THE CHORD GAP DIVERGENCE AND A GENERALIZATION OF THE BHATTACHARYYA DISTANCE

Frank Nielsen

École Polytechnique
Sony Computer Science Laboratories, Inc
Frank.Nielsen@acm.org

ABSTRACT

We introduce a novel family of distances, called the chord gap divergences, that generalizes the Jensen/Burbea-Rao distances and study its properties. It follows a generalization of the statistical Bhattacharyya distance that is frequently met in applications. We then report an iterative concave-convex procedure for computing centroids, and analyze the performance of the k-means++ clustering with respect to that new dissimilarity measure by introducing the Taylor-Lagrange remainder form of skew Jensen divergences.

***Index Terms*—** Jensen/Burbea-Rao divergence, Bregman divergence, Jensen-Bregman divergence, centroid, k-means++

1. INTRODUCTION

In many applications, one faces the crucial dilemma of choosing an appropriate distance $D(\cdot,\cdot)$ between data. In some cases, those distances can be picked up *a priori* from well-grounded principles (e.g., Kullback-Leibler distance in statistical estimation [1]). In other cases, one is rather left at testing several distances [2], and choose *a posteriori* the distance that yielded the best performance. For the latter cases, it is judicious to consider a family of parametric distances $D_\alpha(\cdot,\cdot)$, and learn [3] the hyperparameter α according to the application at hand and potentially the dataset (distance selection). Thus it is interesting to consider parametric generalizations of common distances [4] to improve performance in applications.

Some distances can be designed from inequality gaps [5, 6]. For example, the Jensen divergence $J_F(p,q)$ (also called the Burbea-Rao divergence [5, 7]) is designed from the inequality gap of Jensen inequality $F\left(\frac{p+q}{2}\right) \leq \frac{F(p)+F(q)}{2}$ that holds for a strictly convex function F: $J_F(p,q) = \frac{F(p)+F(q)}{2} - F\left(\frac{p+q}{2}\right)$. We can extend the Jensen divergence to a parametric family of skew Jensen divergences J_F^α (with $\alpha \in (0,1)$) built on the convex inequality gap $F((1-\alpha)p+\alpha q) \leq (1-\alpha)F(p)+\alpha F(q)$:

$$J_F^\alpha(p:q) = (1-\alpha)F(p)+\alpha F(q) - F((1-\alpha)p+\alpha q), \quad (1)$$

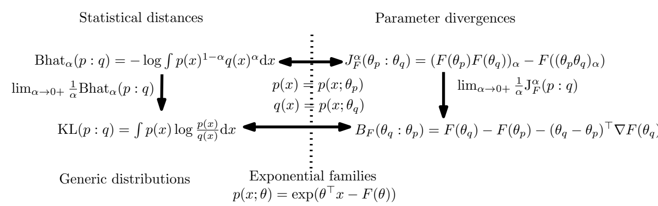

Fig. 1. Links between the statistical skew Bhattacharyya distances and parametric skew Jensen divergences when distributions belong to the same exponential family.

satisfying $J_F^\alpha(q:p) = J_F^{1-\alpha}(p:q)$ and $J_F^{\frac{1}{2}}(p:q) = J_F(p,q)$. Here the ':' notation emphasizes the fact that the distance is potentially asymmetric $J_F^\alpha(p:q) \neq J_F^\alpha(q:p)$. The term divergence is used in information geometry [9] to refer to the smoothness of the distance that yields an information-geometric structure of the space induced by the divergence. Let $[p,q] = \{(pq)_\lambda := (1-\lambda)p+\lambda q, \lambda\in[0,1]\}$ denote the line segment with endpoints p and q. Then we can rewrite [8] Eq.1:

$$J_F^\alpha(p:q) = (F(p)F(q))_\alpha - F((pq)_\alpha). \quad (2)$$

In applications, it is rather the relative comparisons of distances rather than their absolute values that is important. Thus we may multiply a distance by any positive scaling factor and include it in the class of that distance. When F is strictly convex and differentiable, the class of Jensen divergences include in the limit cases the Bregman divergences [7, 10, 11]: $\lim_{\alpha\to 0^+}\frac{J_F^\alpha(p:q)}{\alpha} = B_F(q:p)$, and $\lim_{\alpha\to 1^-}\frac{J_F^\alpha(p:q)}{1-\alpha} = B_F(p:q)$, where

$$B_F(p:q) = F(p) - F(q) - (p-q)^\top \nabla F(q), \quad (3)$$

is the Bregman divergence [13]. Overall, one may define the smooth parametric family of scaled skew Jensen divergences: $\mathrm{sJ}_F^\alpha(p:q) = \frac{1}{\alpha(1-\alpha)}J_F^\alpha(p:q)$ that encompasses the Bregman divergence $B_F(p:q)$ and the reverse Bregman divergence $B_F(q:p)$ in limit cases (with $\alpha\in\mathbb{R}$).

There is a nice relationship between Jensen divergences operating on parameters (e.g., vectors, matrices) and a class

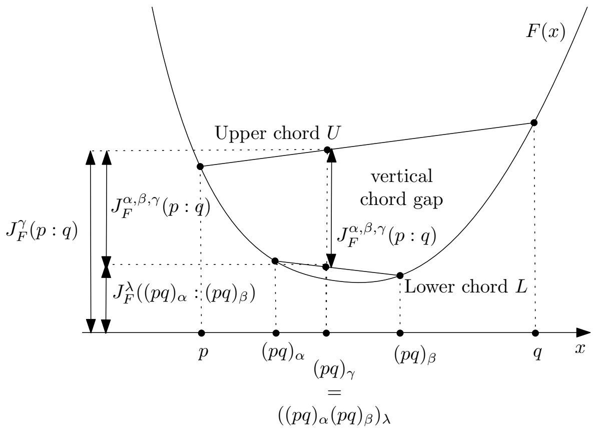

Fig. 2. The triparametric chord gap divergence.

of statistical distances between probability distributions (Figure 1): Let $\{p(x;\theta)\}_\theta$ be an exponential family [11] (includes the Gaussian family and the finite discrete "multinoulli" family) with convex cumulant function $F(\theta)$. Then the skew Bhattacharryya distance [14]

$$\mathrm{Bhat}_\alpha(p:q) = -\log \int p(x)^{1-\alpha} q(x)^\alpha \mathrm{d}x, \quad (4)$$

between two distributions belonging to the same exponential family amounts to a skew Jensen divergence [7]:

$$\mathrm{Bhat}(p(x;\theta_1):p(x;\theta_2)) = J_F^\alpha(\theta_1:\theta_2). \quad (5)$$

We further check that $\lim_{\alpha\rightarrow 0^+} \frac{1}{\alpha}\mathrm{Bhat}_\alpha(p:q) = \mathrm{KL}(p:q)$ and $\lim_{\alpha\rightarrow 1^-} \frac{1}{1-\alpha}\mathrm{Bhat}_\alpha(p:q) = \mathrm{KL}(q:p)$ where $\mathrm{KL}(p:q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x$ is the Kullback-Leibler divergence.

In statistical signal processing, information fusion and machine learning, one often considers the skew Bhattacharryya distance [15, 16, 17] or the Chernoff distance [18, 19, 20] for exponential families (e.g., Gaussian/multinoulli): This highlights the important role in disguise of the equivalent skew Jensen divergences (see Eq. 5).

The paper is organized as follows: Section 2 introduces the novel triparametric family of chord gap divergences that generalizes the skew Jensen divergences (§2.1), describes several properties (§2.2), and deduces a generalization of the statistical Bhattacharyya distance (§ 2.3). Section 3 considers the calculation of the centroid (§3.1) for the chord gap divergences, and report probabilistic guarantee of the k-means++ seeding (§3.2) by highlighting the Taylor-Lagrange forms of those divergences.

2. THE CHORD GAP DIVERGENCE

2.1. Definition

Let $F:\mathcal{X}\rightarrow\mathbb{R}$ be a *strictly convex* function. For $\alpha,\beta\in(0,1)$ with $\alpha\neq\beta$, the chord $L = [((pq)_\alpha, F((pq)_\alpha))((pq)_\beta, F((pq)_\beta))]$ is *below* the distinct chord $U = [(p,F(p))(q,F(q))]$. Thus we can define a divergence [21] as the vertical gap between these two chords for a given coordinate $x\in[(pq)_\alpha,(pq)_\beta]$:

$$J_F^{\alpha,\beta,\gamma}(p:q) = (F(p)F(q))_\gamma - (F((pq)_\alpha)F((pq)_\beta))_\lambda, \quad (6)$$

such that $((pq)_\alpha(pq)_\beta)_\lambda = (pq)_\gamma$ with $\gamma\in(\alpha,\beta)$ (Figure 2). A calculation shows that $\lambda = \lambda(\alpha,\beta,\gamma) = \frac{\gamma-\alpha}{\beta-\alpha}$ or $\gamma = \lambda(\beta-\alpha)+\alpha$ for $\lambda\in[0,1]$ when $\alpha\neq\beta$, so that we get $J_F^{\alpha,\beta,\gamma}(p:q) = (F(p)F(q))_\gamma - (F((pq)_\alpha)F((pq)_\beta))_{\frac{\gamma-\alpha}{\beta-\alpha}}$.

2.2. Properties of the chord gap divergence

We have $J_F^{\alpha,\alpha,\alpha}(p:q) = J_F^\alpha(p:q)$, $J_F^{0,1,\gamma}(p:q) = J^\gamma(p:q)$ and $J_F^{\alpha,\beta,\gamma}(q:p) = J_F^{1-\alpha,1-\beta,1-\gamma}(p:q)$ since $\lambda(1-\alpha,1-\beta,1-\gamma) = \frac{\gamma-\alpha}{\beta-\alpha} = \lambda(\alpha,\beta,\gamma)$ using the fact that $(ab)_{1-\delta} = (ba)_\delta$ for $\delta\in[0,1]$. Thus we have $J_F^{1-\alpha,1-\alpha,1-\alpha}(p:q) = J_F^\alpha(q:p)$. For $\beta = 1-\alpha$, $\gamma=\frac{1}{2}$ (and $\lambda=\frac{1}{2}$), the chord gap divergence amounts to a scaled symmetrized skew Jensen divergences [12] (Eq. 35).

We can express the chord gap divergence as the difference of two skew Jensen divergences (Figure 2):

$$J_F^{\alpha,\beta,\gamma}(p:q) = J_F^\gamma(p:q) - J_F^\lambda((pq)_\alpha:(pq)_\beta), \quad (7)$$

with $\lambda=\frac{\gamma-\alpha}{\beta-\alpha}$ or $\gamma = \lambda(\beta-\alpha)+\alpha$ for $\lambda\in[0,1]$ and $\gamma\in[\alpha,\beta]$. Thus the chord gap divergence can be interpreted as a *truncated* skew Jensen divergence.

A biparametric subfamily $J_F^{\beta,\gamma}$ of $J_F^{\alpha,\beta,\gamma}$ is obtained by setting $\alpha=0$ so that $(pq)_\alpha = p$, so that the two upper/lower chords L and U coincide at extremity p:

$$\begin{aligned} J_F^{\beta,\gamma}(p:q) &= (F(p)F(q))_\gamma - (F(p)F((pq)_\beta))_{\frac{\gamma}{\beta}}, \quad (8)\\ &= \gamma\left(\left(\frac{1}{\beta}-1\right)F(p) + F(q) - \frac{1}{\beta}F((pq)_\beta)\right). \end{aligned}$$

When $\beta=\frac{1}{2}$, we find that $J_F^{\frac{1}{2},\gamma}(p:q) = 2\gamma J_F(p:q)$, the ordinary ($\gamma$-scaled) Jensen divergence. When $\beta\rightarrow 0$, we have $\lim_{\beta\rightarrow 0}\frac{1}{\gamma}J_F^{\beta,\gamma}(p:q) = B_F(q:p)$ (with $\gamma\in(0,\beta)$) since $-\frac{1}{\beta}F((pq)_\beta) \simeq -\frac{1}{\beta} - (q-p)^\top\nabla F(p)$ using a first-order Taylor expansion.

Matrix chord gap divergences can be obtained by taking strictly convex matrix generators [13] (e.g., $F(X) = -\log\det|X|$) for symmetric positive definite matrices $X\in\mathbb{P}_{++}$, $\mathbb{P}_{++} = \{X : X\succ 0\}$ denote the space of positive definite matrices, a convex cone. This may be useful in applications based on covariance matrices [13].

2.3. Generalized Bhattacharyya distances

The interpretation given in Eq. 7 yields a triparametric family of Bhattacharyya statistical distances [14] between members $p(x)=p(x;\theta_p)$ and $q(x)=p(x;\theta_q)$ of the *same* exponential family: $\mathrm{Bhat}_{\alpha,\beta,\gamma}(\theta_p:\theta_q) = \mathrm{Bhat}_\gamma(\theta_p:\theta_q) -$

$\mathrm{Bhat}_\lambda((\theta_p\theta_q)_\alpha : (\theta_p\theta_q)_\beta)$. It follows that

$$\mathrm{Bhat}_{\alpha,\beta,\gamma}(\theta_p : \theta_q) = \\ -\log \frac{\int p(x;\theta_p)^{1-\gamma} p(x;\theta_q)^\gamma \mathrm{d}x}{\int p(x;(\theta_p\theta_q)_\alpha)^{1-\lambda} p(x;(\theta_p\theta_q)_\beta)^\lambda \mathrm{d}x}. \tag{9}$$

Note that when $\alpha = \beta$, we have $p(x;(\theta_p\theta_q)_\alpha)^{1-\lambda} p(x;(\theta_p\theta_q)_\beta)^\lambda = p(x;(\theta_p\theta_q)_\alpha)$ and therefore the denominator is $\int p(x;(\theta_p\theta_q)_\alpha)\mathrm{d}x = 1$, and we recover the skew Bhattacharyya distance, as expected.

We shall extend the generalized Bhattacharrya divergence of Eq. 9 to arbitrary distributions by generalizing the notion of interpolated distribution $p(x;(\theta_p\theta_q)_\delta) = \Gamma_\delta(p(x;\theta_p), p(x;\theta_q))$. When δ ranges from 0 to 1, we obtain a Bhattacharyya arc linking $p(x;\theta_p)$ to $p(x;\theta_q)$ (exponential or e-geodesic in information geometry [9]). We define:

$$\Gamma_\delta(p(x), q(x)) = \frac{p(x)^{1-\delta} q(x)^\delta}{Z_\delta(p(x) : q(x))},$$

with $Z_\delta(p(x) : q(x)) = \int p(x)^{1-\delta} q(x)^\delta \mathrm{d}\nu(x)$. Note that we need the integral to converge properly in order to define $\Gamma_\delta(p(x), q(x))$. This always holds for distributions belonging to the same exponential families since $(\theta_p\theta_q)_\delta$ is guaranteed to belong to the natural parameter space, and

$$Z_\delta(p(x;\theta_p) : p(x;\theta_q)) = \exp(-J_F^\delta(\theta_p : \theta_q)). \tag{10}$$

By extension, the triparametric Bhattacharyya distance can be defined by:

$$\mathrm{Bhat}^{\alpha,\beta,\gamma}(p(x) : q(x)) = \\ -\log \left(\frac{\int p(x)^{1-\gamma} q(x)^\gamma \mathrm{d}\nu(x)}{\Gamma_\alpha(p(x), q(x))^{1-\gamma} \Gamma_\beta(p(x), q(x))^\gamma} \right). \tag{11}$$

Thus we explicitly define the generalized Bhattacharrya distance by:

$$\mathrm{Bhat}^{\alpha,\beta,\gamma}(p(x) : q(x)) = \\ -\log \left(\frac{\int p(x)^{1-\gamma} q(x)^\gamma \mathrm{d}\nu(x)}{\int \left(\frac{p(x)^{1-\alpha} q(x)^\alpha \mathrm{d}\nu(x)}{\int p(x)^{1-\alpha} q(x)^\alpha \mathrm{d}\nu(x)} \right)^{1-\lambda} \left(\frac{p(x)^{1-\beta} q(x)^\beta \mathrm{d}\nu(x)}{\int p(x)^{1-\beta} q(x)^\beta \mathrm{d}\nu(x)} \right)^\lambda \mathrm{d}\nu(x)} \right)$$

Notice that when $\alpha = \beta$, for any $\lambda \in [0,1]$, the denominator collapses to one, and we find that $\mathrm{Bhat}_{\alpha,\beta,\gamma}(p(x) : q(x)) = \mathrm{Bhat}_\alpha(p(x) : q(x))$, as expected.

For multivariate gaussians/normals belonging to the family $\{\mathcal{N}(\mu,\Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathbb{P}^d_{++}\}$, we have the natural parameter [22] $\theta = (v, M) = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})$, and the cumulant function $F(v, M) = \frac{d}{2}\log 2\pi - \frac{1}{2}\log|-2M| - \frac{1}{4}v^\top M^{-1} v$ that can also be expressed in the usual parameters $F(\mu,\Sigma) = \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{1}{2}\mu^\top\Sigma^{-1}\mu$. We have $(\theta_p\theta_q)_\delta = ((1-\delta)\Sigma_p^{-1}\mu_p + \delta\Sigma_q^{-1}\mu_q, -\frac{1-\delta}{2}\Sigma_p^{-1} - \frac{\delta}{2}\Sigma_q^{-1})$ so that we get [23]: $J_F^\alpha(p(x;\mu_p,\Sigma_p) : p(x;\mu_q,\Sigma_q)) = \frac{\alpha(1-\alpha)}{2}\Delta\mu^\top((1-\alpha)\Sigma_p + \alpha\Sigma_q)^{-1}\Delta\mu + \frac{1}{2}\log\frac{|(1-\alpha)\Sigma_p + \alpha\Sigma_q|}{|\Sigma_p|^{1-\alpha}|\Sigma_q|^\alpha}$ with $|.|$ the determinant and $\Delta\mu = \mu_q - \mu_p$. This gives a closed-form formula for $\mathrm{Bhat}^{\alpha,\beta,\gamma}$ for multivariate Gaussians.

3. CENTROID-BASED CLUSTERING

Bhattacharrya clustering is often used in statistical signal processing, information fusion, and machine learning (see [15, 24, 25, 26] for some illustrative examples). Popular clustering algorithms are center-based clustering, where each cluster stores a prototype (a representative element), and each datum is assigned to the cluster with the closest prototype wrt. a distance function. The cluster prototypes are then updated, and the algorithm iterates until (local) convergence. This scheme includes the k-means and the k-medians [27]. Lloyd k-means heuristic updates the prototype c of a cluster X by choosing its center of mass $c = \frac{1}{|X|}\sum_{x\in X} x$ that minimizes the cluster variance: $\min_c \sum_{x\in X}\|x - c\|^2$ (this holds for any Bregman divergence too [11]).

3.1. Chord gap divergence centroid

We extend k-means for a weighted point set

$$\mathcal{P} = \{(w_1, p_2), \ldots, (w_n, p_n)\},$$

with $w_i > 0$ and $\sum_i w_i = 1$, using the chord gap divergence by solving the following minimization problem: $\min_x E(x) = \sum_{i=1}^n w_i J_F^{\alpha,\beta,\gamma}(p_i : x)$. By expanding the chord gap divergence formula and removing all terms independent of x, we obtain an equivalent minimization problem as a difference of convex function programming [28]: $\min_x E(x) = \min_x A(x) - B(x)$ with $A(x) = \sum_{i=1}^n (F(p_i)F(x))_\gamma$ and $B(x) = \sum_{i=1}^n (F((p_i x)_\alpha) F((p_i x)_\beta))_\lambda$, both strictly convex functions. It follows a concave-convex procedure [29] (CCCP) solving locally $\min_x A(x) - B(x)$: initialize $x_0 = p_1$ and then iteratively update $\nabla A(x_{t+1}) = \nabla B(x_t)$. When the reciprocal gradient ∇A^{-1} is available in closed form, we end up with $x_{t+1} = \nabla A^{-1}(\nabla B(x_t))$. Since we have $\nabla A(x) = n\gamma\nabla F(x)$ and $\nabla B(x) = \sum_i (1-\lambda)\alpha\nabla F(F((p_i x)_\alpha) + \lambda\beta\nabla F((p_i x)_\beta)$, the update rule is

$$x_{t+1} = \\ \nabla F^{-1}\left(\frac{1}{\gamma}\sum_i w_i\left((1-\lambda)\alpha\nabla F((p_i x_t)_\alpha) + \lambda\beta\nabla F((p_i x_t)_\beta)\right)\right).$$

When $\alpha = \beta = \gamma$, we find the simplified update rule $x_{t+1} = \nabla F^{-1}\left(\sum_i w_i \nabla F((p_i x_t)_\alpha)\right)$ corresponding to the skew Jensen divergences [7]. Note that it is enough to improve iteratively the prototypes to get a variational Lloyd's k-means [30].

3.2. Performance analysis of k-means++

For high-performance clustering, one may use k-means++ [31] that is a guaranteed probabilistic initialization of the cluster prototypes. To get an expected competitive ratio [31] of $2U^2(1+V)(2+\log k)$ [30], we need to upper bound: (i) U such that the divergence $D = J_F^{\alpha,\beta\gamma}$ satisfies the U-triangular inequality $D(x : z) \leq U(D(x : y) + D(y : z))$, and (ii) V such that the divergence satisfies the symmetric inequality $D(y : x) \leq V D(x : y)$. The proof follows the proof reported in [30] for total Jensen divergences once we can express the divergences in their Taylor-Lagrange forms $D(p : q) = (p-q)^\top H_D(p : q)(p-q)$ where $H_D(p : q) \succ 0$. For example, the Taylor-Lagrange form of the Bregman divergence [32] is obtained from a first-order Taylor expansion with the exact Lagrange remainder:

$$B_F(p : q) = \frac{1}{2}(p-q)^\top \nabla^2 F(\xi)(p-q), \qquad (12)$$

for some $\xi \in [p, q]$. This expression can be interpreted as a squared Mahalanobis distance $M_Q(p, q) = (p-q)^\top Q(p-q)$ with precision matrix $Q = \frac{1}{2}\nabla^2 F(\xi)$ depending on p and q. Any squared Mahalanobis distance satisfies $U = 2$ (see [33]) and $V = 1$, and can be interpreted as a squared norm-induced distance: $M_Q(p, q) = \|Q^{\frac{1}{2}}(p-q)\|_2^2$.

We report the Taylor-Lagrange form of the skew Jensen divergences: There exists $\xi_1, \xi_2 \in [p, q]$, such that the skew Jensen divergence can be expressed as $J_F^\alpha(p : q) = (p-q)^\top H_F^\alpha(p : q)(p-q)$, with

$$H_F^\alpha(p : q) = \frac{1}{2}\alpha(1-\alpha)(\alpha\nabla^2 F(\xi_1) + (1-\alpha)\nabla^2 F(\xi_2)). \qquad (13)$$

The proof relies on introducing the skew Jensen-Bregman (JB) divergence [7] defined by

$$\mathrm{JB}_F^\alpha(p : q) = (1-\alpha)B_F(p : (pq)_\alpha) + \alpha B_F(q : (pq)_\alpha), \qquad (14)$$

and observing the $\mathrm{JB}_F^\alpha(p : q) = J_F^\alpha(p : q)$ since $p - (pq)_\alpha = \alpha(p-q)$ and $q - (pq)_\alpha = (1-\alpha)(q-p)$ (and therefore the $\nabla F((pq)_\alpha)$-terms cancel out). Then we apply the Taylor-Lagrange form of Bregman divergences of Eq. 12 to get the result. Notice that when $\alpha \to 0$ or $\alpha \to 1$, the scaled skew Jensen difference tend to Bregman divergences, and we have $\lim_{\alpha\to 1} \frac{J_F^{(\alpha)}(p:q)}{\alpha(1-\alpha)} = \frac{1}{2}(p-q)^\top \nabla^2 F(\xi_1)(p-q) = B_F(p : q)$ for $\xi_1 \in [p, q]$, and $\lim_{\alpha\to 0} \frac{J_F^{(\alpha)}(p:q)}{\alpha(1-\alpha)} = \frac{1}{2}(p-q)^\top \nabla^2 F(\xi_2)(p-q) = B_F(q : p)$ for $\xi_2 \in [p, q]$, as expected.

Using expression of Eq. 7 for the chord gap divergence, and the fact that $(pq)_\alpha - (pq)_\beta = (\alpha-\beta)(q-p)$, we get the Taylor-Lagrange form of the chord gap divergence $J_F^{\alpha,\beta,\gamma} = (p-q)^\top H_F^{\alpha,\beta,\gamma}(p : q)(p-q)$ with

$$\begin{aligned} H_F^{\alpha,\beta,\gamma}(p : q) &= \frac{1}{2}\gamma(1-\gamma)\nabla^2 F(\xi') - \frac{1}{2}\lambda(1-\lambda)(\alpha-\beta)^2\nabla^2 F(\xi''), && (15) \\ &= \frac{1}{2}\left(\gamma(1-\gamma)\nabla^2 F(\xi') - (\gamma-\alpha)(\gamma-\beta)\nabla^2 F(\xi'')\right), && (16) \end{aligned}$$

for $\xi', \xi'' \in \mathcal{X}$.

When dealing with a finite (weighted) point set $\mathcal{P}$, let

$$\rho = \frac{\sup_{\xi',\xi'',p,q\in\mathrm{co}(\mathcal{P})} \|(\nabla^2 F(\xi'))^{\frac{1}{2}}(p-q)\|}{\inf_{\xi',\xi'',p,q\in\mathrm{co}(\mathcal{P})} \|(\nabla^2 F(\xi''))^{\frac{1}{2}}(p-q)\|} < \infty, \qquad (17)$$

where $\mathrm{co}(\mathcal{P})$ denotes the convex closure of $\mathcal{P}$ Then it comes that $U = O_\rho(1)$ and $V = O_\rho(1)$ so that k-means++ probabilistic seeding is $\bar{O}_\rho(\log k)$ competitive for the chord gap divergence.

4. CONCLUDING REMARKS

We introduced the chord gap divergence as a generalization of the skew Jensen divergences [7, 13], studied its properties and obtained a generalization of the skew Bhattacharrya divergences. We showed that the chord gap divergence centroid can be obtained using a convex-concave iterative procedure [7], and analyzed the k-means++ [31] performance by giving the Taylor-Lagrange forms of the skew Jensen and chord gap divergences. We expect our contributions to be useful for the signal processing, information fusion and machine learning communities where the Bhattacharrya [34, 35] or Chernoff information [2, 18] is often used. In practice, the triparametric chord gap divergence shall be tuned according to the application at hand (and the dataset for supervised tasks using cross-validation for example). Java™ source code is available for reproducible research:`www.lix.polytechnique.fr/~nielsen/CGD/`

5. REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[2] F. A. Sadjadi, "Performance evaluations of correlations of digital images using different separability measures," *IEEE TPAMI*, vol. 4, no. 4, pp. 436–441, 1982.

[3] H. Choi, S. Choi, A. Katake, and Y. Choe, "Learning α-integration with partially-labeled data," in *IEEE ICASSP*, 2010, pp. 2058–2061.

[4] M. M. Deza and E. Deza, *Encyclopedia of Distances*, Springer, 2009.

[5] J. Burbea and C Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Transactions on Information Theory*, vol. 28, no. 3, pp. 489–495, 1982.

[6] F. Nielsen, K. Sun, and S. Marchand-Maillet, "On Hölder projective divergences," *Entropy*, vol. 19, no. 3, pp. 122, 2017.

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

[8] F. Nielsen and R. Nock, "Generalizing Skew Jensen Divergences and Bregman Divergences With Comparative Convexity," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1123-1127, 2017.

[9] S.-i. Amari, *Information geometry and its applications*, Springer, 2016.

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

[11] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *JMLR*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[12] F. Nielsen, "A family of statistical symmetric divergences based on Jensen's inequality," arXiv:1009.4004, 2010.

[13] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Jensen-Bregman logdet divergence with application to efficient similarity search for covariance matrices," *IEEE TPAMI*, vol. 35, no. 9, pp. 2161–2174, 2013.

[14] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhyā*, pp. 401–406, 1946.

[15] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," in *International Conference on Spoken Language (ICSLP)*, 1996, vol. 4, pp. 2005–2008.

[16] C. Sheth and R V. Babu, "Object saliency using a background prior," in *IEEE ICASSP*, 2016, pp. 1931–1935.

[17] D. Kawai, K. Yamamoto, and S. Nakagawa, "Lyric recognition in monophonic singing using pitch-dependent DNN," in *IEEE ICASSP*, 2017, pp. 326–330.

[18] F. Nielsen, "An information-geometric characterization of Chernoff information," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 269–272, 2013.

[19] R. Boyer and F. Nielsen, "Information geometry metric for random signal detection in large random sensing systems," in *IEEE ICASSP*, 2017, pp. 4471–4475.

[20] J. Dong, D. Niu, and E. Song, "An approach based on Chernoff distance to sparse sensing for distributed detection," in *IEEE Information Fusion (Fusion)*, 2017, pp. 1–6.

[21] F. Nielsen, "A generalization of the Jensen divergence: The chord gap divergence," *CoRR*, vol. abs/1709.10498, 2017.

[22] F. Nielsen and R. Nock, "A closed-form expression for the Sharma-Mittal entropy of exponential families," *Journal of Physics A: Mathematical and Theoretical*, vol. 45, no. 3, pp. 032003, 2011.

[23] A. Hero, B. Ma, O. Michel, and J. Gorman, "Alpha-divergence for classification, indexing and retrieval," *Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. Mich*, 2001.

[24] G. B. Coleman and H. C. Andrews, "Image segmentation by clustering," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 773–785, 1979.

[25] X. Li, W. Hu, and W. Hu, "A coarse-to-fine strategy for vehicle motion trajectory clustering," in *ICPR*, 2006, vol. 1, pp. 591–594.

[26] D.-C; Park, Y. Lee, and D.-M. Woo, "Classification of audio signals using a Bhattacharyya kernel-based centroid neural network," *Advances in Knowledge Discovery and Data Mining*, pp. 604–611, 2009.

[27] S. Har-Peled and S. Mazumdar, "On coresets for k-means and k-median clustering," in *ACM Symposium on Theory Of Computing (STOC)*, 2004, pp. 291–300.

[28] P. D. Tao, "The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems," *Annals of operations research*, vol. 133, no. 1-4, pp. 23–46, 2005.

[29] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (CCCP)," in *Advances in neural information processing systems (NIPS)*, 2002, pp. 1033–1040.

[30] F. Nielsen and R. Nock, "Total Jensen divergences: definition, properties and clustering," in *IEEE ICASSP*, 2015, pp. 2016–2020.

[31] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *ACM-SIAM symposium on Discrete algorithms (SODA)*, 2007, pp. 1027–1035.

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

[33] M. R. Ackermann and J. Blömer, "Bregman clustering for separable instances," in *Scandinavian Symposium and Workshops on Algorithm Theory (SWAT)*, pp. 212–223, 2010.

[34] H. Sun, Z. Liu, and L. Kong, "A document clustering method based on hierarchical algorithm with model clustering," in *Advanced Information Networking and Applications-Workshops*. IEEE, 2008, pp. 1229–1233.

[35] Z. Yu, P. Luo, J.e You, H.-S. Wong, H. Leung, S. Wu, J. Zhang, and G. Han, "Incremental semi-supervised clustering ensemble for high dimensional data clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 701–714, 2016.