Total Jensen divergences: Definition, Properties and k-Means++ Clustering

Frank Nielsen[1] Richard Nock[2]

www.informationgeometry.org

[1]Sony Computer Science Laboratories, Inc.

[2]UAG-CEREGMIA

September 2013

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Divergences: Distortion measures

F a smooth convex function, the generator.

- Skew Jensen divergences:

$$J'_\alpha(p:q) = \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q),$$
$$= (F(p)F(q))_\alpha - F((pq)_\alpha),$$

where $(pq)_\gamma = \gamma p + (1-\gamma)q = q + \gamma(p-q)$ and $(F(p)F(q))_\gamma = \gamma F(p) + (1-\gamma)F(q) = F(q) + \gamma(F(p) - F(q))$.

- Bregman divergences:

$$B(p:q) = F(p) - F(q) - \langle p-q, \nabla F(q)\rangle,$$

$$\lim_{\alpha\rightarrow 0} J_\alpha(p:q) = B(p:q),$$
$$\lim_{\alpha\rightarrow 1} J_\alpha(p:q) = B(q:p).$$

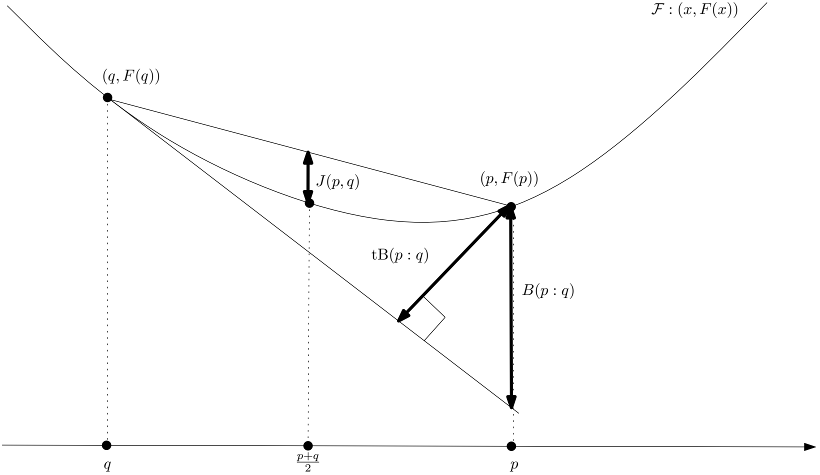
- Statistical Bhattacharrya divergence:

$$\mathrm{Bhat}(p_1:p_2) = -\log \int p_1(x)^\alpha p_2(x)^{1-\alpha} \mathrm{d}\nu(x) = J'_\alpha(\theta_1:\theta_2)$$

for exponential families [5].

Geometrically designed divergences

Plot of the convex generator F.



© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Total Bregman divergences

Conformal divergence, conformal factor ρ:

$$D'(p:q) = \rho(p,q) D(p:q)$$

plays the rôle of "regularizer" [8]

Invariance by rotation of the axes of the design space

$$\begin{aligned} \mathrm{tB}(p:q) &= \frac{B(p:q)}{\sqrt{1+\langle \nabla F(q), \nabla F(q)\rangle}} = \rho_B(q) B(p:q), \\ \rho_B(q) &= \frac{1}{\sqrt{1+\langle \nabla F(q), \nabla F(q)\rangle}}. \end{aligned}$$

Total squared Euclidean divergence:

$$tE(p,q) = \frac{1}{2} \frac{\langle p-q, p-q\rangle}{\sqrt{1+\langle q, q\rangle}}.$$

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Total Jensen divergences

$$\mathrm{tB}(p:q) = \rho_B(q)B(p:q), \quad \rho_B(q) = \sqrt{\frac{1}{1+\langle \nabla F(q), \nabla F(q)\rangle}}$$

$$\mathrm{tJ}_\alpha(p:q) = \rho_J(p,q)J_\alpha(p:q), \quad \rho_J(p,q) = \sqrt{\frac{1}{1+\frac{(F(p)-F(q))^2}{\langle p-q, p-q\rangle}}}$$

Jensen-Shannon divergence, square root is a metric [2]:

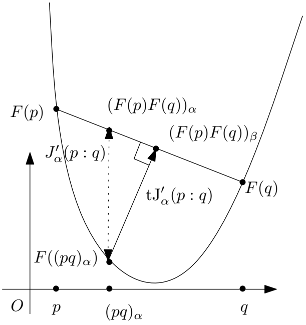
$$\mathrm{JS}(p,q) = \frac{1}{2}\sum_{i=1}^d p_i \log \frac{2p_i}{p_i+q_i} + \frac{1}{2}\sum_{i=1}^d q_i \log \frac{2q_i}{p_i+q_i}$$

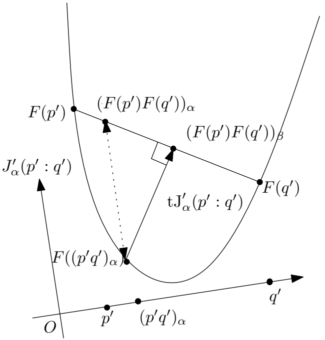
Lemma

The square root of the total Jensen-Shannon divergence is not a metric.

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

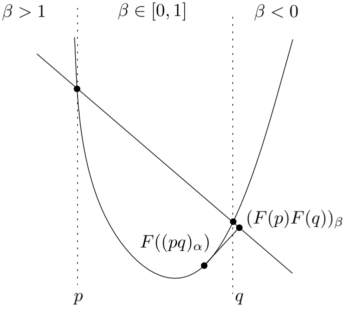
Total Jensen divergence: Illustration

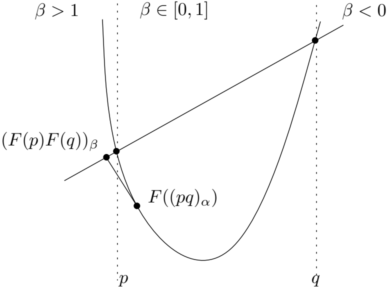




Total Jensen divergence: Illustration

α on graph plot, β on interpolated segment
Two kinds of total Jensen divergences (but one always yields closed-form)





Total Jensen divergences/Total Bregman divergences

Total Jensen is not a generalization of total Bregman.
limit cases $\alpha \in \{0,1\}$, we have:

$$
\begin{aligned}
\lim_{\alpha \rightarrow 0} \mathrm{tJ}_\alpha(p:q) &= \rho_J(p,q) B(p:q) \neq \rho_B(q) B(p:q), \\
\lim_{\alpha \rightarrow 1} \mathrm{tJ}_\alpha(p:q) &= \rho_J(p,q) B(q:p) \neq \rho_B(p) B(q:p),
\end{aligned}
$$

since $\rho_J(p,q) \neq \rho_B(q)$.

Squared chord slope index in ρ_J:

$$
s^2 = \frac{\Delta_F^2}{\|\Delta\|^2} = \frac{\Delta^\top \nabla F(\epsilon) \Delta^\top \nabla F(\epsilon)}{\Delta^\top \Delta} = \langle \nabla F(\epsilon), \nabla F(\epsilon) \rangle = \|\nabla F(\epsilon)\|^2.
$$

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Conformal factor from mean value theorem

When $p \simeq q$, $\rho_J(p,q) \simeq \rho_B(q)$, and the total Jensen divergence tends to the total Bregman divergence for any value of α.

$$\rho_J(p,q) = \frac{1}{\sqrt{1 + \langle \nabla F(\epsilon), \nabla F(\epsilon) \rangle}} = \rho_B(\epsilon),$$

for $\epsilon \in [p,q]$.

For univariate generators, explicitly the value of ϵ:

$$\epsilon = \nabla F^{-1}\left(\frac{\Delta_F}{\Delta}\right) = \nabla F^{*}\left(\frac{\Delta_F}{\Delta}\right),$$

where F^* is the Legendre convex conjugate [5].
Stolarsky mean [7]:

$$\mathrm{tJ}_\alpha(p:q) = \rho_B(\epsilon) J(p:q)$$

Centroids and statistical robustness

Centroids (barycenters) are minimizers of average (weighted) divergences:

$$
\begin{aligned}
L(x; w) &= \sum_{i=1}^{n} w_i \times \mathrm{tJ}_\alpha(p_i : x), \\
c_\alpha &= \arg\min_{x \in \mathcal{X}} L(x; w),
\end{aligned}
$$

- Is it unique?
- Is it robust to outliers [3]?

Iterative convex-concave procedure (CCCP) [5]

Robustness of Jensen centroids (univariate generator)

Theorem

The Jensen centroid is robust for a strictly convex and smooth generator f if $|f'(\frac{p+y}{2})|$ is bounded on the domain $\mathcal{X}$ for any prescribed p.

- Jensen-Shannon: $\mathcal{X} = \mathbb{R}^+$, $f(x) = x \log x - x$,$f'(x) = \log(x)$, $f''(x) = 1/x$.
  $|f'(\frac{p+y}{2})| = |\log \frac{p+y}{2}|$ is unbounded when $y \to +\infty$.
  JS centroid is not robust
- Jensen-Burg: $\mathcal{X} = \mathbb{R}^+$, $f(x) = -\log x$, $f'(x) = -1/x$, $f''(x) = \frac{1}{x^2}$
  $|f'(\frac{p+y}{2})| = |\frac{2}{p+y}|$ is always bounded for $y \in (0, +\infty)$.

$$z(y) = 2p^2 \left(\frac{1}{p} - \frac{2}{p+y} \right)$$

When $y \to \infty$, we have $|z(y)| \to 2p < \infty$.
JB centroid is robust.

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Clustering: No closed-form centroid, no cry!

k-means++ [1] picks up randomly seeds, no centroid calculation.

Algorithm 1: Total Jensen k-means++ seeding

Input: Number of clusters $k \geq 1$;
Let $\mathcal{C} \leftarrow \{h_j\}$ with uniform probability ;
for $i = 2, 3, ..., k$ **do**
  Pick at random $h \in \mathcal{H}$ with probability:

$$\pi_{\mathcal{H}}(h) = \frac{\mathrm{tJ}_\alpha(c_h : h)}{\sum_{y \in \mathcal{H}} \mathrm{tJ}_\alpha(c_y : y)}$$

  where $c_h = \arg\min_{z \in \mathcal{C}} \mathrm{tJ}_\alpha(z : h)$;
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{h\}$;
Output: Set of initial cluster centers $\mathcal{C}$;

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Divergence-based k-means++

Theorem

Suppose there exist some U and V such that, $\forall x, y, z$:

$$
\begin{aligned}
\mathrm{tJ}_\alpha(x:z) &\leq U(\mathrm{tJ}_\alpha(x:y) + \mathrm{tJ}_\alpha(y:z)) \ , \ \textit{(triangular inequality)} \\
\mathrm{tJ}_\alpha(x:z) &\leq V\mathrm{tJ}_\alpha(z:x) \ , \ \textit{(symmetric inequality)}
\end{aligned}
$$

Then the average potential of total Jensen seeding with k clusters satisfies

$$E[\mathrm{tJ}_\alpha] \leq 2U^2(1+V)(2+\log k)\mathrm{tJ}_{\mathrm{opt},\alpha},$$

where $\mathrm{tJ}_{\mathrm{opt},\alpha}$ is the minimal total Jensen potential achieved by a clustering in k clusters.

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Divergence-based k-means++: Two assumptions H

H:

- First, the maximal condition number of the Hessian of F, that is, the ratio between the maximal and minimal eigenvalue (> 0) of the Hessian of F, is upperbounded by K_1.
- Second, we assume the Lipschitz condition on F that $\Delta_F^2 / \langle \Delta, \Delta \rangle \leq K_2$, for some $K_2 > 0$.

Lemma

Assume $0 < \alpha < 1$. *Then, under assumption H, for any* $p, q, r \in \mathcal{S}$, *there exists* $\epsilon > 0$ *such that:*

$$\mathrm{tJ}_\alpha(p:r) \leq \frac{2(1+K_2)K_1^2}{\epsilon} \left(\frac{1}{1-\alpha} \mathrm{tJ}_\alpha(p:q) + \frac{1}{\alpha} \mathrm{tJ}_\alpha(q:r) \right) .$$

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Divergence-based k-means++

Corollary

The total skew Jensen divergence satisfies the following triangular inequality:

$$\mathrm{tJ}_\alpha(p:r) \quad \leq \quad \frac{2(1+K_2)K_1^2}{\epsilon\alpha(1-\alpha)} \left(\mathrm{tJ}_\alpha(p:q) + \mathrm{tJ}_\alpha(q:r)\right) \quad .$$

$$U = \frac{2(1+K_2)K_1^2}{\epsilon}$$

Lemma

Symmetric inequality condition holds for $V = K_1^2(1+K_2)/\epsilon$*, for some* $0 < \epsilon < 1$*.*

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Total Jensen divergences: Recap

Total Jensen divergence = conformal divergence with non-separable double-sided conformal factor.

- Invariant to axis rotation of "design space"
- Equivalent to total Bregman divergences [8, 4] only when $p \simeq q$
- Square root of total Jensen-Shannon divergence is not a metric (square root of total JS is a metric).
- Jensen centroids are not always robust (*e.g.*, Jensen-Shannon centroid)
- Total Jensen k-means++ do not require centroid computations and guaranteed approximation

Interest of conformal divergences in SVM [9] (double-sided separable), in information geometry [6] (flattening).

Thank you.

```
@article{totalJensen-arXiv1309.7109 ,
author="Frank Nielsen and Richard Nock",
title="Total {J}ensen divergences: {D}efinition, Properties and $k$-Means++ Clustering",
year="2013",
eprint="arXiv/1309.7109"
}
```

www.informationgeometry.org

Bibliographic references I

David Arthur and Sergei Vassilvitskii.

k-means++: the advantages of careful seeding.

In *Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Bent Fuglede and Flemming Topsoe.

Jensen-Shannon divergence and Hilbert space embedding.

In *IEEE International Symposium on Information Theory*, pages 31–31, 2004.

F. R. Hampel, P. J. Rousseeuw, E. Ronchetti, and W. A. Stahel.

Robust Statistics: The Approach Based on Influence Functions.

Wiley Series in Probability and Mathematical Statistics, 1986.

Meizhu Liu, Baba C. Vemuri, Shun-ichi Amari, and Frank Nielsen.

Shape retrieval using hierarchical total Bregman soft clustering.

Transactions on Pattern Analysis and Machine Intelligence, 34(12):2407–2419, 2012.

Frank Nielsen and Sylvain Boltz.

The Burbea-Rao and Bhattacharyya centroids.

IEEE Transactions on Information Theory, 57(8):5455–5466, August 2011.

Atsumi Ohara, Hiroshi Matsuzoe, and Shun-ichi Amari.

A dually flat structure on the space of escort distributions.

Journal of Physics: Conference Series, 201(1):012012, 2010.

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.

Bibliographic references II

Kenneth B Stolarsky.

Generalizations of the logarithmic mean.

Mathematics Magazine, 48(2):87–92, 1975.

Baba Vemuri, Meizhu Liu, Shun-ichi Amari, and Frank Nielsen.

Total Bregman divergence and its applications to DTI analysis.

IEEE Transactions on Medical Imaging, pages 475–483, 2011.

Si Wu and Shun-ichi Amari.

Conformal transformation of kernel functions a data dependent way to improve support vector machine classifiers.

Neural Processing Letters, 15(1):59–67, 2002.

© 2013 Frank Nielsen, Sony Computer Science Laboratories, Inc.