Journal of Machine Learning Research 8 (2007) 65-102 Submitted 1/06; Revised 9/06; Published 1/07

A Unified Continuous Optimization Framework for Center-Based Clustering Methods

Marc Teboulle TEBOULLE@MATH.TAU.AC.IL
School of Mathematical Sciences
Tel-Aviv University
Tel-Aviv 69978, Israel

Editor: Charles Elkan

Abstract

Center-based partitioning clustering algorithms rely on minimizing an appropriately formulated objective function, and different formulations suggest different possible algorithms. In this paper, we start with the standard nonconvex and nonsmooth formulation of the partitioning clustering problem. We demonstrate that within this elementary formulation, convex analysis tools and optimization theory provide a unifying language and framework to design, analyze and extend hard and soft center-based clustering algorithms, through a generic algorithm which retains the computational simplicity of the popular k-means scheme. We show that several well known and more recent center-based clustering algorithms, which have been derived either heuristically, or/and have emerged from intuitive analogies in physics, statistical techniques and information theoretic perspectives can be recovered as special cases of the proposed analysis and we streamline their relationships.

Keywords: clustering, k-means algorithm, convex analysis, support and asymptotic functions, distance-like functions, Bregman and Csiszar divergences, nonlinear means, nonsmooth optimization, smoothing algorithms, fixed point methods, deterministic annealing, expectation maximization, information theory and entropy methods

1. Introduction

The clustering problem is to partition a given data set into similar subsets or clusters, so that objects/points in a cluster are more similar to each other than to points in another cluster. This is one of the fundamental problem in unsupervised machine learning, and it arises in a wide scope of applications such as astrophysics, medicine, information retrieval, and data mining to name just a few. A closely related problem is the one of vector quantization, mainly developed in the field of communication/information theory. The interdisciplinary nature of clustering is evident through its vast literature which includes many clustering problem formulations, and even more algorithms. For a survey on clustering approaches, see Jain et al. (1999), and for vector quantization the review paper of Gray and Neuhoff (1998), and the references therein.

Basically, the two main approaches to clustering are hierarchical clustering and partitioning clustering. In this paper we focus on partitioning clustering, where the number of clusters is known in advance. Most well known partitioning clustering methods iteratively update the so-called centroids or cluster centers, and as such, are often referred as *center-based* clustering methods. These clustering methods are usually classified as: *hard* and *soft*, depending on the way data points are assigned to clusters. Hard clustering produces a disjoint partition of the data, that is, a binary strat-

egy is used so that each data point belongs exactly to one of the partitions. In that class, one of the most celebrated and widely used hard clustering algorithm is the classical k-means algorithm, which basic ingredients can already be traced back to an earlier work of Steinhaus (1956). The algorithm has been derived in the statistical literature by Forgy (1965) and MacQueen (1967). Another well known variant includes the Linear Vector Quantization (LVQ) algorithm of Lloyd (1982). Soft clustering is a *relaxation* of the binary strategy used in hard clustering, and allows for overlapping of the partitions. In that class, there exists a plethora of algorithms which will be shortly outlined below.

1.1 Motivation

Finding a true optimal partition of a fixed number of sets in a given n-dimensional space is known to be an NP-hard problem (Garey and Johnson, 1979), and thus is in general out of reach. As a result, the current literature abounds in approximation algorithms for the partitioning clustering problem.

This paper is a theoretical study of center-based clustering methods from a continuous optimization theory viewpoint. One of the most basic and well-known formulation of the partitioning clustering problem consists of minimizing the sum of a finite collection of "min"-functions, which is a nonsmooth and nonconvex optimization problem. Building on convex and nonlinear analysis techniques, we present a generic way to design, analyze and extend center-based clustering algorithms by replacing the nonsmooth clustering problem with a smooth optimization problem. The general framework and formalism has the advantage to provide a rigorous analysis for center-based clustering algorithms, as well as to reveal the underlying difficulties, and it paves the way for building new schemes. Moreover, as we shall see below, this provides a closure and unification to many disparate approaches that have led to center-based algorithms which have been widely used in applications, for example, fuzzy k-means, deterministic annealing, clustering with general divergences, and which are shown to be special cases of the proposed framework.

We give now a brief summary of some of the relevant works that have motivated the present study. All current center-based algorithms seek to minimize a particular objective function with an attempt to improve upon the standard k-means algorithm. The latter is very attractive due to its computational simplicity. It begins with an initial guess of the centers (usually at random), and consists of two main steps: the first is the cluster assignment, which assigns each data point to the closest cluster center; the second step re-compute the cluster centers as a weighted arithmetic mean of all points assigned to each cluster centers. The algorithm stops when there is no more changes in the partitions, that is, cluster centers no longer move. The simplicity of the k-means has also a price, and it is well known that besides the fact that it does not find an optimal partition, a fact which is surely not too surprising given the alluded difficulty of the partitioning problem, the algorithm also suffers from several drawbacks, for example, it is highly sensitive to initial choice of cluster centers, it might produce empty clusters, it does not provide flexibility to model and measure the influence of specific data types arising in different applications etc...All these difficulties have motivated the search for "better quality" clustering algorithms that could possibly cope with the listed drawbacks.

To achieve this goal, various approaches and techniques have been advocated and a large body of literature has emerged, in particular in the statistic, computer science and engineering research related disciplines, while in the continuous optimization community, clustering analysis has received limited attention, with some of the early studies including for example Rao (1971) and Gordon and Henderson (1977). One direction of research in optimization has been to consider heuristic exten-

sions of the k-means algorithm for the classical minimum sum of squares clustering problem, and which have been shown to be quite successful in practice. For instance, the work of Hansen and Mladenovic (2001) suggests a new descent local search heuristic and reports experimental results showing that it outperforms other known local search methods, quite substantially. A more recent account of optimization approaches to clustering can be found in the excellent and extensive survey paper of Bagirov et al. (2003), and references therein, which among various nonsmooth and global optimization methods, includes the application of the discrete gradient method. The latter is a technique for local optimization that can escape from stationary points which are not local minimizers (see also for instance the more recent work Bagirov and Ugon, 2005).

Another direction of research has been to consider objective functions involving proximity measures other than the usual squared Euclidean distance that is used in k-means. Indeed, different data types arising in many applications have justified the search for more meaningful proximity measures that can better model a given data set. In hard clustering algorithms, several researchers have considered such an approach to extend the k-means algorithm. To mention just a few of the recent studies in that direction (see, for example, Modha and Spanger, 2003; Banerjee et al., 2005; Teboulle et al., 2006), and the extensive relevant bibliography given in these papers.

More intensive research activities have focused on developing soft clustering algorithms. In that context, the literature on iterative methods for clustering is wide, and includes a large number of works and approaches that have been motivated by different fields of applications, and often use different tools and terminology. Many soft clustering algorithms have emerged and have been developed from heuristic considerations, axiomatic approaches, or/and are based on statistical techniques, physical analogies, and information theoretic perspectives. For example, well known soft clustering methods include the Fuzzy k-means (FKM), (see, for example, Bezdek, 1981), the Expectation Maximization algorithm (EM), (see, for example, Duda et al., 2001), Maximum Entropy Clustering Algorithms (MECA), (see, for example, Rose, 1998), the Deterministic Annealing (DA) (Rose et al., 1990), and the closely related similar technique with the same name proposed by Ueda and Nakano (1998). The latter technique is very useful in practice, see for instance its application to documents clustering in the recent work of Elkan (2006). More recent and other soft clustering methods include for example the work of Zhang et al. (1999) which proposes a clustering algorithm called *k*-harmonic means, and which relies on optimizing an objective function defined as the harmonic mean of the squared Euclidean distance from each data points to all centers. An extension of this method has also been further developed in Hamerly and Elkan (2002). All the cited studies have also reported many experimental results to demonstrate the potential benefits of modifying, and extending the aforementioned classes of center-based algorithms, and to show their promise, and/or advantage over the standard k-means, as well as their relevance in several practical and real-life application contexts.

1.2 Main Contributions

Motivated by all these works, this paper has three main goals: (a) to reveal the underlying mathematical tools that explain and enables us to design, analyze and extend center-based clustering algorithms, (b) to develop a generic iterative scheme that keeps the simplicity of the k-means, allows for a rigorous analysis of center-based clustering methods, and reveals their potential advantages and limitations; (c) to provide a closure and unification to a long list of disparate motivations and ap-

proaches that have been proposed for center-based clustering methods, and which as alluded above, have been widely used in practice.

To achieve these goals, in this paper we develop a systematic and theoretical study of center-based clustering methods from a continuous optimization theory perspective, which leads to a common language, and a unifying framework for building and analyzing a broad class of hard and soft center-based clustering algorithms. This provides the basis for significantly extend the scope of center-based partitioning clustering algorithms, to bridge the gap between previous works that were relying on heuristics and experimental methods, and to bring new insights into their potential. The proposed framework also shows that all current center-based algorithms are capable of handling only the nonsmoothness difficulty inherent in the clustering problem, but do not provide a cure to the nonconvexity difficulty which remains a challenging one.

A brief summary of our results and the organization of the paper is a follows. In Section 2 we begin with the standard optimization formulation of the partitioning clustering problem that focuses on the nonsmooth nonconvex optimization formulation in a finite dimensional Euclidean space. An obvious key factor in modelling a clustering problem is the choice of the distance measure involved, and which depends on the nature of the problem's data. To handle this situation, we will consider a broad concept of distance-like functions (Auslender and Teboulle, 2006), which extends (and includes) the usual quadratic Euclidean distance setting. We outline their basic properties, and give two generic examples which include the useful and important Bregman and Csiszar based divergences. In Section 3, we furnish the necessary background and known results from convex analysis, with a particular focus on two central mathematical objects: *support and asymptotic functions*, which will play a primary role in the forthcoming analysis of clustering problems. The connection between support and asymptotic functions has been used in past optimization studies to develop a general approach to smoothing nonsmooth optimization problems (Ben-Tal and Teboulle, 1989; Auslender, 1999; Auslender and Teboulle, 2003). Building on these ideas, in Section 4 we first describe an *exact smoothing* mechanism, which provides a very simple way to design and analyze a wide class of center-based hard clustering algorithms. In turns, the support function formulation of the clustering problem also provides the starting point for developing a new and general *approximate smoothing* approach to clustering problems. This is achieved by combining the notion of asymptotic functions with another fundamental mathematical object: the concept of *nonlinear means* of Hardy et al. (1934). We study the relationships and properties of both concepts, and demonstrate that their combination provides a natural and useful framework in the context of clustering. This enables us to arrive at a unified approach for the formulation and rigorous analysis of soft center-based clustering methods. Building on these results, and thanks to the specific form of the objective function one has to minimize in the resulting smooth reformulation of the clustering problem, in Section 5 we introduce a simple generic fixed point algorithm, analyze its properties and establish its convergence to a stationary point. The generic algorithm is computationally as simple as the k-means method, and thus appears suitable for practical purposes. Finally, in Section 6, we show that all the aforementioned hard and soft center-based clustering methods, which have been proposed in the literature from different motivations and approaches can be realized as special cases of the proposed analysis, and we streamline their relationships.

Notations. We use boldface notation for a finite collection of k vectors in a given finite dimensional Euclidean space $\mathbb{R}^n$, that is, $\mathbb{R}^{kn} \ni \mathbf{x} := (x^1, \ldots, x^k)$, with $\mathbb{R}^n \ni x^l := (x^l_1, \ldots x^l_n)$, $l = 1, \ldots, k$. The inner product for two vectors u, v in $\mathbb{R}^n$ is denoted by $u^T v \equiv \langle u, v \rangle$. For an open set $S \subset \mathbb{R}^n$, the notation $\bar{S}$ stands for the topological closure of S, and we also use the notation $\mathbf{S}, (\bar{\mathbf{S}})$ to denote the k-

fold Cartesian product $S \times \ldots \times S$, $(\bar{S} \times \ldots \times \bar{S})$. For any nonempty convex set $C \subset \mathbb{R}^n$, δ_C denotes the indicator function of C, $\operatorname{int} C$ ($\operatorname{ri} S$) its interior (relative interior). The convex hull of a set $\mathcal{A}$ is denoted by $\operatorname{conv} \mathcal{A}$. The set of vectors in $\mathbb{R}^n$ with nonnegative (positive, negative) components is denoted by $\mathbb{R}^n_+$ ($\mathbb{R}^n_{++}, \mathbb{R}^n_{--}$). For any function g defined on $\mathbb{R}^n$, we also use the notation $g(z) \equiv g(z_1, \ldots, z_n)$.

2. The Clustering Problem with General Distance-Like Functions

In this paper, we focus on the basic nonsmooth nonconvex optimization formulation of the partitioning clustering problem which uses a broad class of distance-like functions (proximity measures) that replaces (and includes) the usual squared Euclidean norm. As we shall see later on in Sections 4 and 5, this formulation provides a source of explanations to design and analyze center-based clustering iterative algorithms.

2.1 Nonsmooth Optimization Formulation of Clustering

Let $\mathcal{A} = \{a^1, \ldots, a^m\}$ be a given set of points in the subset S of a finite dimensional Euclidean space $\mathbb{R}^n$, and let $1 < k < m$ be a fixed given number of clusters. The clustering problem consists of partitioning the data $\mathcal{A}$ into k subsets $\{A_1, \ldots, A_k\}$, called clusters. The common approach to formulate the clustering problem is as follows. For each $l = 1, \ldots, k$, the cluster A_l is represented by its center (centroid) x^l, and we want to determine k cluster centers $\{x^1, \ldots x^k\}$ such that the sum of proximity measures from each point a^i to a nearest cluster center x^l is minimized. Suppose for the moment that we are given a proximity measure $d(\cdot,\cdot)$ that satisfies the following basic properties:

$$d(u,v) \geq 0, \ \forall (u,v) \in S \ \text{ and } \ d(u,v) = 0 \iff u = v.$$

We will call $d(\cdot,\cdot)$ a *distance-like* function, since we are not necessarily asking for $d(\cdot,\cdot)$ to be symmetric or to satisfy the triangle inequality, (a more precise definition for d is given later in § 2.2). Then, the distance from each $a^i \in \mathcal{A}$ to the closest cluster center is:

$$D(\mathbf{x}, a^i) = \min_{1 \leq l \leq k} d(x^l, a^i).$$

The clustering problem seeks to minimize the average over the entire data set $\mathcal{A}$. Thus, assigning a positive weight ν_i to each $D(\mathbf{x}, a^i)$, such that $\sum_{i=1}^m \nu_i = 1$, (for example, each ν_i can be used to model the relative importance of each point a^i), the clustering problem consists of finding the set of k centers $\{x^1, \ldots, x^k\}$ that solves

$$(NS) \quad \min_{x^1,\ldots,x^k \in S} F(x^1, \ldots, x^k) := \sum_{i=1}^m \nu_i \min_{1 \leq l \leq k} d(x^l, a^i).$$

When the distance is the square of the Euclidean norm, that is, $d(u,v) = \|u - v\|^2$, and the average is the special uniform case, that is, $\nu_i = m^{-1}$ for all i, this formulation can be traced back to the work of Lloyd (1982) when related to vector quantization algorithms (Linde et al., 1980).

For $k = 1$, problem (NS) can either be analytically solved (e.g., when d is the quadratic norm) or is just an easy convex problem, when the proximity measure d is given convex. Likewise, for $k = m$, the (NS) problem is trivial, (i.e., each point is assigned to each cluster), while for $k > m$ the problem is infeasible. Thus, the interesting situation is when $1 < k < m$, for which the problem (NS) is nonsmooth and nonconvex. Furthermore, the number of variables is $n \times k$, and since n is typically

very large, even with a moderate number of clusters k, the clustering problem yields a very large scale optimization problem to be solved. Therefore, the clustering problem in this most elementary formulation (i.e., where k, the number of clusters is known) combines three of the most difficult and challenging characteristics one encounters in an optimization problem: *nonsmooth, nonconvex and large scale*.

2.2 Clustering with General Distance-Like Functions

We introduce a broad class of distance-like functions d that is used to formulate the clustering problem in its general form, and we provide two generic families of such distance-like functions for clustering. To measure the proximity of two given vectors in some subset S of $\mathbb{R}^n$, we consider the following concept of distance-like functions, as defined in the recent work of Auslender and Teboulle (2006). The later concept has been widely used in several optimization algorithms, and for more details and results we refer the reader to that work and the references therein.

First, we need to recall some basic notations and definitions in convex analysis, (see, for example, Rockafellar, 1970). For a convex function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, its effective domain is defined by $\operatorname{dom} g = \{u \mid g(u) < +\infty\}$, and the function is called proper if $\operatorname{dom} g \neq \emptyset$, and $g(u) > -\infty$, $\forall u \in \operatorname{dom} g$. For a proper, convex and lower semicontinuous function (lsc) g, (that is to say, that the epigraph of g is a closed set in $\mathbb{R}^n \times \mathbb{R}$) its subdifferential at x is defined by $\partial g(x) = \{\gamma \in \mathbb{R}^n \mid g(z) \geq g(x) + \langle \gamma, z - x \rangle,\ \forall z \in \mathbb{R}^n\}$ and we set $\operatorname{dom} \partial g = \{x \in \mathbb{R}^n \mid \partial g(x) \neq \emptyset\}$. When g is differentiable, the subdifferential set reduces to the singleton $\nabla g(x)$, the gradient of g at x. Equipped with these notations, we define now the notion of distance-like function.

Definition 1 *A function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+ \cup \{+\infty\}$ is called a distance-like function with respect to an open nonempty convex set $S \subset \mathbb{R}^n$ if for each $y \in S$ it satisfies the following properties:*
(d_1) *$d(\cdot, y)$ is proper, lsc, convex, and C^2 on S with a positive definite symmetric matrix Hessian*[1] *denoted by $\nabla^2 d(\cdot, y)$;*
(d_2) $\operatorname{dom} d(\cdot, y) \subset \overline{S}$, *and* $\operatorname{dom} \partial_1 d(\cdot, y) = S$ *where $\partial_1 d(\cdot, y)$ denotes the subgradient map of the function $d(\cdot, y)$ with respect to the first variable;*
(d_3) *$d(\cdot, y)$ is level bounded on $\mathbb{R}^n$, that is,* $\lim_{\|u\| \to \infty} d(u, y) = +\infty$*;*
(d_4) $d(y, y) = 0$, *(which also implies that $\nabla_1 d(y, y) = 0$, where $\nabla_1 d(\cdot, y)$, is the gradient with respect to the first variable.)*

We denote by $\mathcal{D}(S)$ the family of functions d satisfying the premises of Definition 1. It should be emphasized that the triangle inequality, and symmetry is not required in the definition of d, hence the use of the "distance-like" terminology.

To understand the motivation behind the technical assumptions of Definition 1, and the forthcoming mathematical developments given in Sections 4 and 5, let us already announce as an appetizer, that the main (essentially the only one) computational step that will be needed in the generic algorithm we developed in this paper for solving the clustering problem (NS), reduces to the solution of an optimization problem which admits the simple form:

$$\min\{\sum_{i=1}^{m} \gamma_i d(x, a^i) \mid x \in \overline{S}\}, \tag{1}$$

1. We recall that the positive definiteness of the Hessian matrix implies that $d(\cdot, y)$ is strictly convex.

with some given $\gamma_i > 0,\ i = 1,\ldots,m$. The properties (d_1), and (d_3) are needed to guarantee the existence of a unique global minimizer x^* to the optimization problem of the form (1), while property (d_2) plays the role of a "barrier", and enforces the optimal solution x^* to be in the open set S, (see Section 5.1, and Lemma 14 for details). Property (d_4) is just for normalization of d.

The class $\mathcal{D}(S)$ has been chosen in our setup, since it is broad and includes two useful generic families that produces a wide variety of distance-like functions which are discussed next. However, it should be noticed that as long as we can guarantee the existence of a unique global minimizer to problem (1), the use of other classes of distance-like functions is possible; this situation will be further illustrated below in Example 3.

2.3 Generic Families of Distance-Like Functions

As special cases of the class $\mathcal{D}(S)$, we briefly recall two particularly useful generic families that produce a wide variety of distance-like functions which have been already shown to be relevant for clustering, see for instance the recent work of Teboulle et al. (2006), and references therein.

• **Bregman Based Distances Type**[2] Originally proposed by Bregman (1967), this class of distances have been widely extended and analyzed in several optimization contexts and methods by Censor and his co-authors, (see, for example, Censor and Lent, 1981), and the more recent comprehensive monograph of Censor and Zenios (1997).

Let $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper, lsc, strictly convex function, with $\operatorname{dom}\psi \subset \overline{S}$, and such that ψ is continuously differentiable on $S := \operatorname{int}(\operatorname{dom}\psi)$. The Bregman based distance associated with ψ is the function $d_\psi : \mathbb{R}^n \times \mathbb{R}^n \to [0,+\infty]$ defined by

$$d_\psi(x,y) := \begin{cases} \psi(x) - \psi(y) - \langle x-y, \nabla\psi(y)\rangle & \text{if } y \in S \\ +\infty & \text{otherwise.} \end{cases}$$

Most interesting and useful Bregman based distances functions are separable, and can be generated with the choice

$$\psi(x) = \sum_{j=1}^{n} \omega(x_j), \tag{2}$$

with ω being some appropriate *scalar* twice differentiable convex function with $\omega''(\cdot) > 0$ on $\operatorname{int}(\operatorname{dom}\omega)$.

Example 1 Typical choices for $\omega(\cdot)$ include $\omega(t) = t^2, t\log t, -\log t, -(1-t^2)^{1/2}$ with domain $\operatorname{dom}\omega = \mathbb{R}, \mathbb{R}_+, \mathbb{R}_{++}, [-1,1]$, respectively. Substituting these functions in d_ψ, with ψ as defined in (2), the first three choices yields respectively the squared Euclidean distance, the Kullback-Liebler based relative entropy and the Burg based relative entropy (also called the Itakura Saito distance). For more examples and details in the context of clustering, see Teboulle et al. (2006), and references therein.

• **φ-Divergence Based Distances Type**. Originally introduced by Csiszar (1967) in the context of information theory to provide a notion generalizing divergence between probability measures, (e.g., the Kullback-Liebler relative entropy). It has been considered for optimization and algorithmic purposes by Teboulle (1987, 1992, 1997).

2. Another terminology is Bregman divergence. In this paper we will freely use/exchange both terminologies.

Let $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a proper, lsc, convex function such that $\operatorname{dom}\varphi \subset \mathbb{R}_+$, $\operatorname{dom}\partial\varphi = \mathbb{R}_{++}$, and such that φ is C^2, strictly convex, and nonnegative on $\mathbb{R}_{++}$, with $\varphi(1) = \varphi'(1) = 0$. Then, the φ-divergence based distance is defined by,

$$d_\varphi(x,y) := \sum_{j=1}^{n} y_j \varphi\left(\frac{x_j}{y_j}\right),$$

and which by its definition is already separable.

Example 2 Typical examples for the φ-divergence, d_φ with $S = \mathbb{R}^n_{++}$, include

$$\begin{aligned}
\varphi_1(t) &= -\log t + t - 1, \text{ (Arithmetic Mean)},\\
\varphi_2(t) &= t\log t - t + 1, \text{ (Geometric Mean)},\\
\varphi_3(t) &= 2(\sqrt{t}-1)^2, \text{ (Root Mean Square)}.
\end{aligned}$$

The names in parenthesis indicate the type of *means* that results by solving problem of the form (1) in these cases, (Teboulle et al., 2006).

For the above choices, and other appropriate ψ or/and φ, and S, both functionals d_ψ, d_φ can be shown to be in the class $\mathcal{D}(S)$ as defined in Definition 1.[3] In particular, it can be easily seen that $d \equiv d_\psi$ or d_φ enjoy the required basic property of a distance-like function, namely one can verify that:

$$\forall (u,v) \in S \times S \quad d(u,v) \geq 0 \text{ and } d(u,v) = 0 \text{ iff } u = v.$$

Moreover, note that the distance-like function d_φ is jointly convex on $S \times S$, and hence in particular, for any $u \in S$, the function $v \to d_\varphi(u,v)$ is convex in its *second argument*, a property which is not shared in general by the Bregman-based distance.

As previously noticed, symmetry is not requested in the definition of d, and as a result, as long as we can guarantee the existence of a global minimizer for problem of the form (1), some of the properties requested in Definition 1 with respect to the first argument, can be exchanged with respect to the second argument, or even relaxed. This situation is further exemplified now.

Example 3 In a very recent paper, Banerjee et al. (2005) have proposed to use Bregman distance for clustering by considering it as function with respect to its second argument (i.e., by changing the order of the variables), and have derived an interesting and somewhat surprising result. More precisely, consider problem (1) with

$$d(x,a) := d_\psi(a,x) := \psi(a) - \psi(x) - \langle a - x, \nabla\psi(x)\rangle, \ \forall (a,x) \in \bar{S} \times S,$$

where $S = \operatorname{int}(\operatorname{dom}\psi)$. In that case, in general, $x \to d_\psi(a,x)$ is not necessarily convex,[4] and hence the corresponding problem (1) is *nonconvex*. Even though problem (1) is nonconvex, it has been recently shown (Banerjee et al., 2005), that if a minimizer of (1) exists in S, then it is a *global*

3. Note that when no confusion occurs, and with some abuse of notations, throughout the paper, d_φ (d_ψ) always stands for the φ-divergence (Bregman divergence).
4. Exceptions include the well known special cases, when the Bregman divergence is the squared Euclidean distance and the relative entropy. Otherwise, the convexity of $x \to d_\psi(a,x)$ is not warranted without imposing further conditions on ψ, which unfortunately precludes the use of relevant or/and useful ψ for defining a Bregman divergence.

minimizer, and is always the (weighted) *arithmetic mean* of the data. This can be seen as follows. A very simple, but fundamental property satisfied by any Bregman distance, is the following three point identity revealed in Chen and Teboulle (1993), (see Lemma 3.1 there), and which naturally generalizes Pythagoras theorem. For any three points $u, v \in \mathrm{int}(\mathrm{dom}\,\psi)$ and $w \in \mathrm{dom}\,\psi$ the following identity holds true

$$d_\psi(w,v) - d_\psi(w,u) = d_\psi(u,v) + \langle \nabla\psi(u) - \nabla\psi(v), w-u\rangle. \tag{3}$$

Thanks to the identity (3), and the fact that $d_\psi(\cdot,\cdot) \geq 0$, one has for any $i = 1,\ldots,m$:

$$\begin{aligned} d_\psi(a^i,x) - d_\psi(a^i,z) &= d_\psi(z,x) + \langle \nabla\psi(z) - \nabla\psi(x), a^i - z\rangle, \\ &\geq \langle \nabla\psi(z) - \nabla\psi(x), a^i - z\rangle. \end{aligned}$$

Thus, multiplying the last inequality by $\gamma_i > 0$, and summing over $i = 1,\ldots,m$, it immediately follows that with $z := \sum_{i=1}^n \gamma_i a^i$, the right hand side of the inequality vanishes, and hence

$$\sum_{i=1}^m \gamma_i d_\psi(a^i,x) \geq \sum_{i=1}^m \gamma_i d_\psi(a^i,z),\ \forall x \in S,$$

showing that $z \in S$ is the global minimizer of problem (1).

Note that it is also possible to consider other classes of distance-like functions, which are not necessarily based on Bregman or/and φ divergences (as long as a unique global optimal solution of (1) is warranted). An interesting recent study can be found for example, in the work of Modha and Spanger (2003) which have considered convex-k-means clustering algorithms based on some other proximity measures that are convex in the second argument.

Finally, note also that one could easily enrich the model (NS) by considering for example a more general formulation that associates with each l a different distance $d_l \in \mathcal{D}(S_l)$ (and which can be useful in applications to accommodate different types of data, Modha and Spanger, 2003), so that the more general model would consist of solving

$$\min\left\{ \sum_{i=1}^m \nu_i \min_{1\leq l\leq k} d_l(x^l,a^i) \mid (x^1,\ldots,x^k) \in S_1 \times \cdots \times S_k \right\}.$$

The analysis and theoretical results that we developed below also hold for such more general formulations as well.

3. Convex Analysis Background: Support and Asymptotic Functions

The main approach in this paper is based on considering ways to replace the nonsmooth clustering problem (NS) via a *smooth* optimization problem, and to study and derive a corresponding generic algorithm solving the nonsmooth problem (NS) via its smoothed counterpart. To develop this approach, this section furnishes some of the key concepts and results of convex analysis that will be used throughout this paper. For further details and proofs of the material presented in this section (see for example Rockafellar 1970, Sections 12 and 13, and Auslender and Teboulle 2003, Chapter 2). Readers familiar with these concepts may skip directly to Section 4, perhaps after reading the examples of the present section which provide motivation to the forthcoming developments.

3.1 Support Functions

A fundamental concept for dealing with properties of closed convex sets is the concept of support function. It allows to transfer properties about sets via functions, and as a result turns out to play a central role in optimization problems and facilitate their analysis.

Definition 2 *For any set $C \subset \mathbb{R}^k$, the function $\sigma_C : \mathbb{R}^k \to [-\infty, +\infty]$ defined by*

$$\sigma_C(v) := \sup\{\langle u, v\rangle \mid u \in C\}, \tag{4}$$

is called the support function of C.

Geometrically, the support function of a set C describes the closed half-spaces which contain C, namely

$$C \subset \{u \mid \langle u, v\rangle \leq \alpha\} \iff \sigma_C(v) \leq \alpha.$$

Note that the supremum in (4) may be finite or infinite; attained on C or not. If $C = \emptyset$, we set $\sigma_C \equiv -\infty$, while if $C \neq \emptyset$, one has $\sigma_C > -\infty$ and $\sigma_C(0) = 0$.

For the forthcoming analysis, we consider the case when C is a closed convex set in $\mathbb{R}^k$. The support function can be computed for many interesting geometric convex sets C, (Rockafellar, 1970, Section 13, p. 113). Let us give here two particularly interesting examples.

Example 4 Let $C := \mathbf{B} = \{u \in \mathbb{R}^k \mid \|u\| \leq 1\}$ be the unit Euclidean ball. Then, applying Cauchy-Schwartz inequality, it is easy to verify that

$$\sigma_{\mathbf{B}}(v) = \|v\|,$$

that is, the Euclidean norm is the support function of the unit ball.

Example 5 Let C be the unit simplex in $\mathbb{R}^k$, that is,

$$C := \Delta = \{u \in \mathbb{R}^k \mid \sum_{j=1}^{k} u_j = 1, u_j \geq 0,\ j = 1, \ldots, k\}.$$

Then, a simple computation shows that

$$\sigma_\Delta(v) = \sup\{\langle u, v\rangle \mid u \in \Delta\} = \max_{1 \leq j \leq k} v_j,$$

the supremum being attained on the compact set Δ, at $\{u_l^* :\ l = 1, \ldots, k\}$ given by:

$$u_l^* = \begin{cases} 1 & \text{if } l = \operatorname*{argmax}_{1 \leq j \leq k} v_j \\ 0 & \text{otherwise.} \end{cases}$$

At this stage, and to motivate the reader for the forthcoming technical results and analysis, let us already emphasize that support functions are essentially "built-in" for most optimization problems. More precisely, most smooth and nonsmooth optimization problems can be modelled via the following generic abstract optimization model

$$\inf\{c_0(u) + \sigma_Y(c(u)) \mid c(u) \in \operatorname{dom} \sigma_Y\},$$

where $c(u) = (c_1(u), \ldots, c_m(u)) \in \mathbb{R}^m$, where all $\{c_i\}_{i=0}^m$ are real valued on $\mathbb{R}^k$, and $Y \subset \mathbb{R}^m$ is adequately defined. The re-formulation of optimization problems via their support functions provides an alternative way to view and tackle optimization problems by exploiting mathematical properties of support functions, and this is the line of analysis that will be used here for the clustering problem. The following examples illustrate the support function formulations of some generic classes of optimization problems. More details can be found in Auslender and Teboulle (2003, Chapter 2).

Example 6 (a) (Nonlinear programming). Consider the standard optimization problem

$$v_{NLP} := \inf\{c_0(u) \mid c_i(u) \leq 0,\ i = 1, \ldots, m,\ u \in \mathbb{R}^k\}.$$

Then, it is easy to see that with $Y = \mathbb{R}^m_+$, one has

$$v_{NLP} = \inf_{u \in \mathbb{R}^k} \{c_0(u) + \sup\{\langle y, c(u)\rangle \mid y \in \mathbb{R}^m_+\}\} = \inf_{u \in \mathbb{R}^k} \{c_0(u) + \sigma_{\mathbb{R}^m_+}(c(u))\}.$$

Note that the first equation above is nothing else but the Lagrangian representation of (NLP), with $y \in \mathbb{R}^m_+$ being the Lagrangian multiplier associated with the constraints.
(b) (l_p-norm optimization problems). Consider the problem

$$v_{LP} := \inf\{\|c(u)\|_p,\ u \in \mathbb{R}^k\},\ \ (p \geq 1),$$

where $\|z\|_p := (\sum_{i=1}^m |z_i|^p)^{1/p}$ is the usual l_p-norm of $z \in \mathbb{R}^m$. Let Y be the l_q-unit ball in $\mathbb{R}^m$, that is,

$$Y := \{y \in \mathbb{R}^m \ \|y\|_q \leq 1\},\ \text{with } p + q = pq.$$

Then, invoking Hölder inequality, it follows that $\sigma_Y(c(u)) = \|c(u)\|_p$, and hence,

$$v_{LP} = \inf\{\sigma_Y(c(u)) \mid u \in \mathbb{R}^k\}.$$

(c) (Finite minimax problems). Consider the finite minimax problem

$$v_{MM} := \inf_{u \in \mathbb{R}^k} \max_{1 \leq i \leq m} c_i(u).$$

Then, using Example 5 with $Y = \Delta$, the unit simplex in $\mathbb{R}^m$, one obtains

$$v_{MM} := \inf\{\sigma_Y(c(u)) \mid u \in \mathbb{R}^k\}.$$

The above examples have illustrated the primary role of support functions in formulating optimization problems. Our goal will be to exploit the special structure and further properties of support functions for deriving useful equivalent reformulations of the clustering problem. The next result recorded below in Theorem 4 reveals an important and relevant property of support functions.

First, let us recall that the support function of a given set C is intimately connected to the well-known indicator of the set C, through another fundamental operation in convex analysis, which is the conjugacy operation.

Definition 3 *For any function $g : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$, the convex conjugate of g is the function $g^* : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ defined by*

$$g^*(z) = \sup_{y \in \mathbb{R}^k} \{\langle y, z\rangle - g(y)\} = \sup_{y \in \operatorname{dom} g} \{\langle y, z\rangle - g(y)\}.$$

In addition, if g is proper, lower semicontinuous (lsc) and convex on $\mathbb{R}^k$, then so is its conjugate g^*, and $g^{**} = g$.

Now, consider the indicator function of C defined by

$$\delta_C(u) = \begin{cases} 0 & \text{if } u \in C \\ +\infty & \text{otherwise.} \end{cases}$$

Then, using the definition of the support function, one has

$$\sigma_C(v) = \sup_{u \in C} \langle u, v \rangle = \sup_{u \in \mathbb{R}^k} \{\langle u, v \rangle - \delta_C(u)\}.$$

Thus, we see that σ_C is nothing else but the conjugate of the indicator function δ_C, that is, $\sigma_C = \delta_C^*$. Moreover, if C is also closed convex, then

$$\delta_C^{**} = \delta_C = \sigma_C^*.$$

Therefore, the indicator function and the support function of a closed convex set are conjugate to each other. In fact, the support function provides a remarkable one-to-one correspondence between nonempty closed convex subsets of $\mathbb{R}^k$ and the class of positively homogeneous lsc proper convex functions through the conjugacy operation.[5] This result is formally cited below.

Theorem 4 *(Rockafellar, 1970, Theorem 13.2). The functions which are the support functions of non-empty convex sets are the lsc proper convex functions which are positively homogeneous.*

3.2 Asymptotic Functions

We now look at the relationship between the support function and another important mathematical object, called the asymptotic function. For that purpose, we first need to define the notion of asymptotic cone.[6]

The asymptotic cone of a nonempty convex set $C \subset \mathbb{R}^k$ is a convex cone containing the origin defined by,

$$C_\infty := \{v \in \mathbb{R}^k : v + C \subset C\}.$$

Geometrically, this means that the asymptotic convex cone C_∞ includes the origin and consists of all directions $v \in \mathbb{R}^k$, such that for each $u \in C$, the halfline $\{u + tv \mid t \geq 0\}$ is contained in C. This notion is useful for dealing with unbounded sets, namely when we are concerned in specifying directions in which a set is unbounded. For example, one can show that a nonempty closed convex subset C of $\mathbb{R}^k$ is bounded if and only if $C_\infty = \{0\}$. Here, we are interested in the behavior of convex functions "in the large", that is, in the way convex functions vary, as their argument move along halflines in $\mathbb{R}^k$. A convenient way to achieve this is through the notion of asymptotic function which is just the asymptotic cone of the epigraph of that function. Intuitively speaking, the result given below says that the asymptotic behavior of g along halflines depends only on the direction of the halfline, and not on its location.

5. Recall that a function p is positively homogeneous on $\mathbb{R}^k$ if $0 \in \operatorname{dom} p$ and $p(tu) = tp(u)$ for all $u \in \mathbb{R}^k$ and all $t > 0$.
6. In the convex setting (i.e., when working with convex sets and convex functions, the asymptotic cone (function) is often called recession cone (function). In fact for closed convex sets (functions), the two concepts coincide (Auslender and Teboulle, 2003).

Definition 5 *Let $g : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex and lower semicontinuous (lsc) function. The asymptotic function g_∞ of g is the function $g_\infty : \mathbb{R}^k \to \mathbb{R}$ defined by*

$$epi\ (g_\infty) = (epi\, g)_\infty,$$

where $epi\, g = \{(x,r) \in \mathbb{R}^k \times \mathbb{R} : g(x) \leq r\} \subset \mathbb{R}^{k+1}$ is the epigraph of g.

The next result shows that this definition makes sense, and collect some basic properties of the asymptotic function g_∞.

Proposition 6 *(Rockafellar, 1970, Theorem 8.5) Let $g : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$, be a proper, convex function. Then its asymptotic function g_∞ is a proper convex function on $\mathbb{R}^k$ that is positively homogeneous with $g_\infty(0) = 0$. For any $z \in \mathbb{R}^k$, one has*

$$g_\infty(z) = \sup\{g(u+z) - g(u) \mid u \in \mathrm{dom}\, g\}.$$

Furthermore, if g is also assumed lsc on $\mathbb{R}^k$, then g_∞ is also lsc, and for any $u \in \mathrm{dom}\, g$ and any $z \in \mathbb{R}^k$,

$$g_\infty(z) = \sup_{t>0} t^{-1}[g(u+tz) - g(u)] = \lim_{t\to+\infty} t^{-1}[g(u+tz) - g(u)].$$

In particular, one also has

$$g_\infty(z) = \lim_{s\to 0^+} \left\{ g_s(z) := sg(s^{-1}z) \right\},\ \forall\, \mathbb{R}^k \ni z \in \mathrm{dom}\, g. \tag{5}$$

The last property (5) is useful to compute asymptotic functions. To illustrate this, let us give a few interesting examples (see, for example Example 2.5.1, page 51 in Auslender and Teboulle, 2003).

Example 7 In the following three examples, it can be verified that $g(\cdot)$ is a proper lsc convex function on $\mathbb{R}^k$, and thus we can use (5) to compute $g_\infty(\cdot)$.
(a) Let $g(u) = \sqrt{1 + \|u\|^2}$. Then, using (5) one has $g_\infty(z) = \|z\|$.
(b) Let $g(u) = \sum_{j=1}^k e^{u_j}$. Then one obtains $g_\infty(z) = \delta_{\mathbb{R}^k_-}$, that is, the indicator of the negative orthant in $\mathbb{R}^k$.
(c) Let $g(u) = \log \sum_{j=1}^k e^{u_j}$. Then, $g_\infty(z) = \max_{1\leq j\leq k} z_j$.

To expand our ability of computing the asymptotic function g_∞ of a given function g, it turns out that it is often easier to work with the conjugate of g, which is always lsc, and therefore, by Proposition 6 so is its asymptotic function. Since this asymptotic function is lsc, proper, positively homogeneous and convex, Theorem 4 guarantees that it must be the support function of some nonempty closed convex set. The next convex analytic result shows that this set should be precisely the effective domain of the conjugate g^*.

Proposition 7 *(Rockafellar, 1970, Theorem 13.3) Let $g : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a proper convex and lsc function, and let $C = \mathrm{dom}\, g^*$ be the effective domain of the conjugate g^*. Then, $\sigma_C = g_\infty$.*

This last result connecting support and asymptotic functions, together with (5) in Proposition 6, provides the basis and motivation for developing a general approach to smoothing nonsmooth optimization problems. This approach was introduced by Ben-Tal and Teboulle (1989), and for more general results and details, the reader is refereed to Auslender (1999), and the recent monograph of Auslender and Teboulle (2003). Building on these ideas, we now develop smoothing approaches to the clustering problem.

4. Smoothing Methodologies for Clustering

We first describe a very simple *exact* smoothing mechanism which provides a novel and simple way to view and design all center-based *hard* clustering algorithms from an optimization perspective. In turns, the support function formulation also provides the starting point for developing a new and general smoothing approach for clustering problems, which is based on combining asymptotic functions, and the fundamental notion of *nonlinear means* of Hardy et al. (1934). The resulting smoothing approach lends to devise a simple generic algorithm, which computationally is as simple as the popular k-means scheme, (see Section 5), and encompasses and extend most well known *soft* clustering algorithms, see Section 6.

4.1 Exact Smoothing: The Support Function Approach and Hard Clustering

Given $\{a^1,\ldots,a^m\}$ in the subset $S\subset \mathbb{R}^n$, and a distance-like function $d\in\mathcal{D}(S)$, the general clustering problem as formulated in Section 2 through (NS) is to solve

$$\min_{x^1,\ldots,x^k\in S} F(x^1,\ldots,x^k)\equiv \min_{\mathbf{x}\in\mathbf{S}} F(\mathbf{x}) := \sum_{i=1}^m \nu_i \min_{1\le l\le k} d(x^l,a^i),$$

where we use the notation $\mathbf{x}=(x^1,\ldots,x^k)$, for the $k\times n$ dimensional vector $\mathbf{x}$, and $\mathbf{S}\subseteq\mathbb{R}^N$, with $N:=kn$, for the k-fold Cartesian product of S.

In the context of the clustering problem, we now briefly show that the support function allows to derive an equivalent smooth formulation of the clustering problem, and in fact provides the foundation to the design and analysis of hard clustering algorithms. Fix any $i\in\{1,\ldots,m\}$, and let

$$d^i(\mathbf{x}) := (d(x^1,a^i),\ldots,d(x^k,a^i))\in\mathbb{R}^k.$$

The nonsmooth term $\min_{1\le l\le k} d(x^l,a^i)$ can be replaced by a smooth one, by using the support function. Indeed, using Example 5, it follows that for any $i=1,\ldots,m$,

$$\min_{1\le l\le k} d(x^l,a^i) = -\sigma_{\Delta^i}(-d^i(\mathbf{x})) = \min\{\langle w^i,d^i(\mathbf{x})\rangle :\ w^i\in\Delta^i\}, \tag{6}$$

where Δ^i is the unit simplex in $\mathbb{R}^k$ given by

$$\Delta^i = \left\{ w^i\in\mathbb{R}^k \mid \sum_{l=1}^k w^i_l = 1, w^i_l\ge 0,\ l=1,\ldots,k\right\},$$

and where w^i_l is the "membership" variables associated to cluster A_l, which satisfies: $w^i_l=1$ if the point a^i is closest to cluster A_l, and $w^i_l=0$ otherwise. Thus, substituting (6) in (NS), it follows that the nonsmooth clustering problem (NS) is equivalent to the exact smooth problem:

$$(ES)\quad \min_{x^1,\ldots,x^k\in S}\ \min_{w^1,\ldots,w^m\in\mathbb{R}^k} \left\{ \sum_{i=1}^m \nu_i \sum_{l=1}^k w^i_l d(x^l,a^i) \mid w^i\in\Delta^i,\ i=1,\ldots,m\right\}.$$

The smooth formulation (ES) explains precisely the mechanism of all well known, old and more recent, hard center-based hard clustering algorithms. Indeed, applying the Gauss-Seidel (GS) minimization algorithm to problem (ES), (which is also often called alternative minimization or coordinate descent method, (see, for instance, Auslender, 1976; Bertsekas and Tsitsiklis, 1989; Bertsekas,

99), that is, at each iteration, first minimize with respect to w^i with x^l fixed, then minimize with respect to x^l with the membership variable fixed, yields a general hard clustering algorithm with distance-like functions, which for short is denoted (**HCD**). The algorithm **HCD** includes as special cases, not only the popular k-means algorithm, but also many others hard clustering methods mentioned in the introduction. In particular, it includes and extend the Bregman hard clustering algorithm recently derived by Banerjee et al. (2005, Algorithm 1, p. 1715), which was introduced and motivated from a completely different view point, relying on statistical and information theoretic arguments. To make the paper self-contained, the algorithm **HCD** has been discussed in some details in the appendix.

4.2 Approximate Smoothing via Asymptotic Functions and Soft Clustering

The support function approach which has provided an *exact* smoothed reformulation of the nonsmooth problem (NS) and the corresponding generic hard clustering method **HCD**, lends itself to another systematic way to obtain an *approximate smoothed* reformulation of the problem (NS), which in turn will provide the basis for producing a generic *soft* clustering algorithm.

We have seen in §4.1, that the nonsmooth clustering problem (NS) is equivalent to the exact smooth formulation (ES). Using (6), an equivalent representation of the clustering problem (ES) can also be written as

$$(NS) \quad \min_{x^1,\ldots,x^k \in S} \sum_{i=1}^{m} -\nu_i \sigma_{\Delta_i}(-d(x^1,a^i),\ldots,-d(x^k,a^i))$$

where $\Delta^i := \{w^i \in \mathbb{R}^k : e^T w^i = 1, w^i \geq 0\}$, $i = 1,\ldots,m$, and $e \equiv (1,\ldots 1) \in \mathbb{R}^k$.

Fix any $i \in \{1,\ldots,m\}$. Thanks to Proposition 7, we know that the support function of the set Δ^i corresponds to an asymptotic convex function, say $(g^i)_\infty(\cdot)$. From Proposition 6, this asymptotic function can be approximated (cf. (5)) via:

$$g^i_\infty(z) = \lim_{s \to 0^+} \left\{ g^i_s(z) := s g^i(s^{-1} z) \right\}, \ \forall\, \mathbb{R}^k \ni z \in \operatorname{dom} g^i, \ \forall i = 1,\ldots m,$$

where g^i is some given convex function, such that $\operatorname{dom}(g^i)^* = \Delta^i$. This naturally suggests to replace the support function $-\sigma_{\Delta_i}(\cdot)$ in (NS) by an approximate function $g^i_s(\cdot)$, and thus to consider for each $s > 0$, the following *approximate* problem for (NS):

$$(NS)_s \qquad \min_{x^1,\ldots,x^k \in S} \sum_{i=1}^{m} -\nu_i g^i_s(-d(x^1,a^i),\ldots,-d(x^k,a^i)). \tag{7}$$

Thus, with a function g^i_s smooth enough, this approach leads to a generic smoothed approximate reformulation of the nonsmooth problem (NS) which depends on the parameter $s > 0$ that plays the role of a *smoothing* parameter.

A key question is then to find appropriate candidates for the function g^i for the clustering problem. Answer to this question will be developed in the next two subsections. But first, let us illustrate the above approach with an important example.

Example 8 *(The Log-Sum Exponential Smoothing for Clustering)* The Log-Sum exponential function is an important and very well known operation which has been widely used in optimization

contexts, (Bertsekas, 1982; Ben-Tal and Teboulle, 1989). For the clustering problem, it will lead to a family of important methods.

Consider a slight variant, and more general form of the function considered in Example 7(c) given by

$$g(y) = \log \sum_{l=1}^{k} \pi_l e^{y_l}, \tag{8}$$

where π_l are some given weights, that is, $\pi_l \geq 0$, $\forall l = 1, \ldots, k$ and $\sum_{l=1}^{k} \pi_l = 1$. This function is convex on $\mathbb{R}^k$, and from Example 7(c) one obtains [7]

$$g_\infty(y) = \lim_{s \to 0} g_s(y) = \max_{1 \leq l \leq k} y_l.$$

Thus, using (8), the clustering problem (NS) can be approximated by solving the smoothed problem $(NS)_s$ which in this case reads:

$$(NS)_s \qquad \min_{x^1, \ldots, x^k \in S} F_s(\mathbf{x}) := -s \sum_{i=1}^{m} \nu_i \log \left(\sum_{l=1}^{k} \pi_l e^{-\frac{d(x^l, a^i)}{s}} \right).$$

When d is the squared Euclidean distance, (with $\nu_i = m^{-1}, \forall i$, $\pi_l = k^{-1}, \forall l$), the objective function just derived from the proposed smoothing optimization approach, is in fact exactly the objective function arising in some well known clustering methods, such as the so-called (EM)-algorithm for normal mixtures, (Duda et al., 2001), and the deterministic annealing (Rose et al., 1990), which were motivated by, and derived from, a statistical/probabilistic framework and statistical physics analogies. Further, when d is a Bregman function, as given in Example 3, the approximation model $(NS)_s$ yields precisely the Bregman soft clustering method recently derived by Banerjee et al. (2005) from an information theory view point. This will be further discussed in Section 6.

As we shall see next, another natural way to smooth the clustering problem, and which later on will reconcile with the asymptotic function approach, is by using the so-called concept of nonlinear means.

4.3 Approximate Smoothing via Nonlinear Means

The concept of nonlinear means defined below, was introduced in 1934 by Hardy et al. (1934, Chapter III) as a natural generalization of the well known power means, that is, the weighted l_p-norm of a positive vector z.

Definition 8 *Let $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a strictly increasing and convex function, and let h^{-1} be its inverse, which is thus strictly increasing and concave. The nonlinear mean of k real numbers $z_1, \ldots, z_k$ associated to h is defined by*

$$G_h(z) = h^{-1} \left(\sum_{l=1}^{k} \pi_l h(z_l) \right),$$

where π_l are weights which are arbitrary positive numbers whose sum is one, that is, $\pi \in \mathrm{ri}(\Delta)$.

7. In fact, one also has $\mathrm{dom}\, g^* = \Delta$, (see Lemma 18), and thus as promised by Proposition 7, $\sigma_\Delta = g_\infty$.

In the rest of this paper we use the notation Δ_+ for the relative interior of the simplex in $\mathbb{R}^k$, that is, $\Delta_+ := \{\pi \in \mathbb{R}^k \mid \sum_{l=1}^k \pi_l = 1, \pi > 0\}$. As in Hardy et al. (1934, Chapter III, p. 66) we adopt the convention that $h^{-1}(\infty) = b$, where $b := \sup\{t \mid h(t) < \infty\}$.

More details and many results on nonlinear means can be found in Hardy et al. (1934, Chapter III). At this juncture, it is interesting to note that Karayiannis (1999) has suggested an *axiomatic approach* to re-formulate clustering objective functions that essentially leads him to rediscover the notion of nonlinear means given in Hardy et al. (1934). This interesting axiomatic approach can thus be further viewed as an additional supportive argument to the general smoothing optimization approach we develop here.

The next simple result shows that the nonlinear mean does provide an approximation, more precisely a lower bound for the maximum function $\max\limits_{1\le l\le k} z_l$.

Lemma 9 *For each $z \in \mathbb{R}^k$ and any $\pi \in \Delta_+$ the following inequalities hold,*

$$\sum_{l=1}^k \pi_l z_l \le G_h(z) \le \max_{1\le l\le k} z_l. \tag{9}$$

Proof. By the convexity of the function h one has

$$h\left(\sum_{l=1}^k \pi_l z_l\right) \le \sum_{l=1}^k \pi_l h(z_l),$$

and since h^{-1} is increasing, this proves the left hand side inequality of (9). To prove the right hand side of the inequality, note that since $\sum\limits_{l=1}^k \pi_l z_l \le \max\limits_{1\le l\le k} z_l$, and h is increasing, then

$$\sum_{l=1}^k \pi_l h(z_l) \le \max_{1\le l\le k} h(z_l) = h\left(\max_{1\le l\le k} z_l\right),$$

and an application of h^{-1} to both sides of the inequality completes the proof. ■

Recall that our basic clustering problem (NS) consists of minimizing the objective function F which can be rewritten as:

$$F(\mathbf{x}) = \sum_{i=1}^m \nu_i \min_{1\le l\le k} d(x^l, a^i) = -\sum_{i=1}^m \nu_i \max_{1\le l\le k} \{-d(x^l, a^i)\}. \tag{10}$$

Since $d(\cdot,\cdot) \ge 0$, to approximate $\max_l -d(x^l, a^i)$ we need only to consider nonlinear means with domain containing the negative orthant $\mathbb{R}^k_-$.

Example 9 *(Nonlinear Means on $\mathbb{R}^k_-$.)* Consider the functions $h_i(t), i = 1,2,3$ given respectively by $-\log(-t), -t^{-1}, -\sqrt{-t}$ with domain $(-\infty, 0)$ for the first two, and $(-\infty, 0]$ for the last one. The corresponding nonlinear means G_{h_i} are then respectively the weighted geometric, harmonic, and square root mean of $z \in \mathbb{R}^k_{--}$, for the first two choices of h, and of $z \in \mathbb{R}^k_-$ for the last one, while if $z \notin \mathbb{R}^k_{--}$ ($z \notin \mathbb{R}^k_-$ for the last one), one has $G_h(z) = 0$.

In view of the upper bound of Lemma 9, for each i, we can consider approximating the quantity $\max_{1\le l\le k}(-d(x^l,a^i))$, by its nonlinear mean G_h, and hence the resulting objective F given in (10) by an approximate objective given by:

$$\hat{F}(\mathbf{x}) = -\sum_{i=1}^{m} \nu_i h^{-1}\left(\sum_{l=1}^{k} \pi_l h(-d(x^l,a^i))\right). \tag{11}$$

Let us illustrate this on two specific examples with some h as given in Example 9.

Example 10 *(Harmonic Mean Approximation)* Consider the function $h_2(t) = -t^{-1}$ from Example 9, with $\operatorname{dom} h_2 = \operatorname{int}\operatorname{dom} h_2 = (-\infty, 0)$, that yields the harmonic mean G_{h_2}. Then, using h_2 in (11), to approximate F given in (10), one has to consider minimizing the approximate objective:

$$\hat{F}(\mathbf{x}) = \sum_{i=1}^{m} \nu_i \left(\sum_{i=1}^{k} \frac{\pi_l}{d(x^l,a^i)}\right)^{-1}.$$

This recovers and extends the approximate objective $\hat{F}(\mathbf{x})$, (with $d(\cdot,\cdot)$ the squared Euclidean distance, $\pi_l = k^{-1}$, $\forall l$, $\nu_i = m^{-1}$, $\forall i$), which was recently suggested by Zhang et al. (1999) from heuristic and intuitive considerations, together with a corresponding *k-harmonic means* algorithm, and some interesting numerical results. More recently, Hamerly and Elkan (2002) have further studied new variants of the k-harmonic means algorithm, and have experimentally shown its superiority for finding clustering of high quality in low dimensions. However, no mathematical or/and convergence analysis of the proposed algorithms have been provided in these works. It will be shown later on, that the k-harmonic means algorithm can also be viewed as a particular realization of our generic algorithm for which our convergence result can be applied, see Section 5.

Example 11 *(Geometric Mean Approximation)* Take the function $h_1(t) = -\log(-t)$ given in Example 9 with $\operatorname{dom} h_1 = \operatorname{int}\operatorname{dom} h_1 = (-\infty, 0)$ that yields the geometric means G_{h_1}. Then, using (11), we then obtain as an approximation of F, the resulting approximate objective:

$$\hat{F}(\mathbf{x}) = \sum_{i=1}^{m} \nu_i \prod_{l=1}^{k} d(x^l,a^i)^{\pi_l}.$$

This example provides an apparently new approximate model for clustering, on which one can apply the generic scheme developed in Section 5.

Now, we return to the Log-Sum exponential function described in Example 8. With the choice $h(t) = e^t$, in Definition 8, the resulting nonlinear means G_h precisely recovers the convex Log-Sum exponential function,

$$G_h(z) = \log \sum_{l=1}^{k} \pi_l e^{z_l}.$$

Therefore, since in that case $G_h(\cdot)$ is convex, Proposition 6 can be applied, and one has:

$$\max_{1\le l\le k} z_l = \lim_{s\to 0} sG_h(s^{-1}z) = G_h^{\infty}(z),$$

where the later expression denotes the asymptotic function of the mean $G_h(\cdot)$.

Thus, this specific example shows that the objective function of the clustering problem (NS) (cf. (10)), can be approximated either by a nonlinear (smooth) mean, or, by the corresponding asymptotic function of $G_h(\cdot)$, provided the later can be well defined. This suggests an approach that would combine nonlinear means and asymptotic functions to provide a generic smoothing model which will approximate the original clustering problem (NS), and approach it in the limit, for a broad class of smooth approximations. This can be achieved, provided we can characterize the class of functions h for which $G_h(\cdot)$ remains convex. This is developed next.

4.4 Combining Asymptotic Functions and Convex Nonlinear Means

The mean $G_h(\cdot)$ being by definition the composition of a convex function with a concave one, is not necessarily convex. Thus, defining its corresponding asymptotic function as given in Proposition 6 is not warranted. Furthermore, it turns out that the convexity of $G_h(\cdot)$ plays a crucial role in the convergence proof of the forthcoming generic algorithm (see Section 5). Thus, it is important to characterize the convexity of the nonlinear mean $G_h(\cdot)$. By specializing a general result proven in Ben-Tal and Teboulle (1986, Theorem 2.1), we can identify a wide class of functions h for which $G_h(\cdot)$ is convex.

In the sequel, for convenience we will often use the notation $\Omega := \text{int}(\text{dom}\, h)$, and Ω^k to denote the k-fold Cartesian product of $\text{int}(\text{dom}\, h)$.

Lemma 10 *Let $h:\ \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be C^3 on Ω, strictly increasing and convex, and let $r(t) := -h''(t)/h'(t)$. Then $G_h(z)$ is convex on Ω^k if and only if $1/r(t)$ is convex on Ω.*

Proof. Define $\xi(z) := \sum_{l=1}^k \pi_l h(z_l)$, so that $G_h(z) = h^{-1}(\xi(z))$. Then, G_h is convex if and only if it satisfies the gradient inequality, that is, recalling that $(h^{-1})'(\cdot) > 0$, this is equivalent to say,

$$\frac{h^{-1}(\xi(y)) - h^{-1}(\xi(x))}{(h^{-1})'(\xi(x))} \geq \sum_{l=1}^k \pi_l (y_l - x_l) h'(x_l), \ \ \forall x, y \in \Omega^k. \tag{12}$$

To prove that (12) holds true, define,

$$H(s,t) := \frac{h^{-1}(s) - h^{-1}(t)}{(h^{-1})'(t)}.$$

Invoking Ben-Tal and Teboulle (1986, Lemma 1, p. 1449), one has H is concave in (s,t) if and only if $1/r(t)$ is convex. To complete the proof it thus remains to show that the concavity of H reduces to the validity of (12). Applying Jensen's Inequality for the concave function $H(\cdot,\cdot)$ at $s := \xi(y), t := \xi(x)$, it follows that the function H is concave if and only if,

$$H(\xi(y), \xi(x)) \geq \sum_{l=1}^k \pi_l H(h(y_l), h(x_l)) = \sum_{l=1}^k \pi_l \frac{(y_l - x_l)}{(h^{-1})'(h(x_l))}.$$

Therefore, using the relation $h'\left(h^{-1}(t)\right) = 1/(h^{-1})'(t)$ at $t := h(x_l)$, proves that the last inequality is exactly (12). ■

Define the class of functions,

$$\mathcal{H} := \left\{ h \in C^3(\Omega) \mid h' > 0,\ h'' > 0, \text{ and } \frac{1}{r} \text{ is convex} \right\}.$$

The class of functions $\mathcal{H}$ satisfying the condition $1/r(t)$ is convex is wide, (Ben-Tal and Teboulle, 1986). It includes in particular, all functions h such that $1/r(t)$ is a linear function, that is, $1/r(t) = at+b,\ a,b \in \mathbb{R}$.

Example 12 *(Convex Nonlinear Means)* It can be easily verified that the class of functions $h \in \mathcal{H}$ satisfying $1/r(t) = at+b$ for some $a,b \in \mathbb{R}$ includes in particular (with $a=0,\ b \neq 0$) the function $h(t) = e^{t/b}$, as well as, for example, the functions h with $\Omega \equiv \text{int}(\text{dom}\, h) = (-\infty, \delta),\ (\delta \geq 0))$ given by
(i) $h(t) = -(\delta - t)^p,\ p \in (0,1),\ (a = (p-1)^{-1},\ b = \delta(1-p)^{-1})$.
(ii) $h(t) = (\delta - t)^{-p},\ p \in (0,+\infty),\ (a = -(p+1)^{-1},\ b = \delta(p+1)^{-1})$.
(iii) $h(t) = -\log(\delta - t), (a = -1,\ b = \delta)$.

Clearly these examples for h include and extend all the previous choices $h_i(t),\ i = 1,2,3$ (cf. Example 9), and generate accordingly corresponding *convex* nonlinear means G_h.

Equipped with Lemma 10, the concept of *asymptotic nonlinear mean* associated to a given convex nonlinear mean G_h is now well defined through Proposition 6.

Definition 11 *Let $h \in \mathcal{H}$. For $z \in \mathbb{R}^k$ and $\pi \in \Delta_+$, the asymptotic nonlinear mean is defined by*

$$G_h^{\infty}(z) = \lim_{s \to 0^+} s h^{-1}\left(\sum_{l=1}^{k} \pi_l h\left(\frac{z_l}{s}\right)\right).$$

With a proof identical to that of Lemma 9 we immediately get the following result.

Lemma 12 *If $h \in \mathcal{H}$, then for each $z \in \mathbb{R}^k$ and any $\pi \in \Delta_+$ one has*

$$\sum_{l=1}^{k} \pi_l z_l \leq G_h^{\infty}(z) \leq \max_{1 \leq l \leq k} z_l.$$

This last result, together with Definition 11, and the results developed in Section 4.2 (cf. (7)), provide all the ingredients to combine nonlinear convex means as characterized in Lemma 10 with asymptotic functions, and to formulate a broad class of smooth approximations to the clustering problem (NS) as follows.

For any $h \in \mathcal{H}$, any fixed $s > 0$, and any $\pi \in \Delta_+$, we approximate the nonsmooth objective F of the original clustering problem (NS) by the smooth function:

$$F_s(x^1, \ldots, x^k) \equiv F_s(\mathbf{x}) = -s \sum_{i=1}^{m} \nu_i h^{-1}\left(\sum_{l=1}^{k} \pi_l h\left(\frac{-d(x^l, a^i)}{s}\right)\right).$$

In the rest of the paper, we focus on developing and analyzing a generic algorithm that minimizes the smoothed approximate nonconvex function $F_s(\cdot)$.

Remark 13 Lemma 9 and Lemma 12 show that the nonlinear mean and its asymptotic version always provides a lower bound for $\max_{1 \leq i \leq k} z_i$, and hence when applied to the function F_s it follows that for any $s > 0$,

$$\sum_{i=1}^{m} \nu_i \left\{ \min_{1 \leq l \leq k} d(x^l, a^i) - \sum_{l=1}^{k} \pi_l d(x^l, a^i) \right\} \leq F(\mathbf{x}) - F_s(\mathbf{x}) \leq 0. \tag{13}$$

The quality of the smooth approximation is somewhat hidden in the last inequality. Yet, it is important to note that the Log-Sum exponential mean generated via $h(t) = e^t$ appears to be a sort of *optimal mean* for approximating the finite $\max_{1\le l\le k} z_l$ function. Indeed, take for example $\nu_i = m^{-1}$, $\forall i$, and $\pi_l = k^{-1}$, $\forall k$. In that case,

$$F_s(\mathbf{x}) = s\log k - \frac{s}{m}\sum_{i=1}^{m} s\log\left(\sum_{l=1}^{k} e^{\frac{-d(x^l,a^i)}{s}}\right) := s\log k + L_s(\mathbf{x}).$$

Then, the right inequality (13) produces the well known approximation result for $L_s(\cdot)$ (Bertsekas, 1982; Ben-Tal and Teboulle, 1989):

$$0 \le F(\mathbf{x}) - L_s(\mathbf{x}) \le sm\log k,$$

showing that the Log-Sum exponential mean shares a unique type of uniform approximation, and for which $G_h^{\infty}(z) \equiv \max_{1\le i\le k} z_i$. More on this specific property of the Log-Sum exponential function will be discussed in Section 6.

5. The Smooth k-Means Algorithm: Properties and Convergence

Building on the previously developed results, in this section we present a simple generic center-based algorithm for soft clustering, that we call the Smooth k-means (**SKM**) algorithm, and we study its convergence properties.

5.1 Motivation

Given $d \in \mathcal{D}(S)$, $h \in \mathcal{H}$ and any $s > 0$, to solve the clustering problem we consider a solution method that solves the approximate smoothed minimization problem,

$$\inf\left\{F_s(\mathbf{x}) \mid \mathbf{x} \in \overline{\mathbf{S}}\right\}, \tag{14}$$

where

$$F_s(\mathbf{x}) = -s\sum_{i=1}^{m} \nu_i h^{-1}\left(\sum_{l=1}^{k} \pi_l h\left(-\frac{d(x^l,a^i)}{s}\right)\right). \tag{15}$$

This problem could be solved by some standard optimization algorithms, such as projected gradient/Newton type methods, Lagrangian multipliers, etc. (see, for example, Bertsekas, 99). However, given that clustering problems are usually very large scale, we are interested to devise a simple iterative scheme which does not require any sophisticated computations at each iteration, (e.g., Hessian computations, matrix inversions, or/and line search techniques), which are usually needed in the alluded standard optimization algorithms.

It turns out that the specific form of F_s lends itself to build a simple iterative scheme, by combining the smoothing approach with successive approximations. The idea of such combination is well known in the field of optimization, and can be traced back to the so-called Weiszfeld algorithm derived in 1937 for solving some basic location theory problems (Weiszfeld, 1937). The Weiszfeld algorithm has provided a fertile ground for many other algorithms and problems in a variety of research areas, (see, for example, Ben-Tal et al., 1991; Brimberg and Love, 1993, and many of the references cited therein).

To motivate the generic algorithm SKM described below in §5.2, let us consider for the moment the special case when $\mathbf{S} = \mathbb{R}^N$, with the distance like function between any two points $u, v \in \mathbb{R}^n$ being the usual squared Euclidean distance $d(u,v) = \|u-v\|^2$.

The necessary local optimality condition for solving problem (14) in that case consists of finding $\mathbf{x} \in \mathbb{R}^N$ satisfying

$$\nabla F_s(\mathbf{x}) = 0. \tag{16}$$

Recall that we use the notation $\mathbf{x} = (x^1, \ldots, x^k)$ with $x^l \in \mathbb{R}^n$, $l = 1, \ldots, k$, and $N = kn$. We denote by $\nabla_l F_s(\mathbf{x})$ the gradient of $F_s(\mathbf{x})$ with respect to $x^l \in \mathbb{R}^n$. To express (16) in a compact and informative way we also use the following notations. For any scalar $s > 0$, $i = 1, \ldots, m$ and $l = 1, \ldots, k$, let

$$\delta^i(x^l) \;:=\; -s^{-1} d(x^l, a^i), \text{ with } \delta^i(\mathbf{x}) = (\delta^i(x^1), \ldots \delta^i(x^k)) \in \Omega^k, \tag{17}$$

$$\rho^{il}(\mathbf{x}) \;:=\; \pi_l \frac{h'(\delta^i(x^l))}{h'(G_h(\delta^i(\mathbf{x})))}. \tag{18}$$

With (17), since $h'(\cdot) > 0$ and $\pi_l > 0, \forall l$, the functions $\rho^{il}(\cdot)$ are positive for every i, l. Now, using the definition of F_s given in (15), we obtain for each $l = 1, \ldots, k$

$$\nabla_l F_s(\mathbf{x}) = \pi_l \sum_{i=1}^m \nu_i \left(h^{-1}\right)' \left(\sum_{l=1}^k \pi_l h \left(-\frac{d(x^l, a^i}{s} \right) \right) \cdot h' \left(-\frac{d(x^l, a^i)}{s} \right) \nabla_1 d(x^l, a^i), \tag{19}$$

where $\nabla_1 d(x^l, a^i)$ is the gradient of d with respect to the first variable x^l. Using in (19) the relation

$$(h^{-1})'(t) = \frac{1}{h'\left(h^{-1}(t)\right)},$$

the definition of $G_h(\cdot)$, and (18), simple algebra shows that (16) reduces to

$$\nabla_l F_s(\mathbf{x}) = \sum_{i=1}^m \nu_i \rho^{il}(\mathbf{x}) \nabla_1 d(x^l, a^i) = 0, \; l = 1, \ldots, k. \tag{20}$$

Since for the moment, we assumed $d(u,v) = \|u-v\|^2$, then $\nabla_1 d(u,v) = 2(u-v)$, and (20) simplifies to

$$\sum_{i=1}^m \nu_i \rho^{il}(\mathbf{x})(x^l - a^i) = 0, \; l = 1, \ldots, k. \tag{21}$$

Defining for each $i = 1, \ldots, m$, and $l = 1, \ldots, k$,

$$\lambda^{il}(\mathbf{x}) := \nu_i \rho^{il}(\mathbf{x}) \cdot \left(\sum_{j=1}^m \nu_j \rho^{jl}(\mathbf{x}) \right)^{-1}, \tag{22}$$

one has $\lambda^{il}(\mathbf{x}) > 0$, and $\sum_{i=1}^m \lambda^{il}(\mathbf{x}) = 1$, and (21) reduces to

$$x^l = \sum_{i=1}^m \lambda^{il}(\mathbf{x}) a^i, \qquad l = 1, \ldots, k. \tag{23}$$

Formula (23) suggests that in order to find x^l, we can consider the following fixed point iteration: for $t = 0, 1, \ldots$,

$$x^l(t+1) = \sum_{i=1}^{m} \lambda^{il}(\mathbf{x}(t))\, a^i, \quad l = 1, \ldots, k. \tag{24}$$

The explicit formula for $\mathbf{x}(t+1)$ given in (24) has been achieved thanks to our ability to solve the equation (20) for $\mathbf{x}$, which in turns, follows from the *linearity* of the gradient map $\nabla_1 d(\cdot,\cdot)$, for the special case of the squared Euclidean distance. In fact, the fixed point iteration (24) obtained from solving (20) reads equivalently as:

$$x^l(t+1) = \operatorname*{argmin}_{x^l} \left\{ \sum_{i=1}^{m} \nu_i \rho^{il}(\mathbf{x}(t)) d(x^l, a^i) \right\}, \; l = 1, \ldots, k.$$

This provides the motivation for the extension given next, and which allows to handle the general case with distance-like functions.

5.2 The SKM Algorithm

Mimicking the approach just outlined in the special case of the squared Euclidean distance, this naturally suggests that for handling general distance-like function $d \in \mathcal{D}(S)$, one computes $\mathbf{x}(t+1) = (x^1(t+1), \ldots, x^k(t+1))$ by solving:

$$\mathbf{x}(t+1) = \operatorname*{argmin}_{\mathbf{x}} \left\{ \sum_{i=1}^{m} \sum_{l=1}^{k} \nu_i \rho^{il}(\mathbf{x}(t)) d(x^l, a^i) \right\} \equiv \operatorname*{argmin}_{\mathbf{x} \in \overline{\mathbf{S}}} A_s(\mathbf{x}, \mathbf{x}(t)),$$

where $A_s : \mathbb{R}^N \times \mathbf{S} \to \mathbb{R}_+ \cup \{+\infty\}$ is defined by

$$A_s(\mathbf{x}, \mathbf{u}) = \sum_{i=1}^{m} \sum_{l=1}^{k} \nu_i \rho^{il}(\mathbf{u}) d(x^l, a^i), \tag{25}$$

and with $\{\rho^{il}(\cdot)\}_{i,l} > 0$ as defined in (18)) for some given $h \in \mathcal{H}$, (A_s, depends on $s > 0$ through the definition of ρ^{il}).

This leads us to propose the following simple generic family of iterative algorithms for solving (14).

The SKM Algorithm. For given data points $\{a^1, \ldots, a^m\} \in S \subset \mathbb{R}^n$, pick a distance like function $d \in \mathcal{D}(S)$, a function $h \in \mathcal{H}$ and fix $s > 0$. Set $t = 0$, choose $\mathbf{x}(0) \in \mathbf{S}$ and generate iteratively the sequence $\{\mathbf{x}(t)\}_{t=0}^{\infty}$ by solving:

$$\mathbf{x}(t+1) = \operatorname{argmin} \left\{ A_s(\mathbf{x}, \mathbf{x}(t)) \mid \mathbf{x} \in \overline{\mathbf{S}} \right\},$$

until convergence.

The next result shows that the algorithm **SKM** is well defined.

Lemma 14 *Let $d \in \mathcal{D}(S)$, $h \in \mathcal{H}$ and $s > 0$. For any fixed $\mathbf{u} \in \mathbf{S}$, consider the convex optimization problem*

$$(P_u) \quad v(\mathbf{u}) := \inf\{A_s(\mathbf{x}, \mathbf{u}) \mid \mathbf{x} \in \overline{\mathbf{S}}\},$$

where $A_s(\mathbf{x},\mathbf{u})$ is defined in (25). Then, the following statements hold:
(i) The optimal set $X^*(\mathbf{u})$ of problem (P_u) is nonempty and compact.
(ii) There exists a unique minimizer $\mathbf{y}(\mathbf{u}) := (y^1(\mathbf{u}),\ldots,y^k(\mathbf{u})) \in \mathbf{S}$ solving (P_u) and satisfying

$$\sum_{i=1}^{m} \nu_i \rho^{il}(\mathbf{u}) \nabla_1 d(y^l(\mathbf{u}), a^i) = 0, \;\; l = 1,\ldots,k. \tag{26}$$

Proof. (i) For any fixed $\mathbf{u} \in \mathbf{S}$, and $s > 0$, let $\Phi_u(\mathbf{x}) := A_s(\mathbf{x},\mathbf{u}) + \sum_{l=1}^{k} \delta\left(x^l|\overline{S}\right)$, where $\delta(\cdot|\overline{S})$ denotes the indicator of $\overline{S}$. Then, by the definition of $A_s(\cdot,\mathbf{u})$ given in (25), and property (d2) in Definition 1, the minimization problem (P_u) can be written as

$$v(\mathbf{u}) := \inf\{\Phi_u(\mathbf{x}) : \; \mathbf{x} \in \mathbb{R}^N\}.$$

Since here $v(\mathbf{u})$ is finite, recalling the definition of the distance-like function d (cf. Definition 1), it follows by (d_3) that $\Phi_u(\cdot)$ is level bounded. Therefore with $\Phi_u(\cdot)$ being a proper, lsc, and convex function, it follows that the optimal set $X^*(\mathbf{u})$ of problem (P_u) is nonempty and compact, and hence the existence of a minimizer is warranted.
(ii) The minimizer is unique thanks to the strict convexity of $A_s(\cdot,\mathbf{u})$ (which is implied from (d_1) of Definition 1, recalling that $\nu_i > 0$ and $\rho^{il}(\cdot) > 0$ for every i,l). From the optimality conditions, for each $y(\mathbf{u}) \in X^*(\mathbf{u})$ we have $0 \in \partial\Phi_u(y(\mathbf{u}))$, where $\partial\Phi_u$ stands for the subdifferential of Φ_u. Then, applying Rockafellar (1970, Theorem 23.8), it follows that for each $l = 1,\ldots,k$ the optimality of $\mathbf{y}(\mathbf{u})$ yields

$$0 \in \sum_{i=1}^{m} \nu_i \partial_1 d(y^l(\mathbf{u}), a^i) \rho^{il}(\mathbf{u}) + N_{\overline{S}}(y^l(\mathbf{u})), \tag{27}$$

where $N_{\overline{S}}(y^l(\mathbf{u}))$ stands for the normal cone[8] to $\overline{S}$ at $y^l(\mathbf{u})$. Since by definition, $\nu_i > 0$, $\rho^{il}(\mathbf{u}) > 0$, $\forall i,l$, and since by (d_2) of Definition 1, for each $i \in [1,m]$, one has $\operatorname{dom}\partial_1 d(\cdot,a^i) = S$ a nonempty open convex set, it follows that $\mathbf{y}(\mathbf{u}) \in \mathbf{S}$, and $N_{\overline{S}}(y^l(\mathbf{u})) = \{0\}$, and hence (27) reduces to the desired equation (26). ■

The main computational step of the algorithm **SKM** consists of solving for $x^l(t+1)$ the equation

$$\sum_{i=1}^{m} \nu_i \rho^{il}(\mathbf{x}(t)) \nabla_1 d(x^l(t+1), a^i) = 0, \qquad l = 1,\ldots,k. \tag{28}$$

As already mentioned (cf. §2.2), the class of distance-like functions which are separable includes most interesting and useful examples based on Bregman divergences, while φ-divergences are by definition given separable. More generally, let us consider now what will be called the class of *separable* distance-like functions, with d defined by

$$d(x,y) = \sum_{j=1}^{n} \theta(x_j, y_j), \tag{29}$$

8. For a closed convex set $C \subset \mathbb{R}^n$, recall Rockafellar (1970) that the normal cone to C at $x \in C$ is defined by $N_C(x) = \partial\delta_C(x) = \{\nu \in \mathbb{R}^n \mid \langle \nu, z-x\rangle \le 0, \forall z \in C\}$.

where $\theta : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ satisfies the premises of Definition 1 (i.e., with $S := I$, and I being an open interval). For such a separable d the equation (28) reduces to solving for each $l = l, \ldots k$, the simple *scalar* equation in the variable $x_j^l(\cdot)$:

$$\sum_{i=1}^{m} \nu_i \rho^{il}(\mathbf{x}(t)) \theta'(x_j^l(t+1), a_j^i) = 0, \; j = 1, \ldots, n, \tag{30}$$

where $\theta'(\cdot, a_j^i)$ is the derivative with respect to the first argument. Both separable Bregman divergences, and φ-divergences are then recovered as special cases of (29) with

$$\theta(\alpha, \beta) := \psi(\alpha) - \psi(\beta) - (\alpha - \beta)\psi'(\beta) \;\text{ and }\; \theta(\alpha, \beta) := \beta\varphi(\frac{\alpha}{\beta}),$$

respectively, and with the appropriate scalar functions ψ, φ (cf. §2.2). Note that the equation (30) can be solved *analytically* for $x^l(t+1)$ in the case of Bregman divergences, as well as for the case of φ-divergences for many interesting choices of φ, see, for example, Example 2, and for more examples and details, the recent work (Teboulle et al., 2006), and references therein.

Thanks to the strict monotonicity of $\theta'(\cdot, a_j^i)$ in its first variable (inherited from Definition 1), we can establish for the class of separable distance-like functions defined in (29), the following property of **SKM**, which simple proof is left to the reader.

Proposition 15 *Let $d \in \mathcal{D}(S)$ be separable and let $\{\mathbf{x}(t)\}_{t=0}^{\infty}$ be the sequence generated by* **SKM**. *Then, for each $l = 1, \ldots, k$, the iterates $S \ni x^l(t+1)$, $t = 0, 1, \ldots$ that solves (30), satisfy:*

$$\min_{1 \le i \le m} a_j^i \le x_j^l(t+1) \le \max_{1 \le i \le m} a_j^i, \; j = 1, \ldots, n,$$

that is, the sequence $\{\mathbf{x}(t)\}_{t=0}^{\infty}$ lies in a bounded hypercube in **S**.

We end this section with two additional remarks on the computational aspects of **SKM**.
(a) The computational complexity of **SKM** is as simple as the standard k-means algorithm, which makes **SKM** viable for large scale clustering problems, and allows to significantly extend the scope of soft center-based iterative clustering methods.
(b) Although this paper is concerned with theoretical issues, it should also be noted that in a more practical implementation of **SKM**, one could also start at $t = 0$ with a fixed positive value for s_t and decrease iteratively the parameter s_t. For that purpose, various strategies from standard optimization techniques, such as for example within penalty/barrier methods can be considered, (see, for example, Bertsekas, 1982, 99). Yet, recall that any fixed $s > 0$ do provide an approximate solution as well, as explained in §4.3, see, for example, Example 10.

5.3 Convergence Analysis

We are now in the position to state and prove the main convergence result for **SKM**. Note that the key element in the proof strongly relies on the convexity result established in Lemma 10 for the nonlinear means $G_h(\cdot)$.

Theorem 16 *Let $\{\mathbf{x}(t)\}_{t=0}^{\infty}$ be the sequence generated by the* **SKM** *algorithm. Then,*
(i) $F_s(\mathbf{x}(t+1)) < F_s(\mathbf{x}(t))$, for all $\mathbf{x}(t+1) \neq \mathbf{x}(t)$.
(ii) Let $d \in \mathcal{D}(S)$ be separable. Then, the sequence $\{\mathbf{x}(t)\}_{t=0}^{\infty}$ is bounded and each limit point $\mathbf{x} \in \mathbf{S}$ of this sequence is a stationary point for F_s.

Proof (i) By definition of the algorithm **SKM** and Lemma 14, $\mathbf{x}(t+1) \in \mathbf{S}$ is the unique minimizer solving $\mathbf{x}(t+1) = \underset{\mathbf{x}}{\operatorname{argmin}} A_s(\mathbf{x}, \mathbf{x}(t))$, and hence

$$A_s(\mathbf{x}(t+1), \mathbf{x}(t)) < A_s(\mathbf{x}(t), \mathbf{x}(t)), \ \forall\, \mathbf{x}(t+1) \neq \mathbf{x}(t). \tag{31}$$

Let

$$\begin{aligned} v_t^{il} &:= -s^{-1} d(x^l(t), a^i) \in \Omega, \ i = 1, \ldots, m, \ l = 1, \ldots, k, \\ \text{and } v_t^i &:= (v_t^{i1}, \ldots, v_t^{ik}) \in \Omega^k, \ i = 1, \ldots, m. \end{aligned} \tag{32}$$

Then,

$$F_s(\mathbf{x}(t)) = -s \sum_{i=1}^m \nu_i h^{-1} \left(\sum_{l=1}^k \pi_l h(v_t^{il}) \right) = -s \sum_{i=1}^m \nu_i G_h(v_t^i). \tag{33}$$

Since $h \in \mathcal{H}$, $s > 0$ and $\nu_i > 0$, then by Lemma 10, $G_h(\cdot)$ is convex on Ω^k. Therefore, applying the gradient inequality to the convex function $G_h(\cdot)$ one has:

$$G_h(z) - G_h(y) \geq \langle z - y, \nabla G_h(y) \rangle, \ \ \forall y, z \in \Omega^k. \tag{34}$$

Using the points given by $z := v_{t+1}^i$, and $y := v_t^i$, and noting that the l-th component of the gradient of $G_h(\cdot)$ is given by

$$\left(\nabla G_h(v_t^i) \right)_l = \pi_l \frac{h'(v_t^{il})}{h'\left(G_h(v_t^i)\right)} = \rho^{il}(\mathbf{x}(t)), \ l = 1, \ldots, k,$$

(where the last equality follows from (18)), one obtains after substituting these expressions in (34),

$$G_h(v_t^i) - G_h(v_{t+1}^i) \leq \sum_{l=1}^k \nu_i \left(v_t^{il} - v_{t+1}^{il} \right) \rho^{il}(\mathbf{x}(t)), \ \ i = 1, \ldots, m.$$

Multiplying by $s > 0$ the above inequality, summing over $i = 1, \ldots, m$, using (33) and (32) it follows that for all $\mathbf{x}(t+1) \neq \mathbf{x}(t)$,

$$\begin{aligned} F_s(\mathbf{x}(t+1)) - F_s(\mathbf{x}(t)) &\leq s \sum_{i=1}^m \sum_{l=1}^k \nu_i (v_t^{il} - v_{t+1}^{il}) \rho^{il}(\mathbf{x}(t)), \\ &= \sum_{i=1}^m \sum_{l=1}^k \nu_i d(x^l(t+1), a^i) \rho^{il}(\mathbf{x}(t)) - \sum_{i=1}^m \sum_{l=1}^k \nu_i d(x^l(t), a^i) \rho^{il}(\mathbf{x}(t)), \\ &= A_s(\mathbf{x}(t+1), \mathbf{x}(t)) - A_s(\mathbf{x}(t), \mathbf{x}(t)) < 0, \end{aligned}$$

where the last inequality is from (31), and (i) is proved.
(ii) Let $d \in \mathcal{D}(S)$ be separable. Then, by Proposition 15, the sequence $\{\mathbf{x}(t)\}_{t=0}^\infty$ is bounded with limit points in $\mathbf{S}$. Thus, there exists $\bar{\mathbf{x}} \in \mathbf{S}$ and a convergent subsequence $\{\mathbf{x}(t_j)\}$ to $\bar{\mathbf{x}}$, namely $\lim_{j \to \infty} \mathbf{x}(t_j) = \bar{\mathbf{x}}$ with $\lim_{j \to \infty} t_j = +\infty$. Let $\mathbf{x}^* \in \mathbf{S}$ be a stationary point of F_s, then one has,

$$\sum_{i=1}^m \nu_i \nabla_1 d((x^l)^*, a^i) \rho^{il}(\mathbf{x}^*)) = 0, \ l = 1, \ldots, k. \tag{35}$$

We need to show that $\overline{\mathbf{x}} = \mathbf{x}^*$. Thanks to Lemma 14(ii), and by definition of algorithm **SKM**, given $\mathbf{x}(t) \in \mathbf{S}$, there exists a unique $\mathbf{x}(t+1) \in \mathbf{S}$ which satisfies

$$\sum_{i=1}^{m} \nu_i \nabla_1 d(x^l(t+1), a^i) \rho^{il}(\mathbf{x}(t)) = 0, \; l = 1, \ldots, k. \tag{36}$$

Let us denote the solution of (36) by $x^l(t+1) := T^l(\mathbf{x}(t)), \; l = 1, \ldots, k$, and for any $\mathbf{u} \in \mathbf{S}$, set $T(\mathbf{u}) := \left(T^1(\mathbf{u}), \ldots, T^k(\mathbf{u})\right)$. Since $\mathcal{D}(S) \ni d(\cdot, a_i)$ is given C^2 on S, with positive definite matrix Hessian $\nabla_1^2 d(\cdot, a^i)$, and since $\nu_i \rho^{il}(\mathbf{x}(t)) > 0$, then $\sum_{i=1}^{m} \nu_i \nabla_1^2 d(\cdot, a^i) \rho^{il}(\mathbf{x}(t))$ is also a positive definite matrix. Therefore, by invoking the implicit function theorem, it follows that T is continuous on $\mathbf{S}$. Now, by definition of **SKM**, using (36) and (35), in terms of T, it follows that:

$$\mathbf{x}(t) = \mathbf{x}^* \text{ if and only if } \; T(\mathbf{x}(t)) = \mathbf{x}(t). \tag{37}$$

To complete the proof, we need to consider two cases. First, if for some $t \geq 0$, $\mathbf{x}(t+1) = T(\mathbf{x}(t)) = \mathbf{x}(t)$, then in that case the sequence repeats itself from that point, and one has $\overline{\mathbf{x}} = \mathbf{x}(t)$, which by (37) implies that $\overline{\mathbf{x}} = \mathbf{x}^*$. In the other case, with $\mathbf{x}(t+1) \neq \mathbf{x}(t)$, by (i) one has for all $t \geq 0$, $F_s(\mathbf{x}(t+1) < F_s(\mathbf{x}(t))$. Therefore, the sequence $\{F_s(\mathbf{x}(t)\}$ is monotone decreasing, and since $F_s(\mathbf{x}) \geq F(\mathbf{x}) \geq 0$, being also bounded below, it must converge to some limit, and it follows that,

$$\lim_{j \to \infty} \left[F_s(\mathbf{x}(t_j)) - F_s(T(\mathbf{x}(t_j)))\right] = 0. \tag{38}$$

Since T is continuous, one also have $\lim_{j \to \infty} T(\mathbf{x}(t_j)) = T(\overline{\mathbf{x}})$, and hence together with (38) we obtain $F_s(\overline{\mathbf{x}}) = F_s(T(\overline{\mathbf{x}}))$. Therefore, by (i), the last equation implies that $\overline{\mathbf{x}} = T(\overline{\mathbf{x}})$, and hence by (37), $\overline{\mathbf{x}} = \mathbf{x}^*$. ■

Remark 17 It should be noted (as already explained in §2.3), that as long as a unique global minimizer of $\mathbf{x} \to A(\mathbf{x}, \mathbf{u})$ exists, and the continuity of the map $T(\cdot)$ on $\mathbf{S}$ can be ensured, a close inspection of the proof reveals that the convergence result established in Theorem 16 could also be used for other classes of distance-like functions, and in particular for the important class of Bregman divergences considered by Banerjee et al. (2005), and discussed in Example 3.

6. Relations with Known Center-Based Clustering Algorithms and Extensions

The purpose of this section is not an intent to survey all the current existing approaches and center-based clustering methods. Rather, our aim is to briefly demonstrate that many of the seemingly different methods cited in the introduction, and which have emerged from various view points, can be derived, analyzed and extended under the unified smoothing optimization approach we have developed in this paper. This is now illustrated below, with a particular focus on the Deterministic Annealing algorithm (DA) and its possible extensions.

Before proceeding, we recall our setting. As outlined in Section 2, there exists essentially two equivalent ways to formulate the clustering problem: the nonsmooth formulation and its equivalent exact smooth re-formulation, given respectively by

$$(NS) \;\; \min\{F(\mathbf{x}) \mid \mathbf{x} \in \overline{\mathbf{S}}\} \qquad \Longleftrightarrow \qquad (SF) \;\; \min_{\mathbf{x},\mathbf{w}} \left\{\mathbf{C}_1(\mathbf{x}, \mathbf{w}) \mid w^i \in \Delta^i\right\}$$

where

$$F(\mathbf{x}) = \sum_{i=1}^{m} \nu_i \min_{1\le l\le k} d(x^l, a^i); \qquad C_1(\mathbf{x},\mathbf{w}) := \sum_{i=1}^{m}\sum_{l=1}^{k} \nu_i w_l^i d(x^l, a^i),$$

and with $\mathbf{w} := (w^1, \ldots w^m) \in \Delta = \Delta^1 \times \ldots \times \Delta^m$.

6.1 The Fuzzy k-Means Algorithm (FKM)

This method (Bezdek, 1981) was originally devised to relax the solution of problem (SF), by introducing the objective function

$$\mathbf{C}_\beta(\mathbf{x},\mathbf{w}) := \sum_{i=1}^{m}\sum_{l=1}^{k} \nu_i (w_l^i)^\beta d(x^l, a^i),$$

where $\beta > 1$ is the parameter that governs the "fuzzy partition" through $\mathbf{w}$. Indeed, the *nonlinearity* of the function $\mathbf{w} \to \mathbf{C}_\beta(\mathbf{x},\mathbf{w})$, (as opposed to the standard k-means objective which corresponds to $\beta = 1$) yields a solution $\mathbf{w}$ which is not anymore of the binary type as in hard clustering, hence the "fuzzy" terminology, (which also corresponds to the *soft* terminology). Applying the Gauss-Seidel algorithm described in the appendix to problem (SF) with the objective $\mathbf{C}_\beta(\mathbf{x},\mathbf{w})$ yields the FKM, (see, for example, Duda et al., 2001, page. 528).

Alternatively, keeping $\mathbf{x}$ fixed, and minimizing with respect to $\mathbf{w}$, that is, solving the strictly convex problem in $\mathbf{w}$:

$$\mathbf{w}^*(\mathbf{x}) = \operatorname{argmin}\{\mathbf{C}_\beta(\mathbf{x},\mathbf{w}) \mid w^i \in \Delta^i, i = 1, \ldots, m\},$$

one obtains the optimal solution

$$(w_l^i)^*(\mathbf{x}) = d(x^l, a^i)^{\frac{1}{1-\beta}} \left(\sum_{j=1}^{k} d(x^j, a^i)^{\frac{1}{1-\beta}} \right)^{-1}, \quad i = 1, \ldots, m,\ l = 1, \ldots k.$$

Plugging-in the optimal solution $\mathbf{w}^*(\mathbf{x})$ into the objective $\mathbf{C}_\beta(\mathbf{x},\mathbf{w})$, an easy computation shows that the remaining optimization problem to be solved in the variable $\mathbf{x}$ reduces to:

$$\min \left\{ \sum_{i=1}^{m} \nu_i \left(\sum_{l=1}^{k} d(x^l, a^i)^{\frac{1}{1-\beta}} \right)^{1-\beta} \mid \mathbf{x} \in \overline{\mathbf{S}} \right\}, \ (\beta > 1). \tag{39}$$

Therefore, with the choice $h(t) = (-t)^{1/(1-\beta)}$ (which is in the class $\mathcal{H}$, see Example 12) in the definition of $F_s(\cdot)$ as given in (15), one obtains that with the particular choice $\pi_l = k^{-1}, \forall l$, the objective function (39), and hence the resulting FKM algorithm, are recovered as a special case of the smoothing approach and of algorithm **SKM**, for which our convergence result applies for any distance $d \in \mathcal{D}(S)$, thus also broadening the scope of FKM based method. Note that with the special choice $\beta = 2$, the Harmonic Mean algorithm (cf. Example 10) is recovered.

6.2 The Deterministic Annealing (DA)

In Rose et al. (1990), building on statistical physics analogies, the authors have introduced the Deterministic Annealing (DA) algorithm for clustering problems. In the recent work (Teboulle

and Kogan, 2005) we already announced that DA can be derived and interpreted as a smoothing optimization method. Indeed, the DA algorithm simply corresponds to the choice $\mathcal{H} \ni h(t) = e^t$, in the nonlinear mean G_h, and when substituted in (15) yields to solve the smooth nonconvex optimization problem:

$$\min \left\{ -s \sum_{i=1}^{m} \nu_i \log \sum_{l=1}^{k} \pi_l e^{-d(x^l, a^i)/s} \mid \mathbf{x} \in \overline{\mathbf{S}} \right\}, \tag{40}$$

where the smoothing parameter $s > 0$ plays the role of the inverse temperature used in the DA formulation (Rose et al., 1990). Thus, applying **SKM** to problem (40), we obtain the classical DA algorithm whenever $d(\cdot,\cdot)$ is the usual squared Euclidean distance, as well as its extension with $d \in \mathcal{D}(S)$.

The DA algorithm is thus also a smoothing optimization method. It has been claimed in the literature (Rose et al., 1990; Ueda and Nakano, 1998; Rose, 1998), but, to the best of our knowledge, not mathematically proven, that by suitably tuning the temperature, namely in our language, the smoothing parameter, the DA can deliver "global" optimal solutions. However, our current analysis demonstrates that only the nonsmoothness difficulty appears to be eliminated via the DA approach, yet the nonconvexity difficulty remains. Nevertheless, deterministic annealing based algorithms continue to be successfully used in practice (Elkan, 2006) and appear to share two particularly interesting and unique features:
(i) As reported in several studies, the DA converges very quickly to"good" solutions (as compared to the k-means algorithm).
(ii) The DA algorithm which is obtained from our framework with the special choice $h(t) = e^t$, is also a source of many other seemingly different methods, in particular when we consider its extension with distance-like functions other than the usual squared Euclidean distance.

In view of the combined smoothing and successive approximation approach we have developed, these two features are perhaps not too surprising in the following sense. The quick delivery of a reasonable approximate solution relies on the gradient descent property of the **SKM** algorithm developed in Section 5, and hence of the DA algorithm in particular. Moreover, the Log-Sum exponential function appears to be *optimal* in the sense we previously explained in Section 4 and in Remark 13. Below, we further exemplify the point (ii) by briefly showing how the methods mentioned in the introduction, such as, Maximum Entropy Clustering Algorithms (MECA), Expectation Maximization (EM), and the Bregman soft clustering algorithm are essentially equivalent smoothing methods.

6.3 Deterministic Annealing, Entropy Methods and Information Theory

A remarkable mathematical property of the Log-Sum exponential function is that it is just the conjugate of the entropy function on the unit simplex, and vice-versa. More precisely, the following result holds.

Lemma 18 *For any given $\pi \in \Delta$,*

$$\log \sum_{l=1}^{k} \pi_l e^{z_l} = \max_{y \in \Delta} \left\{ \langle y, z \rangle - \sum_{l=1}^{k} y_l \log \frac{y_l}{\pi_l} \right\},$$

where $\Delta = \{y \in \mathbb{R}^k : \sum_{l=1}^{k} y_l = 1, y \geq 0\}$. Moreover, with $g(z) = \log \sum_{l=1}^{k} \pi_l e^{z_l}$, one has

$$g^*(y) = \sum_{l=1}^{k} y_l \log \frac{y_l}{\pi_l}, \quad \textit{with} \quad \operatorname{dom} g^* = \Delta.$$

Proof. By direct computation, or see for example Rockafellar (1970, p. 148). ■

Using the dual representation of the Log-Sum exponential function given in Lemma 18 into the objective function of (40), some algebra shows that the smooth optimization problem (40) is equivalent to:

$$\min_{\mathbf{x},\mathbf{w}} \left\{ \mathbf{C}_1(\mathbf{x},\mathbf{w}) + s\sum_{i=1}^{m}\sum_{l=1}^{k} \nu_i w_l^i \log \frac{w_l^i}{\pi_l} \mid w^i \in \Delta^i,\ i=1,\ldots,m \right\}. \tag{41}$$

This equivalent reformulation of the smooth optimization model (40) allows to show connections with other approaches that we now discuss.

Problem (40) recovers the basic formulation of what is called in the literature *Maximum Entropy Clustering Algorithms*, (MECA) (Rose, 1998). Of course, this shows that maximum entropy methods applied to the clustering problem, are thus a special case of our smoothing approach.

Furthermore, it is interesting to notice that in MECA models one usually assume that $\pi_l = k^{-1}, \forall l = 1,\ldots k.$, that is, a uniform distribution. We can enrich the model by considering π_l as weights (probabilities) associated to each cluster center l, and ask to find the "best" possible distribution for π, namely for given $(\mathbf{x},\mathbf{w})$ in problem (41), we need to solve:

$$\min \left\{ -s\sum_{i=1} \nu_i \sum_{l=1}^{k} w_l^i \log \pi_l \mid \pi \in \Delta_+ \right\}.$$

Clearly, the objective function in the later problem is convex in π, and a straightforward application of Karush-Khun-Tucker (KKT) optimality conditions (Bertsekas, 99) to the latter problem yields the optimal choice for π:

$$\pi_l^* = \sum_{i=1}^{m} \nu_i w_l^i,\ l=1,\ldots k. \tag{42}$$

Another interpretation is to view MECA as follows. Going back to the formulation (SF), the problem (41) can in fact be viewed from various angles via the classical penalty-barrier optimization method, which is also a smoothing approach (Auslender, 1999), whereby the entropy is used to penalize the simplex constraints on w^i, and s would play the role of the penalty-barrier parameter. Namely by defining,

$$E(\mathbf{w},\pi) := \sum_{i=1}^{m}\sum_{l=1}^{k} \nu_i w_l^i \log \frac{w_l^i}{\pi_l}$$

with $\mathbf{w} \in \Delta$, where $\mathbf{w} := (w^1,\ldots w^m)$, $w^i \in \mathbb{R}^k$, and $\Delta = \Delta^1 \times \ldots \times \Delta^m$, problem (41) can be viewed as a family of penalized problems, with $s > 0$, being the penalty parameter for solving the constrained problem:

$$\min_{\mathbf{x},\mathbf{w}} \{ C_1(\mathbf{x},\mathbf{w}) \mid E(\mathbf{w},\pi) \le \varepsilon, \mathbf{w} \in \Delta \},$$

where $\varepsilon > 0$, is preassigned. An interesting interpretation of the latter problem was described by Rose (1998) as follows. The smoothing parameter s can be viewed as a Lagrange multiplier to an entropy constraint which would measure the level of randomness in the following sense: the first term in the objective of (41) is a predefined "expected distortion", and thus we are trading "entropy", the second term in (41), for reduction in the distortion as $s \to 0$. Alternatively, problem (41) can

also be seen directly related to the fundamental Shannon rate-distortion function (Berger, 1971). Suppose the knowledge of the "expected distortion" $\mathbf{C}_1(\mathbf{x},\mathbf{w})$ is pre-assumed at a certain level, say $\delta > 0$, that is, one has $C_1(\mathbf{x},\mathbf{w}) = \sum_{i,l} \nu_i w_l^i d(x^l, a^i) \leq \delta$. Fix any $s > 0$, say $s = 1$, and let π be as given in (42). Then, for a fixed given $\mathbf{x}$ one has to solve,

$$R(\delta) = \min_{\mathbf{w}} \{E(\mathbf{w},\pi) \mid \mathbf{C}_1(\mathbf{x},\mathbf{w}) \leq \delta,\ \mathbf{w} \in \Delta\}. \tag{43}$$

Following information theory concepts (Thomas and Cover, 1991), a close inspection of the last problem reveals that the objective function $E(\mathbf{w},\pi)$ in (43) is the so-called average mutual information functional, and the optimal value $R(\delta)$ of problem (43) is nothing else but the mathematical description of the rate distortion function. The later problem is a convex optimization problem in $\mathbf{w}$ (a probabilistic/soft assignment variable), that can be solved via the so-called Blahut-Arimoto algorithm, (Berger, 1971), which is an iterative fixed point type convex dual optimization method.

6.4 The EM algorithm and Bregman Soft Clustering

The Expectation Maximization (EM) algorithm is a workhorse in statistical estimation problems for learning mixtures of distributions, (see, for example, Duda et al., 2001). In a very recent paper, Banerjee et al. (2005) have shown that there exists a bijective correspondence between regular exponential distributions and Bregman divergences. This result enables them to show (see, Banerjee et al., 2005, Section 5), that the EM algorithm for learning mixtures of regular exponential family distributions is in fact equivalent to Bregman soft clustering.

Without recourse to any probability/statistical arguments, we provide below, yet another interpretation and realization of this result, by showing that it corresponds to a special case of our generic scheme, with the special choices $h(t) = e^t$, and $d(x,a) := d_\psi(a,x)$.

Fix any $s > 0$, say $s = 1$ in (40). Then, one has to solve the equivalent problem (after an obvious change of sign to pass to maximization):

$$\max_{\mathbf{x}} \mathcal{F}(x,\pi) \equiv \max \left\{ \sum_{i=1}^{m} \nu_i \log \sum_{l=1}^{k} \pi_l e^{-d(x^l,a^i)} \mid \mathbf{x} \in \mathbf{S} \right\},$$

where $\pi \in \Delta_+$. Applying **SKM**, given $\pi \in \Delta_+$, and $\overline{x}^l \in \mathbf{S}$ we first need to compute:
The E-Step: compute "conditional probabilities" (cf. (18), and 22)):

$$\rho^{il}(\overline{\mathbf{x}}) := \frac{\pi_l e^{-d(\overline{x}^l,a^i)}}{\sum_{j=1}^{k} \pi_j e^{-d(\overline{x}^l,a^i)}}. \tag{44}$$

Now, the second step in **SKM** consists of solving

$$\min_{\mathbf{x}\in\mathbf{S}} \left\{ \sum_{i=1}^{m} \sum_{l=1}^{k} \nu_i d(x^l,a^i) \rho^{il}(\overline{\mathbf{x}}) \right\},$$

which admits a unique global minimizer (cf. Example 3), and yields
The M-step:

$$x^l = \frac{\sum_{i=1}^{m} \nu_i \rho^{il}(\overline{\mathbf{x}}) a^i}{\sum_{j=1}^{m} \nu_j \rho^{jl}(\overline{\mathbf{x}})},\ l = 1,\ldots,k. \tag{45}$$

Now, if we assume that π^l is also considered as a variable in the maximization of $\mathcal{F}(x,\pi)$, (e.g., π_l gives the fraction of points representing optimal clusters l), then given $\overline{\mathbf{x}} \in \mathbf{S}$, one has to solve

$$\max_{\pi} \mathcal{F}(x,\pi) \equiv \max \left\{ \sum_{i=1}^{m} \nu_i \log \sum_{l=1}^{k} \pi_l e^{-d(\vec{x}^l, a^i)} \mid \pi \in \Delta_+ \right\}. \tag{46}$$

It is easy to see that the objective function in problem (46) is a concave function in the variable π, and hence a direct application of the KKT optimality conditions to this convex problem (maximizing a concave objective subject to a simplex constraint) yields using the notations in (44), the global optimal solution:

$$\pi_l^* = \sum_{i=1}^{m} \nu_i \rho^{il}(\overline{\mathbf{x}}),\ l = 1, \ldots, k. \tag{47}$$

Therefore, with the equations (44), (45) and (47), we have recovered through a special realization of **SKM**, the EM algorithm for learning mixtures model of exponential family distributions or equivalently the Bregman soft clustering method, as recently derived in Banerjee et al. (2005) from a completely different perspective.

6.5 Discussion

There exists many other related clustering algorithms not discussed here, that can be designed, analyzed and extended through our framework, and we refer the reader to the relevant cited references throughout this paper and their bibliography therein. The above comparisons were just briefly outlined to demonstrate the fundamental and useful role that convex analysis and optimization theory can play in the analysis and interpretation of iterative center-based clustering algorithms. As such, the general framework we have proposed should be viewed as complementary to alternative formulations and approaches. Indeed, it is also important to mention that for specific application domains which often involve particular input data representation, such as in statistics and information theory, alternative approaches and formulations should not be ignored, as they can at times provide ways for better/new insights or/and solution methods. For example, in the problem of learning mixture models, an alternative approach is via spectral projection techniques, which provide algorithms with theoretical guarantee for learning mixtures of log-concave distributions (Kannan et al., 2005). Another example is the information bottleneck method (Tishby et al., 1999) which provides a useful formalism and principle to extract relevant information in a given data set. The recent interesting study (Banerjee et al., 2005) connecting Bregman clustering and lossy compression schemes through an information theoretic formalism, and which allows for extending the information bottleneck method, further demonstrates the usefulness of considering alternative formulations.

7. Concluding Remarks

This paper is a theoretical contribution to clustering analysis, and has three messages. First, the proposed optimization framework and formalism provides a systematic and simple way to design and analyze a broad class of hard and soft center-based clustering algorithms, which retain the computational simplicity of the k-means algorithm. Secondly, the proposed formalism has provided a closure and unification to a long list of disparate motivations and approaches that have been proposed for center-based clustering methods. As discussed in the paper, many of these algorithms which have

been widely used in applications are special cases of our analysis. Third, the common optimization language and the fundamental tools we have used, which rely on the combination of convex asymptotic functions, nonlinear means and distance-like functions, and from which our generic scheme has emerged, enables for a rigorous analysis of center-based clustering algorithms, have revealed the advantages and limitations of such methods, and have provided the basis for significantly extend the scope of partitioning clustering algorithms.

As a final remark, we hope that the current study will further stimulate the use and application of convex analysis and optimization in data analysis. Indeed, there are several theoretical and practical challenges that need to be met in future research works in clustering analysis. Let us mention just a few questions that naturally emerged from the present analysis. As already pointed out throughout this study, for a given specific data set to cluster, current experimental results indicate that with the choice of the Euclidean squared distance, the deterministic annealing algorithm (based on the log-exponential mean) and the harmonic k-means can produce better quality clustering (see, for example, Rose, 1998; Ueda and Nakano, 1998; Zhang et al., 1999; Hamerly and Elkan, 2002; Elkan, 2006). Thus, future experimentation based on these methods, but using other proximity measures that could model various data types, deserves to be considered. Furthermore, as pointed out by a referee, it would be interesting to identify the "best" choice of the function h to be used in the broader family of convex nonlinear means. Similarly, can we characterize the classes of functions h or/and the classes distance-like functions d that would allow us to eliminate or/and control the inherent nonconvexity difficulty which is present in the clustering problem? Can we rigorously measure the quality of clustering produced by the generic scheme, or some other possible refined variants, in terms of the problem data and the couple (h,d)? Even partial answers to such theoretical questions would have a significant practical impact, and deserve further investigations.

Acknowledgments

We thank an anonymous referee for his constructive comments that prompted us to streamline our results, and have helped to greatly improve the presentation of the paper. We also thank the area editor Charles Elkan for his valuable comments and for pointing out additional relevant references. This research was partially supported by the United States-Israel Binational Science Foundation under BSF grant 2002-010.

Appendix A.

In this appendix we briefly describe the basic mechanism of hard clustering center-based algorithms.

A.1 The Gauss-Seidel Algorithm-GSA

The Gauss-Seidel algorithm, also called coordinate descent or alternative optimization method, proceeds as follows to solve the generic minimization problem:

$$\min\min\{F(x,y) : x \in X, y \in Y\}.$$

- At iteration t, fix $x(t) \in X$, and minimize with respect to y the function $F(x(t),y)$, to get $y(t)$.
- Update x by minimizing $F(x,y(t))$ with respect to x.

- Continue this process iteratively until some declared stopping criteria is satisfied.

Convergence of GSA to a stationary point can be established under suitable conditions on the problem's data, (Auslender, 1976; Bertsekas, 99).

A.2 Hard Clustering with Distance-Like Functions

The popular k-means algorithm with d being the squared of the Euclidean distance, is nothing else but the Gauss-Seidel method, when applied to the exact smooth formulation (ES) of the clustering problem.

$$\min_{\mathbf{x},\mathbf{w}} \left\{ \mathbf{C}_1(\mathbf{x},\mathbf{w}) = \sum_{i=1}^{m} \sum_{l=1}^{k} \nu_i w_l^i d(x^l, a^i) \right\}.$$

More generally, applying GSA on problem (ES) with distance-like functions d, we can obtain a broad class of general Hard Clustering algorithms **HCD**. Note that the main computational step in the generic **HCD** algorithm keeps the computational simplicity of the k-means algorithm, yet allows for significantly expand the scope of hard clustering center-based methods.

Algorithm HCD–Hard Clustering with Distance-Like Functions

- **Step 0-Initialization** Set $t = 0$ and let $\{x^l(0) : l = 1, \ldots, k\}$ be the k initial centers in S. (These can be picked randomly).

- **Step 1-Cluster Assignment** For $i = 1, \ldots, m$ solve $w^i(t) = \operatorname*{argmin}_{w \in \Delta_i} \sum_{l=1}^{k} w_l d(x^l(t), a^i)$.

- **Step 2-Update Cluster Centers** For each $l = 1, \ldots, k$ solve

$$\left(x^1(t+1), \ldots, x^k(t+1)\right) = \operatorname*{argmin}_{x^1,\ldots,x^k} \left\{ \sum_{i=1}^{m} \sum_{l=1}^{k} \nu_i w_l^i(t) d(x^l, a^i) \right\}.$$

- **Step 3-Stopping Criteria** Stop when some stopping criteria is satisfied, (e.g., $\mathbf{x}(t+1) = \mathbf{x}(t)$), else set $t \longleftarrow t+1$ and goto **Step 1**.

Algorithm HCD clearly implies that the objective function of (ES) is nonincreasing at the successive iterations. The resulting stationary point obtained by this procedure satisfies the Karush-Khun-Tucker (KKT) necessary optimality conditions for problem (ES) (Bertsekas, 99).

The remarkable simplicity of algorithm HCD relies on the fact that Step 1 is trivially solved, while step 2 can be solved *analytically* for a wide class of distances d. Indeed, at any given iteration t, to solve Step 1, for all $i = 1, \ldots, m$, let $l(i) = \operatorname*{argmin}_{1 \le l \le k} d(x^l(t), a^i)$. Then, an optimal w^i is simply given by

$$w_{l(i)}^i(t) = 1, \text{ that is, when } x^l \text{ is the center closest to } a^i, \text{and } w_l^i(t) = 0, \forall l \neq l(i).$$

To solve step 2, noting that the objective is *separable* in each variable x^l, it reduces to solve for each x^l:

$$x^l(t+1) = \operatorname*{argmin}_{x} \left\{ \sum_{i=1}^{m} w_{l(i)}^i(t) d(x, a^i) \right\}.$$

For the class $d \in \mathcal{D}(S)$, as well as for other distance-like functions as discussed in Section 2.2, this problem admits a unique global optimal solution. Furthermore, for distance-like functions which are given separable, this problem even reduces to solve a *one dimensional* optimization problem, which can often be solved analytically for many examples (Teboulle et al., 2006). It is easy to see that algorithm **HCD** includes as special cases, not only the popular k-means algorithm, but also many others hard clustering methods mentioned in the introduction. In particular, it includes and extend the Bregman hard clustering algorithm recently derived in Banerjee et al. (2005, Algorithm 1, page 1715).

References

A. Auslender. *Optimisation: Methodes Numériques*. Masson, Paris, 1976.

A. Auslender. Penalty and barrier methods: a unified framework. *SIAM J. Optimization* 10(1), 211-230, 1999.

A. Auslender and M. Teboulle. *Asymptotic Cones and Functions in Optimization and Variational Inequalities*. Springer–Verlag, New York, 2003.

A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optimization*, 16(3):697-725, 2006.

A. M. Bagirov, A. M. Rubinov, N.V. Soukhoroukova, and J. Yearwood. Unsupervised and supervised data classification via nonsmooth and global optimization. *TOP*, (Formerly Trabajos Investigación Operativa) 11(1):1–93, 2003.

A. M. Bagirov and J. Ugon. An algorithm for minimizing clustering functions. *Optimization*, 54(4-5): 351–368, 2005.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6, 1705–1749, 2005.

A. Ben-Tal and M. Teboulle. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Sciences*, 32(11):1445–1466, 1986.

A. Ben-Tal and M. Teboulle. A smoothing technique for nondifferentiable optimization problems. In *Springer Verlag Lecture Notes in Mathematics*, volume 1405, pages 1–11, Berlin, 1989.

A. Ben-Tal, M. Teboulle, and W. H. Yang, A least-squares based method for a class of nonsmooth minimization problems with applications in plasticity. *Applied Mathematics and Optimization*, 24(3):273–288, 1991.

T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice Hall, Englewood Cliffs, New Jersey, 1971.

D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, Belmont, Massachusetts, second edition, 1996.

D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Massachusetts, second edition, 1999.

D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, New Jersey, 1989.

J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

L. M. Bregman. A relaxation method of finding a 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.

J. Brimberg and Love R.F. Global convergence of a generalized iterative procedure for the minisum location problem with l_p distances. *Operations Research*, 41:1153–1163, 1993.

Y. Censor and A. Lent. An interval row action method for interval convex programming. *J. of Optimization Theory and Applications*, 34:321–353, 1981.

Y. Censor and S. A. Zenios, *Parallel Optimization*, Oxford University Press, Oxford, 1997.

G. Chen and M. Teboulle. Convergence analysis of a proximal-like algorithm using Bregman functions. *SIAM J. on Optimization* 3:538–543, 1993.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

I. Csiszar. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Mathematica Hungarica*, 2:299–318, 1967.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., second edition, 2001.

C. Elkan. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of the 23rd International Conference on Machine Learning ICML 06*, 289-296, 2006.

E. Forgy. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics*, Abstract, 21(3):768, 1965.

M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, San Francisco, CA, 1979.

A.D. Gordon and J.T. Henderson. An algorithm for Euclidean sum of squares classification. *Biometrics*, 33:355-362, 1977.

R.M. Gray and D.L. Neuhoff. Quantization. *IEEE Transaction on Information Theory*, 44(6):2325–2382, 1998.

G. Hamerly and C. Elkan. Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02)*, pages 600-607, 2002.

J P. Hansen and N. Mladenovic. J-Means: A New Heuristic for Minimum Sum-of-Squares Clustering. *Pattern Recognition* 34:405–413, 2001.

G. Hardy, J.E. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, Cambridge, 1934.

A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review, *ACM Computing Surveys*, 31(3):264-323, 1999.

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the 18th Annual Conference on Learning Theory*, 444-457, 2005.

N.B. Karayiannis. An axiomatic approach to soft learning vector quantization and clustering. *IEEE Transactions on Neural Networks*, 10(5):1153-1165, 1999.

S.P. Lloyd. Least squares quantization in PCM. Bell Telephone Laboratories Paper, Murray Hill, NJ, 1957. Also in, *IEEE Transactions on Information Theory*, 28:127-135, 1982.

Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84-95, 1980.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Math., Stat. and Probability*, pages 281–296, 1967.

D. Modha and S. Spangler. Feature weighting in k-means clustering. *Machine Learning*, 52(3):217-237, 2003.

M.R. Rao. Cluster analysis and mathematical programming. *J. American Statistical Association*, 66:622–626, 1971.

R. T. Rockafellar. *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

K. Rose, E. Gurewitz, and C.G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594, 1990.

K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.

H. Steinhaus. Sur la division des corps materiels en parties. *Bull. Acad. Polon. Sci.*, C1. III, vol. IV, 801–804, 1956.

M. Teboulle. "On φ-divergence and its applications". In *Systems and Management Science by Extremal Methods* (F.Y. Phillips, J. Rousseau, eds.), Kluwer Academic Press, chap. 17, pages 255–273, 1992.

M. Teboulle. Entropic proximal mappings with application to nonlinear programming. *Mathematics of Operations Research*, 17:670–690, 1992.

M. Teboulle. Convergence of proximal-like algorithms. *SIAM J. of Optimization*, 7:1069-1083, 1997.

M. Teboulle and J. Kogan. Deterministic annealing and a k-means type smoothing optimization algorithm. In *Proceedings of the Workshop on Clustering High Dimensional Data and its Applications* (held in conjunction with the Fifth SIAM International Conference on Data Mining). I. Dhillon, J. Ghosh and J. Kogan (eds.), pages 13-22, 2005.

M. Teboulle, P. Berkhin, I. Dhillon, Y. Guan, and J. Kogan. Clustering with entropy-like k-means algorithms. In *Grouping Multidimensional Data: Recent Advances in Clustering*. J. Kogan, C. Nicholas, and M. Teboulle, (Eds.), Springer Verlag, NY, pages 127–160, 2006.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11(2): 271–282, 1998.

E. Weiszfeld. Sur le point pour lequel la somme des distances de *n* points donnés est minimum. *Tohoku Mathematical Journal*, 43:355–386, 1937.

B. Zhang, M. Hsu, and U. Dayal. K-harmonic means - a data clustering algorithm. Technical Report HPL-1999-124 991029, HP Labs, Palo Alto, CA, 1999.