Mixed Bregman Clustering with Approximation Guarantees

Richard Nock[1], Panu Luosto[2], and Jyrki Kivinen[2]

[1] Ceregmia — Université Antilles-Guyane, Schoelcher, France
rnock@martinique.univ-ag.fr
[2] Department of Computer Science, University of Helsinki, Finland
{panu.luosto,jyrki.kivinen}@cs.helsinki.fi

Abstract. Two recent breakthroughs have dramatically improved the scope and performance of k-means clustering: squared Euclidean seeding for the initialization step, and Bregman clustering for the iterative step. In this paper, we first unite the two frameworks by generalizing the former improvement to *Bregman seeding* — a biased randomized seeding technique using Bregman divergences — while generalizing its important theoretical approximation guarantees as well. We end up with a complete Bregman hard clustering algorithm integrating the distortion at hand in both the initialization and iterative steps. Our second contribution is to further generalize this algorithm to handle *mixed Bregman distortions*, which smooth out the asymetricity of Bregman divergences. In contrast to some other symmetrization approaches, our approach keeps the algorithm simple and allows us to generalize theoretical guarantees from regular Bregman clustering. Preliminary experiments show that using the proposed seeding with a suitable Bregman divergence can help us discover the underlying structure of the data.

1 Introduction

Intuitively, the goal of clustering is to partition a set of data points into *clusters* so that similar points end up in the same cluster while points in different clusters are dissimilar. (This is sometimes called *hard clustering*, since each data point is assigned to a unique cluster. In this paper we do not consider so-called soft clustering.) One of the most influential contributions to the field has been Lloyd's k-means algorithm [Llo82]. It is beyond our scope to survey the vast literature on the theory and applications of the k-means algorithm. For our purposes, it is sufficient to note three key features of the basic algorithm that can serve as starting points for further development: (i) Each cluster is represented by its *centroid*. (ii) Initial *seeding* chooses random data points as centroids. (iii) Subsequently the algorithm improves the quality of the clustering by locally optimizing its *potential*, defined as the sum of the squared Euclidean distances between each data point and its nearest centroid.

Our starting points are two recent major improvements that address points (ii) and (iii) above. First, Banerjee et al. [BMDG05] have generalized the k-means

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 154–169, 2008.

algorithm to allow, instead of just squared Euclidean distance, any *Bregman divergence* [Bre67] as a distortion measure in computing the potential. Bregman divergences are closely associated with exponential families of distributions and include such popular distortion measures as Kullback-Leibler divergence and Itakura-Saito divergence. As these divergences are in general not symmetrical, they introduce nontrivial technical problems. On the other hand, they give us a lot of freedom in fitting the performance measure of our algorithm to the nature of the data (say, an exponential family of distributions we feel might be appropriate) which should lead to qualitatively better clusterings. Bregman divergences have found many applications in other types of machine learning (see *e.g.* [AW01]) and in other fields such as computational geometry [NBN07], as well.

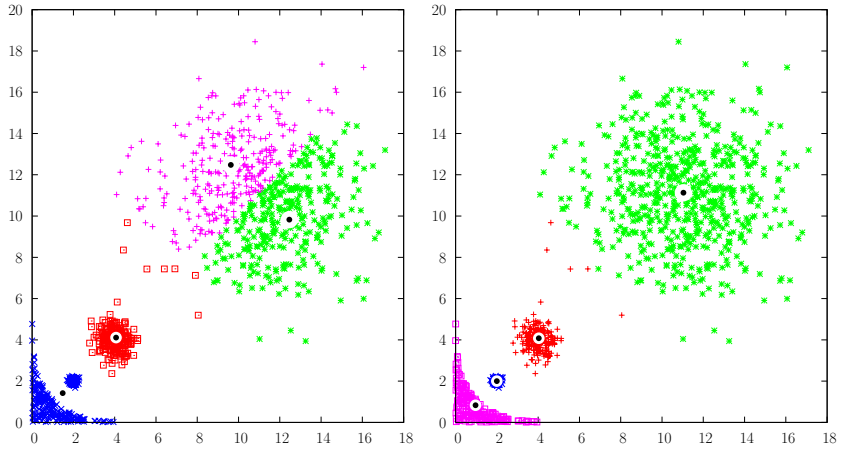

Fig. 1. Clusterings obtained by minimizing Euclidean (left) and Kullback-Leibler (right) potential. The centroids are shown as black dots.

To appreciate the effect of the distortion measure on clustering results, consider the exaggerated toy example in Figure 1. Visually, the data consists of four clusters. The first one is centered around the origin and spreads along the x and y axes. It can be seen as an approximate Itakura-Saito ball [NN05]. The other three clusters come from isotropic Gaussian distributions with different variances and centers at $(2, 2)$, $(4, 4)$ and $(11, 11)$. The cluster around $(11, 11)$ has 600 data points, while the other three clusters have 200 data points each. For $k = 4$, a Euclidean distortion measure favors clusterings that use a single centroid to cover the two small clusters close to the origin and uses two centroids to cover the one big cluster. In contrast, Kullback-Leibler divergence gives a better score to solutions that correspond to the visually distinguishable clusters.

A second recent improvement to k-means clustering is D^2 seeding by Arthur and Vassilvitskii [AV07]. Instead of choosing the k initial cluster centroids uniformly from all the data points, we choose them in sequence so that in choosing

the next initial centroid we give a higher probability to points that are not close to any of the already chosen centroids. Intuitively, this helps to get centroids that cover the data set better; in particular, if the data does consist of several clearly separated clusters, we are more likely to get at least one representative from each cluster. Surprisingly, this simple change in the seeding guarantees that the squared Euclidean potential of the resulting clustering is in expectation within $O(\log k)$ of optimal. For another recent result that obtains approximation guarantees by modifying the seeding of the k-means algorithm, see [ORSS06].

The first contribution in this paper is to combine the previous two advances by replacing squared Euclidean distance in the D^2 seeding by an arbitrary Bregman divergence. The resulting *Bregman seeding* gives a similar approximation guarantee as the original D^2 seeding, except that the approximation factor contains an extra factor $\rho_\psi \geq 1$ that depends on the chosen Bregman divergence and the location of the data in the divergence domain. For Mahalanobis divergence this factor is always 1; for others, such as Itakura-Saito, it can be quite large or quite close to 1 depending on the data. The key technique allowing this generalization is a relaxed form of the triangle inequality that holds for Bregman divergences; this inequality is a sharper form of a recent bound by Crammer et al. [CKW07]. Empirically, for different artificial data sets we have found that choosing the appropriate Bregman divergence can noticeably improve the chances of the seeding including a centroid from each actual cluster. Again, for an exaggerated example consider that data in Figure 1. Experimentally, Kullback-Leibler seeding picks exactly one point from each of the four visible clusters about 15% of the time, while the original D^2 seeding achieves a rate of only 2%. It should be noted, however, that while the proof of the approximation guarantee in [AV07] relies crucially on a successful seeding, in practice it seems that the iteration phase of the algorithm can quite often recover from a bad seeding.

Our second contribution concerns point (i) above, the representation of clusters by a centroid. Since Bregman divergences are asymmetric, it is very significant whether our potential function considers divergence from a data point to the centroid, or from the centroid to a data point. One choice keeps the arithmetic mean of the data points in a cluster as its optimal centroid, the other does not [BMDG05, BGW05]. The strong asymmetricity of Bregman divergences may seem undesirable in some situations, so a natural thought is to symmetrize the divergence by considering the average of the divergences in the two different directions. However, this makes finding the optimal centroid quite a nontrivial optimization problem [Vel02] and makes the statistical interpretation of the centroid less clear. As a solution, we suggest using *two* centroids per cluster, one for each direction of the Bregman divergence. This makes the centroid computations easy and allows a nice statistical interpretation. We call this symmetrized version with two centroids a *mixed Bregman divergence*.

Previously, approximation bounds for Bregman clustering algorithms have been given by [CM08] and [ABS08]. Chadhuri and McGregor [CM08] consider the KL divergence, which is a particularly interesting case as the KL divergence between two members of the same exponential family is a Bregman divergence

between their natural parameters [BMDG05]. Ackermann *et al.* [ABS08] consider a statistically defined class of distortion measures which includes the KL divergence and other Bregman divergences. In both of these cases, the algorithms achieve $(1+\varepsilon)$-approximation for arbitrary $\varepsilon > 0$. This is a much stronger guarantee than the logarithmic factor achieved here using the technique of [AV07]. On the other hand, the $(1+\varepsilon)$-approximation algorithms are fairly complex, whereas our algorithm based on [AV07] is quite easy to implement and runs in time $O(nkd)$.

Section 2 presents definitions; Section 3 presents our seeding and clustering algorithm. Section 4 discusses some results. Section 5 provides experiments, and Section 6 concludes the paper with open problems.

2 Definitions

Divergences. Let $\psi : \mathbb{X} \to \mathbb{R}$ be a strictly convex function defined on a convex set $\mathbb{X} \subseteq \mathbb{R}^d$, with the gradient $\nabla\psi$ defined in the interior of $\mathbb{X}$. (Hereafter, for the sake of simplicity, we do not make the difference between a set and its interior.) We denote by $\psi^\star(\boldsymbol{x}) \doteq \langle \boldsymbol{x}, (\boldsymbol{\nabla}\psi)^{-1}(\boldsymbol{x})\rangle - \psi((\boldsymbol{\nabla}\psi)^{-1}(\boldsymbol{x}))$ its convex conjugate. The *Bregman divergence* $\Delta_\psi(\boldsymbol{x}\|\boldsymbol{y})$ between any two point $\boldsymbol{x}$ and $\boldsymbol{y}$ of $\mathbb{X}$ is [Bre67]:

$$\Delta_\psi(\boldsymbol{x}\|\boldsymbol{y}) \doteq \psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{\nabla}\psi(\boldsymbol{y})\rangle \ .$$

Popular examples of Bregman divergences include Mahalanobis divergence with $D_{\mathrm{M}}(\boldsymbol{x}\|\boldsymbol{y}) \doteq (\boldsymbol{x}-\boldsymbol{y})^\top \mathrm{M}(\boldsymbol{x}-\boldsymbol{y})$ ($\mathbb{X} = \mathbb{R}^d$, M symmetric positive definite), Kullback-Leibler divergence, $D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{y}) \doteq \sum_{i=1}^d (x_i \log(x_i/y_i) - x_i + y_i)$ ($\mathbb{X} = \mathbb{R}^d_{+*}$), Itakura-Saito divergence, $D_{\mathrm{IS}}(\boldsymbol{x}\|\boldsymbol{y}) \doteq \sum_{i=1}^d ((x_i/y_i) - \log(x_i/y_i) - 1)$ ($\mathbb{X} = \mathbb{R}^d_{+*}$), and many others [NBN07, BMDG05]. It is not hard to prove that Mahalanobis divergence is the *only* symmetric Bregman divergence. This general asymmetry, which arises naturally from the links with the exponential families of distributions [BMDG05], is not really convenient for clustering. Thus, we let:

$$\triangle_{\psi,\alpha}(\boldsymbol{x}\|\boldsymbol{y}\|\boldsymbol{z}) \doteq (1-\alpha)\Delta_\psi(\boldsymbol{x}\|\boldsymbol{y}) + \alpha\Delta_\psi(\boldsymbol{y}\|\boldsymbol{z}) \tag{1}$$

denote the *mixed* Bregman divergence of parameters (ψ, α), with $0 \leq \alpha \leq 1$. When $\alpha = 0, 1$, this is just a regular Bregman divergence. The special case $\alpha = 1/2$ and $\boldsymbol{x} = \boldsymbol{z}$ is known as a symmetric Bregman divergence [Vel02].

Clustering. We are given a set $\mathcal{S} \subseteq \mathbb{X}$. For some $\mathcal{A} \subseteq \mathcal{S}$ and $\boldsymbol{y} \in \mathbb{X}$, let

$$\begin{aligned} \psi_\alpha(\mathcal{A}, \boldsymbol{y}) &\doteq \sum_{\boldsymbol{x}\in\mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{y}\|\boldsymbol{x}\|\boldsymbol{y}) \ , \\ \psi^\star_\alpha(\mathcal{A}, \boldsymbol{y}) &\doteq \psi_{1-\alpha}(\mathcal{A}, \boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{x}\|\boldsymbol{y}\|\boldsymbol{x}) \ . \end{aligned} \tag{2}$$

Let $\mathcal{C} \subset \mathbb{X}^2$. The *potential* for Bregman clustering with the centroids of $\mathcal{C}$ is:

$$\psi_\alpha(\mathcal{C}) \doteq \sum_{\boldsymbol{x}\in\mathcal{S}} \min_{(\boldsymbol{c},\boldsymbol{c}^\star)\in\mathcal{C}} \triangle_{\psi,\alpha}(\boldsymbol{c}^\star\|\boldsymbol{x}\|\boldsymbol{c}) \ . \tag{3}$$

When $\alpha = 0$ or $\alpha = 1$, we can pick $\mathcal{C} \subset \mathbb{X}$, and we return to regular Bregman clustering [BMDG05]. The contribution to this potential of some subset $\mathcal{A}$, not necessarily defining a cluster, is noted $\psi_\alpha(\mathcal{A})$, omitting the clustering that shall be implicit and clear from context. An optimal clustering, $\mathcal{C}_{\rm opt}$, can be defined either as its set of centroids, or the partition of $\mathcal{S}$ induced. It achieves:

$$\psi_{{\rm opt},\alpha} \doteq \min_{\mathcal{C} \subset \mathbb{X}^2, |\mathcal{C}|=k} \psi_\alpha(\mathcal{C}) \ . \tag{4}$$

In this clustering, the contribution of some cluster $\mathcal{A}$ is:

$$\psi_{{\rm opt},\alpha}(\mathcal{A}) \doteq \sum_{\boldsymbol{x} \in \mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{c}^\star_{\mathcal{A}} \| \boldsymbol{x} \| \boldsymbol{c}_{\mathcal{A}}) \ ,$$

where $(\boldsymbol{c}_{\mathcal{A}}, \boldsymbol{c}^\star_{\mathcal{A}}) \in \mathcal{C}_{\rm opt}$ is the pair of centroids which minimizes $\psi_\alpha(\mathcal{A})$ over all possible choices of $(\boldsymbol{c}, \boldsymbol{c}^\star)$ in (3). It turns out that these two centroids are always respectively the arithmetic and *Bregman* averages of $\mathcal{A}$:

$$\boldsymbol{c}_{\mathcal{A}} \doteq \frac{1}{|\mathcal{A}|} \sum_{\boldsymbol{x} \in \mathcal{A}} \boldsymbol{x} \ , \tag{5}$$

$$\boldsymbol{c}^\star_{\mathcal{A}} \doteq (\boldsymbol{\nabla}\psi)^{-1} \left(\frac{1}{|\mathcal{A}|} \sum_{\boldsymbol{x} \in \mathcal{A}} \boldsymbol{\nabla}\psi(\boldsymbol{x}) \right) \ . \tag{6}$$

To see that it holds for the arithmetic average, we may write:

$$\forall \boldsymbol{c} \in \mathcal{A}, \sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{x} \| \boldsymbol{c}) - \sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{x} \| \boldsymbol{c}_{\mathcal{A}}) = |\mathcal{A}| \Delta_\psi(\boldsymbol{c}_{\mathcal{A}} \| \boldsymbol{c}) \ . \tag{7}$$

Since the right hand side is not negative and zero only when $\boldsymbol{c} = \boldsymbol{c}_{\mathcal{A}}$, (5) is the best choice for $\boldsymbol{c}$. On the other hand, if we compute (7) on $\psi^\star$ and then use the following well-known dual symmetry relationship which holds for any Bregman divergence,

$$\Delta_\psi(\boldsymbol{x} \| \boldsymbol{y}) = \Delta_{\psi^\star}(\boldsymbol{\nabla}\psi(\boldsymbol{y}) \| \boldsymbol{\nabla}\psi(\boldsymbol{x})) \ ,$$

then we obtain:

$$\forall \boldsymbol{c} \in \mathcal{A}, \sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{c} \| \boldsymbol{x}) - \sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{c}^\star_{\mathcal{A}} \| \boldsymbol{x}) = |\mathcal{A}| \Delta_\psi(\boldsymbol{c} \| \boldsymbol{c}^\star_{\mathcal{A}}) \ , \tag{8}$$

and we conclude that (6) is the best choice for $\boldsymbol{c}^\star$. Since $\boldsymbol{c}^\star_{\mathcal{A}} \neq \boldsymbol{c}_{\mathcal{A}}$ except when $\triangle_{\psi,\alpha}$ is proportional to Mahalanobis divergence, the mixed divergence (1) is only a partial symmetrization of the Bregman divergence with respect to approaches like *e.g.* [Vel02] that enforce $\boldsymbol{c}^\star_{\mathcal{A}} = \boldsymbol{c}_{\mathcal{A}}$. There are at least two good reasons for this symmetrization to remain partial for Bregman divergences. The first is statistical: up to additive and multiplicative factors that would play no role in its optimization, (1) is an exponential family's log-likelihood in which α tempers the probability to fit in the expectation parameter's space versus the natural

Algorithm 1. MBS($\mathcal{S}$, k, α, ψ)

Input: Dataset $\mathcal{S}$, integer $k > 0$, real $\alpha \in [0, 1]$, strictly convex ψ;
Let $\mathcal{C} \leftarrow \{(\boldsymbol{x}, \boldsymbol{x})\}$;
//where $\boldsymbol{x}$ is chosen uniformly at random in $\mathcal{S}$;
for $i = 1, 2, ..., k-1$ **do**
  Pick at random point $\boldsymbol{x} \in \mathcal{S}$ with probability:

$$\pi_{\mathcal{S}}(\boldsymbol{x}) \doteq \frac{\triangle_{\psi,\alpha}(\boldsymbol{c_x} \| \boldsymbol{x} \| \boldsymbol{c_x})}{\sum_{\boldsymbol{y} \in \mathcal{S}} \triangle_{\psi,\alpha}(\boldsymbol{c_y} \| \boldsymbol{y} \| \boldsymbol{c_y})} \ , \tag{9}$$

  //where $(\boldsymbol{c_x}, \boldsymbol{c_x}) \doteq \arg\min_{(\boldsymbol{z},\boldsymbol{z}) \in \mathcal{C}} \triangle_{\psi,\alpha}(\boldsymbol{z} \| \boldsymbol{x} \| \boldsymbol{z})$;
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\boldsymbol{x}, \boldsymbol{x})\}$;
Output: Set of initial centroids $\mathcal{C}$;

parameter's space [BMDG05]. This adds a twist in the likelihood for the uncertainty of the data to model which is, in the context of clustering, desirable even against regular Bregman divergences. However, it does not hold for approaches like [Vel02]. The second reason is algorithmic: mixed divergences incur no complexity counterpart if we except the computation of the inverse gradient for $\boldsymbol{c}^\star_{\mathcal{A}}$; in the complete symmetric approaches, there is no known general expression for the centroid, and it may be time consuming to get approximations even when it is trivial to compute $\boldsymbol{c}^\star_{\mathcal{A}}$ [Vel02]. Finally, we define a dual potential for the optimal clustering, obtained by permuting the parameters of the divergences:

$$\psi^\star_{\mathrm{opt},\alpha}(\mathcal{A}) \doteq \sum_{\boldsymbol{x} \in \mathcal{A}} \triangle_{\psi,1-\alpha}(\boldsymbol{c}_{\mathcal{A}} \| \boldsymbol{x} \| \boldsymbol{c}^\star_{\mathcal{A}}) = \sum_{\boldsymbol{x} \in \mathcal{A}} (\alpha \Delta_\psi(\boldsymbol{c}_{\mathcal{A}} \| \boldsymbol{x}) + (1-\alpha) \Delta_\psi(\boldsymbol{x} \| \boldsymbol{c}^\star_{\mathcal{A}})) \ .$$

3 Mixed Bregman Clustering

3.1 Mixed Bregman Seeding

Algorithm 1 (Mixed Bregman Seeding) shows how we seed the initial cluster centroids. It is generalizes the approach of [AV07] and gives their D^2 seeding as a special case when using the squared Euclidean distance as distortion measure. Since the Bregman divergence between two points can usually be computed in the same $O(d)$ time as the Euclidean distance, our algorithm has the same $O(nkd)$ running time as the original one by [AV07]. The main result of [AV07] is an approximation bound for the squared Euclidean case:

Theorem 1. *[AV07] The average initial potential resulting from D^2 seeding satisfies $\boldsymbol{E}[\psi] \leq 8(2 + \log k)\psi_{\mathrm{opt}}$, where ψ_{opt} is the smallest squared Euclidean potential possible by partitioning $\mathcal{S}$ in k clusters.*

We prove a generalization of Theorem 1 by generalizing each of the lemmas used by [AV07] in their proof.

Lemma 1. *Let $\mathcal{A}$ be an arbitrary cluster of $\mathcal{C}_{\rm opt}$. Then:*

$$\boldsymbol{E}_{\boldsymbol{c}\sim U_{\mathcal{A}}}[\psi_\alpha(\mathcal{A},\boldsymbol{c})] = \psi_{{\rm opt},\alpha}(\mathcal{A}) + \psi^\star_{{\rm opt},\alpha}(\mathcal{A})) \ , \tag{10}$$

$$\boldsymbol{E}_{\boldsymbol{c}\sim U_{\mathcal{A}}}[\psi^\star_\alpha(\mathcal{A},\boldsymbol{c})] = \psi_{{\rm opt},1-\alpha}(\mathcal{A}) + \psi^\star_{{\rm opt},1-\alpha}(\mathcal{A})) \ , \tag{11}$$

where $U_{\mathcal{A}}$ is the uniform distribution over $\mathcal{A}$.

Proof. We use (7) and (8) in (13) below and obtain:

$$\mathbf{E}_{\boldsymbol{c}\sim U_{\mathcal{A}}}[\psi_\alpha(\mathcal{A},\boldsymbol{c})] = \frac{1}{|\mathcal{A}|}\sum_{\boldsymbol{c}\in\mathcal{A}}\sum_{\boldsymbol{x}\in\mathcal{A}}\{\alpha\Delta_\psi(\boldsymbol{x}\|\boldsymbol{c}) + (1-\alpha)\Delta_\psi(\boldsymbol{c}\|\boldsymbol{x})\} \tag{12}$$

$$= \frac{1}{|\mathcal{A}|}\sum_{\boldsymbol{c}\in\mathcal{A}}\left\{\alpha\left(\sum_{\boldsymbol{x}\in\mathcal{A}}\Delta_\psi(\boldsymbol{x}\|\boldsymbol{c}_{\mathcal{A}}) + |\mathcal{A}|\Delta_\psi(\boldsymbol{c}_{\mathcal{A}}\|\boldsymbol{c})\right)\right.$$
$$\left. + (1-\alpha)\left(\sum_{\boldsymbol{x}\in\mathcal{A}}\Delta_\psi(\boldsymbol{c}^\star_{\mathcal{A}}\|\boldsymbol{x}) + |\mathcal{A}|\Delta_\psi(\boldsymbol{c}\|\boldsymbol{c}^\star_{\mathcal{A}})\right)\right\} \tag{13}$$

$$= \alpha\sum_{\boldsymbol{x}\in\mathcal{A}}\Delta_\psi(\boldsymbol{x}\|\boldsymbol{c}_{\mathcal{A}}) + \alpha\sum_{\boldsymbol{x}\in\mathcal{A}}\Delta_\psi(\boldsymbol{c}_{\mathcal{A}}\|\boldsymbol{x})$$
$$+(1-\alpha)\sum_{\boldsymbol{x}\in\mathcal{A}}\Delta_\psi(\boldsymbol{c}^\star_{\mathcal{A}}\|\boldsymbol{x}) + (1-\alpha)\sum_{\boldsymbol{x}\in\mathcal{A}}\Delta_\psi(\boldsymbol{x}\|\boldsymbol{c}^\star_{\mathcal{A}})$$

$$= \alpha\psi_{{\rm opt},1}(\mathcal{A}) + (1-\alpha)\psi_{{\rm opt},0}(\mathcal{A}) + \alpha\psi^\star_{{\rm opt},1}(\mathcal{A}) \tag{14}$$
$$+(1-\alpha)\psi^\star_{{\rm opt},0}(\mathcal{A})$$
$$= \psi_{{\rm opt},\alpha}(\mathcal{A}) + \psi^\star_{{\rm opt},\alpha}(\mathcal{A}) \ .$$

This gives (10). Applying (2) to (10) gives (11). □

Analyzing the biased distribution case requires a triangle inequality for Bregman divergences, stated below. For any positive semidefinite matrix M, $\mathrm{M}^{1/2}$ denotes the positive semidefinite matrix such that $\mathrm{M}^{1/2}\mathrm{M}^{1/2} = \mathrm{M}$.

Lemma 2. *For any three points $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}$ of ${\rm co}(\mathcal{S})$, the convex closure of $\mathcal{S}$,*

$$\Delta_\psi(\boldsymbol{x},\boldsymbol{z}) \leq 2\rho^2_\psi(\Delta_\psi(\boldsymbol{x},\boldsymbol{y}) + \Delta_\psi(\boldsymbol{y},\boldsymbol{z})) \ , \tag{15}$$

where ρ_ψ is defined as:

$$\rho_\psi \doteq \sup_{\boldsymbol{s},\boldsymbol{t},\boldsymbol{u},\boldsymbol{v}\in{\rm co}(\mathcal{S})} \frac{\|\mathrm{H}^{1/2}_{\boldsymbol{s}}(\boldsymbol{u}-\boldsymbol{v})\|_2}{\|\mathrm{H}^{1/2}_{\boldsymbol{t}}(\boldsymbol{u}-\boldsymbol{v})\|_2} \ , \tag{16}$$

where $\mathrm{H}_{\boldsymbol{s}}$ denotes the Hessian of ψ in $\boldsymbol{s}$.

Proof. The key to the proof is the Bregman triangle equality:

$$\Delta_\psi(\boldsymbol{x}\|\boldsymbol{z}) = \Delta_\psi(\boldsymbol{x}\|\boldsymbol{y}) + \Delta_\psi(\boldsymbol{y}\|\boldsymbol{z}) + (\boldsymbol{\nabla}\psi(\boldsymbol{z}) - \boldsymbol{\nabla}\psi(\boldsymbol{y}))^\top(\boldsymbol{y}-\boldsymbol{x}) \ . \tag{17}$$

A Taylor-Lagrange expansion on Bregman divergence Δ_ψ yields:

$$\Delta_\psi(\boldsymbol{a}\|\boldsymbol{b}) = \frac{1}{2}(\boldsymbol{a}-\boldsymbol{b})^\top\mathrm{H}_{\boldsymbol{ab}}(\boldsymbol{a}-\boldsymbol{b}) = \frac{1}{2}\|\mathrm{H}^{1/2}_{\boldsymbol{ab}}(\boldsymbol{a}-\boldsymbol{b})\|^2_2 \ , \tag{18}$$

for some value $\mathrm{H}_{\boldsymbol{ab}}$ of the Hessian of ψ in the segment $\boldsymbol{ab} \subseteq \mathrm{co}(\mathcal{S})$. Another expansion on the gradient part of (17) yields:

$$\boldsymbol{\nabla}\psi(\boldsymbol{z}) - \boldsymbol{\nabla}\psi(\boldsymbol{y}) = \mathrm{H}_{\boldsymbol{zy}}(\boldsymbol{z}-\boldsymbol{y}) \ . \tag{19}$$

Putting this altogether, (17) becomes:

$$\begin{aligned}
\Delta_\psi(\boldsymbol{x}\|\boldsymbol{z}) &\overset{(19)}{=} \Delta_\psi(\boldsymbol{x}\|\boldsymbol{y}) + \Delta_\psi(\boldsymbol{y}\|\boldsymbol{z}) + (\mathrm{H}^{1/2}_{\boldsymbol{zy}}(\boldsymbol{z}-\boldsymbol{y}))^\top(\mathrm{H}^{1/2}_{\boldsymbol{zy}}(\boldsymbol{y}-\boldsymbol{x})) \\
&\leq \Delta_\psi(\boldsymbol{x}\|\boldsymbol{y}) + \Delta_\psi(\boldsymbol{y}\|\boldsymbol{z}) + \|\mathrm{H}^{1/2}_{\boldsymbol{zy}}(\boldsymbol{z}-\boldsymbol{y})\|_2\|\mathrm{H}^{1/2}_{\boldsymbol{zy}}(\boldsymbol{y}-\boldsymbol{x})\|_2 \qquad (20) \\
&\leq \Delta_\psi(\boldsymbol{x}\|\boldsymbol{y}) + \Delta_\psi(\boldsymbol{y}\|\boldsymbol{z}) + \rho_\psi^2\left(\|\mathrm{H}^{1/2}_{\boldsymbol{yz}}(\boldsymbol{z}-\boldsymbol{y})\|_2\|\mathrm{H}^{1/2}_{\boldsymbol{xy}}(\boldsymbol{y}-\boldsymbol{x})\|_2\right) \\
&\overset{(18)}{=} \Delta_\psi(\boldsymbol{x}\|\boldsymbol{y}) + \Delta_\psi(\boldsymbol{y}\|\boldsymbol{z}) + 2\rho_\psi^2\sqrt{\Delta_\psi(\boldsymbol{x}\|\boldsymbol{y})\Delta_\psi(\boldsymbol{y}\|\boldsymbol{z})}
\end{aligned}$$

where (20) makes use of Cauchy-Schwartz inequality. Since $\rho_\psi \geq 1$, the right-hand side of the last inequality is of the form $a+b+2\rho_\psi^2\sqrt{ab} \leq \rho_\psi^2(a+b+2\sqrt{ab}) \leq \rho_\psi^2(2a+2b) = 2\rho_\psi^2(a+b)$. Since $a = \Delta_\psi(\boldsymbol{x}\|\boldsymbol{y})$ and $b = \Delta_\psi(\boldsymbol{y}\|\boldsymbol{z})$, we obtain the statement of the Lemma. □

Lemma 2 is a sharper version of the bound used by [CKW07]. The improvement is basically that we use the same vector $\boldsymbol{u}-\boldsymbol{v}$ in the numerator and denominator in (16), so we are not automatically hurt by anisotropy in the divergence. In particular, we have $\rho_\psi = 1$ for any Mahalanobis distance.

The following lemma generalizes [AV07, Lemma 3.2]. We use Lemmas 1 and 2 instead of special properties of the squared Euclidean distance. Otherwise the proof is essentially the same.

Lemma 3. *Let $\mathcal{A}$ be an arbitrary cluster of $\mathcal{C}_{\mathrm{opt}}$, and $\mathcal{C}$ an arbitrary clustering. If we add a random pair $(\boldsymbol{y},\boldsymbol{y})$ from $\mathcal{A}^2$ to $\mathcal{C}$ in Algorithm 1, then*

$$\boldsymbol{E}_{\boldsymbol{y}\sim\boldsymbol{\pi}_{\mathcal{S}}}[\psi_\alpha(\mathcal{A},\boldsymbol{y})|\boldsymbol{y}\in\mathcal{A}] = \boldsymbol{E}_{\boldsymbol{y}\sim\boldsymbol{\pi}_{\mathcal{A}}}[\psi_\alpha(\mathcal{A},\boldsymbol{y})] \leq 4\rho_\psi^2(\psi_{\mathrm{opt},\alpha}(\mathcal{A}) + \psi^\star_{\mathrm{opt},\alpha}(\mathcal{A})) \ .$$

Proof. The equality comes from the fact that the expectation is constrained to the choice of $\boldsymbol{y}$ in $\mathcal{A}$. The contribution of $\mathcal{A}$ to the potential is thus:

$$\begin{aligned}
&\mathbf{E}_{\boldsymbol{y}\sim\boldsymbol{\pi}_{\mathcal{A}}}[\psi_\alpha(\mathcal{A},\boldsymbol{y})] \\
&= \sum_{\boldsymbol{y}\in\mathcal{A}}\left\{\frac{\triangle_{\psi,\alpha}(\boldsymbol{c_y}\|\boldsymbol{y}\|\boldsymbol{c_y})}{\sum_{\boldsymbol{x}\in\mathcal{A}}\triangle_{\psi,\alpha}(\boldsymbol{c_x}\|\boldsymbol{x}\|\boldsymbol{c_x})}\sum_{\boldsymbol{x}\in\mathcal{A}}\min\left\{\triangle_{\psi,\alpha}(\boldsymbol{c_x}\|\boldsymbol{x}\|\boldsymbol{c_x}),\triangle_{\psi,\alpha}(\boldsymbol{y}\|\boldsymbol{x}\|\boldsymbol{y})\right\}\right\} \qquad (21)
\end{aligned}$$

We also have:

$$\begin{aligned}
&\triangle_{\psi,\alpha}(\boldsymbol{c_y}\|\boldsymbol{y}\|\boldsymbol{c_y}) \\
&= \alpha\Delta_\psi(\boldsymbol{y}\|\boldsymbol{c_y}) + (1-\alpha)\Delta_\psi(\boldsymbol{c_y}\|\boldsymbol{y}) \\
&\leq \alpha\Delta_\psi(\boldsymbol{y}\|\boldsymbol{c_x}) + (1-\alpha)\Delta_\psi(\boldsymbol{c_x}\|\boldsymbol{y}) \\
&\leq 2\rho_\psi^2(\alpha\Delta_\psi(\boldsymbol{y}\|\boldsymbol{x}) + \alpha\Delta_\psi(\boldsymbol{x}\|\boldsymbol{c_x}) + (1-\alpha)\Delta_\psi(\boldsymbol{c_x}\|\boldsymbol{x}) + (1-\alpha)\Delta_\psi(\boldsymbol{x}\|\boldsymbol{y})) \\
&= 2\rho_\psi^2(\triangle_{\psi,\alpha}(\boldsymbol{c_x}\|\boldsymbol{x}\|\boldsymbol{c_x}) + \triangle_{\psi,\alpha}(\boldsymbol{x}\|\boldsymbol{y}\|\boldsymbol{x})) \ ,
\end{aligned}$$

where we have used Lemma 2 on the last inequality. Summing over $\boldsymbol{x} \in \mathcal{A}$ yields:

$$\triangle_{\psi,\alpha}(\boldsymbol{c_y}\|\boldsymbol{y}\|\boldsymbol{c_y}) \leq 2\rho_\psi^2 \left(\frac{1}{|\mathcal{A}|} \sum_{\boldsymbol{x}\in\mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{c_x}\|\boldsymbol{x}\|\boldsymbol{c_x}) + \frac{1}{|\mathcal{A}|} \sum_{\boldsymbol{x}\in\mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{x}\|\boldsymbol{y}\|\boldsymbol{x}) \right) ;$$

plugging this into (21) and replacing the min by its left or right member in the two sums yields:

$$\begin{aligned} \mathbf{E}_{\boldsymbol{y}\sim\boldsymbol{\pi}_\mathcal{A}}[\psi_\alpha(\mathcal{A},\boldsymbol{y})] &\leq 4\rho_\psi^2 \frac{1}{|\mathcal{A}|} \sum_{\boldsymbol{y}\in\mathcal{A}} \sum_{\boldsymbol{x}\in\mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{x}\|\boldsymbol{y}\|\boldsymbol{x}) \\ &= 4\rho_\psi^2 (\psi_{\mathrm{opt},\alpha}(\mathcal{A}) + \psi^\star_{\mathrm{opt},\alpha}(\mathcal{A})) \ , \end{aligned}$$

where we have used (12). □

For any subset of clusters $\mathcal{A}$ of some optimal clustering $\mathcal{C}_{\mathrm{opt}}$, let $\tilde{\psi}_{\mathrm{opt},\alpha}(\mathcal{A}) \doteq (1/2)(\psi_{\mathrm{opt},\alpha}(\mathcal{A}) + \psi^\star_{\mathrm{opt},\alpha}(\mathcal{A}))$. We remark that:

$$\mathbf{E}_{\boldsymbol{y}\sim\pi_\mathcal{A}}[\psi_\alpha(\mathcal{A},\boldsymbol{y})] \leq 8\rho_\psi^2 \tilde{\psi}_{\mathrm{opt},\alpha}(\mathcal{A}) \ , \tag{22}$$

$$\forall \mathcal{A},\mathcal{B} : \mathcal{A}\cap\mathcal{B} = \emptyset, \tilde{\psi}_{\mathrm{opt},\alpha}(\mathcal{A}\cup\mathcal{B}) = \tilde{\psi}_{\mathrm{opt},\alpha}(\mathcal{A}) + \tilde{\psi}_{\mathrm{opt},\alpha}(\mathcal{B}) \ . \tag{23}$$

Lemma 4. *Let $\mathcal{C}$ be an arbitrary clustering. Choose $u > 0$ clusters from $\mathcal{C}_{\mathrm{opt}}$ that are still not covered by $\mathcal{C}$, and let $\mathcal{S}_u$ denote the set of points in these clusters. Also, let $\mathcal{S}_c \doteq \mathcal{S} - \mathcal{S}_u$. Now suppose that we add $t \leq u$ random pairs of centroids, chosen according to $\pi_\mathcal{S}$ as in Algorithm 1. Let $\mathcal{C}'$ denote the resulting clustering. Define $H_t \doteq 1 + (1/2) + ... + (1/t)$. Then*

$$\boldsymbol{E}_{\boldsymbol{c}\sim\pi_\mathcal{S}}[\psi_\alpha(\mathcal{C}')] \leq (1+H_t)\left(\psi_\alpha(\mathcal{S}_c) + 8\rho_\psi^2 \tilde{\psi}_{\mathrm{opt},\alpha}(\mathcal{S}_u)\right) + \left(\frac{u-t}{u}\right)\psi_\alpha(\mathcal{S}_u) \ .$$

Again, the proof is obtained from the proof of [AV07, Lemma 3.3] by just applying (22) and (23) to handle $\tilde{\psi}$. We omit the details. As in [AV07] we now obtain the main approximation bound as a special case of Lemma 4.

Theorem 2. *The average initial potential obtained by Mixed Bregman Seeding (Algorithm 1) satisfies $\boldsymbol{E}[\psi_\alpha] \leq 8\rho_\psi^2(2+\log k)\tilde{\psi}_{\mathrm{opt},\alpha}$, where $\tilde{\psi}_{\mathrm{opt},\alpha}$ is the minimal mixed Bregman divergence possible by partitioning $\mathcal{S}$ into k clusters as defined in (4).*

When the Hessian of ψ satisfies $\mathrm{H}_\cdot = \sigma\mathrm{I}$ for $\sigma > 0$, we return to regular k-means and the bound of Theorem 1 [AV07]. Interestingly, the bound remains the same for general Mahalanobis divergence ($\rho_\psi = 1$, $\psi_{\mathrm{opt},\alpha}(\mathcal{S}) = \psi^\star_{\mathrm{opt},\alpha}(\mathcal{S}) = \tilde{\psi}_{\mathrm{opt}}(\mathcal{S})$).

3.2 Integrating Mixed Bregman Seeding into Clustering

Bregman seeding in the special case $\alpha = 1$ can be integrated with Bregman clustering [BMDG05] to provide a complete clustering algorithm in which the

Algorithm 2. MBC($\mathcal{S}$, k, α, ψ)

Input: Dataset $\mathcal{S}$, integer $k > 0$, real $\alpha \in [0, 1]$, strictly convex ψ;
Let $\mathcal{C} = \{(\boldsymbol{c}_{\mathcal{A}_i}, \boldsymbol{c}^\star_{\mathcal{A}_i})\}_{i=1}^k \leftarrow \mathrm{MBS}(\mathcal{S}, k, \alpha, \psi)$;
repeat
  //Assignment
  for $i = 1, 2, ..., k$ **do**
    $\mathcal{A}_i \leftarrow \{\boldsymbol{s} \in \mathcal{S} : i = \arg\min_j \triangle_{\psi,\alpha}(\boldsymbol{c}^\star_{\mathcal{A}_j} \| \boldsymbol{s} \| \boldsymbol{c}_{\mathcal{A}_j})\}$;
  //Re-estimation
  for $i = 1, 2, ..., k$ **do**
    $\boldsymbol{c}_{\mathcal{A}_i} \leftarrow \frac{1}{|\mathcal{A}_i|} \sum_{\boldsymbol{s} \in \mathcal{A}_i} \boldsymbol{s}$;
    $\boldsymbol{c}^\star_{\mathcal{A}_i} \leftarrow (\boldsymbol{\nabla}\psi)^{-1} \left(\frac{1}{|\mathcal{A}_i|} \sum_{\boldsymbol{s} \in \mathcal{A}_i} \boldsymbol{\nabla}\psi(\boldsymbol{s}) \right)$;
until *convergence* ;
Output: Partition of $\mathcal{S}$ in k clusters following $\mathcal{C}$;

divergence at hand is integrated in all steps of clustering. What remains to do is take this algorithm as a whole and lift it further to handle mixed Bregman divergences, that is, generalize the Bregman clustering of [BMDG05] to hold for any $0 \leq \alpha \leq 1$. This is presented in Algorithm 2 (Mixed Bregman Clustering). This algorithm is conceptually as simple as Bregman clustering [BMDG05], and departs from the complexity of approaches that would be inspired by fully symmetrized Bregman divergences [Vel02]. However, for this algorithm to be a suitable generalization of Bregman clustering, we have to ensure that it monotonously achieves a local minimum of the mixed potential in finite time. This is done in the following Lemma, whose proof, omitted to save space, follows similar steps as in [BMDG05] while making use of (7) and (8).

Lemma 5. *Algorithm 2 monotonically decreases the function in (3). Furthermore, it terminates in a finite number of steps at a locally optimal partition.*

4 Discussion

One question arises on such a scheme, namely how the choice of the main free parameter, the generator ψ, impacts on the final output. This question is less relevant to the clustering phase, where the optimization is local and all that may be required is explicitly given in Lemma 5, independently of ψ. It is more relevant to the seeding phase, and all the more interesting as the upper bound in Theorem 2 exhibits two additional penalties that depend on ψ: one relies on the way we measure the potential and seed centroids ($\tilde{\psi}$), the other relies on convexity (ρ_ψ). The analysis of D^2 seeding by [AV07] is tight on average, as they show that for some clusterings the upper bound of 1 is within a constant factor of the actual performance of the algorithm.

Beyond D^2 seeding, it is not hard to show that the analysis of [AV07] is in fact tight for Mahalanobis divergence. To see this, we only have to make a variable change, and set $\tilde{\boldsymbol{x}} \doteq \mathrm{M}^{-1/2}\boldsymbol{x}$ for any point $\boldsymbol{x}$ in the lower bound proof of [AV07].

Mahalanobis divergence on the new points equals the k-means potential on the initial points, the optimal centroids do not change, and the proof remains as is. For arbitrary divergences, the upper bound of Theorem 2 gets unfastened in stronger convex regimes of the generator, that is when ρ_ψ increases. Some non-metric analysis of seeding that would avoid the use of a triangle inequality might keep it tighter, as stronger convex regimes do not necessarily penalize that much seeding. Sometimes, artificial improvements are even possible. The following Lemma, whose proofsketch is available in an appendix at the end of the paper, gives a lower bound for the uniform approximation that seeding achieves in some cluster.

Lemma 6. *Let $\mathcal{A}$ be an arbitrary cluster. Then:*

$$\boldsymbol{E}_{\boldsymbol{c}\sim U_{\mathcal{A}}}[\psi_\alpha(\mathcal{A},\boldsymbol{c})] \geq \frac{2\rho_\psi^2}{2\rho_\psi^2-1}\psi_\alpha(\mathcal{A}) \ . \tag{24}$$

(24) matches ratio 2 that follows from Lemma 1 for Mahalanobis divergence. The average participation of the seeds in (24) hides large discrepancies, as there do exist seeds whose clustering potential come arbitrarily close to the lower bound (24) as ρ_ψ increases. In other words, since this lower bound is decreasing with ρ_ψ, increasing ρ_ψ may make seeding artificially more efficient if we manage to catch these seeds, a fact that Theorem 2 cannot show. A toy example shows that we can indeed catch such seeds with high probability: we consider $k = 2$ clusters on $n > 2$ points with $\alpha = 1$. The first cluster contains two points p and q as in Figure 2, with p located at abscissa 0 and q at abscissa δ (for the sake of simplicity, ψ is assumed defined on $[0, +\infty)$). Add $n-2$ points $x_1, x_2, ..., x_{n-2}$, all at abscissa $\Delta > \delta$, and pick Δ sufficiently large to ensure that these $n-2$ points define a single cluster, while p and q are grouped altogether in cluster $\mathcal{A}$. It follows that $\psi_{\text{opt},1} = 2\text{BR}_\psi(\{0,\delta\}) = \psi_{\text{opt},1}(\mathcal{A})$. The probability to seed one of the x_i in the two centers is at least $(n-2)/n + (2/n)\cdot(n-2)/(n-1) > 1 - 4/n^2$, which makes that the expected potential is driven by the event that we seed exactly one of the x_i. The associated potential is then either $\Delta_\psi(\delta\|0)$ (we seed p with x_i) or $\Delta_\psi(0\|\delta)$ (we seed q with x_i). Take $\psi(x) = (x+1)^K$ for $K \notin [0,1]$. Then the ratio between these seeding potentials and $\psi_{\text{opt},1}$ respectively satisfy $\rho_p \leq 2^{K-1}/(2^{K-1}-1)$ and $\rho_q = \theta(K)$, while $\rho_\psi^2 = (1+\Delta)^K$. When $K \to +\infty$, we have (i) $\rho_p \to 1$, and so seeding p rapidly approaches the lower bound (24); (ii) $\rho_q \to +\infty$, and so seeding q drives $\mathbf{E}[\psi_1]$; (iii) ratio ρ_ψ^2 is extremely large compared to ρ_q.

5 Experiments

Empirical tests were made with synthetic point sets which had a distinctive structure with high probability. The number of clusters was always 20 and every cluster had 100 points in $\mathbb{R}^{50}_{+*}$. Each of the 20 distributions for the clusters was generated in two phases. In the first phase, for each coordinate lots were drawn independently with a fixed probability p whether the coordinate value is a

Table 1. Percentage of seeding runs in which one center was picked from every original cluster. Numbers in the labels KL-0, KL-0.25 etc. refer to different values of α.

p	unif.	D^2	KL-0	KL-0.25	KL-0.5	KL-0.75	KL-1	IS-0	IS-0.25	IS-0.5	IS-0.75	IS 1
0.1	0	9.70	56.8	75.5	77.1	76.1	57.2	34.4	94.3	95.4	**96.0**	48.4
0.5	0	24.0	77.8	83.1	81.8	80.7	79.3	94.9	95.9	**96.5**	95.8	94.4
0.9	0	7.10	34.6	38.8	42.2	39.4	29.5	28.1	72.6	**75.8**	68.6	20.5
1.0	0	4.10	7.20	**10.0**	7.90	9.30	5.90	0	0	0	0	0.100

Table 2. Percentage of original clusters from which no point was picked in the seeding phase. Average over 1000 runs.

p	unif.	D^2	KL-0	KL-0.25	KL-0.5	KL-0.75	KL-1	IS-0	IS-0.25	IS-0.5	IS-0.75	IS-1
0.1	35.8	7.60	2.48	1.32	1.19	1.24	2.32	5.50	0.285	0.230	**0.200**	3.39
0.5	35.5	5.47	1.18	0.880	0.945	0.975	1.09	0.260	0.205	**0.180**	0.210	0.285
0.9	35.8	8.54	4.15	3.75	3.45	3.74	4.68	4.42	1.46	**1.31**	1.67	6.45
1.0	35.9	9.81	8.27	**7.86**	8.27	8.27	9.00	27.8	23.1	23.4	25.1	21.7

Table 3. Bregman clustering potentials with Kullback-Leibler divergence ($\alpha = 0.25$)

p	unif.	D^2	KL-0	KL-0.25	KL-0.5	KL-0.75	KL-1	IS-0	IS-0.25	IS-0.5	IS-0.75	IS-1
0.1	31.2	4.59	2.64	1.59	1.46	1.46	1.85	6.45	1.10	1.09	**1.06**	2.33
0.5	19.5	4.55	1.74	1.59	1.50	1.66	1.74	1.14	1.14	**1.08**	1.10	1.15
0.9	7.29	2.92	1.97	1.82	1.73	1.92	2.05	1.90	1.34	**1.29**	1.39	2.57
1.0	4.13	2.21	**1.97**	1.99	2.01	2.05	2.11	3.93	3.64	3.69	3.89	3.54

Poisson distributed random variable or the constant 0. Then in the second phase the expectations of the Poisson random variables were chosen independently and uniformly from range $]0, 100[$. 100 points were generated from each distribution and after that the value $\epsilon = 10^{-6}$ was added to every coordinate of every point in order to move the points to the domain of Kullback-Leibler and Itakura-Saito divergences. Because not only the seeding methods but also the datasets were random, 10 datasets were generated for each value of p. The seeding and clustering test were repeated 100 times for each dataset.

When p was less than 1, each cluster was characterized with high probability by the position of coordinates whose value was not ϵ. That made the mixed Itakura-Saito divergences between two points belonging to different clusters very high, and picking one point from every original cluster was strikingly easy using those distortion measures (Tables 1 and 2). However, when all the coordinates were Poisson distributed, the task of finding a center candidate from every cluster was far more difficult. In that case the Kullback-Leibler seeding performed best.

In the clustering tests uniform, D^2, Kullback-Leibler and Itakura-Saito seeding ($\alpha \in \{0, 0.25, \ldots, 1\}$) were used with KL-divergences (same set of values for α as in the seeding) in the iterative phase and by evaluation of the clustering potential. Table 3 illustrates the situation when value $\alpha = 0.25$ is used in the iterative phase. All the values in the table were normalized using the clustering

potentials which were achieved by refining the known centers of the distributions with Bregman clustering. When p was 0.1 uniform seeding brought an over twenty-eight times and D^2 seeding over four times larger average potential than the mixed versions of Itakura-Saito seeding.

In general, there was a clear correlation between the quality of the seeding and the final clustering potential, even if the relative differences in the final potentials tended to diminish gradually when p increased. That means the mixed versions of Bregman seeding algorithms led to low clustering potentials also when a regular Bregman divergence was used in the clustering phase. Additional tests were run with Poisson random variables replaced by Binomial(n, r) distributed variables, so that $n = 100$ and r was taken uniformly at random from range $]0, 1[$. The results were quite similar to those shown here.

6 Conclusions and Open Problems

We have seen that the D^2 seeding of [AV07] can be generalized for Bregman clustering while maintaining some form of approximation guarantee. Our other main contribution was symmetrization of Bregman clustering by using pairs of centroids. Experiments suggest that the resulting new algorithm can significantly improve the quality of both the seeding and the final clustering. The experiments are somewhat preliminary, though, and should be extended to cover more realistic data sets. We also need a better understanding of how much the seeding affects the end result in practice.

On theoretical side, it is not clear if the factor ρ_ψ really is necessary in the bounds. Conceivably, some proof technique not relying on the triangle inequality could give a sharper bound. Alternatively, one could perhaps prove a lower bound that shows the ρ_ψ factor necessary. It would also be interesting to consider other divergences. One possibility would be the p-norm divergences [Gen03] which in some other learning context give results similar to Kullback-Leibler divergence but do not have similar extreme behavior at the boundary of the domain.

Acknowledgments

R. Nock gratefully acknowledges support from the University of Helsinki for a stay during which this work was done. R. Nock was supported by ANR "Blanc" project ANR-07-BLAN-0328-01 "Computational Information Geometry and Applications." P. Luosto and J. Kivinen were supported by Academy of Finland grants 118653 (ALGODAN) and 210796 (ALEA) and the PASCAL Network of Excellence. We thank the anonymous referees for helpful comments.

References

[ABS08] Ackermann, M.R., Blömer, J., Sohler, C.: Clustering for metric and non-metric distance measures. In: Proc. of the 19th ACM-SIAM Symposium on Discrete Algorithms, pp. 799–808 (2008)

[AV07] Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proc. of the 18th ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)

[AW01] Azoury, K.S., Warmuth, M.K.: Relative loss bounds for on-line density estimation with the exponential family of distributions. Machine Learning Journal 43(3), 211–246 (2001)

[BGW05] Banerjee, A., Guo, X., Wang, H.: On the optimality of conditional expectation as a Bregman predictor. IEEE Trans. on Information Theory 51, 2664–2669 (2005)

[BMDG05] Banerjee, A., Merugu, S., Dhillon, I., Ghosh, J.: Clustering with Bregman divergences. Journal of Machine Learning Research 6, 1705–1749 (2005)

[Bre67] Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comp. Math. and Math. Phys. 7, 200–217 (1967)

[CKW07] Crammer, K., Kearns, M., Wortman, J.: Learning from multiple sources. In: Advances in Neural Information Processing Systems 19, pp. 321–328. MIT Press, Cambridge (2007)

[CM08] Chaudhuri, K., McGregor, A.: Finding metric structure in information-theoretic clustering. In: Proc. of the 21st Conference on Learning Theory (2008)

[DD06] Deza, E., Deza, M.-M.: Dictionary of distances. Elsevier, Amsterdam (2006)

[Gen03] Gentile, C.: The robustness of the p-norm algorithms. Machine Learning Journal 53(3), 265–299 (2003)

[Llo82] Lloyd, S.: Least squares quantization in PCM. IEEE Trans. on Information Theory 28, 129–136 (1982)

[NBN07] Nielsen, F., Boissonnat, J.-D., Nock, R.: On Bregman Voronoi diagrams. In: Proc. of the 18th ACM-SIAM Symposium on Discrete Algorithms, pp. 746–755 (2007)

[NN05] Nock, R., Nielsen, F.: Fitting the smallest enclosing Bregman ball. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 649–656. Springer, Heidelberg (2005)

[ORSS06] Ostrovsky, R., Rabani, Y., Schulman, L.J., Swamy, C.: The effectiveness of Lloyd-type methods for the k-means problem. In: Proc. of the 47th IEEE Symposium on the Foundations of Computer Science, pp. 165–176. IEEE Computer Society Press, Los Alamitos (2006)

[Vel02] Veldhuis, R.: The centroid of the symmetrical Kullback-Leibler distance. IEEE Signal Processing Letters 9, 96–99 (2002)

Appendix: Proofsketch of Lemma 6

Fix $\mathcal{A} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_K\}$ and let

$$\mathrm{BR}_\psi(\mathcal{A}) \doteq \frac{\sum_{i=1}^K \psi(\boldsymbol{x}_i)}{K} - \psi\left(\frac{\sum_{i=1}^K \boldsymbol{x}_i}{K}\right)$$

be the Burbea-Rao divergence generated by ψ on $\mathcal{A}$, that is, the non negative remainder of Jensen's inequality [DD06]. It shall be convenient to abbreviate the

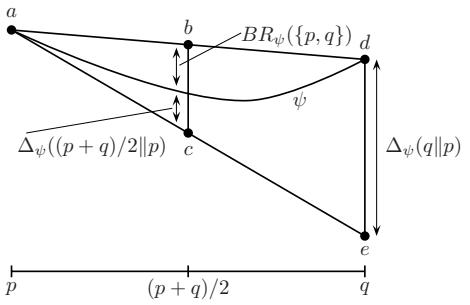

Fig. 2. Plot of some convex function ψ defined on some segment $[p, q]$. Here, $a = (p, \psi(p))$, $b = (q, \psi(q))$ and segment ae is tangent to ψ in a. Thales theorem in triangles (a, b, c) and (a, d, e) proves (29), as it gives indeed $|de|/|bc| = |ad|/|ab| = |ae|/|ac|$ (here, $|.|$ denotes the Euclidean length).

arithmetic and Bregman averages of $\mathcal{A}$ as $\boldsymbol{c}$ and $\boldsymbol{c}^\star$ respectively. Then we want to estimate the ratio between the uniform seeding potential and the optimal potential for $\mathcal{A}$:

$$\rho_{\mathcal{A}} \doteq \frac{1}{K} \sum_{i=1}^{K} \frac{\sum_{\boldsymbol{x} \in \mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{x}_i \| \boldsymbol{x} \| \boldsymbol{x}_i)}{\sum_{\boldsymbol{x} \in \mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{c}^\star \| \boldsymbol{x} \| \boldsymbol{c})} \ . \tag{25}$$

Fix $i \in \{1, 2, ..., K\}$. First, it comes from (7) that $\sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{x} \| \boldsymbol{x}_i) - \sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{x} \| \boldsymbol{c}) = K \Delta_\psi(\boldsymbol{c} \| \boldsymbol{x}_i)$, and we also get from (8) that $\sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{x}_i \| \boldsymbol{x}) - \sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{c}^\star \| \boldsymbol{x}) = K \Delta_\psi(\boldsymbol{x}_i \| \boldsymbol{c}^\star)$. The numerator of (25) becomes:

$$\begin{aligned} \sum_{\boldsymbol{x} \in \mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{x}_i \| \boldsymbol{x} \| \boldsymbol{x}_i) &= \alpha \left(\sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{x} \| \boldsymbol{c}) + K \Delta_\psi(\boldsymbol{c} \| \boldsymbol{x}_i) \right) \\ &\quad + (1-\alpha) \left(\sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_\psi(\boldsymbol{c}^\star \| \boldsymbol{x}_i) + K \Delta_\psi(\boldsymbol{x}_i \| \boldsymbol{c}^\star) \right) \qquad (26) \\ &= \sum_{\boldsymbol{x} \in \mathcal{A}} \triangle_{\psi,\alpha}(\boldsymbol{c}^\star \| \boldsymbol{x} \| \boldsymbol{c}) + K \triangle_{\psi,1-\alpha}(\boldsymbol{c} \| \boldsymbol{x}_i \| \boldsymbol{c}^\star) \ . \qquad (27) \end{aligned}$$

The left summand in (27) is the optimal potential for the cluster. Finally, the denominator of (25) can be rewritten as $\sum_{i=1}^{K} \triangle_{\psi,\alpha}(\boldsymbol{c}^\star \| \boldsymbol{x}_i \| \boldsymbol{c}) = K(\alpha \mathrm{BR}_\psi(\mathcal{A}) + (1-\alpha)\mathrm{BR}_{\psi^\star}(\boldsymbol{\nabla}\psi\mathcal{A}))$, where $\boldsymbol{\nabla}\psi\mathcal{A}$ is the set of gradient images of the elements of $\mathcal{A}$. We get:

$$\rho_{\mathcal{A}} = 1 + \frac{1}{K} \sum_{i=1}^{K} \frac{\triangle_{\psi,1-\alpha}(\boldsymbol{c} \| \boldsymbol{x}_i \| \boldsymbol{c}^\star)}{\alpha \mathrm{BR}_\psi(\mathcal{A}) + (1-\alpha)\mathrm{BR}_{\psi^\star}(\boldsymbol{\nabla}\psi\mathcal{A})} \ . \tag{28}$$

For the sake of simplicity, let $\boldsymbol{c}_{i,j}$ denote the weighted arithmetic average of $\mathcal{A}$ in which the weight of each $\boldsymbol{x}_k$ is $\frac{1}{2^j K}$ for $k \neq i$, and the weight of $\boldsymbol{x}_i$ is $\frac{1}{2^j K} + 1 - \frac{1}{2^j}$

$(\forall j \geq 0)$. We also let $\boldsymbol{c}^\star_{i,j}$ denote the weighted Bregman average of $\mathcal{A}$ under this same distribution. Thus, as j increases, the averages get progressively close to $\boldsymbol{x}_i$. Then, we have $\forall j \geq 0$:

$$\begin{aligned}
\Delta_\psi(\boldsymbol{c}_{i,j} \| \boldsymbol{x}_i) &= 2(\mathrm{BR}_\psi(\{\boldsymbol{c}_{i,j}, \boldsymbol{x}_i\}) + \Delta_\psi((\boldsymbol{c}_{i,j} + \boldsymbol{x}_i)/2 \| \boldsymbol{x}_i)) && (29)\\
&= 2(\mathrm{BR}_\psi(\{\boldsymbol{c}_{i,j}, \boldsymbol{x}_i\}) + \Delta_\psi(\boldsymbol{c}_{i,j+1} \| \boldsymbol{x}_i)) \ . \\
\Delta_\psi(\boldsymbol{x}_i \| \boldsymbol{c}^\star_{i,j}) &= \Delta_{\psi^\star}(\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,j}) \| \boldsymbol{\nabla}\psi(\boldsymbol{x}_i)) && (30)\\
&= 2(\mathrm{BR}_{\psi^\star}(\{\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,j}), \boldsymbol{\nabla}\psi(\boldsymbol{x}_i)\}) && (31)\\
&\quad + \Delta_{\psi^\star}((\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,j}) + \boldsymbol{\nabla}\psi(\boldsymbol{x}_i))/2 \| \boldsymbol{\nabla}\psi(\boldsymbol{x}_i))) && (32)\\
&= 2(\mathrm{BR}_{\psi^\star}(\{\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,j}), \boldsymbol{\nabla}\psi(\boldsymbol{x}_i)\}) + \Delta_\psi(\boldsymbol{x}_i \| \boldsymbol{c}^\star_{i,j+1})) \ .
\end{aligned}$$

While (30) is just stating the convex conjugate, Thales Theorem proves both (29) and (32). Figure 2 presents a simple graphical view to state this result in the context of Bregman and Burbea-Rao divergences. We get:

$$\begin{aligned}
\triangle_{\psi,1-\alpha}(\boldsymbol{c}_{i,j} \| \boldsymbol{x}_i \| \boldsymbol{c}^\star_{i,j}) &= \alpha\Delta_\psi(\boldsymbol{c}_{i,j} \| \boldsymbol{x}_i) + (1-\alpha)\Delta_\psi(\boldsymbol{x}_i \| \boldsymbol{c}^\star_{i,j}) \\
&= 2(\alpha \mathrm{BR}_\psi(\{\boldsymbol{c}_{i,j}, \boldsymbol{x}_i\}) && (33)\\
&\quad + (1-\alpha)\mathrm{BR}_{\psi^\star}(\{\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,j}), \boldsymbol{\nabla}\psi(\boldsymbol{x}_i)\})) \\
&\quad + \triangle_{\psi,1-\alpha}(\boldsymbol{c}_{i,j+1} \| \boldsymbol{x}_i \| \boldsymbol{c}^\star_{i,j+1}) \ . && (34)
\end{aligned}$$

Note that $\boldsymbol{c} = \boldsymbol{c}_{i,0}$ and $\boldsymbol{c}^\star = \boldsymbol{c}^\star_{i,0}$, $\forall i = 1, 2, ..., K$. We let:

$$b_0 \doteq \frac{2}{K} \cdot \frac{\sum_{i=1}^K \left\{\alpha \mathrm{BR}_\psi(\{\boldsymbol{c}_{i,0}, \boldsymbol{x}_i\}) + (1-\alpha)\mathrm{BR}_{\psi^\star}(\{\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,0}), \boldsymbol{\nabla}\psi(\boldsymbol{x}_i)\})\right\}}{\alpha \mathrm{BR}_\psi(\mathcal{A}) + (1-\alpha)\mathrm{BR}_{\psi^\star}(\boldsymbol{\nabla}\psi\mathcal{A})}$$

and for all $j > 0$

$$b_j \doteq 2 \cdot \frac{\sum_{i=1}^K \left\{\alpha \mathrm{BR}_\psi(\{\boldsymbol{c}_{i,j}, \boldsymbol{x}_i\}) + (1-\alpha)\mathrm{BR}_{\psi^\star}(\{\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,j}), \boldsymbol{\nabla}\psi(\boldsymbol{x}_i)\})\right\}}{\sum_{i=1}^K \left\{\alpha \mathrm{BR}_\psi(\{\boldsymbol{c}_{i,j-1}, \boldsymbol{x}_i\}) + (1-\alpha)\mathrm{BR}_{\psi^\star}(\{\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,j-1}), \boldsymbol{\nabla}\psi(\boldsymbol{x}_i)\})\right\}} \ .$$

Furthermore, $\forall j \geq 0$, we let:

$$r_j \doteq \frac{\sum_{i=1}^K \triangle_{\psi,1-\alpha}(\boldsymbol{c}_{i,j} \| \boldsymbol{x}_i \| \boldsymbol{c}^\star_{i,j})}{2\sum_{i=1}^K \left\{\alpha \mathrm{BR}_\psi(\{\boldsymbol{c}_{i,j}, \boldsymbol{x}_i\}) + (1-\alpha)\mathrm{BR}_{\psi^\star}(\{\boldsymbol{\nabla}\psi(\boldsymbol{c}^\star_{i,j}), \boldsymbol{\nabla}\psi(\boldsymbol{x}_i)\})\right\}} \ .$$

Plugging (34) into (28) and using the last notations yields:

$$\begin{aligned}
\rho_{\mathcal{A}} &= 1 + b_0(1 + b_1(...(1 + b_J r_J))) \\
&\geq 1 + b_0(1 + b_1(...(1 + b_{J-1}))) \\
&\geq \sum_{j=0}^{J-1} \left(\frac{1}{2\rho_\psi^2}\right)^j = \frac{(2\rho_\psi^2)^J - 1}{(2\rho_\psi^2)^J} \cdot \frac{2\rho_\psi^2}{2\rho_\psi^2 - 1} \ , \forall J \geq 0 \ .
\end{aligned}$$

The last inequality is obtained after various suitable Taylor expansions of ψ are used for $b_i, i \geq 0$, which gives $b_i \geq 1/(2\rho_\psi^2)$ (not shown to save space).