysis of multiple financial time series. *Journal of Physics: Conference Series* 201(012008).

Scholköpf, B., and Smola, A. J. 2002. *Learning with Kernels*. MIT Press.

Smola, A. J., and Schölkopf, B. 2004. A tutorial on support vector regression. *Statistics and computing* 14(3):199–222.

Smola, A. J.; Schölkopf, B.; and Müller, K. 1998. The connection between regularization operators and support vector kernels. *Neural Networks* 11:637–649.

Steinwart, I.; Hush, D. R.; and Scovel, C. 2006. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory* 52(10):4635–4643.

Suyari, H. 2004. Generalization of Shannon-Khinchin axioms to nonextensive systems and the uniqueness theorem for the nonextensive entropy. *IEEE Transactions on Information Theory* 50:1783–1787.

Tsallis, C.; Mendes, R. S.; and Plastino, A. R. 1998. The role of constraints within generalized nonextensive statistics. *Physica A: Statistical Mechanics and its Applications* 261(3):534–554.

Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statiscal Physics* 52(1-2):479–487.

Turán, P., ed. 1976. *Selected Papers of Alfréd Rényi*. Budapest: Akademia Kiado.