Nonextensive Entropic Kernels

André Martins¹,³  Pedro Aguiar²  Mário Figueiredo³  
Noah Smith¹  Eric Xing¹

¹Language Technologies Institute  
Carnegie Mellon University  
Pittsburgh, PA, USA

²Instituto de Sistemas e Robótica  
Instituto Superior Técnico  
Lisboa, Portugal

³Instituto de Telecomunicações  
Instituto Superior Técnico  
Lisboa, Portugal
Summary

1. Outline
2. Kernels
3. Shannon, Rényi, and Tsallis entropies
4. Jensen differences and divergences
5. Jensen $q$-differences
6. Jensen-Tsallis kernels
7. Experiments
8. Conclusions
Outline

1. Outline
2. Kernels
3. Shannon, Rényi, and Tsallis entropies
4. Jensen differences and divergences
5. Jensen $q$-differences
6. Jensen-Tsallis kernels
7. Experiments
8. Conclusions
We want to classify structured objects (strings, trees, graphs, ...)

This work: A new family of kernels between distributions

Grounded on nonextensive (Tsallis) information theory

Contains known kernels as particular cases

Experiments in text classification
We want to classify structured objects (strings, trees, graphs, ...)

- **Generative** methods allow modeling data generation
- **Discriminative** methods directly discriminate data

This work: A new family of kernels between distributions

- Grounded on nonextensive (Tsallis) information theory
- Contains known kernels as particular cases
- Experiments in text classification
We want to classify **structured objects** (strings, trees, graphs, ...)  
- **Generative** methods allow modeling data generation  
- **Discriminative** methods directly discriminate data

How to get the best of both worlds?
We want to classify structured objects (strings, trees, graphs, ...)
- Generative methods allow modeling data generation
- Discriminative methods directly discriminate data

How to get the best of both worlds?
- Represent objects $x, y$ as probability distributions $p_x(.), p_y(.)$
- Use a kernel between distributions, $k(p_x, p_y)$
We want to classify **structured objects** (strings, trees, graphs, ...)  
- **Generative** methods allow modeling data generation  
- **Discriminative** methods directly discriminate data

**How to get the best of both worlds?**  
- Represent objects $x$, $y$ as probability distributions $p_x(\cdot), p_y(\cdot)$  
- Use a **kernel between distributions**, $k(p_x, p_y)$

**This work**: A new family of kernels between distributions
We want to classify **structured objects** (strings, trees, graphs, ...)

- **Generative** methods allow modeling data generation
- **Discriminative** methods directly discriminate data

How to get the best of both worlds?

- Represent objects $x, y$ as probability distributions $p_x(\cdot), p_y(\cdot)$
- Use a **kernel between distributions**, $k(p_x, p_y)$

This work: A new family of kernels between distributions

- Grounded on **nonextensive (Tsallis) information theory**
- Contains known kernels as particular cases
- Experiments in text classification
Nonextensive Entropic Kernels

Outline

1. Outline
2. Kernels
3. Shannon, Rényi, and Tsallis entropies
4. Jensen differences and divergences
5. Jensen $q$-differences
6. Jensen-Tsallis kernels
7. Experiments
8. Conclusions
Theorem: \( k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is a positive definite (pd) kernel iff there is a feature space \( \mathcal{F} \) and a map \( \Phi : \mathcal{X} \to \mathcal{F} \), such that \( k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{F}} \)

A kernel induces a similarity measure
Kernels for structured data

- What if \( \mathcal{X} \) is structured?

- Extract features and use a linear kernel (Joachims, 1997)
- Decompose objects into subparts (convolution kernels, Haussler, 1999)
- Generative approach through Fisher kernel (Jaakkola, 1999)

Our approach: Map each object to a probability distribution, and devise kernels on probability distributions:

\[
\begin{align*}
\mathbf{x} &\mapsto p_x(.), \\
\mathbf{y} &\mapsto p_y(.), \\
\iff K(\mathbf{x}, \mathbf{y}) &\equiv k(p_x, p_y)
\end{align*}
\]
Kernels for structured data

- What if $X$ is structured?
  - Extract features and use a linear kernel (Joachims, 1997)
  - Decompose objects into subparts (convolution kernels, Haussler, 1999)
  - Generative approach through Fisher kernel (Jaakkola, 1999)
Kernels for structured data

- What if $\mathcal{X}$ is structured?
  - Extract features and use a linear kernel (Joachims, 1997)
  - Decompose objects into subparts (convolution kernels, Haussler, 1999)
  - Generative approach through Fisher kernel (Jaakkola, 1999)

- Our approach: Map each object to a probability distribution, and devise kernels on probability distributions:

  \[ x \mapsto p_x(.) \quad y \mapsto p_y(.) \iff K(x, y) \equiv k(p_x, p_y) \]
Kernels on probability distributions

- **Inner product kernels** (Jebara, Kondor, Howard, 2004)
  \[ k_{JKH}(p_1, p_2) \triangleq \langle p_1^\alpha, p_2^\alpha \rangle \]

- (†) **Information geometry** of the multinomial (Lafferty, Lebanon, 2005),
  \[ k_{\text{heat}}(p_1, p_2) \approx \exp (-\lambda d_g^2(p_1, p_2)) \]

- (†) **KL divergence** (Moreno, Ho, Vasconcelos, 2003),
  \[ k_{\text{MHV}}(p_1, p_2) \triangleq \exp (-\lambda (KL(p_1, p_2) + KL(p_2, p_1))) \]

(†) not pd
Kernels on probability distributions (c’ed)

- **Jensen-Shannon (JS) divergence** (Burbea, Rao, 1982; Lin, 1991)

\[
JS(p_1, p_2) \triangleq \frac{1}{2} KL \left( p_1, \frac{p_1 + p_2}{2} \right) + \frac{1}{2} KL \left( p_2, \frac{p_1 + p_2}{2} \right) \\
= H \left( \frac{p_1 + p_2}{2} \right) - \frac{H(p_1) + H(p_2)}{2}
\]

- Replace KL by JS divergence ⇒ pd (Cuturi, Fukumizu, Vert, 2005; Hein, Bousquet, 2005):

\[
k_{CFV}(p_1, p_2) = \exp (-\lambda JS(p_1, p_2)) \\
k_{HB}(p_1, p_2) = \ln 2 - JS(p_1, p_2)
\]

- We subsume some of these kernels by going from classic to nonextensive (Tsallis) information theory!
Outline

1. Outline
2. Kernels
3. Shannon, Rényi, and Tsallis entropies
4. Jensen differences and divergences
5. Jensen $q$-differences
6. Jensen-Tsallis kernels
7. Experiments
8. Conclusions
Shannon entropy (1948)

- Random variable $X \in \mathcal{X} = \{x_1, \ldots, x_n\}$

$$H(X) = -\sum_{i=1}^{n} P(x_i) \ln P(x_i)$$

$$= -\mathbb{E}[\ln P(X)]$$

- **Extensivity:** for $X$ and $Y$ independent,

$$H(X, Y) = H(X) + H(Y)$$

- “Independent systems add their entropies”—cf. Boltzmann-Gibbs entropy in statistical thermodynamics
Rényi entropies (1961)

- A family parameterized by $q \geq 0$,

$$R_q(X) = \frac{1}{1-q} \ln \sum_{i=1}^{n} P(x_i)^q$$

- Shannon’s entropy as a limit:

$$\lim_{q \to 1} R_q(X) = H(X)$$

- Still extensive: for $X$ and $Y$ independent,

$$R_q(X, Y) = R_q(X) + R_q(Y)$$
Tsallis entropies (1988)

- A family parameterized by $q \geq 0$ (the entropic index),

$$S_q(X) = \frac{1}{q-1} \left( 1 - \sum_{i=1}^{n} P(x_i)^q \right)$$

- Shannon’s entropy as a limit:

$$\lim_{q \to 1} S_q(X) = H(X)$$

- Not extensive! For $X$ and $Y$ independent,

$$S_q(X, Y) = S_q(X) + S_q(Y) - (q - 1)S_q(X)S_q(Y)$$

- Nonextensive thermodynamics—claimed to better model some physical phenomena (e.g. long range interactions, heavy-tailed distributions)
Tsallis entropies can be written as:

\[ S_q(X) = -\mathbb{E}_q[\ln_q P(X)] \]
Tsallis entropies

- Tsallis entropies can be written as:
  \[ S_q(X) = -E_q[\ln_q P(X)] \]

- \( q \)-expectation:
  \[ E_q[f(X)] = \sum_i P(x_i)^q f(x_i), \quad \lim_{q \to 1} E_q[f(X)] = E[f(X)] \]

- \( q \)-logarithm:
  \[ \ln_q(x) = \frac{x^{1-q} - 1}{1 - q}, \quad \lim_{q \to 1} \ln_q(x) = \ln(x) \]
Tsallis entropies

Entropies of a Bernoulli

- Shannon
- Rényi $q=1/2$
- Rényi $q=2$
- Tsallis $q=1/2$
- Tsallis $q=2$
Tsallis entropies

- **Joint Tsallis entropy:**
  \[ S_q(X, Y) = -\mathbb{E}_q[\ln_q P(X, Y)] \]

- **Conditional Tsallis entropy:**
  \[ S_q(X|Y) = -\mathbb{E}_q[\ln_q P(X|Y)] \]

- **Chain rule:**
  \[ S_q(X, Y) = S_q(X|Y) + S_q(Y) \]

- **Tsallis mutual information** (Furuichi, 2006):
  \[ I_q(X; Y) = S_q(X) - S_q(X|Y) \]
Jensen differences

- **Jensen’s inequality**: for a concave function $f$ (e.g., Shannon, Rényi, or Tsallis entropies),

$$f(\mathbb{E}[Z]) \geq \mathbb{E}[f(Z)]$$

- **Weighted Jensen-Shannon divergence** of $m$ distributions

$$J^\pi_H(p_1, \ldots, p_m) \triangleq H \left( \sum_{j=1}^{m} \pi_j p_j \right) - \sum_{j=1}^{m} \pi_j H(p_j)$$

$$= I(X; Y),$$

where $Y \sim (\pi_1, \ldots, \pi_m)$ and $P(X|Y = j) = p_j$
Jensen differences

- **Jensen-Rényi divergences:**

  \[
  J_{R_q}^{\pi}(p_1, \ldots, p_m) \triangleq R_q \left( \sum_{j=1}^{m} \pi_j \ p_j \right) - \sum_{j=1}^{m} \pi_j R_q(p_j)
  \]

- **Jensen-Tsallis divergences:**

  \[
  J_{S_q}^{\pi}(p_1, \ldots, p_m) \triangleq S_q \left( \sum_{j=1}^{m} \pi_j \ p_j \right) - \sum_{j=1}^{m} \pi_j S_q(p_j)
  \]

No mutual information interpretation!
Jensen differences

- **Jensen-Rényi divergences:**
  \[
  J^\pi_{R_q}(p_1, \ldots, p_m) \triangleq R_q \left( \sum_{j=1}^{m} \pi_j p_j \right) - \sum_{j=1}^{m} \pi_j R_q(p_j)
  \]

- **Jensen-Tsallis divergences:**
  \[
  J^\pi_{S_q}(p_1, \ldots, p_m) \triangleq S_q \left( \sum_{j=1}^{m} \pi_j p_j \right) - \sum_{j=1}^{m} \pi_j S_q(p_j)
  \]

- **No mutual information interpretation!**
Outline

1. Outline
2. Kernels
3. Shannon, Rényi, and Tsallis entropies
4. Jensen differences and divergences
5. Jensen $q$-differences
6. Jensen-Tsallis kernels
7. Experiments
8. Conclusions
Jensen $q$-differences

- $f : \mathcal{X} \to \mathbb{R}$ is $q$-convex iff, for any $x, y \in \mathcal{X}$ and $\lambda \in [0, 1]$,
  $$f(\lambda x + (1 - \lambda)y) \leq \lambda^q f(x) + (1 - \lambda)^q f(y)$$

- $q$-Jensen's inequality: for $f$ $q$-concave (e.g., $f = S_q$, $q \geq 1$),
  $$f(\mathbb{E}[Z]) \geq \mathbb{E}_q[f(Z)]$$

- Jensen-Tsallis $q$-difference:
  $$T_q^\pi(p_1, \ldots, p_m) \triangleq S_q \left( \sum_{j=1}^{m} \pi_j p_j \right) - \sum_{j=1}^{m} \pi_j^q S_q(p_j) = I_q(X; Y)$$
Let’s focus on \( m = 2, \pi = (\frac{1}{2}, \frac{1}{2}) \) (“fair coin”).

Balanced JS divergence

\[
JS(p_1, p_2) \triangleq H\left(\frac{p_1 + p_2}{2}\right) - \frac{H(p_1) + H(p_2)}{2}.
\]

Balanced JT \( q \)-difference

\[
T_q(p_1, p_2) \triangleq S_q\left(\frac{p_1 + p_2}{2}\right) - \frac{S_q(p_1) + S_q(p_2)}{2^q}.
\]
Jensen-Tsallis $q$-differences

Jensen Tsallis $q$–Difference to a fixed Bernoulli ($p_0=0.3$)
Outline

1. Outline
2. Kernels
3. Shannon, Rényi, and Tsallis entropies
4. Jensen differences and divergences
5. Jensen $q$-differences
6. Jensen-Tsallis kernels
7. Experiments
8. Conclusions
Jensen-Tsallis kernels

- **Definition:**

\[ k_q(p_1, p_2) \triangleq \ln_q(2) - T_q(p_1, p_2) \]

\[ = \frac{1}{2q(q - 1)} \sum_i \left( (p_{1i} + p_{2i})^q - p_{1i}^q - p_{2i}^q \right) \]

- These kernels can be extended to unnormalized distributions (see paper)

- **Proposition:** the kernel \( k_q \) is pd for \( q \in [0, 2] \).
Special cases

- $q = 0$: Boolean kernel,
  \[ k_{\text{Bool}}(p_1, p_2) = \| p_1 \odot p_2 \|_0 \]

- $q = 1$: JS kernel (Hein & Bousquet, 2005),
  \[ k_{\text{JS}}(p_1, p_2) = \ln(2) - \text{JS}(p_1, p_2) \]

- $q = 2$: linear kernel,
  \[ k_{\text{lin}}(p_1, p_2) = \frac{1}{2} \langle p_1, p_2 \rangle \]

**Corollary**: All these kernels are pd.
Outline

1. Outline
2. Kernels
3. Shannon, Rényi, and Tsallis entropies
4. Jensen differences and divergences
5. Jensen $q$-differences
6. Jensen-Tsallis kernels
7. Experiments
8. Conclusions
Text classification experiments

- WebKB dataset (student vs faculty homepages)
  - 400 documents for training, 450 for testing
- Each document mapped into bag-of-words unigram model
- Baselines: linear kernel with $\ell_2$ normalization (Joachims, 1999) and heat kernel (not pd, Lafferty, Lebanon, 2004)
Text classification experiments (c’ed)

- WebKB dataset (student vs faculty homepages)
  - 400 documents for training, 450 for testing
- Each document mapped into bag-of-5-grams (string kernel)
- Baselines: \( p \)-spectrum kernel (Leslie, 2002) and all-substrings kernel (Vishwanathan, Smola, 2003) with \( \ell_2 \) normalization

![Graph showing average error rate vs entropic index q for different kernels.](image-url)
Outline

1. Outline
2. Kernels
3. Shannon, Rényi, and Tsallis entropies
4. Jensen differences and divergences
5. Jensen $q$-differences
6. Jensen-Tsallis kernels
7. Experiments
8. Conclusions
Conclusions and future work

- A new family of kernels on distributions
- based on nonextensive (Tsallis) information theory
- contains some previously known kernels
- defined on possibly unnormalized distributions (see paper)
- shown to be positive definite
- proofs, kernels between stationary stochastic processes, etc.:
- preliminary experiments on text classification
- future work: exploit nonextensivity in other problems
- future work: when is $q < 1$ best? When is $q > 1$ best?
- future work: multi-kernels