On Power-law Kernels, corresponding Reproducing Kernel Hilbert Space and Applications

Debarghya Ghoshdastidar and Ambedkar Dukkipati

Department of Computer Science & Automation
Indian Institute of Science, Bangalore - 560012.
Outline

1 Power-laws and heavy-tails for learning
   • Motivation and contributions

2 Tsallis distributions
   • Nonextensive Information Theory
   • Maximum Tsallis entropy distributions

3 Proposed kernels
   • The nonextensive kernels
   • Positive definiteness

4 Reproducing kernel Hilbert space
   • RKHS of translation invariant kernels
   • Inverse Fourier transform of proposed kernels

5 Performance comparison (numerical)
   • Kernel SVM
   • Kernel SVR

6 Concluding remarks
Outline

1. **Power-laws and heavy-tails for learning**
   - Motivation and contributions

2. **Tsallis distributions**
   - Nonextensive Information Theory
   - Maximum Tsallis entropy distributions

3. **Proposed kernels**
   - The nonextensive kernels
   - Positive definiteness

4. **Reproducing kernel Hilbert space**
   - RKHS of translation invariant kernels
   - Inverse Fourier transform of proposed kernels

5. **Performance comparison (numerical)**
   - Kernel SVM
   - Kernel SVR

6. **Concluding remarks**
Ubiquitous power-laws:

- Economics: Pareto principle (Pareto, 1906)
- Data analysis: Zipf’s law (Zipf, 1935)
- Environment: GR law in seismology (Gutenberg and Richter, 1954)
- Statistical physics: Lévy flight (Mandelbrot, 1982)
- Random graphs: Barabási-Albert model (Barabási and Albert, 1999)

Power-laws in learning:

- Correlation models for data assimilation (Gneiting, 1999)
- Robust mixture modeling (MacLachlan and Peel, 2000)
- Language modeling (Goldwater, Griffiths, Johnson, 2011)
Our contribution:

- Kernels exhibiting power-law nature, that generalize existing kernels.

Motivation:

- Popular kernels mostly exhibit exponential nature (Gaussian, Laplace etc.)
- Even heavy-tailed improvements\(^1\) of Gaussian kernel decay exponentially.
- Popular kernels (Gaussian, Laplace, Cauchy) correspond to maximum Shannon entropy distributions.
- Generalizations of Shannon entropy associated with power-law distributions.

---

Outline

1. Power-laws and heavy-tails for learning
   - Motivation and contributions

2. Tsallis distributions
   - Nonextensive Information Theory
   - Maximum Tsallis entropy distributions

3. Proposed kernels
   - The nonextensive kernels
   - Positive definiteness

4. Reproducing kernel Hilbert space
   - RKHS of translation invariant kernels
   - Inverse Fourier transform of proposed kernels

5. Performance comparison (numerical)
   - Kernel SVM
   - Kernel SVR

6. Concluding remarks
Boltzmann-Gibbs-Shannon entropy

For a probability density function $p$, 

$$H(p) = \mathbb{E}_p \left[ \ln \left( \frac{1}{p(X)} \right) \right].$$

- Uncertainty of random variable.
- Maximum entropy principle: choose distribution, satisfying given (moment) constraints, that maximizes entropy.
- Maximum entropy distributions: uniform, exponential, Gaussian, Laplace, Cauchy, $t$-distribution, etc.
Tsallis entropy

\[ H_q(p) = E_p \left[ \ln_q \left( \frac{1}{p(X)} \right) \right] \]

where \( q \)-logarithm \( \ln_q(x) = \frac{x^{1-q-1}}{1-q} \), \( q \in \mathbb{R}, q \neq 1 \).

- Proposed in context of thermodynamics\(^2\).
- Retrieves BGS entropy in the limit of \( q \to 1 \).
- Pseudo-additive in nature, \( i.e. \), for \( X \) and \( Y \) independent,

\[ H_q(X, Y) = H_q(X) + H_q(Y) + (1 - q)H_q(X)H_q(Y) \]

Maximum Tsallis entropy principle

maximize $H_q(p)$
subject to $\langle f_i \rangle_q = \alpha_i$ (constant), \quad i = 1, \ldots, l.$

- $q$-expectation, $\langle f \rangle_q = \frac{\int f(x)p(x)^q \, dx}{\int p(x)^q \, dx}$ is a generalization of usual expectation.

- Above principle results in the $q$-exponential family$^3$,

$$p(x) = \exp_q \left(-\lambda_0 - \sum_{i=1}^{l} \lambda_i f_i(x) \right),$$

where $\exp_q(z) = (1 + (1 - q)z)^{\frac{1}{1-q}}$ is called the $q$-exponential function.

- Exponential family retrieved as $q \to 1$.

---

**q-exponential distribution:**

\[ p(x) = \frac{1}{\mu} \exp_q \left( - \frac{x}{(2-q)\mu} \right) \text{ for } \mu > 0 \]

- Constraint: \( q \)-mean, \( \langle X \rangle_q = \mu \).

**q-Laplace distribution:**

\[ p(x) = \frac{1}{2\beta} \exp_q \left( - \frac{|x-\mu|}{(2-q)\beta} \right) \text{ for } \mu \in \mathbb{R}, \beta > 0 \]

- Constraints: \( \langle X \rangle_q = \mu \) and \( \langle |X - \mu| \rangle_q = \beta \).

**q-Gaussian distribution:**

\[ p(x) = \frac{\Lambda_q}{\sigma} \exp_q \left( - \frac{(x-\mu)^2}{(3-q)\sigma^2} \right) \text{ for } \mu \in \mathbb{R}, \sigma > 0 \]

- Constraints: \( \langle X \rangle_q = \mu \) and \( q \)-variance, \( \langle (X - \mu)^2 \rangle_q = \sigma^2 \).  
  - \( \Lambda_q \) is normalizing constant (function of \( q \) only).
Outline

1. Power-laws and heavy-tails for learning
   - Motivation and contributions

2. Tsallis distributions
   - Nonextensive Information Theory
   - Maximum Tsallis entropy distributions

3. Proposed kernels
   - The nonextensive kernels
   - Positive definiteness

4. Reproducing kernel Hilbert space
   - RKHS of translation invariant kernels
   - Inverse Fourier transform of proposed kernels

5. Performance comparison (numerical)
   - Kernel SVM
   - Kernel SVR

6. Concluding remarks
Facts about Tsallis distributions:

For $q > 1$:
- Distributions exhibit power-law nature (decay slower than exponential counterparts).
- $q$-exponential form in distributions can be simplified as
  \[
  \exp_q(-z) = (1 + (q - 1)z)^{\frac{1}{1-q}} \text{ for all } z \geq 0.
  \]
- Distributions non-zero over entire vector space.
- Normalizing constants do not matter when used as kernels.

For $q < 1$:
- Distributions have bounded support.
- Not suitable for defining positive definite kernels, since compact support kernels cannot be positive definite on every $\mathbb{R}^N$. 
\section*{\textbf{q}-Gaussian kernel}

For $\mathcal{X} \subset \mathbb{R}^N$, the q-Gaussian kernel $G_q : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is

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

for all $x, y \in \mathcal{X}$, where $1 < q < 3$ and $\sigma \neq 0$.

\section*{\textbf{q}-Laplacian kernel}

The q-Gaussian kernel $L_q : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is given by

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

for all $x, y \in \mathcal{X}$, where $1 < q < 2$ and $\beta > 0$. 
Figure: Example plots for (a) $q$-Gaussian and (b) $q$-Laplacian kernels with $\sigma = \beta = 1$.

Connections to existing kernels:
- Gaussian kernel obtained from $q$-Gaussian kernel as $q \rightarrow 1$.
- Rational quadratic and Cauchy kernels are special cases of $q$-Gaussian kernel for $q = 2$.
- Inverse multi-quadratic kernel is a limiting case of $q$-Gaussian kernel as $q \rightarrow 3$ for appropriate $\sigma$.
- Laplacian kernel retrieved from $q$-Laplacian kernel as $q \rightarrow 1$. 
Positive definiteness of power-law kernels

Given a p.d. kernel \( \varphi : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \) of the form
\[
\varphi(x, y) = \exp\left( -f(x, y) \right),
\]
where \( f(x, y) \geq 0 \) for all \( x, y \in \mathcal{X} \), the kernel \( \phi : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \) given by
\[
\phi(x, y) = \left( 1 + cf(x, y) \right)^k,
\]
for all \( x, y \in \mathcal{X} \),
is p.d., provided the constants \( c \) and \( k \) satisfy the conditions \( c > 0 \) and \( k < 0 \).

Consequence:

- The \( q \)-Gaussian kernel is positive definite for all \( \sigma \neq 0 \) and \( 1 < q < 3 \).
- The \( q \)-Laplacian kernel is positive definite for all \( \beta > 0 \) and \( 1 < q < 2 \).
- Here, we use the positive definiteness of Gaussian and Laplacian kernels.
Outline

1. Power-laws and heavy-tails for learning
   - Motivation and contributions

2. Tsallis distributions
   - Nonextensive Information Theory
   - Maximum Tsallis entropy distributions

3. Proposed kernels
   - The nonextensive kernels
   - Positive definiteness

4. Reproducing kernel Hilbert space
   - RKHS of translation invariant kernels
   - Inverse Fourier transform of proposed kernels

5. Performance comparison (numerical)
   - Kernel SVM
   - Kernel SVR

6. Concluding remarks
Bochner’s theorem

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 $\rho \geq 0$ such that $\rho(\omega)$ is the inverse Fourier transform of $\varphi(t)$.

Consequence: RKHS of the kernel $\varphi$ is given by

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

with the inner product defined as

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

where $\hat{f}(\omega)$ is the Fourier transform of $f(t)$. 
- Inverse Fourier transform, $\rho$, completely specifies RKHS.
- Positivity of $\rho$ ensured by positive definiteness of proposed kernels.

**Inverse Fourier transforms**

For $q$-Gaussian kernel,

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

For $q$-Laplacian kernel,

$$\rho_L(\omega) = \frac{1}{\left(\frac{\sqrt{\pi(q-1)}}{(2-q)\beta \sqrt{2}} \right)^N \Gamma\left(\frac{1}{q-1}\right)} \times 
\sum_{b=0}^{\infty} \frac{(-1)^b}{b!} \Gamma\left(\frac{1}{q-1} - N - 2b\right) \left(\frac{(2-q)\beta}{(q-1)}\right)^{2b} \sum_{m_1,\ldots,m_N \in \mathbb{N}} \omega_1^{2m_1} \omega_2^{2m_2} \ldots \omega_N^{2m_N}.$$
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>Power-laws and heavy-tails for learning</strong></td>
</tr>
<tr>
<td>- Motivation and contributions</td>
</tr>
<tr>
<td>2. <strong>Tsallis distributions</strong></td>
</tr>
<tr>
<td>- Nonextensive Information Theory</td>
</tr>
<tr>
<td>- Maximum Tsallis entropy distributions</td>
</tr>
<tr>
<td>3. <strong>Proposed kernels</strong></td>
</tr>
<tr>
<td>- The nonextensive kernels</td>
</tr>
<tr>
<td>- Positive definiteness</td>
</tr>
<tr>
<td>4. <strong>Reproducing kernel Hilbert space</strong></td>
</tr>
<tr>
<td>- RKHS of translation invariant kernels</td>
</tr>
<tr>
<td>- Inverse Fourier transform of proposed kernels</td>
</tr>
<tr>
<td>5. <strong>Performance comparison (numerical)</strong></td>
</tr>
<tr>
<td>- Kernel SVM</td>
</tr>
<tr>
<td>- Kernel SVR</td>
</tr>
<tr>
<td>6. <strong>Concluding remarks</strong></td>
</tr>
</tbody>
</table>
Dual problem for kernel SVM

\[
\minimize_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j),
\]

subject to \( \alpha_i \geq 0, i = 1, \ldots, n, \) and \( \sum_{i=1}^{n} \alpha_i y_i = 0 \)

where \( \{(x_j, y_j)\}, j = 1, \ldots, n \}\subset \mathcal{X} \times \{\pm 1\} \) given training data.

- For numerical comparison, \( K \) considered to be \( q \)-Gaussian, \( q \)-Laplacian and polynomial kernels.
- \( d \)-degree polynomial kernel given by \( P_d(x, y) = (x^T y + c)^d \).
**Figure:** Decision boundaries using (a) Gaussian, (b) $q$-Gaussian ($q = 2.95$), (c) Laplacian, and (d) $q$-Laplacian ($q = 1.95$) kernel SVMs.
Data sets for SVM performance comparison\textsuperscript{4}.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Data Set & Class & Attribute & Instance \\
\hline
1 & Acute Inflammations & 2 & 6 & 120 \\
2 & Australian Credit\textsuperscript{*} & 2 & 14 & 690 \\
3 & Blood Transfusion & 2 & 4 & 748 \\
4 & Breast Cancer\textsuperscript{*} & 2 & 9 & 699 \\
5 & Iris & 3 & 4 & 150 \\
6 & Mammographic Mass & 2 & 5 & 830 \\
7 & Statlog (Heart)\textsuperscript{*} & 2 & 13 & 270 \\
8 & Tic-Tac-Toe & 2 & 9 & 958 \\
9 & Vertebral Column & 3 & 6 & 310 \\
10 & Wine\textsuperscript{*} & 3 & 13 & 178 \\
\hline
\end{tabular}
\caption{Data sets (sets marked * have been normalized).}
\end{table}

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

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

\[ f(x) = w_0 + \sum_{j=1}^{n} w_j K(x, x_j), \]

where \( \{(x_j, f(x_j)), j = 1, \ldots, n\} \) given.

- \( K \) considered to be \( q \)-Gaussian, \( q \)-Laplacian and polynomial kernels.
- Problem tackled as \( \epsilon \)-Support Vector problem.

**Figure**: Reconstructed sine curve from 20 uniformly spaced points using \( \epsilon \)-SVR with Gaussian, \( q \)-Gaussian (\( q = 2.95 \)), Laplacian and \( q \)-Laplacian (\( q = 1.95 \)) kernels with \( \sigma = \sqrt{\beta} = 2 \) and \( \epsilon = 0.01 \).
### Data Sets

<table>
<thead>
<tr>
<th>Parameter ($\sigma = \sqrt{\beta}$)</th>
<th>Auto MPG</th>
<th>Servo</th>
<th>Wine quality (red)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.1630</td>
<td>0.9655</td>
<td>0.4916</td>
</tr>
</tbody>
</table>

#### Gaussian ($q \rightarrow 1$)

<table>
<thead>
<tr>
<th>$q$</th>
<th>Auto MPG</th>
<th>Servo</th>
<th>Wine quality (red)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>11.0694</td>
<td>0.9218</td>
<td>0.4883</td>
</tr>
<tr>
<td>1.50</td>
<td>10.9674</td>
<td>0.9035</td>
<td>0.4853</td>
</tr>
<tr>
<td>1.75</td>
<td>10.8826</td>
<td>\textbf{0.8986}</td>
<td>0.4823</td>
</tr>
<tr>
<td>2.00</td>
<td>10.7406</td>
<td>0.9005</td>
<td>0.4781</td>
</tr>
<tr>
<td>2.25</td>
<td>10.5661</td>
<td>0.9072</td>
<td>0.4734</td>
</tr>
<tr>
<td>2.50</td>
<td>\textbf{10.4428}</td>
<td>0.9424</td>
<td>0.4661</td>
</tr>
<tr>
<td>2.75</td>
<td>10.4796</td>
<td>1.0698</td>
<td>0.4595</td>
</tr>
<tr>
<td>2.95</td>
<td>12.2427</td>
<td>1.5439</td>
<td>\textbf{0.4419}</td>
</tr>
</tbody>
</table>

#### Laplacian ($q \rightarrow 1$)

<table>
<thead>
<tr>
<th>$q$</th>
<th>Auto MPG</th>
<th>Servo</th>
<th>Wine quality (red)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>10.2052</td>
<td>0.5532</td>
<td>0.4223</td>
</tr>
<tr>
<td>1.50</td>
<td>10.9578</td>
<td>0.6055</td>
<td>0.4123</td>
</tr>
<tr>
<td>1.75</td>
<td>13.2213</td>
<td>0.7910</td>
<td>0.3961</td>
</tr>
<tr>
<td>1.95</td>
<td>17.7303</td>
<td>1.6934</td>
<td>\textbf{0.3784}</td>
</tr>
</tbody>
</table>

#### Polynomial

<table>
<thead>
<tr>
<th>$d$</th>
<th>Auto MPG</th>
<th>Servo</th>
<th>Wine quality (red)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (linear)</td>
<td>13.3765</td>
<td>1.9047</td>
<td>0.4357</td>
</tr>
<tr>
<td>2</td>
<td>10.5835</td>
<td>2.2740</td>
<td>\textbf{0.4268}</td>
</tr>
<tr>
<td>5</td>
<td>16.8173</td>
<td>2.3305</td>
<td>0.5485</td>
</tr>
<tr>
<td>10</td>
<td>52.4609</td>
<td>2.7358</td>
<td>10.5518</td>
</tr>
</tbody>
</table>

**Table:** Mean Squared Error in kernel SVR.
Outline

1. Power-laws and heavy-tails for learning
   - Motivation and contributions

2. Tsallis distributions
   - Nonextensive Information Theory
   - Maximum Tsallis entropy distributions

3. Proposed kernels
   - The nonextensive kernels
   - Positive definiteness

4. Reproducing kernel Hilbert space
   - RKHS of translation invariant kernels
   - Inverse Fourier transform of proposed kernels

5. Performance comparison (numerical)
   - Kernel SVM
   - Kernel SVR

6. Concluding remarks
Concluding remarks

- Proposed positive definite power-law generalizations of Gaussian and Laplacian kernels.
- Studied associated RKHS for the kernels.
- Numerical illustrations indicate improvements in performance for power-law kernels.
- To summarize, this work introduces notion of power-law kernels, and demonstrates its importance.

Further directions of work:

- Learning the optimal power-law nature ($q$) using theoretical methods similar to those for learning spread parameters.
- Combinations of proposed kernels in a multiple kernel learning framework.
- A theoretical comparison of RKHS between power-law and exponential kernels.
- Positive definiteness properties of similar kernels for $q < 1$. 