TOTAL JENSEN DIVERGENCES: DEFINITION, PROPERTIES AND CLUSTERING

Frank Nielsen

École Polytechnique, France
Sony Computer Science Laboratories Inc., Japan

Richard Nock

NICTA, Australia
ANU, Australia

ABSTRACT

We present a novel class of divergences induced by a smooth convex function called *total Jensen divergences* that are invariant by construction to rotations, a feature inducing a conformal factor on ordinary Jensen divergences. We analyze the relationships between this novel class of total Jensen divergences and the total Bregman divergences. We then define *total Jensen centroids*, analyze their robustness, and prove that the k-means++ initialization that bypasses explicit centroid computations is good enough in practice to guarantee probabilistically a constant approximation factor to the optimal k-means clustering.

Index Terms— Clustering, centroids, k-means++, Jensen-Shannon divergence, Burbea-Rao divergences.

1. INTRODUCTION AND PRIOR WORK

A divergence $D(p : q) \geq 0$ is a smooth distortion measure that quantifies the dissimilarity between any two data points p and q (with $D(p : q) = 0$ iff. $p = q$). A statistical divergence is a divergence between probability (or positive) measures. One motivation to design new divergence families, like the proposed total Jensen divergences, is to elicit some statistical robustness property that allows to bypass the use of costly M-estimators [1].

1.1. Skew Jensen and Bregman divergences

For a strictly convex and differentiable function F, called the *generator* (or potential function), we define a family of parameterized distortion measures with $\alpha \notin \{0, 1\}$ by:

$$J'^{F}_{\alpha}(p : q) = (F(p)F(q))_\alpha - F((pq)_\alpha),$$

where $(pq)_\gamma = \gamma p + (1 - \gamma)q = q + \gamma(p - q)$ and $(F(p)F(q))_\gamma = \gamma F(p) + (1 - \gamma)F(q) = F(q) + \gamma(F(p) - F(q))$. The skew Jensen divergences are *asymmetric* (when $\alpha \neq \frac{1}{2}$) and does *not* satisfy the triangular inequality of metrics. For $\alpha = \frac{1}{2}$, we get a symmetric divergence $J'^{F}_{\frac{1}{2}}(p : q) = J'^{F}_{\frac{1}{2}}(q : p)$, also called Burbea-Rao divergence [2]. It follows from the strict convexity of the generator that $J'^{F}_{\alpha}(p : q) \geq 0$ with equality if and only if $p = q$ (*identity of indiscernibles*). The skew Jensen divergences *may not* be convex divergences: they are convex iff. $F''(x) \geq \frac{1}{2}F''((x + y)/2)$, $\forall x, y \in \mathcal{X}$. Note that the generator may be defined up to a constant c, and that $J'^{\lambda F+c}_{\alpha}(p : q) = \lambda J'^{F}_{\alpha}(p : q)$ for $\lambda > 0$. By rescaling those divergences by a fixed factor $\frac{1}{\alpha(1-\alpha)}$, we obtain a continuous 1-parameter family of divergences, called the α-*skew Jensen divergences*, defined over the *full* real line $\alpha \in \mathbb{R}$ as follows [4, 2]:

$$J^{F}_{\alpha}(p : q) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J'^{F}_{\alpha}(p : q) & \alpha \neq \{0, 1\}, \\ B_F(p : q) & \alpha = 0, \\ B_F(q : p) & \alpha = 1. \end{cases}$$

where $B_F(\cdot : \cdot)$ denotes the Bregman divergence [6, 3]: $B_F(p : q) = F(p) - F(q) - \langle p - q, \nabla F(q)\rangle$, with $\langle x, y\rangle = x^\top y$ denoting the scalar product for vectors. Indeed, the limit cases of Jensen divergences $J^{F}_{\alpha}(p : q) = \frac{1}{\alpha(1-\alpha)} J'^{F}_{\alpha}(p : q)$ when $\alpha = 0$ or $\alpha = 1$ tend to a Bregman divergence [3, 2]:

$$\lim_{\alpha\to 0} J^{F}_{\alpha}(p : q) = B_F(p : q), \ \lim_{\alpha\to 1} J^{F}_{\alpha}(p : q) = B_F(q : p).$$

The skew Jensen divergences are related to *statistical divergences* between probability distributions: Namely, the skew Bhattacharrya divergence [2]: $\mathrm{Bhat}(p_1 : p_2) = -\log \int p_1(x)^\alpha p_2(x)^{1-\alpha} \mathrm{d}\nu(x)$, between $p_1 = p_F(x|\theta_1)$ and $p_2 = p_F(x|\theta_2)$ belonging to the *same* exponential family $\{p_F(x|\theta) = \exp(\langle x, \theta\rangle - F(\theta))\}_\theta$ amounts to compute equivalently a skew Jensen divergence on the corresponding natural parameters for the log-normalized function F: $\mathrm{Bhat}(p_F(x|\theta_1) : p_F(x|\theta_2)) = J'^{F}_{\alpha}(\theta_1 : \theta_2)$ (ν is the counting measure for discrete distributions and the Lebesgue measure for continuous distributions).

1.2. Total Bregman divergences

Let us consider an application in medical imaging to motivate the need for a particular kind of invariance when defining divergences: In Diffusion Tensor Magnetic Resonance Imaging (DT-MRI), 3D raw data are captured at voxel positions as 3D ellipsoids denoting the water propagation characteristics [1]. To perform common signal processing tasks like

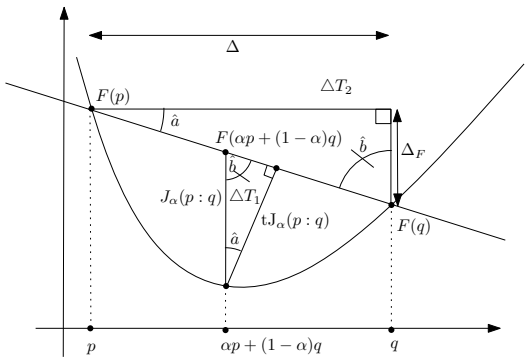

Fig. 1. Geometric proof for the total Jensen divergence: The figure illustrates the two right-angle triangles $\triangle T_1$ and $\triangle T_2$. We deduce that angles $\hat{b}$ and $\hat{a}$ are congruent, and we get the formula on $\mathrm{tJ}_\alpha(p:q) = J_\alpha(p:q)\cos\hat{a}$.

denoising, interpolation or segmentation tasks, one needs to define a proper *dissimilarity measure* between any two such ellipsoids. Those ellipsoids are mathematically handled as *Symmetric Positive Definite* (SPD) matrices [1] that can also be interpreted as centered 3D Gaussian probability distributions. In order not to be biased by the chosen coordinate system for defining those ellipsoids, we ask for a divergence that is invariant to rotations of the coordinate system. For a divergence parameterized by a generator function F derived from the graph of that generator, the invariance under rotations means that the geometric quantity defining the divergence should not change if the original coordinate system is rotated. This is clearly not the case for the skew Jensen divergences that rely on the *vertical axis* to measure the *ordinal* distance. To cope with this drawback, the family of *total Bregman divergences* (tB) have been introduced and shown to be statistically robust [1]. Note that although the traditional Kullback-Leibler divergence (or its symmetrizations like the Jensen-Shannon divergence or the Jeffreys divergence [2]) between two multivariate Gaussians could have been used to provide the desired invariance, the processing tasks are not robust to outliers and perform less well in practice [1].

The *total Bregman divergence* amounts to compute a scaled Bregman divergence: Namely a Bregman divergence multiplied by a *conformal factor* [5] ρ_B: $\mathrm{tB}_F(p:q) = \frac{B_F(p:q)}{\sqrt{1+\langle\nabla F(q),\nabla F(q)\rangle}} = \rho_F(q)B_F(p:q)$:

$$\rho_F(q) = \frac{1}{\sqrt{1+\langle\nabla F(q),\nabla F(q)\rangle}}. \quad (1)$$

For example, choosing the generator $F(x) = \frac{1}{2}\langle x,x\rangle$ with $x \in \mathcal{X} = \mathbb{R}^d$, we get the *total square Euclidean distance*: $tE(p,q) = \frac{1}{2}\frac{\langle p-q,p-q\rangle}{\sqrt{1+\langle q,q\rangle}}$. That is, $\rho_F(q) = \sqrt{\frac{1}{1+\langle q,q\rangle}}$ and $B_F(p:q) = \frac{1}{2}\langle p-q,p-q\rangle = \frac{1}{2}\|p-q\|_2^2$. Total Bregman divergences have proven successful in many applications like Diffusion tensor imaging [1] (DTI) or shape retrieval [7], just to name a few. The total Bregman divergences can be defined over the space of symmetric positive definite (SPD) matrices met in DT-MRI [1]. One key feature of the total Bregman divergence defined over such matrices is its invariance under the special linear group $\mathrm{SL}(d)$ that consists of $d \times d$ matrices of unit determinant: $\mathrm{tB}_F(A^\top PA : A^\top QA) = \mathrm{tB}_F(P:Q),\ \forall A \in \mathrm{SL}(d)$.

2. TOTAL JENSEN DIVERGENCES

2.1. A geometric definition

The skew Jensen divergence J'_α is defined as the "vertical" distance between the interpolated point $((pq)_\alpha, (F(p)F(q))_\alpha)$ lying on the line segment $[(p,F(p)),(q,F(q))]$ and the point $((pq)_\alpha, F((pq)_\alpha))$ lying on the graph of the generator. This measure is therefore *dependent* on the coordinate system chosen for representing the space $\mathcal{X}$ since the notion of "verticality" depends on the coordinate system. To overcome this limitation, we define the *total* Jensen divergence by choosing the *unique orthogonal projection* of $((pq)_\alpha, F((pq)_\alpha)$ onto the line $((p,F(p)),(q,F(q)))$. Let us plot the epigraph of function F restricted to the vertical plane passing through *distinct* points p and q. Let $\Delta_F = F(q) - F(p) \in \mathbb{R}$ and $\Delta = q - p \in \mathcal{X}$ (for $p \neq q$). Consider the two right-angle triangles ΔT_1 and ΔT_2 depicted in Figure 1. Since Jensen divergence J and Δ_F are vertical line segments intersecting the line passing through point $(p,F(p))$ and point $(q,F(q))$, we deduce that the angles $\hat{b}$ are congruent. Thus it follows that angles $\hat{a}$ are also congruent. Now, the cosine of angle $\hat{a}$ measures the ratio of the adjacent side over the hypotenuse of right-angle triangle $\triangle T_2$. Therefore it follows that: $\cos\hat{a} = \frac{\|\Delta\|}{\sqrt{\langle\Delta,\Delta\rangle+\Delta_F^2}} = \sqrt{\frac{1}{1+\frac{\Delta_F^2}{\langle\Delta,\Delta\rangle}}}$, where $\|\cdot\|$ denotes the L_2-norm. In right-triangle $\triangle T_1$, we furthermore deduce that $\mathrm{tJ}'^F_\alpha(p:q) = J'^F_\alpha(p:q)\cos\hat{a} = \rho_F(p,q)J'^F_\alpha(p:q)$. Scaling by factor $\frac{1}{\alpha(1-\alpha)}$, we end up with the following theorem:

Theorem 1 *The total Jensen divergence* tJ^F_α *is invariant to rotations of the coordinate system of* $\mathcal{X}$*. The divergence is mathematically expressed as a scaled skew Jensen divergence* $\mathrm{tJ}^F_\alpha(p:q) = \rho_F(p,q)J^F_\alpha(p:q)$*, where* $\rho_F(p,q) = \sqrt{\frac{1}{1+\frac{\Delta_F^2}{\langle\Delta,\Delta\rangle}}}$ *is symmetric and independent of the skew factor* $\alpha \in \mathbb{R}$.

Observe that the scaling factor $\rho_F(p,q)$ is *independent* of α, symmetric, and is always less or equal to 1. Furthermore, observe that the scaling factor depending on both p and q and is *not separable*: That is, ρ_F cannot be expressed as a product of two terms, one depending only on p and the other depending only on q: $\rho_F(p,q) \neq \rho'_F(p)\rho'_F(q)$. We have $\mathrm{tJ}^F_{1-\alpha}(p:q) = \rho_F(p,q)J^F_{1-\alpha}(p:q) = \rho_F(p,q)J^F_\alpha(q:p)$. Because the conformal factor is independent of α, we have the

following asymmetric ratio equality: $\frac{\mathrm{tJ}_\alpha^F(p:q)}{\mathrm{tJ}_\alpha^F(q:p)} = \frac{J_\alpha^F(p:q)}{J_\alpha^F(q:p)}$. By rewriting $\rho_F(p,q) = \sqrt{\frac{1}{1+s^2}}$, we interpret the non-separable *conformal factor* as a function of the square of the *chord slope* $s = \frac{\Delta_F}{\|\Delta\|}$. The Jensen-Shannon divergence [8] is a separable Jensen divergence for the Shannon information generator $F(x) = x\log x - x$: $\mathrm{JS}(p,q) = \frac{1}{2}\sum_{i=1}^d (p_i \log \frac{2p_i}{p_i+q_i} + \frac{1}{2}\sum_{i=1}^d q_i \log \frac{2q_i}{p_i+q_i})$ that is equivalent to $J^F_{\alpha=\frac{1}{2}}(p:q)$ with $F(x) = \sum_{i=1}^d x_i \log x_i$. Let tJS denotes the total Jensen-Shannon divergence. Although the Jensen-Shannon divergence is symmetric, it is not a metric since it fails the triangular inequality. However, its square root $\sqrt{\mathrm{JS}(p,q)}$ is a metric [9]. But the square root of the total Jensen-Shannon divergence is *not* a metric. It suffices to report a counter-example as follows: Consider the three points of the 1-probability simplex $p = (0.98, 0.02)$, $q = (0.52, 0.48)$ and $r = (0.006, 0.994)$. We have $d_1 = \sqrt{\mathrm{tJS}(p,q)} \simeq 0.351$, $d_2 = \sqrt{\mathrm{tJS}(q,r)} \simeq 0.396$ and $d_3 = \sqrt{\mathrm{tJS}(p,r)} \simeq 0.790$. The triangular inequality fails because $d_1 + d_2 < d_3$. The triangular inequality deficiency is $d_3 - (d_1 + d_2) \simeq 0.042$.

2.2. Total Jensen/Bregman divergences

Although the underlying rationale for deriving the total Jensen divergences followed the same principle of the total Bregman divergences (*i.e.*, replacing the "vertical" projection by an orthogonal projection), the total Jensen divergences *do not coincide* with the total Bregman divergences in limit cases: Indeed, in the limit cases $\alpha \in \{0, 1\}$, we have:

$$\begin{aligned}\lim_{\alpha\to 0} \mathrm{tJ}_\alpha^F(p:q) &= \rho_F(p,q)B_F(p:q) \neq \rho_F(q)B_F(p:q),\\ \lim_{\alpha\to 1} \mathrm{tJ}_\alpha^F(p:q) &= \rho_F(p,q)B_F(q:p) \neq \rho_F(p)B_F(q:p),\end{aligned}$$

since $\rho_F(p,q) \neq \rho_F(q)$. Thus when $p \neq q$, the total Jensen divergence *does not tend* in limit cases to the total Bregman divergences. However, by using a Taylor expansion with exact Lagrange remainder, we write $F(q) = F(p) + \langle q-p, \nabla F(\epsilon)\rangle$, with $\epsilon \in [p,q]$ (assuming wlog. $p < q$). That is, $\Delta_F = F(q) - F(p) = \langle \nabla F(\epsilon), \Delta\rangle$. For univariate divergences, have the squared slope index: $s^2 = \frac{\Delta_F^2}{\Delta^2} = (F'(\epsilon))^2$ since $\Delta_F = \Delta F'(\epsilon)$. For multi-parameter divergences, $\langle \Delta, \nabla F(\epsilon)\rangle = \|\nabla F(\epsilon)\|\|\Delta\|\cos\phi$ where ϕ denotes the angle between the vector Δ and $\nabla F(\epsilon)$, and the slope is equal to $\|\nabla F(\epsilon)\|^2 \cos^2\phi$.

Therefore, in 1D, when $p \simeq q$, we have $\rho_F(p,q) \simeq \rho_F(q)$, and the total Jensen divergence tends to the total Bregman divergence for any value of α. Indeed, in that case, the 1D Bregman/Jensen conformal factors match: $\rho_F(p,q) = \frac{1}{\sqrt{1+(F'(\epsilon))^2}} = \rho_F(\epsilon)$, for $\epsilon \in [p,q]$. We find explicitly the value of ϵ: $\epsilon = (F')^{-1}\left(\frac{\Delta_F}{\Delta}\right) = (F^*)'\left(\frac{\Delta_F}{\Delta}\right)$, where F^* is the Legendre convex conjugate, see [2]. This is the expression of a *Stolarsky mean* [10] of p and q for the strictly monotonous function F'. Therefore when $p \to q$, we have $\lim_{p\to q} \frac{\Delta_F}{\Delta} = F'(q)$ and the total Jensen divergence converges to the total Bregman divergence.

The total Jensen divergence $\mathrm{tJ}_\alpha^F(p:q)$ is equivalent to a Jensen divergence for the convex generator $G(x) = \rho_F(p,q)F(x)$: $\mathrm{tJ}_\alpha^F(p:q) = J_\alpha^{\rho_J(p,q)F}(p:q)$.

3. CENTROIDS AND ROBUSTNESS ANALYSIS

3.1. Total Jensen centroids

Thanks to the invariance to rotations, total Bregman divergences proved highly useful in applications (see [1, 7]) due to the statistical robustness of their centroids. The conformal factors play the role of regularizers. Robustness of the centroid, defined as a notion of centrality robust to "noisy" perturbations, is studied using the framework of the *influence function* [11]. Similarly, the total skew Jensen (right-sided) *centroid* c_α is defined for a finite weighted point set as the minimizer of the following loss function:

$$L_\alpha(x;w) = \sum_{i=1}^n w_i \mathrm{tJ}_\alpha^F(p_i:x), c_\alpha = \arg\min_{x\in\mathcal{X}} L_\alpha(x;w),$$

where $w_i \geq 0$ are the normalized point weights (with $\sum_{i=1}^n w_i = 1$). The left-sided centroids c'_α are obtained by minimizing the equivalent right-sided centroids for $\alpha' = 1-\alpha$: $c'_\alpha = c_{1-\alpha}$ (recall that the conformal factor does not depend on α). Therefore, we consider the right-sided centroids in the remainder. We consider $c^{(t)}$ (initialized with the barycenter $c^{(0)} = \sum_i w_i p_i$) given. This allows us to consider the following simpler minimization problem: $c = \arg\min_{x\in\mathcal{X}} \sum_{i=1}^n w_i \times \rho_F(p_i, c^{(t)}) J_\alpha^F(p_i:x)$. Let $w_i^{(t)} = \frac{w_i \times \rho_F(p_i, c^{(t)})}{\sum_j w_j \times \rho_F(p_j, c^{(t)})}$, be the updated renormalized weights at stage t. We minimize:

$$c = \arg\min_{x\in\mathcal{X}} \sum_{i=1}^n w_i^{(t)} J_\alpha^F(p_i:x).$$

This is a convex-concave minimization procedure [12] (CCCP) that can be solved iteratively until it reaches convergence [2] at $c^{(t+1)}$ (in practice, we need to implement a threshold). That is, we iterate the following formula [2] a given number of times k:

$$c_{l+1}^{(t+1)} \leftarrow (\nabla F)^{-1}\left(\sum_{i=1}^n w_i^{(t)} \nabla F((1-\alpha)c_l^{(t+1)} + \alpha p_i)\right),$$

with $c_0^{(t+1)} = c^{(t)}$. We repeat the (1) reweighting and (2) CCCP iterations until the loss function improvement $L_\alpha(x;w)$ goes below a prescribed threshold. Although the CCCP algorithm is guaranteed to converge *monotonically* to a local optimum, the two steps weight update/CCCP does not provide anymore the monotonous convergence as we have attested in practice.

3.2. Jensen centroids: Robustness analysis

The centroids defined with respect to the total Bregman divergences have been shown to be robust to outliers whatever the chosen generator [1]. We first analyze the robustness for the symmetric Jensen divergence (for $\alpha = \frac{1}{2}$). We investigate the *influence function* [11] $i(y)$ on the centroid when adding an outlier point y with prescribed weight $\epsilon > 0$. Without loss of generality, it is enough to consider only two points: One *outlier* with ϵ mass and one *inlier* with the remaining mass. Let us add an outlier point y with weight ϵ onto an inliner point p. Let $\bar{x} = p$ and $\tilde{x} = p + \epsilon z$ denote the centroids before adding y and after adding y. $z = z(y)$ denotes the influence function. For sake of simplicity, we drop in the remainder the F notations in divergences. The Jensen centroid minimizes (we can ignore dividing by the renormalizing total weight inlier+outlier: $\frac{1}{1+\epsilon}$): $L(x) \equiv J(p,x) + \epsilon J(x,y)$. The derivative of this energy is $D(x) = L'(x) = J'(p,x) + \epsilon J'(y,x)$. The derivative of the Jensen divergence is given by (not necessarily a convex distance): $J'(h,x) = \frac{1}{2}f'(x) - \frac{1}{2}f'\left(\frac{x+h}{2}\right)$, where f is the univariate convex generator and f' its derivative. For the optimal value of the centroid $\tilde{x}$, we have $D(\tilde{x}) = 0$, yielding: $(1+\epsilon)f'(\tilde{x}) - \left(f'\left(\frac{\tilde{x}+p}{2}\right) + \epsilon f'\left(\frac{\tilde{x}+y}{2}\right)\right) = 0$. Using Taylor expansions on $\tilde{x} = p + \epsilon z$ (where $z = z(y)$ is the influence function) on the derivative f', we get $f'(\tilde{x}) \simeq f'(p) + \epsilon z f''(p)$ and $(1+\epsilon)(f'(p)+\epsilon z f''(p)) - \left(f'(p) + \frac{1}{2}\epsilon z f''(p) + \epsilon f'\left(\frac{p+y}{2}\right)\right)$ (ignoring the term in ϵ^2 for small constant $\epsilon > 0$ in the Taylor expansion term of $\epsilon f'$.) Thus we get the following mathematical equality: $z((1+\epsilon)\epsilon f''(p) - \epsilon z/2 f''(p)) = f'(p) + \epsilon f'\left(\frac{p+y}{2}\right) - (1+\epsilon)f'(p)$. Finally, we get the expression of the influence function $z = z(y) = 2\frac{f'(\frac{p+y}{2}) - f'(p)}{f''(p)}$, for small prescribed $\epsilon > 0$.

Theorem 2 *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.*

To illustrate this theorem, consider the Jensen-Shannon with $\mathcal{X} = \mathbb{R}^+$, $f(x) = x\log x - x$, $f'(x) = \log(x)$, $f''(x) = 1/x$. We check that $|f'(\frac{p+y}{2})| = |\log\frac{p+y}{2}|$ is unbounded when $y \to +\infty$. The influence function $z(y) = 2p\log\frac{p+y}{2p}$ is unbounded when $y \to \infty$, and therefore the centroid is not robust to outliers. Now, consider the Jensen-Burg: $\mathcal{X} = \mathbb{R}^+$, $f(x) = -\log x$, $f'(x) = -1/x$, $f''(x) = \frac{1}{x^2}$: We check that $|f'(\frac{p+y}{2})| = |\frac{2}{p+y}|$ 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 $|z(y)| \to 2p < \infty$. The influence function is bounded and the centroid is robust. We can extend to multi-parameter separable Jensen divergences.

3.3. Clustering: No closed-form centroid, no cry!

The most famous clustering algorithm is k-means [13] that consists in first initializing k distinct seeds and then iteratively assign the points to their closest center, and update the cluster centers by taking the centroids of the clusters. A breakthrough was achieved by proving that a randomized seed selection, k-means++ [14], guarantees probabilistically a constant approximation factor to the optimal loss. The k-means++ initialization may be interpreted as a discrete k-means where the k cluster centers are choosen among the input. This yields $\binom{n}{k}$ combinatorial seed sets. Note that k-means is NP-hard when $k = 2$ and the dimension is not fixed, but not discrete k-means [15]. Thus we do not need to compute centroids to cluster with respect to total Jensen divergences. Skew Jensen centroids can be approximated arbitrarily finely using the concave-convex procedure, as reported in [2]. On a compact domain $\mathcal{X}$, we have $\rho_{\min} J(p:q) \leq \mathrm{tJ}(p:q) \leq \rho_{\max} J(p:q)$, with $\rho_{\min} = \min_{x\in\mathcal{X}} \frac{1}{\sqrt{1+\langle\nabla F(x),\nabla F(x)\rangle}}$ and $\rho_{\max} = \max_{x\in\mathcal{X}} \frac{1}{\sqrt{1+\langle\nabla F(x),\nabla F(x)\rangle}}$. We are given a set $\mathcal{S}$ of points that we wish to cluster in k clusters, following a hard clustering assignment. We let $\mathrm{tJ}_\alpha(A:y) = \sum_{x\in A}\mathrm{tJ}_\alpha(x:y)$ for any $A \subseteq S$. The optimal total hard clustering Jensen potential is $\mathrm{tJ}_\alpha^{\mathrm{opt}} = \min_{C\subseteq S:|C|=k}\mathrm{tJ}_\alpha(C)$, where $\mathrm{tJ}_\alpha(C) = \sum_{x\in S}\min_{c\in C}\mathrm{tJ}_\alpha(x:c)$. Finally, the contribution of some $A \subseteq S$ to the optimal total Jensen potential having centers C is $\mathrm{tJ}_{\mathrm{opt},\alpha}(A) = \sum_{x\in A}\min_{c\in C}\mathrm{tJ}_\alpha(x:c)$.

Total Jensen seeding picks randomly without replacement an element x in S with probability proportional to $\mathrm{tJ}_\alpha(C)$, where C is the current set of centers. When $C = \emptyset$, the distribution is uniform.

Theorem 3 *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, for some constants U and V.*

Proof is reported in [18].

4. CONCLUSION

We described a novel family of divergences that are invariant by rotations: *total skew Jensen divergences*. Those divergences scale the ordinary Jensen divergences by a non-separable conformal factor [5] independent of the skew parameter, and extend naturally the underlying principle of the total Bregman divergences [7].

Acknowledgments: NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

5. REFERENCES

[1] B. Vemuri, M. Liu, Shun-ichi Amari, and F. Nielsen, "Total Bregman divergence and its applications to DTI analysis," *IEEE Transactions on Medical Imaging*, vol. 30, no. 2, pp. 475–483, 2011.

[2] F. Nielsen and S. Boltz, "The Burbea-Rao and Bhattacharyya centroids," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5455–5466, August 2011.

[3] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2882–2904, 2009.

[4] J. Zhang, "Divergence function, duality, and convex analysis," *Neural Computation*, vol. 16, no. 1, pp. 159–195, 2004.

[5] R. Nock, F. Nielsen and Shun-ichi Amari, "On conformal divergences and their population minimizers," *CoRR (arXiv)*, abs/1311.5125, 2013.

[6] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[7] M. Liu, B. C. Vemuri, S. Amari, and F. Nielsen, "Shape retrieval using hierarchical total Bregman soft clustering," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2407–2419, 2012.

[8] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145 – 151, 1991.

[9] B. Fuglede and F. Topsoe, "Jensen-Shannon divergence and Hilbert space embedding," in *IEEE International Symposium on Information Theory*, 2004, pp. 31–31.

[10] K. B. Stolarsky, "Generalizations of the logarithmic mean," *Mathematics Magazine*, vol. 48, no. 2, pp. 87–92, 1975.

[11] 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.

[12] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[13] S. P. Lloyd, "Least squares quantization in PCM," Bell Laboratories, Technical Report, 1957.

[14] D. Arthur and S. Vassilvitskii, "k-means$++$: the advantages of careful seeding," in *Proceedings of the Symposium on Discrete Algorithms (SODA)*. 2007, pp. 1027–1035.

[15] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is NP-hard," *Theoretical Computer Science*, vol. 442, no. 0, pp. 13 – 21, 2012.

[16] D. Arthur and S. Vassilvitskii, "k-means++ : the advantages of careful seeding," in *Proceedings of the Symposium on Discrete algorithms (SODA)*. 2007, pp. 1027–1035.

[17] R. Nock, P. Luosto, and J. Kivinen, "Mixed Bregman clustering with approximation guarantees," in *Machine Learning and Knowledge Discovery in Databases*. 2008, pp. 154–169.

[18] F. Nielsen and R. Nock, "Total Jensen divergences: Definition, Properties and k-Means++ Clustering," *CoRR arXiv*, abs/1309.7109, 2013.