# Fitting the smallest enclosing Bregman balls

Richard Nock[1] and Frank Nielsen[2]

[1] Université Antilles-Guyane
`rnock@martinique.univ-ag.fr`
[2] Sony Computer Science Laboratories, Inc.
`Frank.Nielsen@acm.org`

**Abstract.** Finding a point which minimizes the maximal distortion with respect to a dataset is an important estimation problem that has recently received growing attentions in machine learning, with the advent of one class classification. In this paper, we study the problem from a general standpoint, and suppose that the distortion is a Bregman divergence, without restriction. Applications of this formulation can be found in machine learning, statistics, signal processing and computational geometry. We propose two theoretically founded generalizations of a popular smallest enclosing ball approximation algorithm for Euclidean spaces coined by Bădoiu and Clarkson in 2002. Experiments clearly display the advantages of being able to tune the divergence depending on the data's domain. As an additional result, we unveil an useful bijection between Bregman divergences and a family of popular averages that includes the arithmetic, geometric, harmonic and power means.

## 1 Introduction

Consider the following problem: given a set of observed data $\mathcal{S}$, compute some accurate set of parameters, or simplified descriptions, that *summarize* ("fit well") $\mathcal{S}$ according to some criteria. This problem is well known in various fields of statistics and computer science. In many cases, it admits two different formulations:

(1.) Find a point $\mathbf{c}^*$ which minimizes an *average distortion* with respect to $\mathcal{S}$.
(2.) Find a point $\mathbf{c}^*$ which minimizes a *maximal distortion* with respect to $\mathcal{S}$.

These two problems are cornerstones of different subfields of applied mathematics and computer science, such as (i) parametric estimation and the computation of *exhaustive* statistics for broad classes of distributions in statistics, (ii) one class classification and clustering in machine learning, (iii) the one center problem and its generalizations in computational geometry, among others [1, 2, 5, 9].

The main unknown in both problems is what we mean by *distortion*. Intuitively, for any two elements of $\mathcal{S}$, it should be lower-bounded, attain its minimum when they represent the same element, and it should otherwise give an accurate real-valued appreciation of the way they actually "differ". Maybe the most prominent example is the squared Euclidean distance (abbreviated $L_2^2$ for short) for real-valued vectors, which is the componentwise sum of the squared differences. It is certainly the most commonly used distortion measure in computational geometry, and one of the most favored in machine learning (support

vector machines, support vector domain description / one class clustering, etc.) [1, 5, 9–11].

In fact, many examples of distortion measures found in domains concerned by the problems above (computational geometry, machine learning, signal processing, probabilities and statistics, among others) fall into a *single* family of distortion measures known as Bregman divergences [3]. Informally, each of them is the tail of the Taylor expansion of a strictly convex function. Using a neat result in [2], it can be shown that the solution to problem (1.) above is always the average member of $\mathcal{S}$, *regardless of the Bregman divergence*. This means that problem (1.) can be solved in optimal linear time / space in the size of $\mathcal{S}$: since $\mathcal{S}$ may be huge, this property is crucial. Unfortunately, the solution of (2.) does not seem to be as affordable; tackling the problem with quadratic programming buys an expensive time complexity cubic in the worst case, and the space complexity is quadratic [10]. Notice also that it is mostly used with $L_2^2$. Instead of finding an exact solution, a recent approach due to [1] *approximates* the solution of the problem for $L_2^2$: the user specifies some $\varepsilon > 0$, and the algorithm returns, in time *linear* in the size of $\mathcal{S}$ and $1/\varepsilon$ and in space *linear* in the size of $\mathcal{S}$, the center $\mathbf{c}$ of a ball which is at $L_2^2$ divergence no more than $\varepsilon^2 r^*$ from $\mathbf{c}^*$. Here, $r^*$ is the squared radius of the so-called *smallest enclosing ball* of $\mathcal{S}$, whose center $\mathbf{c}^*$ is obviously the solution to problem (2.). Let us name this algorithm the Bădoiu-Clarkson algorithm, and abbreviate it BC. The key point of the algorithm is its simplicity, which deeply contrasts with quadratic programming approaches: basically, after having initialized $\mathbf{c}$ to a random point of $\mathcal{S}$, we iterate through finding the farthest point away from the current center, and then move along the line between these two points. The main limiting factor of BC is its time complexity's linear dependence in $1/\varepsilon$, which means an exponential dependence in the coding size of the approximation parameter [8]. However, from an experimental standpoint, good approximations may be obtained for reasonable values of $\varepsilon$, and the popularity of the algorithm, initially focused in computational geometry, has begun to spread to machine learning as well, with its adaptation to fast approximations of SVM training [10].

The applications of BC have remained so far focused on $L_2^2$, yet the fact that the algorithm gives a clean and simple approach to problem (2.) for *one* Bregman divergence naturally raises the question of whether it can be tailored to approximating problem (2.) *for any* Bregman divergence as well. Figure 1 highlights the importance of this issue. Suppose that the elements of $\mathcal{S}$ are observed speech spectrums, in which case problem (2.) amounts to finding a *model spectrum* [12] which minimizes the maximal distortion with respect to the elements of $\mathcal{S}$. A popular distortion measure in this case is (discrete) Itakura-Saito [7, 12]. Figure 1 gives the example of an Itakura-Saito ball with its center $\mathbf{c}^*$, the solution to problem (2.). We can consider *e.g.* that the elements of $\mathcal{S}$ are sampled uniformly at random in the ball. As usual, running BC on $\mathcal{S}$ approximates the center of its smallest $L_2^2$ enclosing ball (see Figure 1), and *not* its smallest enclosing Itakura-Saito ball. Regardless of the value of $\varepsilon$, the result is an extremely poor approximation of $\mathbf{c}^*$.
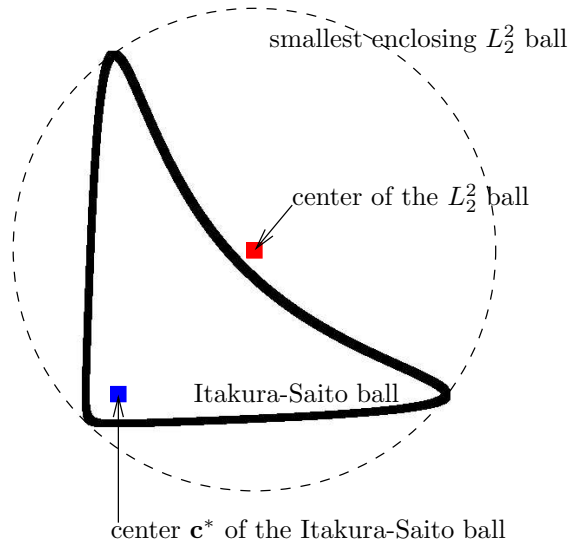
**Fig. 1.** An optimal Itakura-Saito ball and its smallest enclosing $L_2^2$ ball, for $d = 2$. Notice the poor quality of this *optimal* approximation: the center of the $L_2^2$ ball does not even lie inside the Itakura-Saito ball.

In this paper, we propose two theoretically founded generalizations of BC to arbitrary Bregman divergences, along with a bijection property that has a flavor similar to a Theorem of [2]: we show a bijection between the set of Bregman divergences and the set of the most commonly used functional averages, which yields that each element of the latter set encodes the minimax distortion solution for a Bregman divergence. This property is the cornerstone of our modifications to BC. The next Section presents some definitions. Section 3 gives the theoretical foundations and Section 4 the experiments regarding our generalization of BC. A last Section concludes the paper.

## 2 Definitions

Our notations mostly follow those of $[1,2]$. Bold faced variables such as $\mathbf{x}$ and $\boldsymbol{\alpha}$, represent column vectors. Sets are represented by calligraphic upper-case alphabets, *e.g.* $\mathcal{S}$, and enumerated as $\{\mathbf{s}_i : i \geq 1\}$ for vector sets, and $\{s_i : i \geq 1\}$ otherwise. The $j^{th}$ component of vector $\mathbf{s}$ is noted $s_j$, for $j \leq 1$. Vectors are supposed $d$-dimensional. We write $\mathbf{x} \geq \mathbf{y}$ as a shorthand for $x_i \geq y_i, \forall i$. The cardinal of a set $\mathcal{S}$ is written $|\mathcal{S}|$, and $\langle ., . \rangle$ defines the inner product for real valued vectors, *i.e.* the dot product. Norms are $L_2$ for a vector, and Frobenius for a matrix. Bregman divergences are a parameterized family of distortion measures: let $F : \mathcal{X} \to I\!\!R$ be strictly convex and differentiable on the interior $int(\mathcal{X})$ of some convex set $\mathcal{X} \subseteq I\!\!R^d$. Its corresponding Bregman divergence is:

$$D_F(\mathbf{x'}, \mathbf{x}) = F(\mathbf{x'}) - F(\mathbf{x}) - \langle \mathbf{x'} - \mathbf{x}, \boldsymbol{\nabla}_F(\mathbf{x}) \rangle \ . \tag{1}$$
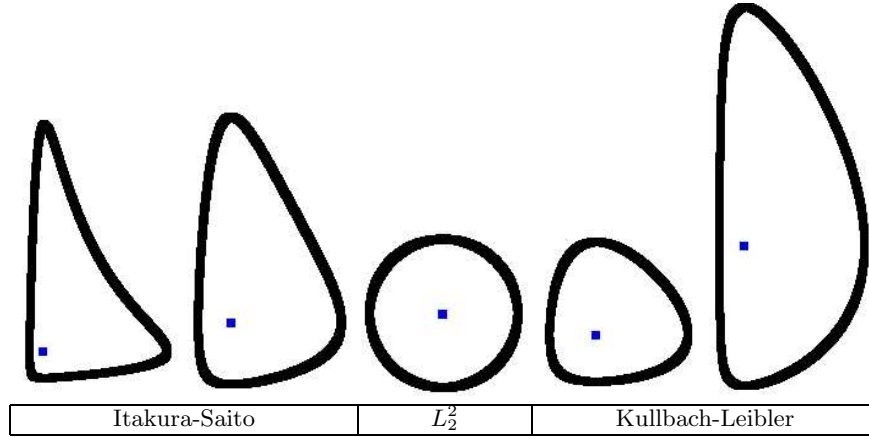
| Itakura-Saito | $L_2^2$ | Kullbach-Leibler |

**Fig. 2.** Examples of Bregman Balls, for $d = 2$. Blue dots are the centers of the balls.

Here, $\boldsymbol{\nabla}_F$ is the gradient operator of $F$. A Bregman divergence has the following properties: it is convex in **x'**, always non negative, and zero iff $\mathbf{x} = \mathbf{x'}$. Whenever $F(\mathbf{x}) = \sum_{i=1}^{d} x_i^2 = \|\mathbf{x}\|_2^2$, the corresponding divergence is the squared Euclidean distance ($L_2^2$): $D_F(\mathbf{x'}, \mathbf{x}) = \|\mathbf{x} - \mathbf{x'}\|_2^2$, with which is associated the common definition of a ball in an Euclidean metric space:

$$\mathcal{B}_{\mathbf{c},r} = \{\mathbf{x} \in \mathcal{X} : \|\mathbf{x} - \mathbf{c}\|_2^2 \le r\} \ , \tag{2}$$

with $\mathbf{c} \in \mathcal{S}$ the center of the ball, and $r \ge 0$ its (squared) radius. Eq. (2) suggests a natural generalization to the definition of balls for arbitrary Bregman divergences. However, since a Bregman divergence is usually not symmetric, any $\mathbf{c} \in \mathcal{S}$ and any $r \ge 0$ define actually two dual *Bregman balls*:

$$\mathcal{B}_{\mathbf{c},r} = \{\mathbf{x} \in \mathcal{X} : D_F(\mathbf{c}, \mathbf{x}) \le r\} \ , \tag{3}$$

$$\mathcal{B}'_{\mathbf{c},r} = \{\mathbf{x} \in \mathcal{X} : D_F(\mathbf{x}, \mathbf{c}) \le r\} \ . \tag{4}$$

Remark that $D_F(\mathbf{c}, \mathbf{x})$ is always convex in $\mathbf{c}$ while $D_F(\mathbf{x}, \mathbf{c})$ is not always, but the *boundary* $\partial\mathcal{B}_{\mathbf{c},r}$ is not always convex (it depends on $\mathbf{x}$, given $\mathbf{c}$), while $\partial\mathcal{B}'_{\mathbf{c},r}$ is always convex. In this paper, we are mainly interested in $\mathcal{B}_{\mathbf{c},r}$ because of the convexity of $D_F$ in $\mathbf{c}$. The conclusion of the paper extends some results to build $\mathcal{B}'_{\mathbf{c},r}$ as well. Figure 2 presents some examples of Bregman balls for three popular Bregman divergences (see Table 1 for the analytic expressions of the divergences). Let $\mathcal{S} \subseteq \mathcal{X}$ be a set of $m$ points that were sampled from $\mathcal{X}$. A *smallest enclosing Bregman ball* (SEBB) for $\mathcal{S}$ is a Bregman ball $\mathcal{B}_{\mathbf{c}^*,r^*}$ with $r^*$ the minimal real such that $\mathcal{S} \subseteq \mathcal{B}_{\mathbf{c}^*,r^*}$. With a slight abuse of language, we will refer to $r^*$ as the *radius* of the ball. Our objective is to approximate as best as possible the SEBB of $\mathcal{S}$, which amounts to minimizing the radius of the enclosing ball we build. As a simple matter of fact indeed, the SEBB is unique.

**Lemma 1.** *The smallest enclosing Bregman ball $\mathcal{B}_{\boldsymbol{c}^*,r^*}$ of $\mathcal{S}$ is unique.*

*Proof.* Suppose that there exists two SEBBs of $\mathcal{S}$, $\mathcal{B}_{\mathbf{c}^*,r^*}$ and $\mathcal{B}_{\mathbf{c'}^*,r^*}$, with $\mathbf{c}^* \neq \mathbf{c'}^*$. Then we have $\mathbf{c''}^* = (\mathbf{c}^* + \mathbf{c'}^*)/2 \in \mathcal{X}$ and $\forall \mathbf{s} \in \mathcal{S}, D_F(\mathbf{c''}^*, \mathbf{s}) < (D_F(\mathbf{c}^*, \mathbf{s}) + D_F(\mathbf{c'}^*, \mathbf{s}))/2 \leq r^*$, the strict inequality following from the strict convexity of $F$. But this shows the existence of an enclosing Bregman ball for $\mathcal{S}$, $\mathcal{B}_{\mathbf{c''}^*,r'^*}$ with $r'^* < r^*$, thereby contradicting the fact that $\mathcal{B}_{\mathbf{c}^*,r^*}$ and $\mathcal{B}_{\mathbf{c'}^*,r^*}$ are SEBBs of $\mathcal{S}$.
□

Algorithm 1 presents Bădoiu-Clarkson's algorithm for the SEBB approximation problem with the $L_2^2$ divergence [1].

---

**Algorithm 1:** BC$(\mathcal{S}, T)$

---
> **Input**: Data $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m\}$;
> **Output**: Center $\mathbf{c}$;
> Choose at random $\mathbf{c} \in \mathcal{S}$;
> **for** $t = 1, 2, ..., T-1$ **do**
> > $\mathbf{s} \leftarrow \arg\max_{\mathbf{s'} \in \mathcal{S}} \|\mathbf{c} - \mathbf{s'}\|_2^2$;
> > $\mathbf{c} \leftarrow \frac{t}{t+1}\mathbf{c} + \frac{1}{t+1}\mathbf{s}$;

---

## 3  Extending BC

The primal SEBB problem is to find:

$$\arg\min_{\mathbf{c}^*,r^*} r^* \quad \text{s.t.} \quad D_F(\mathbf{c}^*, \mathbf{s}_i) \leq r^*, \forall 1 \leq i \leq m \ . \tag{5}$$

Its Lagrangian is:

$$L(\mathcal{S}, \boldsymbol{\alpha}) = r^* - \sum_{i=1}^m \alpha_i(r^* - D_F(\mathbf{c}^*, \mathbf{s}_i)) \ , \tag{6}$$

with the additional KKT condition $\boldsymbol{\alpha} \geq \mathbf{0}$. The solution to (5) is obtained by minimizing $L(\mathcal{S}, \boldsymbol{\alpha})$ for the parameters $\mathbf{c}^*$ and $r^*$, and then maximize the resulting dual for the Lagrange multipliers. We obtain

$$\partial L(\mathcal{S}, \boldsymbol{\alpha})/\partial \mathbf{c}^* = \boldsymbol{\nabla}_F(\mathbf{c}^*) \sum_{i=1}^m \alpha_i - \sum_{i=1}^m \alpha_i \boldsymbol{\nabla}_F(\mathbf{s}_i) \ ,$$

$$\partial L(\mathcal{S}, \boldsymbol{\alpha})/\partial r^* = 1 - \sum_{i=1}^m \alpha_i \ .$$

Setting $\partial L(\mathcal{S}, \boldsymbol{\alpha})/\partial \mathbf{c}^* = \mathbf{0}$ and $\partial L(\mathcal{S}, \boldsymbol{\alpha})/\partial r^* = 0$ yields

$$\sum_{i=1}^m \alpha_i = 1 \ , \tag{7}$$

$$\mathbf{c}^* = \boldsymbol{\nabla}_F^{-1}\left(\sum_{i=1}^m \alpha_i \boldsymbol{\nabla}_F(\mathbf{s}_i)\right) \ . \tag{8}$$

Because $F$ is strictly convex, $\nabla_F$ is bijective, and $\mathbf{c}^*$ lies in the convex closure of $\mathcal{S}$. Finally, we are left with finding:

$$\arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i D_F \left( \boldsymbol{\nabla}_F^{-1} \left( \sum_{j=1}^{m} \alpha_j \boldsymbol{\nabla}_F(\mathbf{s}_j) \right), \mathbf{s}_i \right) \quad \text{s.t.} \quad \boldsymbol{\alpha} \geq \mathbf{0}, \sum_{i=1}^{m} \alpha_i = 1 \ . \quad (9)$$

This problem generalizes the dual of support vector machines: whenever $F(\mathbf{s}) = \sum_{i=1}^{d} s_i^2 = \langle \mathbf{s}, \mathbf{s} \rangle$ (Table 1), we return to their kernel-based formulation [4]. There are essentially two categories of Lagrange multipliers in vector $\boldsymbol{\alpha}$. Those corresponding to points of $\mathcal{S}$ lying on the interior of $\mathcal{B}_{\mathbf{c}^*, r^*}$ are zero, since these points satisfy their respective constraints. The others, corresponding to the *support* points of the ball, are strictly positive. Each $\alpha_i > 0$ represents the contribution of its support point to the computation of the circumcenter of the ball. Eq. (8) is thus some *functional average* of the support points of the ball, to compute $\mathbf{c}^*$.

### 3.1 The Modified Bădoiu-Clarkson algorithm, `MBC`

There is more on eq. (8). A Bregman divergence is not affected by linear terms: $D_{F+q} = D_F$ for any constant $q$ [6]. Thus, the partial derivatives of $F$ in $\boldsymbol{\nabla}_F(.)$ determine entirely the Bregman divergence. The following Lemma is then immediate.

**Lemma 2.** *The set of functional averages (8) is in bijection with the set of Bregman divergences (1).*

The connection between the functional averages and divergences is much interesting because the classical means commonly used in many domains, such as convex analysis, parametric estimation, signal processing, are valid examples of functional averages. A nontrivial consequence of Lemma 2 is that each of them encodes the SEBB solution for an associated Bregman divergence. Apart from the SEBB problem, this is interesting because means are popular statistics, and we give a way to favor the choice of a mean against another one depending on the *domain* of the data and its "natural" distortion measure. Table 1 presents some Bregman divergences and their associated functional averages, for the most commonly encountered. Speaking of bijections, previous results showed the existence of a bijection between Bregman divergences and the family of exponential distributions [2]. This has helped the authors to devise a generalization of the $k$-means algorithm. In our case, Lemma 2 is also of some help to generalize `BC`. Clearly, the dual problem in eq. (9) does not admit the convenient representation of SVMs, and it seems somehow hard to use a kernel trick replacing the elements of $\mathcal{S}$ by local transformations involving $F$ prior to solving problem (9). However, the dual suggests a very simple algorithm to approximate $\mathbf{c}^*$, which consists in making the parallel between $\boldsymbol{\nabla}(\mathbf{c}^*) = \sum_{i=1}^{m} \alpha_i \boldsymbol{\nabla}_F(\mathbf{s}_i)$ (8) and the arithmetic mean in Table 1, and consider (8) as the solution to a minimum distortion problem involving gradients into a $L_2^2$ space. We can thus seek:

$$\arg\min_{\mathbf{g}^*, r'^*} r'^* \quad \text{s.t.} \quad \|\mathbf{g}^* - \boldsymbol{\nabla}_F(\mathbf{s}_i)\|_2^2 \leq r'^*, \forall 1 \leq i \leq m \ . \quad (10)$$

| domain | $F(\mathbf{s})$ | $D_F(\mathbf{c}, \mathbf{s})$ | $c_j\ (1 \leq j \leq d)$ |
|---|---|---|---|
| | | $L_2^2$ norm | arithmetic mean |
| $I\!R^d$ | $\sum_{j=1}^d s_j^2$ | $\sum_{j=1}^d (c_j - s_j)^2$ | $\sum_{i=1}^m \alpha_i s_{i,j}$ |
| $(I\!R^{+,*})^d$ | | (I/KL)-divergence | geometric mean |
| / $d$-simplex | $\sum_{j=1}^d s_j \log s_j - s_j$ | $\sum_{j=1}^d c_j \log(c_j/s_j) - c_j + s_j$ | $\prod_{i=1}^m s_{i,j}^{\alpha_i}$ |
| | | Itakura-Saito distance | harmonic mean |
| $(I\!R^{+,*})^d$ | $-\sum_{j=1}^d \log s_j$ | $\sum_{j=1}^d (c_j/s_j) - \log(c_j/s_j) - 1$ | $1/\sum_{i=1}^m (\alpha_i/s_{i,j})$ |
| | | Mahalanobis distance | arithmetic mean |
| $I\!R^d$ | $\mathbf{s}^T A \mathbf{s}$ | $(\mathbf{c} - \mathbf{s})^T A (\mathbf{c} - \mathbf{s})$ | $\sum_{i=1}^m \alpha_i s_{i,j}$ |
| | $p \in I\!N \backslash \{0, 1\}$ | | weighted power mean |
| $I\!R^d / I\!R^{+d}$ | $(1/p) \sum_{j=1}^d s_j^p$ | $\sum_{j=1}^d \frac{c_j^p}{p} + \frac{(p-1)s_j^p}{p} - c_j s_j^{p-1}$ | $\left( \sum_{i=1}^m \alpha_i s_{i,j}^{p-1} \right)^{1/(p-1)}$ |

**Table 1.** Some common Bregman divergences and their associated functional averages. The second row depicts the general I (information) divergence, also known as Kullbach-Leibler (KL) divergence on the $d$-dimensional probability simplex. On the fourth row, $A$ is the inverse of the covariance matrix [2].

Finally, approximating (5) amounts to running the so-called Modified Bădoiu-Clarkson algorithm, as indicated in algorithm 2 below. Because $\boldsymbol{\nabla}_F$ is bijec-

---

**Algorithm 2:** MBC$(\mathcal{S}, T)$

> **Input**: Data $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m\}$;
> **Output**: Center $\mathbf{c}$;
> $\mathcal{S}' \leftarrow \{\boldsymbol{\nabla}_F(\mathbf{s}_i) : \mathbf{s}_i \in \mathcal{S}\}$;
> $\mathbf{g} \leftarrow$ BC$(\mathcal{S}', T)$;
> $\mathbf{c} \leftarrow \boldsymbol{\nabla}_F^{-1}(\mathbf{g})$;

---

tive, this is guaranteed to yield a solution. The remaining question is whether $\boldsymbol{\nabla}_F^{-1}(\mathbf{g}) = \mathbf{c}$ is close enough from the solution $\mathbf{c}^*$ of (5). The following Lemma upperbounds the sum of the two divergences between $\mathbf{c}$ and any point of $\mathcal{S}$, as a function of $r'^*$. It shows that the two centers can be very close to each other; in fact, they can be *much* closer than with a naive application of Bădoiu-Clarkson directly in $\mathcal{S}$. The Lemma makes the hypothesis that the Hessian of $F$, $H_F$, is non singular. As a matter of fact, it is diagonal (without zero in the diagonal) for all classical examples of Bregman divergences, see Table 1, so this is not a restriction either. In the Lemma, we let $f$ denote the minimal non zero value of the Hessian norm inside the convex closure of $\mathcal{S}$: $f = \min_{\mathbf{x} \in \mathrm{co}(\mathcal{S}): \|H_F(\mathbf{x})\|_2 > 0} \|H_F(\mathbf{x})\|_2$.

**Lemma 3.** $\forall \boldsymbol{s} \in \mathcal{S}$, we have:

$$D_F(\boldsymbol{s}, \boldsymbol{\nabla}_F^{-1}(\boldsymbol{g})) + D_F(\boldsymbol{\nabla}_F^{-1}(\boldsymbol{g}), \boldsymbol{s}) \leq (1+\varepsilon)^2 r'^*/f \ , \tag{11}$$

where $\boldsymbol{g}$ is defined in algorithm 2, $r'^*$ is defined in eq. (10), and $\varepsilon$ is the error parameter of BC.

*Proof.* We know from BC that $\mathbf{g}$ is at Euclidean distance no more than $(1+\varepsilon)\sqrt{r'^*}$ of any other points in the image of $\mathcal{S}$ by $\mathbf{\nabla}_F$. Thus, we have $\|\mathbf{g} - \mathbf{\nabla}_F(\mathbf{s})\|_2 \leq (1+\varepsilon)\sqrt{r'^*}$, and $\mathbf{\nabla}_F(\mathbf{s}) = \mathbf{g} + a(1+\varepsilon)\mathbf{u}$, for some real $a \leq \sqrt{r'^*}$ and $\mathbf{u} \in I\!\!R^d$ for which $\|\mathbf{u}\|_2 = 1$. Fixing $\mathbf{\nabla}_F^{-1}(\mathbf{g}) = \mathbf{c}$ yields for any $\mathbf{s} \in \mathcal{S}$:

$$
\begin{aligned}
D_F(\mathbf{s}, \mathbf{c}) + D_F(\mathbf{c}, \mathbf{s}) &= F(\mathbf{s}) - F(\mathbf{c}) - \langle \mathbf{s} - \mathbf{c}, \mathbf{g} \rangle + F(\mathbf{c}) - F(\mathbf{s}) - \langle \mathbf{c} - \mathbf{s}, \mathbf{\nabla}_F(\mathbf{s}) \rangle \\
&= \langle \mathbf{s} - \mathbf{c}, \mathbf{\nabla}_F(\mathbf{s}) - \mathbf{g} \rangle \\
&= (1+\varepsilon)a\langle \mathbf{s} - \mathbf{c}, \mathbf{u} \rangle \\
&\leq (1+\varepsilon)\sqrt{r'^*}\|\mathbf{s} - \mathbf{c}\|_2 \ .
\end{aligned}
\tag{12}
$$

Provided $F$ is continuous (a mild assumption, given its strict convexity), the mean-value theorem brings that there exists $\hat{\mathbf{c}}$ in the convex closure of $\mathcal{S}$ such that $\mathbf{\nabla}_F(\mathbf{s}) = \mathbf{g} + H_F(\hat{\mathbf{c}}).(\mathbf{s} - \mathbf{c})$. The non singularity of the Hessian yields $\mathbf{s} - \mathbf{c} = H_F^{-1}(\hat{\mathbf{c}}).(\mathbf{\nabla}_F(\mathbf{s}) - \mathbf{g})$, and $\|\mathbf{s} - \mathbf{c}\|_2 \leq \|H_F^{-1}(\hat{\mathbf{c}})\|_2\|\mathbf{\nabla}_F(\mathbf{s}) - \mathbf{g}\|_2 \leq (1+\varepsilon)\sqrt{r'^*}/\|H_F(\hat{\mathbf{c}})\|_2$. Plugging this into ineq. (12) and using the fact that $\|H_F(\hat{\mathbf{c}})\|_2 \geq f$ yields the statement of the Lemma. $\square$

Remark that Lemma 3 is optimal, in the sense that if we consider $D_F = L_2^2$, then each point $\mathbf{s}_i \in \mathcal{S}$ becomes $2\mathbf{s}_i$ in $\mathcal{S}'$. The optimal radii in (5) and (10) satisfy $r'^* = 4r^*$, and we have $f = 2$. Plugging this altogether in eq. (11) yields $2\|\mathbf{c} - \mathbf{s}\|_2^2 \leq (1+\varepsilon)^2 \times 4r^*/2$, *i.e.* $\|\mathbf{c} - \mathbf{s}\|_2 \leq (1+\varepsilon)\sqrt{r^*}$, which is exactly Bǎdiou-Clarkson's bound [1] (here, we have fixed $\mathbf{c} = \mathbf{\nabla}_F^{-1}(\mathbf{g})$, like in Lemma 3). Remark also that Lemma 3 upperbounds the sum of both possible divergences, which is very convenient given the possible assymetry of $D_F$.

### 3.2   The Bregman-Bǎdoiu-Clarkson algorithm, BBC

It is straightforward to check that at the end of BC (algorithm 1), the following holds true:

$$
\begin{cases}
\mathbf{c} = \sum_{i=1}^m \hat{\alpha}_i \mathbf{s}_i \ , \sum_{i=1}^m \hat{\alpha}_i = 1 \ , \hat{\boldsymbol{\alpha}} \geq \mathbf{0} \ , \\
\forall 1 \leq i \leq m, \hat{\alpha}_i \neq 0 \text{ iff } \mathbf{s}_i \text{ is chosen at least once in } \mathtt{BC} \ .
\end{cases}
$$

Since the furthest points chosen by BC ideally belong to $\partial\mathcal{B}_{\mathbf{c}^*, r^*}$, and the final expression of $\mathbf{c}$ matches the arithmetic average of Table 1, it comes that BC *directly* tackles an iterative approximation of eq. (8) for the $L_2^2$ Bregman divergence. If we replace $L_2^2$ by an arbitrary Bregman divergence, then BC can be generalized in a quite natural way to algorithm BBC (for Bregman-Bǎdoiu-Clarkson) below. Again, it is straightforward to check that at the end of BBC, we have generalized the iterative approximation of BC to eq. (8) for any Bregman divergence, as we have:

$$
\begin{cases}
\mathbf{c} = \mathbf{\nabla}_F^{-1}\left(\sum_{i=1}^m \hat{\alpha}_i \mathbf{\nabla}_F(\mathbf{s}_i)\right) \ , \sum_{i=1}^m \hat{\alpha}_i = 1 \ , \hat{\boldsymbol{\alpha}} \geq \mathbf{0} \ , \\
\forall 1 \leq i \leq m, \hat{\alpha}_i \neq 0 \text{ iff } \mathbf{s}_i \text{ is chosen at least once in } \mathtt{BC} \ .
\end{cases}
$$

The main point is whether $\hat{\boldsymbol{\alpha}}$ is a good approximation to the true vector of Lagrange multipliers $\boldsymbol{\alpha}$. From the theoretical standpoint, the proof of BC's approximation ratio becomes tricky when lifted from $L_2^2$ to an arbitrary Bregman

**Algorithm 3:** BBC($\mathcal{S}$)

**Input**: Data $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m\}$;
**Output**: Center $\mathbf{c}$;
Choose at random $\mathbf{c} \in \mathcal{S}$;
**for** $t = 1, 2, ..., T-1$ **do**
$\quad \mathbf{s} \leftarrow \arg\max_{\mathbf{s}' \in \mathcal{S}} D_F(\mathbf{c}, \mathbf{s}')$;
$\quad \mathbf{c} \leftarrow \boldsymbol{\nabla}_F^{-1}\left(\frac{t}{t+1}\boldsymbol{\nabla}_F(\mathbf{c}) + \frac{1}{t+1}\boldsymbol{\nabla}_F(\mathbf{s})\right)$;
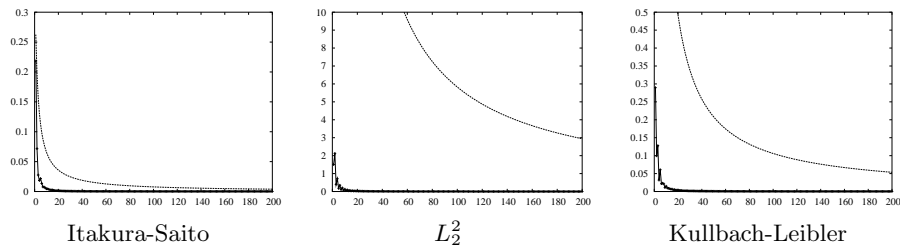


| Itakura-Saito | $L_2^2$ | Kullbach-Leibler |

**Fig. 3.** Average approximation curves for 100 runs of BBC algorithm for three Bregman divergences: Itakura-Saito, $L_2^2$ and KL ($d = 2, m = 1000, T = 200$). The dashed curves are Bădoiu-Clarkson's error bound as a function of the iteration number $t$, and the bottom, plain curves, depict $(D_F(\mathbf{c}^*, \mathbf{c}) + D_F(\mathbf{c}, \mathbf{c}^*))/2$ as a function of $t$ for each divergence, where $\mathbf{c}$ is the output of BBC and $\mathbf{c}^*$ is the optimal center.

divergence, but it can be shown that many of the key properties of the initial proof remain true in this more general setting. An experimental hint that speaks for itself for the existence of such a good approximation ratio is given in the next Section.

## 4 Experimental results

### 4.1 BBC

To evaluate the quality of the approximation of BBC for the SEBB, we have ran the algorithm for three popular representative Bregman divergences. For each of them, averages over a hundred runs were performed for $T = 200$ center updates (see algorithm 3). In each run, a random Bregman ball is generated, and $\mathcal{S}$ is sampled uniformly at random in the ball. Since we know the SEBB, we have a precise idea of the quality of the approximation found by BBC on the SEBB. Figure 3 gives a synthesis of the results for $d = 2$. [1]'s bound is plotted for each divergence, even when it holds formally only for $L_2^2$. The other two curves give an indication of the way this bound behaves with respect to the experimental results. It is easy to see that for each divergence, there is a very fast convergence of the center found, $\mathbf{c}$, to the optimal center $\mathbf{c}^*$. Furthermore, the experimental divergences are always much smaller than [1]'s bound, *for each divergence* (very

often by a factor 100 or more). We have checked this phenomenon for higher dimensions, up to $d = 20$.

To have a visual idea of the way BBC converges for each divergence, Figure 4 plots the Bregman balls found for small number of iterations of BBC with $d = 2$. Notice the differences in the Bregman balls, and the accuracy of BBC to approximate the SEBB on each of them. These three results actually appear to be representative of *all* those obtained for the experiments of Figure 3. Each time, for each divergence and each random Bregman ball, the algorithm displays a very accurate convergence of $\mathbf{c}$ towards $\mathbf{c}^*$. Finally, we have observed that BBC generally selects a very small number of support points for its approximation of $\mathbf{c}^*$. In almost all runs, and for each divergence, the number of support points does not exceed ten (thus, only at most 1% of $\mathcal{S}$). This is certainly a good point as it tends to confirm that the algorithm finds accurate approximations for the vector of Lagrange multipliers in eq. (8). Indeed, a small number of support points (recall that they lie on $\partial \mathcal{B}_{\mathbf{c},r}$) is required to define the center of any Bregman ball: if we could *e.g.* infinitely sample its interior, only two such points would suffice. Using the Pythagorean theorem of Bregman divergences [6], it is quite remarkable that BBC would chose only three points in the worst case for its approximation of $\mathbf{c}^*$: one in the interior of the ball, and two on its boundary.

### 4.2   MBC

Experimentally, MBC is less accurate than BBC. This is quite predictable, as the latter directly optimizes the cost function while the former transforms the problem to fit it into a $L_2^2$ optimization framework. However, when compared to BC, MBC has displayed much better approximation ratios, in particular for "skewed" divergences such as Itakura-Saito. Figure 5 displays an example of both algorithms ran on a ball which closely resembles that of Figure 1. Notice that the final approximation of BC is *so bad* that all the region of $I\!\!R^{+,*^2}$ which intersects the image actually belongs to the Bregman ball found, while BBC finds a center at divergence $< 20\%$ of the optimal radius from $\mathbf{c}^*$. From the experimental standpoint, we have also witnessed an interesting fact when comparing MBC and BBC: even when the approximation ratio of the former is not as good, its rate of convergence towards its final center $\mathbf{c}$ seems sometimes better. From the theoretical standpoint, a partial explanation comes from the way the ball is sampled: the uniform sampling is no longer uniform in the gradient space, and sometimes very few extremal points are available on some regions of the boundary of the Bregman ball. Since fewer boundary points are available, the center found, $\mathbf{c}$, gives sometimes the impression to move inside a smaller region of $\mathcal{X}$.

## 5   Conclusion

In this paper, we bring two non-trivial theoretically founded generalizations of Bǎdoiu and Clarkson's $L_2^2$ smallest enclosing ball algorithm to fit a smallest enclosing ball for an arbitrary Bregman divergence. Experimentally, the algorithms obtain very good results on approximating the SEBB. As a simple
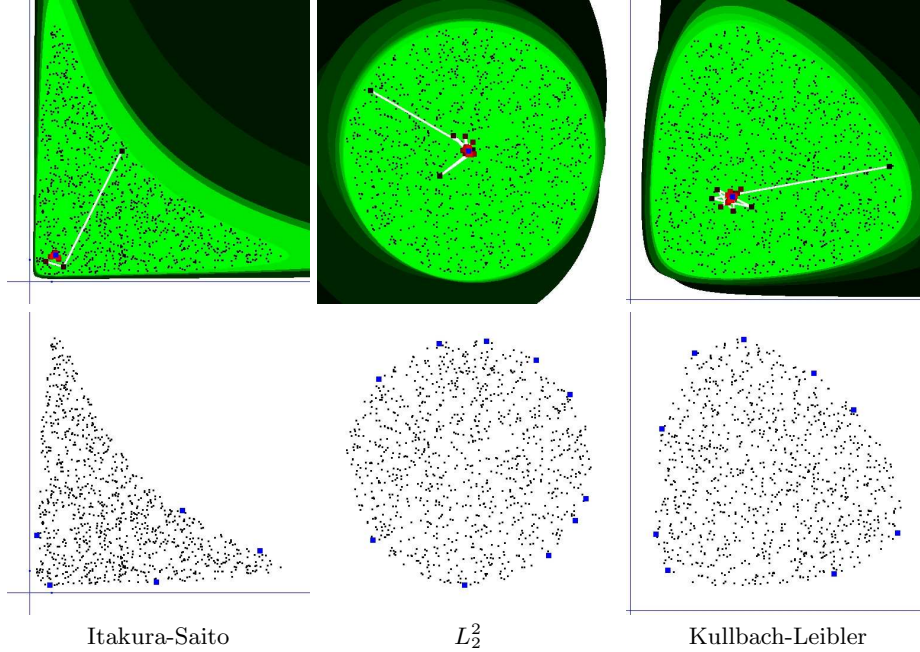
|  Itakura-Saito | $L_2^2$ | Kullbach-Leibler |

**Fig. 4.** (Best viewed in color) Examples of approximation balls found for Itakura-Saito, $L_2^2$ and KL divergences, after respectively 10, 20 and 20 iterations of the BBC algorithm ($d = 2, m = 1000$). *Top plots*: gradient colors, ranging from the darkest to the lightest, depict the balls found from the first to the last iteration: the green gradients are the balls themselves, and the red gradients their centers. The blue points are the optimal centers, and the white lines depict the trajectories of the approximated centers. The blue axes are the $x$ and $y$ axes, and when they are indicated, the small dots on the axes are the points $(50, 0)$ and $(0, 50)$. *Bottom plots*: the blue dots are the support points of the approximated Bregman balls.

matter of fact, the approach can be generalized to build $\mathcal{B}'_{\mathbf{c},r}$ for some Bregman divergences that are convex in their both parameters, such as $L_2^2$ or the (I/KL)-divergences. In this case, the same reasoning may be applied as in Section 3, yet the result obtained is much simpler. Indeed, the primal problem becomes $\arg\min_{\mathbf{c}^*,r^*} r^*$ s.t. $D_F(\mathbf{x}_i, \mathbf{c}^*) \leq r^*, \forall 1 \leq i \leq m$. Differentiating its Lagrangian, we get this time: $\partial L(\mathcal{S}, \boldsymbol{\alpha})/\partial \mathbf{c}^* = H_F(\mathbf{c}^*).(\mathbf{c}^* - \sum_{i=1}^m \alpha_i \mathbf{x}_i)$ and $\partial L(\mathcal{S}, \boldsymbol{\alpha})/\partial r^* = 1 - \sum_{i=1}^m \alpha_i$. Mild assumptions on the Hessian, true for all common Bregman divergences (basically, non singularity), yield that $\partial L(\mathcal{S}, \boldsymbol{\alpha})/\partial \mathbf{c}^* = \mathbf{0}$ implies $\mathbf{c}^* = \sum_{i=1}^m \alpha_i \mathbf{x}_i$, and we have $\sum_{i=1}^m \alpha_i = 1$ as well. Most interestingly, the dual simplifies to finding the observed distribution that maximizes the remainder of Jensen's inequality: $\arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^m \alpha_i F(\mathbf{x}_i) - F\left(\sum_{i=1}^m \alpha_i \mathbf{x}_i\right)$ s.t. $\boldsymbol{\alpha} \geq \mathbf{0}, \sum_{i=1}^m \alpha_i = 1$.
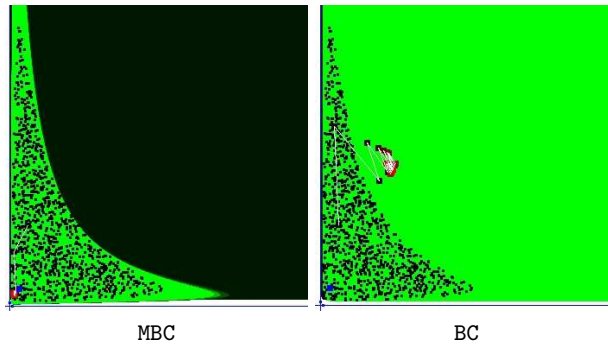
**Fig. 5.** Comparison of MBC and BC on an Itakura-Saito ball. Color conventions follow Figure 4. Notice the poor approximation found by BC on the Bregman ball, compared to MBC's (which lies in the SEBB, near $\mathbf{c}^*$, see text for details).

# References

1. M. Bădoiu and K.-L. Clarkson. Optimal core-sets for balls, 2002. Manuscript.
2. A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with bregman divergences. In *Proc. of the SIAM International Conference on Data Mining*, pages 234–245, 2004.
3. L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.
4. C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
5. K. Crammer and G. Chechik. A needle in a haystack: local one-class optimization. In *Proc. of the 21 $^{th}$ International Conference on Machine Learning*, 2004.
6. C. Gentile and M. Warmuth. Proving relative loss bounds for on-line learning algorithms using Bregman divergences. In *Tutorials of the 13 $^{th}$ International Conference on Computational Learning Theory*, 2000.
7. F. Itakura and S. Saito. A statistical method for estimation of a speech spectral density and formant frequencies. *Electronics and Communications in Japan*, 53:36–43, 1970.
8. F. Nielsen and R. Nock. A fast deterministic smallest enclosing disk approximation algorithm. *Information Processing Letters*, 93:263–268, 2005.
9. D. Tax and R. Duin. Support Vector Domain Description. *Pattern Recognition Letters*, 20:1191–1199, 1999.
10. I. W. Tsang, J. T. Kwok, and P.-M. Cheung. Core Vector Machines: fast SVM training on very large datasets. *Machine Learning Research Journal*, 6:363–392, 2005.
11. V. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.
12. B. Wei and J. D. Gibson. Comparison of distance measures in discrete spectral modeling. In *IEEE Digital Signal Processing Workshop*, 2000.