Skew Jensen-Bregman Voronoi Diagrams*

Frank Nielsen[1] and Richard Nock[2]

[1] École Polytechnique Palaiseau, France
Sony Computer Science Laboratories Inc., Tokyo, Japan
nielsen@lix.polytechnique.fr
[2] CEREGMIA-UAG, University of Antilles-Guyane Martinique, France
rnock@martinique.univ-ag.fr

— Dedicated to the victims of Japan Tohoku earthquake (March 2011).

Abstract. A Jensen-Bregman divergence is a distortion measure defined by a Jensen convexity gap induced by a strictly convex functional generator. Jensen-Bregman divergences unify the squared Euclidean and Mahalanobis distances with the celebrated information-theoretic Jensen-Shannon divergence, and can further be skewed to include Bregman divergences in limit cases. We study the geometric properties and combinatorial complexities of both the Voronoi diagrams and the centroidal Voronoi diagrams induced by such as class of divergences. We show that Jensen-Bregman divergences occur in two contexts: (1) when symmetrizing Bregman divergences, and (2) when computing the Bhattacharyya distances of statistical distributions. Since the Bhattacharyya distance of popular parametric exponential family distributions in statistics can be computed equivalently as Jensen-Bregman divergences, these skew Jensen-Bregman Voronoi diagrams allow one to define a novel family of statistical Voronoi diagrams.

Keywords: Jensen's inequality, Bregman divergences, Jensen-Shannon divergence, Jensen-von Neumann divergence, Bhattacharyya distance, information geometry.

1 Introduction

The Voronoi diagram is one of the most fundamental combinatorial structures studied in computational geometry [2] often used to characterize solutions to geometric problems [3] like the minimum spanning tree, the smallest enclosing ball, motion planning, etc. For a given set of sites, the Voronoi diagram partitions the space into elementary proximity cells denoting portions of space closer to a

* This journal article revises and extends the conference paper [1] presented at the International Symposium on Voronoi Diagrams (ISVD) 2010. This paper includes novel extensions to matrix-based Jensen-Bregman divergences, and present the general framework of skew Jensen-Bregman Voronoi diagrams that include Bregman Voronoi diagrams as particular cases. Supporting materials available at http://www.informationgeometry.org/JensenBregman/

M.L. Gavrilova et al. (Eds.): Trans. on Comput. Sci. XIV, LNCS 6970, pp. 102–128, 2011.

site than to any other one. Voronoi diagrams have been generalized in many ways by considering various types of primitives (points, lines, balls, etc.) and distance functions (L_p Minkowski distances [4], convex distances[1] [5], Bregman divergences [6], etc.) among others.

In this work, we introduce a novel class of information-theoretic distortion measures called *skew Jensen-Bregman divergences* that generalizes the celebrated Jensen-Shannon divergence [7] in information theory [8]. We study both Voronoi diagrams and centroidal Voronoi tessellations [9] with respect to that family of distortion measures. As a by-product, we also show that those skew Jensen-Bregman Voronoi diagrams allow one to characterize statistical Voronoi diagrams induced by the skew Bhattacharyya statistical distance on a given set of probability measures.

Our main contributions are summarized as follows:

- We define the family of Jensen-Bregman divergences extending the concept of Jensen-Shannon divergence, and show that those divergences appear when symmetrizing Bregman divergences,
- By skewing those Jensen-Bregman divergences, we obtain parametric divergences that encapsulate Bregman divergences as limit cases,
- We study the combinatorial complexities of skew Jensen-Bregman Voronoi diagrams (generalizing Bregman Voronoi diagrams [6]),
- We describe an efficient algorithm to arbitrarily finely estimate the Jensen-Bregman centroids, and extend its scope to matrix-valued divergences,
- We show that the statistical Bhattacharyya distance of parametric exponential family distributions amount to compute a Jensen-Bregman divergence on the corresponding parameters.

The paper is organized as follows: Section 2 introduces the class of Jensen-Bregman divergences and described the link with Bregman divergence symmetrization [6]. Section 3 defines the Voronoi diagram with respect to Jensen-Bregman divergences, and bound their complexity by studying the corresponding minimization diagram and investigating properties of the bisectors and level sets. Section 4 presents the Jensen-Bregman centroids, design an efficient iterative estimation algorithm, and provide some experiments on the centroidal Jensen-Bregman Voronoi tessellations. It is followed by Section 5 that extends Jensen-Bregman divergences to matrix-valued data sets. Section 6 introduces a skew factor in the divergence and show how to obtain Bregman Voronoi diagrams [6] as extremal cases. Finally, Section 7 concludes the paper by mentioning the underlying differential geometry.

In order to not overload the paper, Appendix A introduces the class of statistical Bhattacharyya distances, and show how it is equivalent to Jensen-Bregman divergences when distributions belong to the same parametric exponential family.

[1] Convex distances may not necessarily be metrics [5]. A metric satisfies both the symmetry and triangular inequality axioms.

2 Jensen-Bregman Divergences

There is a growing interest in studying *classes* of distortion measures instead of merely choosing a single distance. This trend is attested in many fields of computer science including computational geometry, machine learning, computer vision, and operations research. The goal is to study and design *meta-algorithms* that can provably run correct on a class of distances rather than on a single distance at hand. Among such classes of distances, the Bregman distances [6] appear attractive because this family of dissimilarity measures encapsulate both the geometric (squared) Euclidean distance and the information-theoretic relative entropy. A Bregman distance B_F on an open convex space $\mathcal{X} \subseteq \mathbb{R}^d$ is defined for a strictly convex and differentiable function F as

$$B_F(p,q) = F(p) - F(q) - \langle p-q, \nabla F(q) \rangle, \tag{1}$$

where $\langle p,q \rangle = p^T q$ denotes the inner product, and

$$\nabla F(x) = \left[\frac{\partial F}{\partial x_1} ... \frac{\partial F}{\partial x_d} \right]^T \tag{2}$$

the partial derivatives. Choosing $F(x) = \sum_{i=1}^d x_i^2 = \langle x,x \rangle$ yields the squared Euclidean distance $B_{x^2}(p,q) = \|p-q\|^2$, and choosing $F(x) = \sum_{i=1}^d x_i \log x_i = S(x)$ yields the relative entropy, called the Kullback-Leibler divergence [8]. The Kullback-Leibler divergence is defined for normalized d-dimensional "distribution" points (i.e., points falling in the $(d-1)$-dimensional unit simplex denoting discrete distributions) as:

$$I(p,q) = B_S(p,q) = \sum_{i=1}^d p_i \log \frac{p_i}{q_i}. \tag{3}$$

Handling Bregman divergences instead of the (squared) Euclidean distance brings the opportunity to enlarge the field of applications of geometric algorithms to other settings like statistical contexts.

The generator function F can be interpreted as a measure of *information* (namely a negative entropy, since entropies are usually concave functions [8]). In information theory, the entropy measures the amount of uncertainty of a random variable. For example, one expects that the entropy is maximized for the uniform distribution. Axiomatizing a few behavior properties [8] of entropy yields the unique concave function

$$H(x) = x \log \frac{1}{x} = -x \log x, \tag{4}$$

called the *Shannon entropy* ($-H(x)$ is the convex Shannon information).

Bregman divergences are *never* metrics, and provably only symmetric for the generalized quadratic distances obtained for generator $F(x) = Qx$ for a positive

definite matrix $Q \succ 0$. Thus those distances are preferably technically called divergences instead of distances. Bregman divergences satisfy

$$B_F(p,q) \geq 0, \tag{5}$$

with equality if and only if $p = q$. This is called Gibb's inequality for the particular case of Kullback-Leibler divergence, and can be demonstrated by using a geometric argument as follows: Let $\hat{x} = (x, F(x))$ denote the lifting map of point x to the potential function plot $\mathcal{F} = \{\hat{x} = (x, F(x)) \mid x \in \mathcal{X}\}$. The Bregman divergence measures the vertical distance between two non-vertical hyperplanes: The hyperplane H_q tangent at the potential function $\mathcal{F} = (x, F(x))$ at lifted point $\hat{q}$:

$$H_q(x) = F(q) + \langle x - q, \nabla F(q) \rangle, \tag{6}$$

and its translate H'_q passing through $\hat{p}$:

$$H'_q(x) = F(p) + \langle x - p, \nabla F(q) \rangle. \tag{7}$$

We have

$$B_F(p,q) = H'_q(x) - H_q(x), \tag{8}$$

independent of the position of x. This geometric interpretation is illustrated in Figure 1.

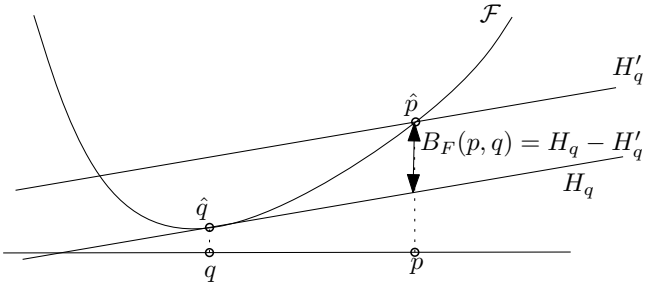


Fig. 1. Interpreting the Bregman divergence as the vertical distance between the tangent plane at q and its translate passing through p (with identical slope $\nabla F(q)$)

Since Bregman divergences are not symmetric, one can naturally symmetrize them as follows:

$$S_F(p,q) = \frac{B_F(p,q) + B_F(q,p)}{2} \tag{9}$$

$$= \frac{1}{2}\langle p - q, \nabla F(p) - \nabla F(q) \rangle. \tag{10}$$

That is indeed what happened historically with Jeffreys [10] considering the J-measure as the sum of sided I measures:

$$J(p,q) = I(p,q) + I(q,p) \tag{11}$$

However, there are two major *drawbacks* for such an information-theoretic divergence:

1. The divergence S_F may be undefined (unbounded): For example, considering the negative Shannon entropy, the symmetrized Kullback-Leibler divergence is undefined if for some coordinate $q_i = 0$ and $p_i \neq 0$.[2]
2. The divergence S_F is not bounded in terms of the variational *metric* distance $V(p,q) = \sum_{i=1}^{d} |p_i - q_i|$ (L_1-metric [4]). For the Kullback-Leibler divergence, it is known that $\mathrm{KL}(p,q) \geq \frac{1}{2}V^2(p,q)$. Such kinds of bounds are called *Pinsker's inequalities* [11].

To overcome those two issues, Lin [7] proposed a new divergence built on the Kullback-Leibler divergence called the *Jensen-Shannon divergence*. The Jensen-Shannon divergence is defined as

$$\mathrm{JS}(p,q) = \mathrm{KL}\left(p, \frac{p+q}{2}\right) + \mathrm{KL}\left(q, \frac{p+q}{2}\right). \tag{12}$$

This divergence is always (1) defined, (2) finite, and furthermore (3) bounded by the variational L_1-metric:

$$V^2(p,q) \leq \mathrm{JS}(p,q) \leq V(p,q) \leq 2 \tag{13}$$

Those two different ways to symmetrize the KL divergence J (S_F for Shannon entropy) and JS are related by the following inequality

$$J(p,q) \geq 4\ \mathrm{JS}(p,q) \geq 0. \tag{14}$$

The Jensen-Shannon divergence can be interpreted as a measure of *diversity* of the source distributions p and q to the *average* distribution $\frac{p+q}{2}$. In the same vein, consider the following Bregman symmetrization [12,13]:

$$J_F(p,q) = \frac{B_F(p, \frac{p+q}{2}) + B_F(q, \frac{p+q}{2})}{,} \tag{15}$$

$$= \frac{F(p)+F(q)}{2} - F\left(\frac{p+q}{2}\right) = J_F(q,p). \tag{16}$$

For d-dimensional multivariate data, we define the corresponding Jensen divergences coordinate-wise as follows:

$$J_F(p,q) = \sum_{i=1}^{d} J_F(p_i,q_i) = \sum_{i=1}^{d} \frac{F(p_i)+F(q_i)}{2} - F\left(\frac{p_i+q_i}{2}\right). \tag{17}$$

[2] We may enforce definiteness by assuming the distributions are mutually absolutely continuous to each others [8].

Jensen-Bregman divergences J_F are always finite ($0 \leq J_F < \infty$) on the domain $\mathcal{X}$ (because entropies F measuring uncertainties are finite quantities: $F(x) < \infty$). J_F denote the Bregman divergence of the source distributions to the average distributions. Another way to interpret this family of divergences is to say that the Jensen-Bregman divergence is the average of the (negative) entropies minus the (negative) entropy of the average. For the negative Shannon entropy, we find the celebrated Jensen-Shannon divergence. Those divergences are *not* translation-invariant, and we require F to be *strictly convex* since for linear generators $L(x) = \langle a, x\rangle + b$, one does not discriminate distributions (i.e., $J_L(p, q) = 0\ \forall p, q$).

This family of divergences can be termed *Jensen-Bregman divergences*.[3] Since F is a strictly convex function, J_F is nonnegative and equal to zero if and only if $p = q$. Figure 2 gives a geometric interpretation of the divergence as the vertical distance between $(\frac{p+q}{2}, F(\frac{p+q}{2}))$ and the midpoint of the segment $[(p, F(p)), (q, F(q))]$. Positive-definiteness follows from the Jensen's inequality.

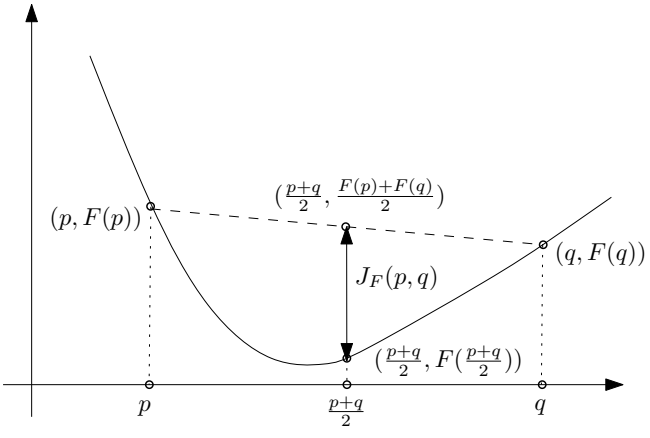


Fig. 2. Interpreting the Jensen-Bregman divergence as the vertical distance between the midpoint of segment $[(p, F(p)), (q, F(q))]$ and the midpoint of the graph plot $\left(\frac{p+q}{2}, F\left(\frac{p+q}{2}\right)\right)$

Note that Jensen-Bregman divergences are defined modulo affine terms $\langle a, x\rangle + b$. (Indeed, let $G(x) = F(x) + \langle a, x\rangle + b$, then one checks that $J_F(p, q) = J_G(p, q)$.) Thus we can choose coefficients $b = -F(0)$ and $a = -\nabla F(0)$ to fix unambiguously the generator. This means that the plot $(x, F(x))$ of the convex function touches the origin at its minimum value.

Jensen-Bregman divergences contain all generalized quadratic distances ($F(x) = \langle Qx, x\rangle$ for a positive definite matrix $Q \succ 0$), well-known in computer

[3] Or Burbea-Rao divergences [14]) or Jensen divergences. We prefer the term Jensen-Bregman because as we shall see (1) skew Jensen-Bregman divergences include Bregman divergences in the limit cases, and (2) they are obtained by symmetrizing Bregman divergences *à la* Jensen-Shannon.

vision as the squared Mahalanobis distances (squared Euclidean distance obtained for $Q = I$, the identity matrix).

$$\begin{aligned} J_F(p,q) &= \frac{F(p)+F(q)}{2} - F\left(\frac{p+q}{2}\right) \\ &= \frac{2\langle Qp,p\rangle + 2\langle Qq,q\rangle - \langle Q(p+q),p+q\rangle}{4} \\ &= \frac{1}{4}(\langle Qp,p\rangle + \langle Qq,q\rangle - 2\langle Qp,q\rangle) \\ &= \frac{1}{4}\langle Q(p-q),p-q\rangle \\ &= \frac{1}{4}\|p-q\|_Q^2. \end{aligned}$$

It is well-known that the square root of the Jensen-Shannon divergence (using Shannon entropy generator $F(x) = -x\log x$) is a metric. However, the square root of Jensen-Bregman divergences are *not* always metric. A Jensen-Bregman divergence is said *separable* if it can be decomposed independently dimension-wise:

$$F(x) = \sum_{i=1}^{d} f_i(x_i), \tag{18}$$

with all f_i's strictly convex functions. Usually all f_i's are taken as an identical univariate function. For example, Shannon (convex) information (the negative of Shannon concave entropy) is defined as

$$I(x) = -\sum_{i=1}^{d} x_i \log x_i, \tag{19}$$

for x belonging to the $(d-1)$-dimensional simplex $\mathbb{S}_{d-1}$ of discrete probabilities ($\sum_{i=1}^{d} x_i = 1$). Shannon information can be extended to the non-normalized *positive measures* $\mathbb{P}_d$ by taking $I(x) = -\sum_{i=1}^{d} x_i \log x_i - x_i$.

Appendix A shows that those Jensen-Bregman distances encapsulate the class of statistical Bhattacharyya distances for a versatile family of probability measures called the exponential families. Let us now consider Jensen-Bregman Voronoi diagrams.

3 The Voronoi Diagram by Jensen Difference

We have shown that Jensen-Bregman divergences are an important class of distortion measures containing both the squared Euclidean/Mahalanobis distance (non-additive quadratic entropy [8]) and the Jensen-Shannon divergence (additive entropy [8]). Note that one key property of the Euclidean distance is that

$D(\lambda p, \lambda q) = \lambda D(p, q)$. That is, it is a distance function of *homogeneous degree* 1. A distance is said of homogeneous degree α if and only if

$$D(\lambda p, \lambda q) = \lambda^{\alpha} D(p, q). \tag{20}$$

Usually, Jensen-Bregman divergences are not homogeneous except for the following three *remarkable* generators:

- Burg entropy ($\alpha = 0$)

$$F(x) = -\log x, \quad J_F(p, q) = \log \frac{p+q}{2\sqrt{pq}} \tag{21}$$

  (logarithm of the ratio of the arithmetic mean over the geometric mean),
- Shannon entropy ($\alpha = 1$)

$$F(x) = x \log x, \quad J_F(p, q) = \frac{1}{2}\left(p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q}\right) \tag{22}$$

- Quadratic entropy ($\alpha = 2$)

$$F(x) = x^2, \quad J_F(p, q) = \frac{1}{4}(p-q)^2 \tag{23}$$

Since Jensen-Bregman Voronoi diagrams include the ordinary Euclidean Voronoi diagram [3],[4] the complexity of those diagrams is at least the complexity of Euclidean diagrams [15,6]: namely, $\Theta(n^{\lceil \frac{d}{2} \rceil})$. In general, the complexity of Voronoi diagrams by an arbitrary distance function (under mild conditions) is at most $O(n^{d+\epsilon})$ for any $\epsilon > 0$, see [16,17]. Thus as the dimension increases there is a potential quadratic gap in the combinatorial complexity between the Euclidean and general distance function diagrams.

Let us analyze the class of Jensen-Bregman diagrams by studying the induced minimization diagram and characterizing the bisector structure.

3.1 Voronoi Diagrams as Minimization Diagrams

Given a point set $\mathcal{P} = \{p_1, ..., p_n\}$ of n sites,[5] the Jensen-Bregman Voronoi diagram partitions the space into elementary *Voronoi cells* such that the Voronoi cell $V(p_i)$ associated to site p_i is the loci of points closer to p_i than to any other point of $\mathcal{P}$ with respect to the Jensen-Bregman divergence:

$$V(p_i) = \{p \mid J_F(p, p_i) < J_F(p, p_j) \ \forall j \neq i\}. \tag{24}$$

[4] Since the Voronoi diagrams by any strictly monotonous increasing function of a distance coincides with the Voronoi diagrams of that distance, the squared Euclidean Voronoi diagram coincides with the ordinary Voronoi diagram.

[5] Without loss of generality, we assumed points distinct and in general position.

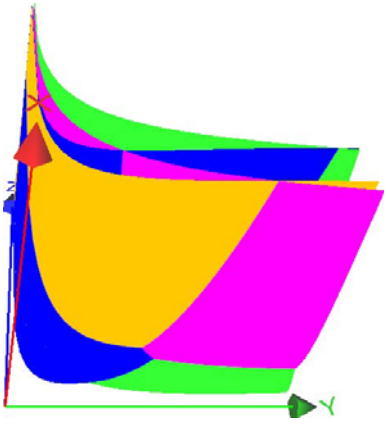


Fig. 3. The 2D Jensen-Burg Voronoi diagram of 4 points from the corresponding lower envelope of corresponding 3D functions

For each Voronoi site p_i, consider the following *anchored distance function* to that site:

$$D_i(x) = J_F(x, p_i) = \frac{F(p_i) + F(x)}{2} - F\left(\frac{p_i + x}{2}\right) \tag{25}$$

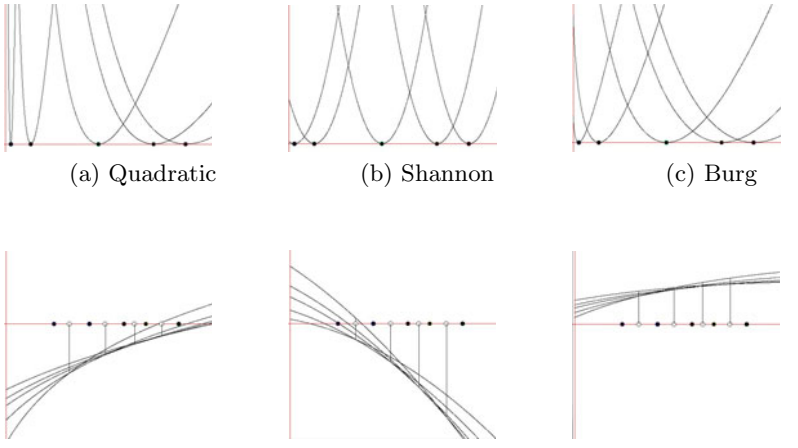
Thus the Voronoi diagram amounts to a minimization diagram. This minimization task can be solved by computing the lower envelope of $(d+1)$-dimensional functions $(x, D_i(x))$. The projection of the lower envelope (resp. upper envelope) yields the Jensen-Bregman Voronoi diagram (resp. farthest Jensen-Bregman Voronoi diagram). Figure 3 displays the lower envelope of four 3D functions for the Burg entropy generator (homogeneous degree $\alpha = 0$).

In general, besides the ordinary Euclidean case with $F(x) = x^2$, the equation of the bisector can be tricky to manipulate, even in the planar case. For example, consider the Burg entropy $(F(x) = -\log x)$. Using $\sum \log \leftrightarrow \log \prod$, the Burg bisector $B(p, q)$ for the corresponding separable Jensen-Burg distance can be written as:

$$B(p, q) : \prod_{i=1}^{d} \frac{p_i + x_i}{\sqrt{p_i}} = \prod_{i=1}^{d} \frac{q_i + x_i}{\sqrt{q_i}}, \tag{26}$$

where $p = (p_1, ..., p_d)$ and $q = (q_1, ..., q_d)$ denote the coordinates of p and q, respectively.

We next concentrate on a *concave-convex structural property* of the Jensen-Bregman bisector. But first, we recall some prior work on Voronoi diagrams. The Voronoi diagrams with respect to convex functions has been studied [18]. However, note that Jensen-Bregman divergences are *not* necessarily convex.



(a) Quadratic (b) Shannon (c) Burg

Fig. 4. (Top) 1D Voronoi diagram from the lower envelope of corresponding anchored distance functions for the (a) quadratic, (b) Shannon and (c) Burg entropies. (Bottom) minimum of the functions $D'_i(\cdot)$ (removing the common term $\frac{1}{2}F(x)$).

Indeed, without loss of generality, consider separable Jensen-Bregman divergences, and let us look at the second-order derivatives of a univariate Jensen-Bregman divergence. We have

$$D'_p(x) = J'_F(x,p) = \frac{1}{2}F'(x) - \frac{1}{2}F'\left(\frac{p+x}{2}\right) \tag{27}$$

and

$$D''_p(x) = J''_F(x,p) = \frac{1}{2}F''(x) - \frac{1}{4}F''\left(\frac{p+x}{2}\right). \tag{28}$$

This second-order derivative is *not* necessarily always strictly positive. For example, consider $F(x) = x^3$ on $\mathbb{R}^+$ (with $F''(x) = 6x$). We have $D''_p(x) = 3(x - \frac{p+x}{4})$; This is non-negative for $x \geq \frac{p}{3}$ only. That means that some Jensen-Bregman divergences are neither convex nor concave either. Another typical example is the Jensen-Burg entropy ($F(x) = -\log x$ and $F''(x) = 1/x^2$). Indeed, the anchored distance $J_F(x,p)$ at p is strictly convex if $x > p(1+\sqrt{2})$ and strictly concave if $x < p(1+\sqrt{2})$. However, Jensen-Shannon divergence (defined for generator $F(x) = x\log x$ with $F''(x) = 1/x$) is convex for all values on the positive orthant (positive measures $\mathbb{P}_2$): Indeed, $D''_F(x) = \frac{1}{2x} - \frac{1}{4}\frac{2}{p+x} = \frac{1}{2}(\frac{1}{x} - \frac{1}{p+x}) > 0$ for all $p > 0$ ($p \in \mathbb{P}$).

In general, a necessary condition is that F is strictly convex: Indeed, choose $x = p$ (for any arbitrary p), it comes that $D''_p(x) = \frac{1}{4}F''(p)$ that is positive if and only if $F''(p) > 0$. That is, it is required that F be strictly convex.

In the general case, J_F is convex if and only if its Hessian is positive definite:[6] $\nabla^2 J_F(\cdot, p) \succ 0 \ \forall p$:

$$\nabla^2 J_F = \begin{bmatrix} \frac{\partial^2 J_F(x,y)}{\partial x^2} & \frac{\partial^2 J_F(x,y)}{\partial x \partial y} \\ \frac{\partial^2 J_F(x,y)}{\partial x \partial y} & \frac{\partial^2 J_F(x,y)}{\partial y^2} \end{bmatrix} \tag{29}$$

$$= \begin{bmatrix} \frac{F''(x)}{2} - \frac{1}{4}F''(\frac{x+y}{2}) & -\frac{1}{4}F''(\frac{x+y}{2}) \\ -\frac{1}{4}F''(\frac{x+y}{2}) & \frac{F''(y)}{2} - \frac{1}{4}F''(\frac{x+y}{2}) \end{bmatrix} \succ 0 \tag{30}$$

Considering separable divergences J_F, the positive definiteness condition of the Hessian becomes

$$F''(x) > F''(\frac{x+y}{2}) - F''(x) \tag{31}$$

It follows that the Jensen-Shannon (separable) divergence ($F(x) = x \log x - x$, $F''(x) = \frac{1}{x}$) is a strictly convex distance function on the set of positive measures $\mathcal{X} = \mathbb{R}_{++}$ since $\frac{1}{x} > \frac{2}{x+y} - \frac{1}{x}$ for all $x, y > 0$.

Lemma 1. *Jensen-Bregman divergences are not necessarily strictly convex nor strictly concave distortion measures. Jensen-Shannon divergence is a strictly convex function on the set $\mathbb{P}_d$ of positive measures. Separable Jensen-Bregman divergences J_F on domain $\mathcal{X}^d$ are strictly convex distance functions if and only if $F''(x) > F''(\frac{x+y}{2}) - F''(x) > 0$ for $x, y \in \mathcal{X}$.*

Lihong [19,18] studied the Voronoi diagrams in 2D and 3D under a *translation-invariant* convex distance function (e.g., a polyhedral convex distance). Translation-invariant means that a convex object C gives a distance profile, and the distance between two points p and q is the smallest scaling factor so that a homothet of C centered at p touches q. Note that Jensen-Bregman divergences are not invariant under translation.

Recently, Dickerson et al. [20] studied the planar Voronoi diagram for *smoothed* separable convex distances. They show that provided that the functions of the minimization diagrams satisfy the constraint $f'''f' < (f'')^2$, then the 2D Voronoi diagram has linear complexity and can be computed using a randomized algorithm in $\tilde{O}(n \log n)$ time. In fact, in that case, the distance *level sets* $\{D_{p_i}(x) = l\}_l$ (iso-distance level) yield pseudo-circles, and the arrangement of bisectors are pseudo-lines. However, if the condition $f'''f' < (f'')^2$ fails, the 3D minimization diagram (and corresponding 2D Voronoi diagram) may have *quadratic* complexity. For example, choosing $f(x) = e^{x^2}$, and $F(x, y) = e^{x^2} + e^{y^2}$ yields potentially a quadratic complexity diagram.

Consider a d-dimensional finite point set $p_1, ..., p_n$, and let $p_{i,1}, ..., p_{i,d}$ denote the coordinates of point p_i for all $i \in \{1, ..., n\}$. We consider separable Jensen-Bregman divergences. Let $x_1, ..., x_d$ denote the coordinates of point x.

[6] A matrix M is said positive definite iff. $x^T M x > 0$ for all $x \neq 0$. A positive definite matrix has all its eigenvalues strictly positive, and hence the trace (sum of eigenvalues) and determinant (product of eigenvalues) are necessarily positive.

Since the term $\frac{F(x)}{2}$ are shared by all D_i's functions, we can remove it equivalently from all anchored distance functions. Therefore the minimization diagram $\min_i D_i(x)$ is equivalent to the minimization diagram of the functions

$$D'_i(x) = \frac{1}{2}F(p_i) - F\left(\frac{p_i + x}{2}\right), \tag{32}$$

or equivalently using separable generator by

$$D'_i(x) = \sum_{k=1}^{d} \frac{1}{2}F(p_{i,k}) - F\left(\frac{p_{i,k} + x_k}{2}\right). \tag{33}$$

This minimization diagram can be viewed as the lower envelope of n concave functions (entropy function $-F$) in dimension $d+1$. The *vertical shift* corresponds to a weight $F(p_i) = \sum_{k=1}^{d} F(p_{i,k})/2$. Let us write the equation of a bisector (p, q):

$$B(p,q) : \frac{F(p)}{2} - F\left(\frac{x+p}{2}\right) = \frac{F(q)}{2} - F\left(\frac{x+q}{2}\right) \tag{34}$$

$$: \sum_{k=1}^{d} \frac{F(p_k)}{2} - F\left(\frac{x_k + p_k}{2}\right) = \sum_{k=1}^{d} \frac{F(q_k)}{2} - F\left(\frac{x_k + q_k}{2}\right). \tag{35}$$

That is, we get

$$B(p,q) : \left(F\left(\frac{x+q}{2}\right) - F\left(\frac{x+p}{2}\right)\right) + \left(\frac{F(p)}{2} - \frac{F(q)}{2}\right) = 0 \tag{36}$$

$$: \sum_{k=1}^{d} \left(F\left(\frac{x_k + q_k}{2}\right) - F\left(\frac{x_k + p_k}{2}\right)\right) + \sum_{k=1}^{d} \left(\frac{F(p_k)}{2} - \frac{F(q_k)}{2}\right) = 0 \tag{37}$$

The bisector is thus interpreted as the sum of a *convex* function

$$\sum_{k=1}^{d} F\left(\frac{x_k + q_k}{2}\right) - \frac{F(p_k)}{2}$$

with a *concave* function

$$\sum_{k=1}^{d} -F\left(\frac{x_k + p_k}{2}\right) - \frac{F(q_k)}{2}.$$

The next section on centroidal Voronoi tessellations show how to handle this concave-convex structural property using a tailored optimization mechanism.

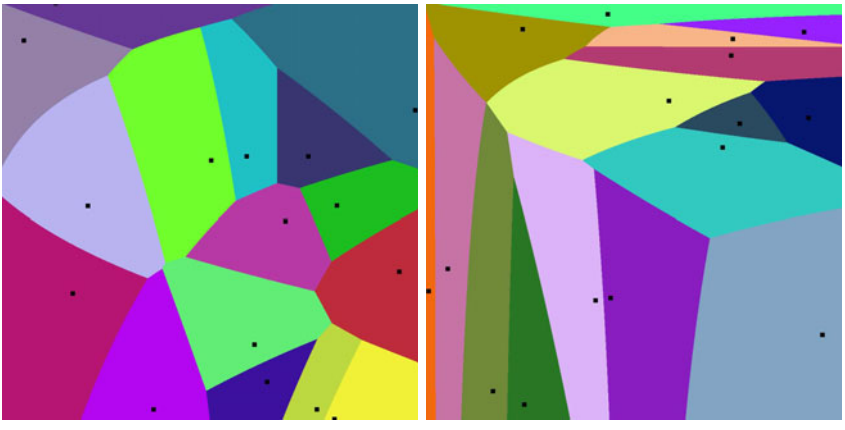


Fig. 5. (Left) The Jensen-Shannon Voronoi diagram for a set of 16 points (positive arrays denoting unnormalized probability distributions). (Right) The Jensen-Burg Voronoi diagram for the Burg entropy $F(x) = -\sum_{i=1}^{d} \log x_i$.

4 Jensen-Bregman Centroidal Voronoi Diagrams

The centroidal Voronoi diagram [9] (or centroidal Voronoi tesselation; CVT for short) is defined as follows: First, we fix a number of generators n, and a compact domain $\mathcal{X}$ (say, a unit square). Then we ask to find the locations of the generators so that the induced Voronoi cells have (approximately) more or less the same area. Figure 6(b) shows a CVT for a set of 16 points. A CVT is computed iteratively by first initializing the generators to arbitrary position (Figure 6(a)), and by iteratively relocating those generators to the center of mass of their Voronoi cell (Lloyd iteration). Proving that such a scheme is always converging is still a difficult open problem of computational geometry [9], although that in practice it is known to converges quickly. Instead of relocating to the center of mass (barycenter) according to a uniform density distribution (i.e., to the centroid or geometric center of the cell), we can relocate those generators to the barycenter of the cell according to an underlying non-uniform density distribution. This is one technique commonly used in non-photorealistic rendering (NPR) called stippling [21], the art of pointillism. Figure 6(c) is a source image representing the underlying grey intensity distribution. Figure 6(d) is the stippling effect produced by computing a CVT with respect to the underlying image density.

To extend the centroidal Voronoi tesselations to Jensen-Bregman divergences, we first need to define centroids (and barycenters) with respect to this dissimilarity measure. Consider a finite set of points $\mathcal{P} = \{p_1, ..., p_n\}$. The *Jensen-Bregman centroid* is defined as the minimizer of the average Jensen-Bregman distance:

$$c^* = \arg\min_{c} \sum_{i=1}^{n} \frac{1}{n} J_F(p_i, c) \tag{38}$$

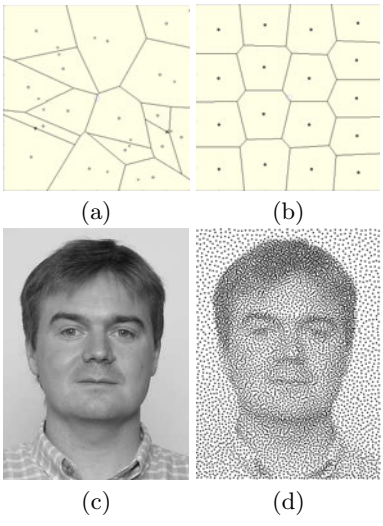


Fig. 6. (Top) Centroidal Voronoi diagram of 16 sites: (a) initialization, and (b) after a few iterations. (Bottom) Application to image stippling: (c) grey density image and (d) centroidal Voronoi diagram according to the underlying density.

By choosing $F(x) = \langle x, x \rangle$, we minimize the sum of the squared Euclidean distances, and find the usual Euclidean centroid.[7] Similarly, the barycenter is defined with respect to (normalized) weights (interpreted as point multiplicities):

$$c^* = \arg\min_c \sum_{i=1}^{n} w_i J_F(p_i, c) = \arg\min_c L(c) \tag{39}$$

Using the structure of the optimization problem, we can use the Convex-ConCave Procedure [22] (CCCP), a general purpose loss function minimizer. Indeed, we can always decompose an arbitrary (non-convex) function as the sum of a *convex* function and *concave* function (or the difference of two convex functions), provided that the Hessian of the loss function function is bounded:[8]

$$L(c) = L_{\text{convex}}(c) + L_{\text{concave}}(c). \tag{40}$$

For the Jensen-Bregman centroid, this decomposition is given *explicitly* as follows:

$$L_{\text{convex}}(c) = \frac{F(c)}{2} \tag{41}$$

[7] If instead of minimizing the squared Euclidean distance, we consider the Euclidean distance, we do not get closed-form solution. This is the so-called Fermat-Weber point.

[8] Always bounded on a compact.

$$L_{\text{concave}}(c) = -\sum_{i=1}^{n} F\left(\frac{p_i + c}{2}\right), \tag{42}$$

since the sum of concave functions is a concave function. The CCCP approach consists in setting the gradient to zero: $\nabla_x L(x) = 0$. We get

$$\frac{1}{2}\nabla F(x) - \sum_{i=1}^{n} \frac{w_i}{2} \nabla F\left(\frac{x + p_i}{2}\right) = 0. \tag{43}$$

That is, we need to solve equivalently for

$$\nabla F(x) = \sum_{i=1}^{n} w_i \nabla F\left(\frac{x + p_i}{2}\right) \tag{44}$$

Since F is *strictly* convex and differentiable, we have ∇F that is *strictly* monotone increasing (because the Hessian is positive definite, i.e. $\nabla^2 F \succ 0$), and the reciprocal gradient ∇F^{-1} is well-defined.

Thus solving Eq. 44 amounts to solve for

$$x = \nabla F^{-1}\left(\sum_{i=1}^{n} w_i \nabla F\left(\frac{x + p_i}{2}\right)\right) \tag{45}$$

Starting from an arbitrary initial value x_0 of x (say, the Euclidean center of mass), the optimization proceeds iteratively as follows:

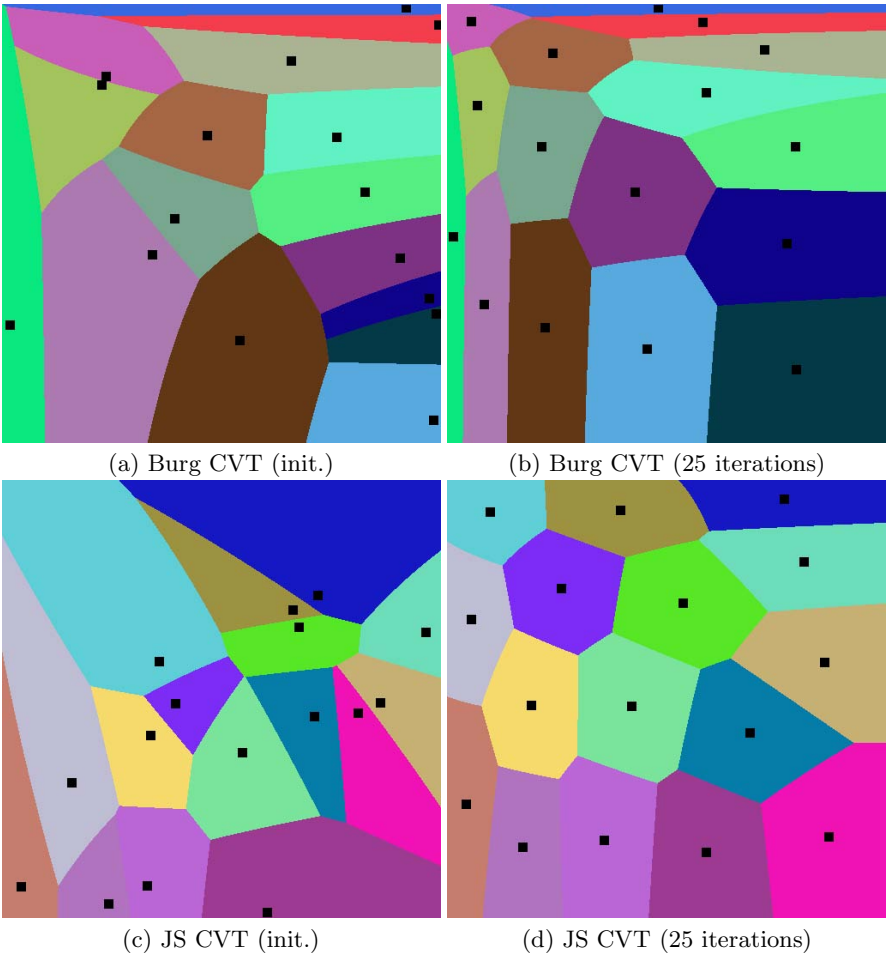
$$x_{t+1} = \nabla F^{-1}\left(\sum_{i=1}^{n} w_i \nabla F\left(\frac{x_t + p_i}{2}\right)\right). \tag{46}$$

For the Jensen-Shannon (separable) divergence defined on positive measures, we thus update the centroid independently on each coordinate by

$$x_{t+1} = \frac{n}{2\sum_{i=1}^{n} \frac{1}{x_t + p_i}}.$$

The CCCP algorithm guarantees monotonicity and convergence to a local minimum or saddle point. For the Jensen-Shannon divergence, this local minimum yields the global minimum since the distance function is strictly convex. Note that for the quadratic entropy $F(x) = \langle x, x \rangle$, we get a closed-form solution (i.e., the center of mass).

However, in general, we do not obtain a closed-form solution, and can only estimate the Jensen-Bregman barycenters up to some arbitrary precision. Thus, we do not have closed-form solutions of computing the Jensen-Bregman centroid of a Voronoi cell. Nevertheless, we can bypass this by finely discretizing the domain, and estimating the centroids using the above generalized mean interations. We implemented and computed the centroidal Jensen-Bregman Voronoi diagrams following such a scheme. Figure 7 presents the Jensen-Bregman centroidal Voronoi tesselations obtained, assuming an underlying uniform density.



(a) Burg CVT (init.) (b) Burg CVT (25 iterations)

(c) JS CVT (init.) (d) JS CVT (25 iterations)

Fig. 7. Centroidal Jensen-Bregman Voronoi diagrams for the Burg and Shannon (JS) entropies. CVTs provide a way to sample uniformly space according to the underlying distance.

The following section shows how to extend those results to matrix-based data sets.

5 Matrix-Based Jensen-Bregman Divergences

A recent trend in data processing is to consider matrix-valued data sets, where each datum is not handled as a scalar or vector but rather as a 2D matrix. Such kind of data sets occurs frequently in many science and engineering application areas where they are termed *tensors*: Gaussian covariance matrices [23] in sound processing, elasticity tensors in mechanical engineering [24], polarimetric

synthetic aperture radar [25], diffusion tensor imaging (DTI) [26], kernel-based machine learning [27], etc. Those matrices M are symmetric and positive definite (SPD) $M \succ 0 : \forall x \in \mathbb{R}^d \neq 0, x^T M x > 0$, and can be visualized as ellipsoids: Each matrix M, also called a tensor, is geometrically represented by an ellipsoid $\{x \mid x^T M x = 1\}$. Let us denote by Sym_{++} the open convex cone of symmetric positive definite matrices [28].

We build a matrix-based Jensen-Bregman divergence from a convex generator $F : \mathrm{Sym}_{++} \to \mathbb{R}^+$ as follows:

$$J_F(P,Q) = \frac{F(P)+F(Q)}{2} - F\left(\frac{P+Q}{2}\right) \geq 0, \tag{47}$$

with equality if and only if $P = Q$.

Typical matrix-based convex generators are :

- $F(X) = \mathrm{tr}(X^T X)$: the quadratic matrix entropy,
- $F(X) = -\log\det X$: the matrix Burg entropy, and
- $F(X) = \mathrm{tr}(X \log X - X)$: the von Neumann entropy.

Interestingly, those generators are invariant by a permutation matrix P, ie. $F(PX) = F(P)$. Choosing $F(X) = \mathrm{tr}(X \log X - X)$, we get the *Jensen-von Neumann* divergence, the matrix counterpart of the celebrated Jensen-Shannon divergence. A $d \times d$-dimensional SPD matrix is represented by $D = \frac{d(d+1)}{2}$ matrix entries. Thus 2×2-matrices are encoded by $D = 3$ scalar values.

The matrix-based centroidal Voronoi tesselation requires to compute the SPD centroid (of discretized matrices $M_1, ..., M_n$) using the CCCP iterative optimization technique mentioned in Eq. 45:

$$C_{t+1} = \nabla F^{-1}\left(\sum_{i=1}^{n} \frac{1}{n} \nabla F\left(\frac{M_i + C_t}{2}\right)\right). \tag{48}$$

Table 1 reports the matrix gradients and reciprocal gradients for common matrix-based generators.

We now present a generalization of those Voronoi diagrams and centroidal Voronoi tessellations when skewing the divergences. We shall see that skewing the Jensen-Bregman divergences allows one to generalize Bregman Voronoi diagrams [6].

Table 1. Characteristics of convex matrix-based functional generators

Entropy name	$F(X)$	$\nabla F(X)$	$(\nabla F)^{-1}(X)$
Quadratic	$\frac{1}{2}\mathrm{tr} X X^T$	X	X
log det	$-\log\det X$	$-X^{-1}$	$-X^{-1}$
von Neumann	$\mathrm{tr}(X \log X - X)$	$\log X$	$\exp X$

6 Skew Jensen-Bregman Voronoi Diagrams

Recall that Jensen-Bregman divergences are divergences defined by a Jensen gap built from a convex generator function. Instead of taking the mid-point (for value $\alpha = \frac{1}{2}$), we may consider skewing the divergence by introducing a parameter α as follows:

$$J_F^{(\alpha)} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$$
$$J_F^{(\alpha)}(p,q) = \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q)$$

We consider the open interval $(0,1)$ since otherwise the divergence has no discriminatory power (indeed, for $\alpha \in \{0,1\}, J_F^{(\alpha)}(p,q) = 0, \ \forall p,q$). Although skewed divergences are asymmetric $J_F^{(\alpha)}(p,q) \neq J_F^{(\alpha)}(q,p)$, we can swap arguments by replacing α by $1-\alpha$:

$$\begin{aligned} J_F^{(\alpha)}(p,q) &= \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q) \\ &= J_F^{(1-\alpha)}(q,p) \end{aligned} \tag{49}$$

Figure 8 illustrates the divergence as a Jensen gap induced by the convex generator.

Those skew Burbea-Rao divergences are similarly found using a skew Jensen-Bregman counterpart (the gradient terms $\nabla F(\alpha p + (1-\alpha)q)$ perfectly cancel in the sum of skew Bregman divergences):

$$\alpha B_F(p, \alpha p + (1-\alpha)q) + (1-\alpha) B_F(q, \alpha p + (1-\alpha)q) = J_F^{(\alpha)}(p,q) \tag{50}$$

In the limit cases, $\alpha \rightarrow 0$ or $\alpha \rightarrow 1$, we have $J_F^{(\alpha)}(p,q) \rightarrow 0 \ \forall p,q$. That is, those divergences loose their discriminatory power at extremities. However, we show that those skew Burbea-Rao divergences tend *asymptotically* to Bregman divergences [29]:

$$B_F(p,q) = \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} J_F^{(\alpha)}(p,q) \tag{51}$$

$$B_F(q,p) = \lim_{\alpha \rightarrow 1} \frac{1}{1-\alpha} J_F^{(\alpha)}(p,q) \tag{52}$$

Let us consider the Voronoi diagram of a finite point set $p_1, ..., p_n$ with respect to ${J'}_F^{(\alpha)}$, a *normalized* skew Jensen difference that matches exactly Bregman or reverse Bregman divergences in limit cases:

$${J'}_F^{(\alpha)}(p,q) = \frac{1}{\alpha(1-\alpha)} J_F^{(\alpha)}(p,q)$$

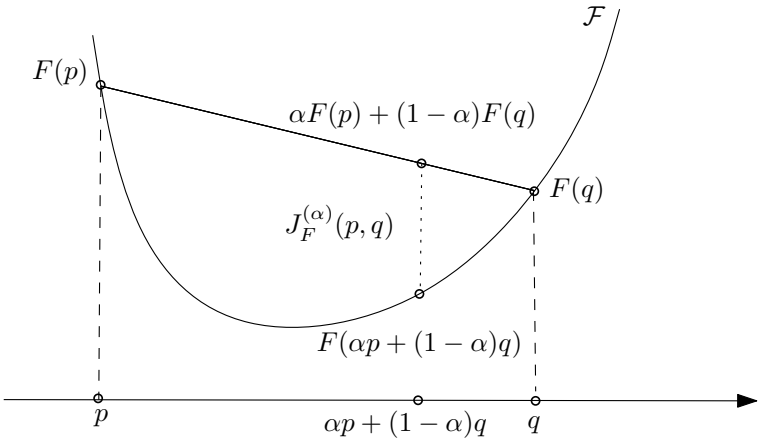


Fig. 8. Skew Jensen-Bregman divergence defined as a Jensen gap induced by a convex generator

The right-sided Voronoi cell associated to site p_i is defined as

$$V_\alpha(p_i) = \{x \mid {J'}_F^{(\alpha)}(p_i, x) \leq {J'}_F^{(\alpha)}(p_j, x) \forall j\} \tag{53}$$

Similarly, the left-sided Voronoi cell

$$V'_\alpha(p_i) = \{x \mid {J'}_F^{(\alpha)}(x, p_i) \leq {J'}_F^{(\alpha)}(x, p_j) \forall j\} \tag{54}$$

is obtained from the right-sided Voronoi cell by changing parameter α to $1-\alpha$:

$$V'_\alpha(p_i) = V_{1-\alpha}(p_i). \tag{55}$$

Thus we restrict ourselves to the right-sided Voronoi cells.

The bisector B of points p_i and p_j is defined by the non-linear equation:

$$B : \alpha(F(p_i) - F(p_j)) + F(\alpha p_j + (1-\alpha)x) - F(\alpha p_i + (1-\alpha)x) = 0 \tag{56}$$

Note that for $\alpha \to 0$ or $\alpha \to 1$, using Gâteaux[9] derivatives [29], we find a bisector either linear in x or in its gradient with respect to the generator (i.e, $\nabla F(x)$). Namely, the (normalized) skew Jensen-Bregman Voronoi diagrams become a regular Bregman Voronoi diagram [6].

Figure 9 depicts several skew left-sided/right-sided Jensen-Bregman Voronoi diagrams. Observe that for $\alpha \in \{0, 1\}$, one of the two sided types of diagrams become affine (meaning bisectors are hyperplanes) since they become a sided Bregman Voronoi diagram [6].

[9] We assume that $\lim_{\lambda \to 0} \frac{F(x+\lambda p) - F(x)}{\lambda}$ exists and is equal to the Gâteaux derivative: $\langle p, \nabla F(x) \rangle$.

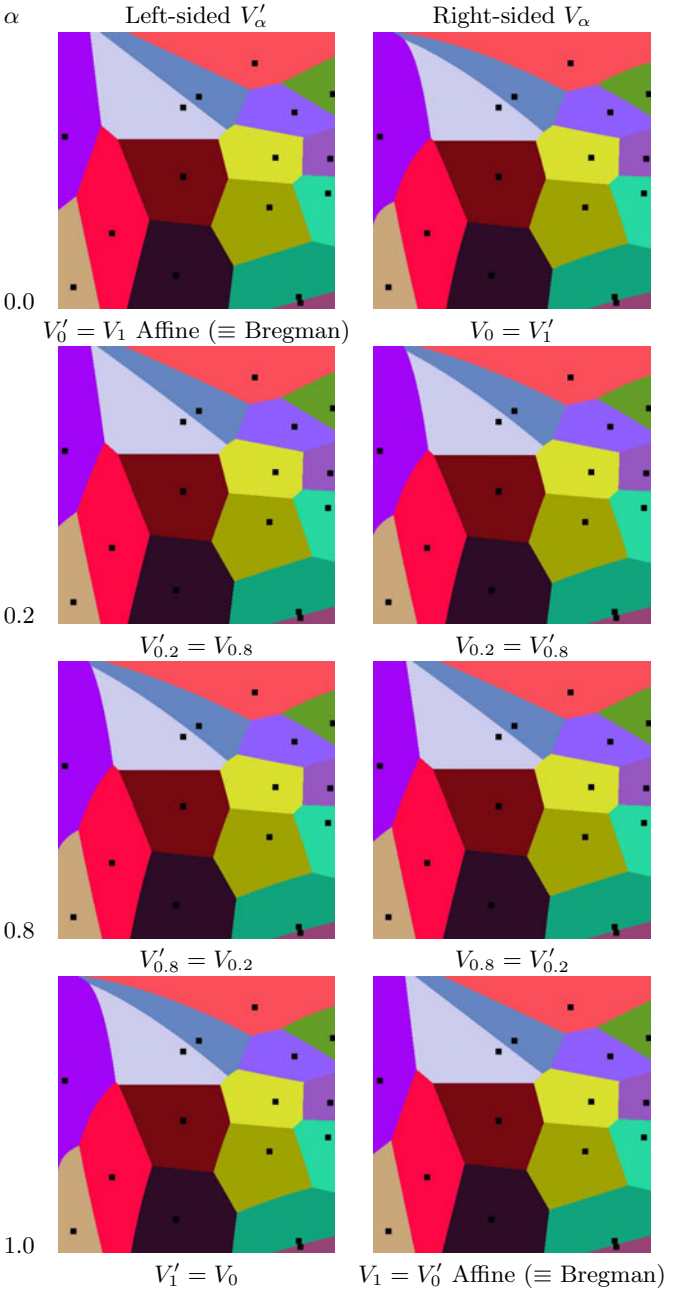


Fig. 9. Skew Jensen-Bregman Voronoi diagrams for various α parameters (Jensen-Shannon divergence). Observe that $V_\alpha = V'_{1-\alpha}$. In the extremal cases $\alpha = 0$ and $\alpha = 1$, the skew Jensen-Bregman diagrams amount to Bregman or reversed Bregman Voronoi diagrams. Note that the left-sided $\alpha = 0$ and right-sided $\alpha = 1$ are affine diagrams since they amount to compute Bregman Voronoi diagrams.

Dealing with skew Jensen-Bregman Voronoi diagrams is interesting for two reasons: (1) it generalizes the Bregman Voronoi diagrams [6] obtained in limit cases, and (2) it allows to consider statistical Voronoi diagrams following [29]. Indeed, consider the parameters of statistical distributions as input, the skew Jensen-Bregman Voronoi diagram amounts to compute equivalently a skew Bhattacharrya Voronoi diagram. Details are left in [29].

7 Concluding Remarks and Discussion

We have introduced a new class of information-theoretic divergences called (skew) Jensen-Bregman divergences that encapsulates both the Jensen-Shannon divergence and the squared Euclidean distance. We showed that those divergences are used when symmetrizing Bregman divergences and computing the Bhattacharyya distance of distributions belonging to the same statistical exponential families (see Appendix). We have studied geometric characteristics of the bisectors. We then introduced the notion of Jensen-Bregman centroid, and described an efficient iterative algorithm to estimate it using the concave-convex optimization framework. This allows one to compute Jensen-Bregman centroidal Voronoi tessellations. We showed how to extend those results to matrix-based Jensen-Bregman divergences, including the Jensen-von Neumann divergence that plays a role in Quantum Information Theory [30] (QIT) dealing with density matrices.

The differential Riemannian geometry induced by such a class of Jensen gaps was studied by Burbea and Rao [14,31] who built quadratic differential metrics on probability spaces using Jensen differences.

The Jensen-Shannon divergence is an instance of a broad class of divergences called the *Csiszár f-divergences*. A f-divergence I_f is a statistical measure of dissimilarity defined by the functional $I_f(p,q) = \int p(x) f(\frac{q(x)}{p(x)}) \mathrm{d}x$. It turns out that the Jensen-Shannon divergence is a f-divergence for generator

$$f(x) = \frac{1}{2}\left((x+1)\log\frac{2}{x+1} + x\log x\right). \tag{57}$$

The class of f-divergences preserves the information monotonicity [32], and their differential geometry was studied by Vos [33]. Note that the squared Euclidean distance *does not* belong to the class of f-divergences although it is a Jensen-Bregman divergence.

To conclude, skew Jensen-Bregman Voronoi diagrams extend naturally Bregman Voronoi diagrams [6], but are not anymore affine diagrams in general. Those diagrams allow one to equivalently compute Voronoi diagrams of statistical distributions with respect to (skew) Bhattacharrya distances. This perspective further opens up the field of computational geometry to statistics and decision theory under uncertainty, where Voronoi bisectors denote decision boundaries [34].

Additional material on Jensen-Bregman divergences including videos are available online at:

`www.informationgeometry.org/JensenBregman/`

Acknowledgments. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. The authors gratefully acknowledge financial support from French funding agency ANR (contract GAIA 07-BLAN-0328-01) and Sony Computer Science Laboratories, Inc.

A Bhattacharyya Distances as Jensen-Bregman Divergences

This appendix proves that (skew) Jensen-Bregman divergences occurs when computing the (skew) Bhattacharrya distance of statistical parametric distributions belonging to the same probability family, called an exponential family. It follows that statistical Voronoi diagrams of members of the same exponential family with respect to the Bhattacharrya distance amount to compute equivalently Jensen-Bregman Voronoi diagrams on the corresponding measure parameters.

A.1 Statistical Exponential Families

Many usual statistical parametric distributions $p(x;\lambda)$ (e.g., Gaussian, Poisson, Bernoulli/multinomial, Gamma/Beta, etc.) share common properties arising from their common canonical decomposition of probability distribution:

$$p(x;\lambda) = p_F(x;\theta) = \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right). \tag{58}$$

Those distributions[10] are said to belong to the exponential families (see [35] for a tutorial). An exponential family is characterized by its *log-normalizer* $F(\theta)$, and a distribution in that family is indexed by its *natural parameter* θ belonging to the *natural space* Θ. The log-normalizer F is strictly convex, infinitely differentiable (C^∞), and can also be expressed using the source coordinate system λ using the bijective map $\tau : \Lambda \to \Theta$ that converts parameters from the source coordinate system to the natural coordinate system:

$$F(\theta) = (F \circ \tau)(\lambda) = F_\lambda(\lambda). \tag{59}$$

The vector $t(x)$ denote the *sufficient statistics*, that is the set of linear independent functions that allows to concentrate without any loss all information about the parameter θ carried in the i.i.d. observation sample $x_1, x_2, ...,$. The

[10] The distributions can either be discrete or continuous. We do not introduce the framework of probability measures here so as to not to burden the paper.

inner product $\langle p, q\rangle$ is a dot product $\langle p, q\rangle = p^T q$ for vectors. Finally, $k(x)$ represents the carrier measure according to the counting measure or the Lebesgue measure. Decompositions for most common exponential family distributions are given in [35]. To give one simple example, consider the family of Poisson distributions with probability mass function:

$$p(x;\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda), \tag{60}$$

for $x \in \mathbb{N}^*$ a non-negative integer. Poisson distributions are univariate exponential families ($x \in \mathbb{N}$) of order 1 (i.e., a single parameter λ). The canonical decomposition yields

- the sufficient statistic $t(x) = x$,
- $\theta = \tau(\lambda) = \log\lambda$, the natural parameter (and $\tau^{-1}(\theta) = \exp\theta$),
- $F(\theta) = \exp\theta$, the log-normalizer,
- and $k(x) = -\log x!$ the carrier measure (with respect to the counting measure).

A.2 Bhattacharyya Distance

For arbitrary probability distributions $p(x)$ and $q(x)$ (parametric or not), we measure the amount of overlap between those distributions using the Bhattacharyya coefficient [36]:

$$B_c(p, q) = \int \sqrt{p(x)q(x)}\mathrm{d}x, \tag{61}$$

where the integral is understood to be multiple if x is multivariate. Clearly, the Bhattacharyya coefficient measures the *affinity* between distributions [37], and falls in the unit range: $0 \leq B_c(p, q) \leq 1$. In fact, we may interpret this coefficient geometrically by considering $\sqrt{p(x)}$ and $\sqrt{q(x)}$ as unit vectors (eventually in infinite-dimensional spaces). The Bhattacharyya coefficient is then the dot product, the cosine of the angle made by the two unit vectors. The Bhattacharyya distance $B : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is derived from its coefficient [36] as

$$B(p, q) = -\ln B_c(p, q). \tag{62}$$

Although the Bhattacharyya distance is symmetric, it is not a metric because it fails the triangle inequality. For distributions belonging to the *same* exponential family, it turns out that the Bhattacharyya distance is always available in closed-form. Namely, the Bhattacharyya distance on probability distributions belonging to the same exponential family is equivalent to a Jensen-Bregman divergence defined for the log-normalizer of the family applied on the natural parameters. This result is not new [38] but seems to have been rediscovered a number of times [39,40]. Let us give a short but insightful proof.

Proof. Consider $p = p_F(x;\theta_p)$ and $q = p_F(x;\theta_q)$ two members of the same exponential families $\mathcal{E}_F$ with natural parameters θ_p and θ_q, respectively. Let us manipulate the Bhattacharyya coefficient $B_c(p,q) = \int \sqrt{p(x)q(x)}\mathrm{d}x$:

$$\begin{aligned}
&= \int \exp\left(\left\langle t(x), \frac{\theta_p+\theta_q}{2}\right\rangle - \frac{F(\theta_p)+F(\theta_q)}{2} + k(x)\right)\mathrm{d}x \\
&= \int \exp\left(\left\langle t(x), \frac{\theta_p+\theta_q}{2}\right\rangle - F\left(\frac{\theta_p+\theta_q}{2}\right) + k(x) + \right. \\
&\quad \left. F\left(\frac{\theta_p+\theta_q}{2}\right) - \frac{F(\theta_p)+F(\theta_q)}{2}\right)\mathrm{d}x \\
&= \exp\left(F\left(\frac{\theta_p+\theta_q}{2}\right) - \frac{F(\theta_p)+F(\theta_q)}{2}\right),
\end{aligned}$$

since $\int p_F(x;\frac{\theta_p+\theta_q}{2})\mathrm{d}x = 1$. We deduce from $B(p,q) = -\ln B_c(p,q)$ that

$$B(p_F(x;\theta_p), p_F(x;\theta_q)) = \frac{F(\theta_p)+F(\theta_q)}{2} - F\left(\frac{\theta_p+\theta_q}{2}\right) \tag{63}$$

It follows that the Bhattacharyya distance for members of the same exponential family is equivalent to a Jensen-Bregman divergence induced by the log-normalizer on the corresponding natural parameters:

$$B(p_F(x;\theta_p), p_F(x;\theta_q)) = J_F(\theta_p;\theta_q), \tag{64}$$

with the Jensen-Bregman divergence defined as the following Jensen difference [41]:

$$J_F(p;q) = \frac{F(p)+F(q)}{2} - F\left(\frac{p+q}{2}\right) \tag{65}$$

For Poisson distributions, we end up with the following Bhattacharyya distance

$$\begin{aligned}
B(p_F(x;\theta_p), p_F(x;\theta_q)) &= J_F(\theta_p,\theta_q) \\
&= J_F(\log\lambda_p, \log\lambda_q), \\
&= \frac{\lambda_p+\lambda_q}{2} - \exp\frac{\log\lambda_p+\log\lambda_q}{2}, \\
&= \frac{\lambda_p+\lambda_q}{2} - \sqrt{\lambda_p\lambda_q} \\
&= \frac{1}{2}(\sqrt{\lambda_p}-\sqrt{\lambda_q})^2
\end{aligned} \tag{66}$$

Exponential families in statistics are mathematically convenient once again. Indeed, the relative entropy of two distributions belonging to the same exponential family, is equal to the Bregman divergence defined for the log-normalizer on swapped natural parameters [6]: $\mathrm{KL}(p_F(x;\theta_p), p_F(x;\theta_q)) = B_F(\theta_q,\theta_p)$.

For skew divergences, we consider the Chernoff divergences

$$C_\alpha(p,q) = -\ln \int p^\alpha(x) q^{1-\alpha}(x) \mathrm{d}x \tag{67}$$

defined for some α (and generalizing the Bhattacharyya divergence for $\alpha = \frac{1}{2}$). The Chernoff α-divergence amounts to compute a weighted asymmetric Jensen-Bregman divergence:

$$\begin{aligned} C_\alpha(p_F(x;\theta_p), p_F(x;\theta_q)) &= J_F^\alpha(\theta_p, \theta_q) && (68)\\ &= \alpha F(\theta_p) + (1-\alpha) F(\theta_q) - F(\alpha\theta_p + (1-\alpha)\theta_q). && (69) \end{aligned}$$

References

1. Nielsen, F., Nock, R.: Jensen-Bregman Voronoi diagrams and centroidal tessellations. In: Proceedings of the 2010 International Symposium on Voronoi Diagrams in Science and Engineering (ISVD), pp. 56–65. IEEE Computer Society, Washington, DC (2010)
2. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial tessellations: Concepts and applications of Voronoi diagrams. In: Probability and Statistics, 2nd edn., 671 pages. Wiley, NYC (2000)
3. de Berg, M., Cheong, O., van Kreveld, M., Overmars, M.: Computational Geometry: Algorithms and Applications, 3rd edn. Springer, Heidelberg (2008)
4. Lee, D.T.: Two-dimensional Voronoi diagrams in the L_p-metric. Journal of the ACM 27, 604–618 (1980)
5. Chew, L.P., Dyrsdale III, R.L.S.: Voronoi diagrams based on convex distance functions. In: Proceedings of the First Annual Symposium on Computational Geometry, SCG 1985, pp. 235–244. ACM, New York (1985)
6. Boissonnat, J.D., Nielsen, F., Nock, R.: Bregman Voronoi diagrams. Discrete and Computational Geometry 44(2), 281–307 (2010)
7. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 37, 145–151 (1991)
8. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience, New York (1991)
9. Du, Q., Faber, V., Gunzburger, M.: Centroidal voronoi tessellations: Applications and algorithms. SIAM Rev. 41, 637–676 (1999)
10. Jeffreys, H.: An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London 186, 453–461 (1946)
11. Reid, M.D., Williamson, R.C.: Generalised Pinsker inequalities. CoRR abs/0906.1244 (2009); published at COLT 2009
12. Chen, P., Chen, Y., Rao, M.: Metrics defined by Bregman divergences: Part I. Commun. Math. Sci. 6, 9915–9926 (2008)
13. Chen, P., Chen, Y., Rao, M.: Metrics defined by Bregman divergences: Part II. Commun. Math. Sci. 6, 927–948 (2008)
14. Burbea, J., Rao, C.R.: On the convexity of some divergence measures based on entropy functions. IEEE Transactions on Information Theory 28, 489–495 (1982)

15. Chazelle, B.: An optimal convex hull algorithm in any fixed dimension. Discrete & Computational Geometry 10, 377–409 (1993)
16. Aurenhammer, F.: Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput. Surv. 23, 345–405 (1991)
17. Sharir, M., Agarwal, P.K.: Davenport-Schinzel Sequences and their Geometric Applications. Cambridge University Press, New York (2010)
18. Icking, C., Ha, L.: A tight bound for the complexity of Voronoi diagrams under polyhedral convex distance functions in 3d. In: STOC 2001: Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing, pp. 316–321. ACM, New York (2001)
19. Ma, L.: Bisectors and Voronoi diagrams for convex distance functions, PhD thesis (2000)
20. Dickerson, M., Eppstein, D., Wortman, K.A.: Dilation, smoothed distance, and minimization diagrams of convex functions, arXiv 0812.0607
21. Balzer, M., Schlömer, T., Deussen, O.: Capacity-constrained point distributions: a variant of Lloyd's method. ACM Trans. Graph. 28 (2009)
22. Yuille, A., Rangarajan, A.: The concave-convex procedure. Neural Computation 15, 915–936 (2003)
23. Arshia Cont, S.D., Assayag, G.: On the information geometry of audio streams with applications to similarity computing. IEEE Transactions on Audio, Speech and Language Processing 19 (2011) (to appear)
24. Cowin, S.C., Yang, G.: Averaging anisotropic elastic constant data. Journal of Elasticity 46, 151–180 (1997), doi:10.1023/A:1007335407097
25. Wang, Y.H., Han, C.Z.: Polsar image segmentation by mean shift clustering in the tensor space. Acta Automatica Sinica 36, 798–806 (2010)
26. Xie, Y., Vemuri, B.C., Ho, J.: Statistical Analysis of Tensor Fields. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6361, pp. 682–689. Springer, Heidelberg (2010)
27. Tsuda, K., Rätsch, G., Warmuth, M.K.: Matrix exponentiated gradient updates for on-line learning and bregman projection. Journal of Machine Learning Research 6, 995–1018 (2005)
28. Bhatia, R., Holbrook, J.: Riemannian geometry and matrix geometric means. Linear Algebra and its Applications 413, 594–618 (2006); Special Issue on the 11th Conference of the International Linear Algebra Society, Coimbra (2004)
29. Nielsen, F., Boltz, S.: The Burbea-Rao and Bhattacharyya centroids. IEEE Transactions on Information Theory (2010)
30. Nielsen, M.A., Chuang, I.L.: Quantum computation and quantum information. Cambridge University Press, New York (2000)
31. Burbea, J., Rao, C.R.: On the convexity of higher order Jensen differences based on entropy functions. IEEE Transactions on Information Theory 28, 961–963 (1982)
32. Csiszár, I.: Information theoretic methods in probability and statistics
33. Vos, P.: Geometry of f-divergence. Annals of the Institute of Statistical Mathematics 43, 515–537 (1991)
34. Hastie, T., Tibshirani, R., Friedman, R.: Elements of Statistical Learning Theory. Springer, Heidelberg (2002)
35. Nielsen, F., Garcia, V.: Statistical exponential families: A digest with flash cards (2009) arXiv.org:0911.4863
36. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of Calcutta Mathematical Society 35, 99–110 (1943)

37. Matusita, K.: Decision rules based on the distance, for problems of fit, two samples, and estimation. Annal of Mathematics and Statistics 26, 631–640 (1955)
38. Huzurbazar, V.S.: Exact forms of some invariants for distributions admitting sufficient statistics. Biometrika 42, 533–573 (1955)
39. Kailath, T.: The divergence and Bhattacharyya distance measures in signal selection. IEEE Transactions on Communications [legacy, pre - 1988] 15, 52–60 (1967)
40. Jebara, T., Kondor, R.: Bhattacharyya and expected likelihood kernels. In: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel, p. 57 (2003)
41. Sahoo, P.K., Wong, A.K.C.: Generalized Jensen difference based on entropy functions. Kybernetika, 241–250 (1988)