A SMOOTHING ALTERNATING MINIMIZATION-BASED ALGORITHM FOR CLUSTERING WITH SUM-MIN OF EUCLIDEAN NORMS

SHOHAM SABACH, MARC TEBOULLE, AND SERGEY VOLDMAN

We are delighted to contribute to this volume in honor of Adi Ben-Israel 85th birthday. Marc Teboulle, wants to express his deepest thanks and appreciation to him. Adi is my academic grand father, and has essentially paved the ways to my academic career. Starting from my early stages at the undergraduate level, Adi has continuously and generously provided me with his enthusiastic strongest support. His fundamental and pioneering works in Linear Algebra, Optimization, and his vast knowledge in mathematical sciences have been a constant source of inspiration and admiration. Adding to that his great sense of humor, the fun in is company is unlimited. Happy Birthday Adi!

Abstract. We consider the problem of minimizing an objective function defined as the finite sum of a minimum collection of nonsmooth and convex functions, which includes the fundamental clustering problem as a particular case. To tackle this nonsmooth and nonconvex problem, we develop a smoothing alternating minimization-based algorithm (SAMBA), which is proven to globally converge to a critical point of the smoothed problem. We then show how it can be applied to the clustering problem with adequate smoothing functions, producing two very simple algorithms resembling the so-called k-means algorithm, with global convergence analysis.

Key words: Nonconvex nonsmooth minimization, clustering methods, smoothing techniques, proximal map, alternating minimization, semialgebraic optimization, Kurdyka-Łojasiewicz property, global convergence.

1. Introduction

Clustering consists of grouping objects according to similarities among them. More precisely, the objective is to partition a given data-set into subsets, called *clusters*, such that the data points in each cluster are similar to one another, and dissimilar to objects in other clusters. Clustering techniques are central to data analysis tasks, and widely used in broad and disparate applications including machine learning, pattern recognition, computational biology and information retrieval to mention just a few. The

Date: April 25, 2018.

2010 *Mathematics Subject Classification.* 90C26, 90C30, 49N45, 65K05.

Key words and phrases. Alternating minimization, Kurdyka- Lojasiewisz property, global convergence, nonconvex-nonsmooth minimization, gradient methods, smoothing, clustering, semi-algebraic functions.

literature is huge, and for a sample, we refer the reader to, e.g., [21, 33, 22] and references therein. Traditionally, clustering techniques can be divided into *hierarchical* and *partitioning*. In this work, we focus on partitioning clustering, where the number of clusters is known in advance. Most partitioning clustering methods iteratively update the cluster centers by solving an optimization problem, and therefore are often referred to as *center-based* clustering methods. Since clusters can formally be seen as subsets of the data-set, these methods are often referred as *hard and soft clustering*. In hard clustering, each data point is assigned to a single cluster, namely a binary strategy is used so that each data point belongs exactly to one partition, while in soft clustering, the binary strategy is relaxed so that each data point may be assigned to more than one cluster, i.e., clusters could overlap. A basic well-known formulation of the partitioning clustering problem can be set up as follows. Let $\mathcal{A} = \left\{a^1, a^2, \ldots, a^m\right\}$ be a given set of points in $\mathbb{R}^n$ and assume that $1 < k < m$. For each $l = 1, 2, \ldots, k$, the cluster is represented by its center $x^l \in \mathbb{R}^n$. The clustering problem can be formulated as the following optimization problem:

$$\text{(CP)} \qquad \min_{x \in \mathbb{R}^{nk}} \sum_{i=1}^{m} \min_{1 \le l \le k} d\left(x^l, a^i\right),$$

where $x = \left(x^1, x^2, \ldots, x^k\right) \in \mathbb{R}^{nk}$ and $d(\cdot, \cdot)$ is some distance-like function which measures the similarity between x^l and a^i. The similarity measure in clustering can be defined using various distance-like functions. The use of different similarity measures allows one to find different cluster structures in a data-set.

A widely used similarity measure is based on the squared Euclidean norm, that is, $d(u, v) = \|u - v\|^2$. In that case, one of the most famous hard clustering algorithm to solve problem (CP) is the so-called *k-means algorithm* [17], which can be traced back to [29]. The k-means algorithm partitions the data-set $\mathcal{A}$ in an iterative way, where it begins with a random initialization of the centers and then alternates between two steps. The first step is the assignment of each data point to the closest center. The second step is the center update as a weighted arithmetic mean of all points assigned to each cluster center. The simplicity of the algorithm, both in the updating rules and implementation aspects, made it very popular. However, it is well-known that it has several drawbacks, e.g., it is highly sensitive to the initial choice of cluster centers, it can produce empty clusters, and it does not properly handle outliers, due to the use of the squared Euclidean distance. The inherent nonconvexity and nonsmooth nature of the problem is a major difficulty which has generated intensive research activities toward the search and design of *approximation* algorithms that could produce better quality clustering. Moreover, different data types arising in many applications have justified the use of other meaningful distance-like functions, including non-Euclidean proximity measures, that can better model a given data-set. As a

result, a large number of hard and soft clustering algorithms have emerged from various and different perspectives, see, e.g., [10, 16, 25, 3, 31, 19, 20], and for a unified framework covering many of these clustering methods, see [30] and references therein.

Motivated by this line of research, in this work we focus on the (CP) model when the similarity measure is the *Euclidean norm*, rather the usual squared norm, which as alluded above might provide a better way to handle the presence of outliers in a given data-set. The resulting optimization model is the nonsmooth and nonconvex problem, given by

$$\text{(CP-N)} \qquad \min_{x\in\mathbb{R}^{nk}} \sum_{i=1}^{m} \min_{1\le l\le k} \left\| x^l - a^i \right\|. \tag{CP-N}$$

While this paper is motivated by the challenging nonsmooth and nonconvex formulation of the clustering problem (CP-N), our developments, algorithm and analysis will focus on a broader class of optimization problems which involves the minimization of a finite sum of a minimum collection of nonsmooth and convex functions described by (see Section 2 for details):

$$\text{(NS)} \qquad \min \left\{ \sum_{i=1}^{m} \min_{1\le l\le k} d^i\left(x^l\right) : x^1, x^2, \ldots, x^k \in \mathbb{R}^n \right\}, \tag{NS}$$

which clearly includes the clustering problem (CP-N) as a particular case. Our main objective is to develop and analyze a new, simple method for tackling problem (NS). In this optimization model, the nonsmoothness occurs in a double way: once due to the "min" operation, the other due to the nonsmoothness of $d^i(\cdot)$, $i = 1, 2, \ldots, m$. Removing the first level of nonsmoothness will be the starting point of our developments. Relying on a simple and well-known observation, we first derive an equivalent reformulation of the problem, which eliminates the nonsmoothness caused by the inner discrete minimum through the use an of additional variable and a simple constraint. The obtained constrained problem remains nonsmooth due to the nonsmoothness of the functions $d^i(\cdot)$, $i = 1, 2, \ldots, m$. This second level of nonsmoothness is then addressed by a simple smoothing technique, to produce a smooth approximate optimization model, see Section 2. As we shall see in Section 3, this naturally paves the way to design a Smoothing Alternating Minimization-Based Algorithm (SAMBA) which involved two simple explicit computational steps. The underlying idea of this approach can be traced back to the so-called Weiszfeld algorithm [32] for solving the Fermat-Weber location theory problem, see [4] for recent developments on this method. Building on some very recent convergence results for semi-algebraic optimization [1, 13], and, in particular, the general proof mechanism developed in [13], the convergence analysis of SAMBA is developed in Section 4, where we prove that the proposed method generates a sequence of iterates which converges globally to a critical point of the smoothed objective function. In Section 5, we applied our results to the clustering problem

(CP-N), deriving two variants of SAMBA which are similar to the k-means algorithm in the sense that they alternates between clusters assignment and centers update. In fact, both versions of SAMBA produces the same clusters assignment as the k-means, while the centers update are given through a closed and computationally inexpensive formula. In the last section we illustrate the performance of both smoothing techniques.

Notation. Our notation and basic definitions are standard and can be found, for example, in [27]. We denote the unit simplex defined by $\Delta = \{w \in \mathbb{R}^k : \sum_{l=1}^k w_l = 1, w \geq 0\}$, and the Cartesian product of m copies of the unit simplex Δ, we denote by $\Delta^m := \Delta \times \Delta \times \cdots \times \Delta$. The orthogonal projection onto the simplex Δ is defined by $P_\Delta(u) := \operatorname{argmin}_{v \in \Delta} \|v - u\|^2$, and δ_Δ stands for the indicator function of Δ.

2. A Constrained Smooth Based Approximation Approach

For the purpose of our developments, we consider the following general nonsmooth and nonconvex optimization model, which naturally captures and extends the basic clustering problem (CP-N):

$$\text{(NS)} \quad \min \left\{ F\left(x^1, x^2, \ldots, x^k\right) \equiv \sum_{i=1}^m \min_{1 \leq l \leq k} d^i\left(x^l\right) : x^1, x^2, \ldots, x^k \in \mathbb{R}^n \right\}.$$

Throughout this paper we assume:

- For each $i = 1, 2, \ldots, m$, the function $d^i : \mathbb{R}^n \to \mathbb{R}$ is continuous, nonsmooth and convex.
- Problem (NS) is solvable, namely, $\operatorname{argmin} F \neq \emptyset$.

Our objective is to tackle the nonsmooth problem (NS) via a smooth approximation counterpart. The nonsmoothness here clearly occurs in two different levels:

(a) through the finite "min" operator (even if $d^i(\cdot)$, $i = 1, 2, \ldots, m$, would be smooth); and
(b) through $d^i(\cdot)$, $i = 1, 2, \ldots, m$, which here is assumed to be nonsmooth.

To handle the first level of nonsmoothness as mentioned in item (a), we first reformulate problem (NS) as a *constrained* minimization, by replacing the inner discrete minimization with a minimum over the unit simplex Δ. To this end, following [30], we use the simple fact that for any $u \in \mathbb{R}^k$,

$$\min_{1 \leq l \leq k} u_l = \min \{\langle u, w \rangle : w \in \Delta\},$$

which for each $l = 1, 2, \ldots, k$ admits the minimizer $w_l^* = 1$ if $l = \operatorname{argmin}_{1 \leq j \leq k} u_j$, else $w_l^* = 0$.

Using this fact, in problem (NS), and introducing new variables $w^i \in \mathbb{R}^k$, $i = 1, 2, \ldots, m$, yields the equivalent reformulation of problem (NS)

$$\min_{x \in \mathbb{R}^{nk}} \sum_{i=1}^{m} \min_{w^i \in \Delta} \sum_{l=1}^{k} w^i d^i \left(x^l \right).$$

Therefore, problem (NS) reduces to the following constrained problem

$$\min_{x \in \mathbb{R}^{nk}, w \in \mathbb{R}^{mk}} \left\{ \sum_{i=1}^{m} \sum_{l=1}^{k} w_l^i d^i (x^l) : \; w \in \Delta^m \right\},$$

where $w = \left(w^1, w^2, \ldots, w^m \right) \in \mathbb{R}^{km}$. This constrained reformulation is now linear in w, but remains nonsmooth in the variable x, as alluded above in item (b), due to the nonsmoothness of each d^i, $i = 1, 2, \ldots, m$. To handle this, we follow the simple idea (see, e.g., [7]) of replacing each nonsmooth function $d^i(\cdot)$, $i = 1, 2, \ldots, m$, by an adequate smooth approximation of it, which leads us to introduce the following notion of a *smooth approximation* of a given convex function which is nondifferentiable.

Definition 2.1 (Smooth approximation)**.** Let $d : \mathbb{R}^n \to \mathbb{R}$ be a continuous and convex function which is nondifferentiable. A function $d_s : \mathbb{R}^n \to \mathbb{R}$ is a *smooth approximation* of $d(\cdot)$ with smoothing parameter $s > 0$, if the following two conditions hold:

(i) $d_s : \mathbb{R}^n \to \mathbb{R}$ is convex, continuously differentiable and satisfies $\lim_{s \to 0^+} d_s(u) = d(u)$ for every $u \in \mathbb{R}^n$.

(ii) There exists a continuous function $l_s : \mathbb{R}^n \to [0, \infty)$, such that the convex function $d_s(\cdot)$ satisfies

$$d_s(v) \leq d_s(u) + \langle \nabla d_s(u), v - u \rangle + \frac{l_s(u)}{2} \|v - u\|^2, \quad \forall \; u, v \in \mathbb{R}^n. \tag{2.1}$$

For convenience, we denote by $\mathcal{S}[\mathbb{R}^n, l_s]$ the class of smooth approximations, namely $d \in \mathcal{S}[\mathbb{R}^n, l_s]$, means that there exists a convex and continuously differentiable function d_s satisfying the premises of Definition 2.1.

In Section 5, we will discuss two different approaches to generate smoothing functions $d_s(\cdot)$ for the clustering problem. Meanwhile, a few remarks are in order regarding the above definition. First, it should be noted that the concept of a smooth approximation of a nondifferentiable function is not new, see e.g., [8, 7] for some earlier works and references therein, and [30, 6] for more recent ones. Our definition bears similarity and is closer to [6] who have introduced a general concept of smoothable convex functions, except that instead of asking Lipschitz continuity of the gradient of the convex function $d_s(\cdot)$ on $\mathbb{R}^n$, which is equivalent to the classical Descent Lemma (see [9]), here the premise of Definition 2.1(ii), allows for more flexibility by considering $l_s(\cdot)$ to be a function of u, rather than, a usual constant $l_s \equiv l > 0$. As we shall see in Section 5, this is motivated by the clustering problem

itself (see problem (CP-N)), where appropriate smoothing techniques will be given (in one scenario this flexibility will allow us to find a better step size in the forthcoming algorithm).

Throughout, we assume that $d^i \in \mathcal{S}[\mathbb{R}^n, l_s]$ for each $i = 1, 2, \ldots, m$.

Equipped with the above, we have now all of the necessary components to consider a smooth approximation model for problem (NS). To that end, and to facilitate the forthcoming development, we first introduce some convenient notations. For each $i = 1, 2, \ldots, m$ we denote

$$\rho_s^i(x) := \left(d_s^i\left(x^1\right), d_s^i\left(x^2\right), \ldots, d_s^i\left(x^k\right)\right) \in \mathbb{R}^k,$$

and

$$H_s(x, w) := \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i d_s^i\left(x^l\right) = \sum_{i=1}^{m}\left\langle w^i, \rho_s^i(x)\right\rangle.$$

We then suggest to consider the following smooth approximation model of problem (NS):

$$\min\left\{H_s(x, w) :\ x \in \mathbb{R}^{nk}, w \in \Delta^m\right\}. \tag{SA-NS}$$

Since the above objective is already smooth (linear) in w, we show below that $H_s(\cdot, w)$ satisfies the premises (i) and (ii) of Definition 2.1. For any $w \in \Delta^m$, and all $l = 1, 2, \ldots, k$, we define, for $u \in \mathbb{R}^n$, the function

$$H_s^l(u, w) := \sum_{i=1}^{m} w_l^i d_s^i(u). \tag{2.2}$$

First, since $H_s(x, w) = \sum_{l=1}^{k} H_s^l\left(x^l, w\right)$, it is easy to see that for any $w \in \Delta^m$, we have that $x \to H_s(x, w)$ is convex, continuously differentiable, and $\lim_{s\to 0^+} H_s(x, w) = H(x, w)$, since by assumption $d_s^i(\cdot)$, $i = 1, 2, \ldots, m$, shares the same properties. Thus, it shows that $H_s(\cdot, w)$ satisfies Definition 2.1(i). The next result shows that $H_s^l(\cdot, w)$, $l = 1, 2, \ldots, k$, satisfies the condition of Definition 2.1(ii), which will be useful to show that $H_s(\cdot, w)$ satisfies this property too.

Lemma 2.1. *Fix $w \in \Delta^m$ and $l = 1, 2, \ldots, k$. Then, for all $u, v \in \mathbb{R}^n$, we have*

$$H_s^l(v, w) \le H_s^l(u, w) + \left\langle \nabla_u H_s^l(u, w), v - u\right\rangle + \frac{L_s^l(u, w)}{2}\|v - u\|^2, \tag{2.3}$$

where

$$L_s^l(u, w) := \sum_{i=1}^{m} w_l^i l_s(u). \tag{2.4}$$

Proof. Since each $d^i(\cdot)$, $i = 1, 2, \ldots, m$, satisfies Definition 2.1(ii), we can multiply (2.1) by the nonnegative parameter w_l^i and summing for all $i = 1, 2, \ldots, m$, which yields that (2.3) holds true, as stated. □

As an immediate consequence, we get for all $x, y \in \mathbb{R}^{nk}$ that

$$H_s(x,w) \leq H_s(y,w) + \langle \nabla_x H_s(y,w), x-y \rangle + \frac{L_s(y,w)}{2}\|x-y\|^2, \tag{2.5}$$

with $L_s(y,w) = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i L_s^l\left(y^l, w\right)$, thus showing that the desired property of Definition 2.1(ii) holds for $H_s(\cdot, w)$.

The particular structure of the smoothed problem (SA-NS) naturally provides the key step towards designing a simple algorithm, and its convergence analysis which is developed next.

3. SAMBA: A Smoothing Alternating Minimization-Based Algorithm

Let $F_s : \mathbb{R}^{nk} \times \mathbb{R}^{mk} \to (-\infty, +\infty]$ be the function defined by

$$F_s(x,w) := H_s(x,w) + \sum_{i=1}^{m} \delta_\Delta\left(w^i\right). \tag{3.1}$$

Using this notation, problem (SA-NS) can be written as

$$\min\left\{F_s(x,w) : x \in \mathbb{R}^{nk},\ w \in \mathbb{R}^{mk}\right\}.$$

Note that $F_s(\cdot,\cdot)$ is nonconvex in (x,w). However, it is convex in each of its arguments, when the other is kept fixed. This two-block structure of the objective function $F_s(\cdot,\cdot)$ can be exploited towards simple computations with respect to each block separately. To this end, we will build on the well-known concept of Alternating Minimization (AM) [2, 9], which allow us to focus on each block separately. That is, starting with $x(0) \in \mathbb{R}^{nk}$ we generate iteratively a sequence $\{(x(t), w(t))\}_{t\in\mathbb{N}}$ via the following

$$w(t+1) = \operatorname{argmin}\left\{F_s(x(t), w) : w \in \mathbb{R}^{mk}\right\},$$
$$x(t+1) = \operatorname{argmin}\left\{F_s(x, w(t+1)) : x \in \mathbb{R}^{nk}\right\}.$$

Using the AM idea, we have split the difficult task of minimizing the nonconvex function $F_s(\cdot,\cdot)$ into two *convex* sub-problems that should be solved in each iteration.

As we shall see below, the minimization subproblem with respect to w is immediate and can be solved analytically. On the other hand, with respect to x we have a convex subproblem which, due to the choice of $d_s^i(\cdot)$, $i = 1, 2, \ldots, m$, is not analytically solvable, but thanks to the property of $H_s(\cdot, w)$, for fixed $w \in \mathbb{R}^{mk}$, suggests approximating its solution via a single gradient step. Very recently, this general technique of combining AM with approximate steps was used in several contexts and scenarios (see, for example, [13, 18, 5] and the references therein). Here we will develop a new algorithm which is designed to tackle problem (SA-NS) via simple computations.

3.1. **The subproblem with respect to the w-block.** It is easy to see that the function $w \to F_s(x,w)$, for fixed $x \in \mathbb{R}^{nk}$, is separable for all $i = 1,2,\ldots,m$ and each part consists of a linear function that should be minimized over the unit simplex Δ. Indeed, for fixed $x \in \mathbb{R}^{nk}$, we have, for all $i = 1,2,\ldots,m$, the following optimization problem:

$$\min_{w^i \in \mathbb{R}^k} \left\{ \left\langle w^i, \rho_s^i(x) \right\rangle : \ w^i \in \Delta \right\}.$$

This subproblem is easy to solve and requires finding the minimal entry of the vector $\rho_s^i(x)$, which we denote by $l(i)$, that is,

$$l(i) := \operatorname{argmin}_{1 \le l \le k} d_s^i\left(x^l\right).$$

Then, an optimal solution w^i is given by

$$w_l^i = \begin{cases} 1, & l = l(i), \\ 0, & \text{otherwise.} \end{cases}$$

3.2. **The subproblem with respect to the x-block.** The convex subproblem of (SA-NS) with respect to x is unconstrained and separable for all $l = 1,2,\ldots,k$, and, therefore, we can minimize $F_s(\cdot, w)$ with respect to each x^l separately. Recall that for fixed $w \in \mathbb{R}^{mk}$ and all $l = 1,2,\ldots,k$,

$$H_s^l(u,w) = \sum_{i=1}^m w_l^i d_s^i(u), \quad u \in \mathbb{R}^n, \tag{3.2}$$

and the function $u \to H_s^l(u,w)$ is convex and satisfies the upper approximation given in (2.3). This obviously suggests to tackle the unconstrained minimization with respect to each x^l, $l = 1,2,\ldots,k$, via a simple gradient descent step on $H_s^l(\cdot,w)$. Therefore, we propose the following generic algorithm to solve the problem (SA-NS).

Smoothing Alternating Minimization-Based Algorithm – SAMBA

(1) Smoothing function: Pick $d_s^i \in \mathcal{S}[\mathbb{R}^n, l_s]$ for each $i = 1,2,\ldots m$.
(2) Initialization: Start with any $x(0) \in \mathbb{R}^{nk}$.
(3) Iterative step: Generate the sequence $\{(x(t), w(t))\}_{t \in \mathbb{N}}$ via:
  (3.1) For all $i = 1,2,\ldots,m$ compute

$$w^i(t+1) = \operatorname{argmin}\left\{ \left\langle w^i, \rho_s^i(x(t)) \right\rangle : \ w^i \in \Delta \right\}. \tag{3.3}$$

  (3.2) For each $l = 1,2,\ldots,k$ compute

$$L_s^l(t) := L_s^l\left(x^l(t), w(t+1)\right) = \sum_{i=1}^m w_l^i(t+1)\, l_s\left(x^l(t)\right), \tag{3.4}$$

and

$$x^l(t+1) = x^l(t) - \frac{1}{L_s^l(t)} \nabla_x H_s^l\left(x^l(t), w(t+1)\right). \tag{3.5}$$

Remark 3.1. We would like to comment about the possibility that at certain iteration $t \in \mathbb{N}$, we will have that $L_s^l(t) = 0$. This obviously can only be if $w_l^i(t+1) = 0$ for all $l = 1, 2, \ldots, k$. In such cases, we can always replace $L_s^l(t)$ with

$$\overline{L_s^l}(t) := \max\left\{L_s^l(t), \bar{\beta}\right\},$$

for some given $\bar{\beta} > 0$, and the upper approximation obtained in Lemma 2.1 (see (2.3)) remains valid. Hence, without the loss of generality we can assume that

$$L_s^l(t) \geq \bar{\beta}, \quad t \in \mathbb{N},$$

and, therefore, the updating rule of x^l, $l = 1, 2, \ldots, k$, is well-defined.

4. Convergence Analysis

We are now ready to derive the global convergence analysis of SAMBA to a critical point of the smoothed objective function $F_s(\cdot, \cdot)$. To this end, we will follow the recent methodology of [13], which provides the foundation to derive global convergence of descent algorithms in the nonconvex setting (see also the very recent work [14] on nonconvex Lagrangian-based schemes for a related approach). Unfortunately, this general mechanism cannot be directly applied in our case, and, therefore, we will present a weaker variant which fits our setting. We first describe this variant in general, since it could be also of interest in other scenarios, and then will provide the convergence analysis of SAMBA.

4.1. Abstract Framework.

Suppose one is interested in the minimization of a certain function $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$. The authors of [13] have formulated a set of three conditions, which are needed to show that a sequence, generated by a generic algorithm, converges globally to a critical point of the function $\Psi(\cdot)$. In this general setting of nonsmooth and nonconvex functions, and following [28], we say that z^* is a critical point of $\Psi(\cdot)$, if $0 \in \partial\Psi(z^*)$, where $\partial\Psi(\cdot)$ denotes here the (limiting) subdifferential of $\Psi(\cdot)$.

The first essential step of the methodology of [13] is proving that the generated sequence is a gradient-like descent sequence for minimizing $\Psi(\cdot)$ in the following sense.

Definition 4.1. A sequence $\{z^t\}_{t\in\mathbb{N}}$ is called a gradient-like descent sequence for minimizing $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$, if the following three conditions hold:

(C1) *Sufficient decrease property.* There exists a positive scalar ρ_1 such that

$$\rho_1 \left\|z^{t+1} - z^t\right\|^2 \leq \Psi\left(z^t\right) - \Psi\left(z^{t+1}\right), \quad \forall\, t \in \mathbb{N}.$$

(C2) *A subgradient lower bound for the iterates gap.* There exist some $u^{t+1} \in \partial\Psi\left(z^{t+1}\right)$ and a positive scalar ρ_2 such that

$$\left\|u^{t+1}\right\| \leq \rho_2 \left\|z^{t+1} - z^t\right\|, \quad \forall\ t \in \mathbb{N}.$$

(C3) Let $\overline{z}$ be a limit point of a subsequence $\left\{z^t\right\}_{t\in\mathcal{T}}$, then

$$\limsup_{t\in\mathcal{T}\subset\mathbb{N}} \Psi\left(z^t\right) \leq \Psi\left(\overline{z}\right).$$

As mentioned above, we will need to consider a variant of this definition, which will be suited to our setting. The reason for studying this weaker variant, is the fact that SAMBA generates a sequence, which satisfies a sufficient decrease property, as in condition (C1), but the descent is not measured in terms of the whole sequence $\{(x(t), w(t))\}_{t\in\mathbb{N}}$ itself, and holds only partially for the x-sequence. Therefore, here we study a weaker version of conditions (C1) and (C2), while condition (C3) remains the same. More precisely, based on Definition 4.1, for our optimization model, we suggest the following version, but keep the same terminology for the sake of simplicity.

Definition 4.2. A sequence $\{z^t\}_{t\in\mathbb{N}}$ is called *a gradient-like descent sequence* for minimizing $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$, if $\{z^t\}_{t\in\mathbb{N}}$ can be split as follows $z^t = \left(x^t, w^t\right)$, $t \in \mathbb{N}$, and the following three conditions hold:

(C1′) *Sufficient decrease property.* There exists a positive scalar ρ_1 such that

$$\rho_1 \left\|x^{t+1} - x^t\right\|^2 \leq \Psi\left(z^t\right) - \Psi\left(z^{t+1}\right), \quad \forall\ t \in \mathbb{N}.$$

(C2′) *A subgradient lower bound for the iterates gap.* There exist some $u^{t+1} \in \partial\Psi\left(z^{t+1}\right)$ and a positive scalar ρ_2 such that

$$\left\|u^{t+1}\right\| \leq \rho_2 \left\|x^{t+1} - x^t\right\|, \quad \forall\ t \in \mathbb{N}.$$

(C3) Let $\overline{z}$ be a limit point of a subsequence $\left\{z^t\right\}_{t\in\mathcal{T}}$, then

$$\limsup_{t\in\mathcal{T}\subset\mathbb{N}} \Psi\left(z^t\right) \leq \Psi\left(\overline{z}\right).$$

Now, in order to obtain global convergence of a gradient-like descent sequence for minimizing $\Psi(\cdot)$, we need the objective function $\Psi(\cdot)$ to satisfy an additional property which is semi-algebraicity (for more information on this property and the relation to optimization we refer the reader to [11, 12] and the references therein). It should be noted that here, since we use the weaker version of gradient-like descent sequence (see Definition 4.2), we are able to prove that only the sequence $\{x^t\}_{t\in\mathbb{N}}$ converges globally and not the whole sequence $\{z^t\}_{t\in\mathbb{N}}$ as done in [13]. More precisely, following [13], we propose now the following general result and postpone its proof to the appendix.

Theorem 4.1. *Let $\left\{z^t = \left(x^t, w^t\right)\right\}_{t\in\mathbb{N}}$ be a bounded gradient-like descent sequence for minimizing $\Psi(\cdot)$. If Ψ is semi-algebraic, then the sequence $\{x^t\}_{t\in\mathbb{N}}$ has finite length, i.e., $\sum_{t=1}^{\infty} \left\|x^{t+1} - x^t\right\| < \infty$, and it converges to*

some x^*. In addition, for any limit point w^* of $\{w^t\}_{t\in\mathbb{N}}$, $z^* = (x^*, w^*)$ is a critical point of Ψ.

4.2. **Convergence of SAMBA.** We will show now that $\{(x(t), w(t))\}_{t\in\mathbb{N}}$ generated by SAMBA is a gradient-like descent sequence for minimizing $F_s(\cdot,\cdot)$ (see Definition 4.2).

We start with the following elementary result which will be essential below, and is an immediate consequence of the property of each $H_s^l(\cdot, w)$, for fixed $w \in \mathbb{R}^{mk}$ and $l = 1, 2, \ldots, k$, as established in Lemma 2.1. For completeness we include the proof.

Lemma 4.1. *Fix $w \in \mathbb{R}^{mk}$ and $l = 1, 2, \ldots, k$. For any $u \in \mathbb{R}^n$, and $u^+ \in \mathbb{R}^n$ defined by*

$$u^+ = u - \frac{1}{L_s^l(u,w)} \nabla_x H_s^l(u,w),$$

we have

$$H_s^l\left(u^+, w\right) \leq H_s^l(u,w) - \frac{L_s^l(u,w)}{2}\left\|u^+ - u\right\|^2. \tag{4.1}$$

Proof. Substituting in (2.3) $v = u^+$, yields that

$$\begin{aligned} H_s^l\left(u^+, w\right) &\leq H_s^l(u,w) + \left\langle \nabla_x H_s^l(u,w), u^+ - u \right\rangle + \frac{L_s^l(u,w)}{2}\left\|u^+ - u\right\|^2 \\ &= H_s^l(u,w) - L_s^l(u,w)\left\langle u^+ - u, u^+ - u\right\rangle + \frac{L_s^l(u,w)}{2}\left\|u^+ - u\right\|^2 \\ &= H_s^l(u,w) - \frac{L_s^l(u,w)}{2}\left\|u^+ - u\right\|^2, \end{aligned}$$

which proves the desired result. □

In the forthcoming analysis, we will use the following simple property of block vectors.

Lemma 4.2. *For any $v = \left(v^1, v^2, \ldots, v^p\right) \in \mathbb{R}^{np}$, the following holds:*

$$\|v\| \leq \sum_{j=1}^{p} \left\|v^j\right\| \leq \sqrt{p}\,\|v\|. \tag{4.2}$$

Proof. For any $v = \left(v^1, v^2, \ldots, v^p\right) \in \mathbb{R}^{np}$, we have that $\|v\| = \sqrt{\sum_{j=1}^p \|v^j\|^2}$. The left-hand side inequality in (4.2) follows, with $\alpha_j := \left\|v^j\right\|$, from the obvious fact that $\sum_{j=1}^p \alpha_j^2 \leq \left(\sum_{j=1}^p \alpha_j\right)^2$, which holds for any $\alpha_j \geq 0$, $j = 1, 2, \ldots, p$. To obtain the right-hand side inequality of (4.2), note that for any $\alpha_j \in \mathbb{R}$, we have (by the convexity of $\alpha \to \alpha^2$) that

$$\left(\sum_{j=1}^{p} \alpha_j\right)^2 \leq p \sum_{j=1}^{p} \alpha_j^2,$$

and the desired result follows with $\alpha_j := \left\|v^j\right\|$, $j = 1, 2, \ldots, p$. $\square$

Equipped with the above, the following result shows that condition (C1') holds true, where the descent is measured in terms of the sequence $\{x(t)\}_{t\in\mathbb{N}}$. For simplicity, from now on, we denote $z(t) = (x(t), w(t))$, $t \in \mathbb{N}$.

Proposition 4.1 (Sufficient decrease property)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be a sequence which is generated by SAMBA. Then, there exists $\rho_1 > 0$ such that*

$$\rho_1 \|x(t+1) - x(t)\|^2 \leq F_s(z(t)) - F_s(z(t+1)), \quad \forall\, t \in \mathbb{N}.$$

Proof. The step with respect to w (see step (3.3)) is actually an exact minimization of the function $F_s(x(t), \cdot)$ with respect to each w^l, $l = 1, 2, \ldots, k$, separately. Therefore, we obviously have that

$$F_s(x(t), w(t+1)) \leq F_s(x(t), w(t)), \quad \forall\, t \in \mathbb{N}. \tag{4.3}$$

From step (3.5) and Lemma 4.1 with $u = x^l(t)$ and $u^+ = x^l(t+1)$, we get for all $l = 1, 2, \ldots, k$ that

$$H_s^l\left(x^l(t+1), w(t+1)\right) \leq H_s^l\left(x^l(t), w(t+1)\right) - \frac{L_s^l(t)}{2}\left\|x^l(t+1) - x^l(t)\right\|^2.$$

Summing the last inequality over $l = 1, 2, \ldots, k$, yields

$$\begin{aligned} H_s(z(t+1)) &\leq H_s(x(t), w(t+1)) - \sum_{l=1}^{k} \frac{L_s^l(t)}{2}\left\|x^l(t+1) - x^l(t)\right\|^2 \\ &\leq H_s(x(t), w(t+1)) - \sum_{l=1}^{k} \frac{\bar{\beta}}{2}\left\|x^l(t+1) - x^l(t)\right\|^2 \\ &= H_s(x(t), w(t+1)) - \frac{\bar{\beta}}{2}\|x(t+1) - x(t)\|^2, \end{aligned}$$

where the second inequality follows from Remark 3.1 and the equality follows from the fact that $\|v\|^2 = \sum_{j=1}^{p}\left\|v^j\right\|^2$ for any $v = \left(v^1, v^2, \ldots, v^p\right) \in \mathbb{R}^{np}$. Now, from the definition of $F_s(\cdot,\cdot)$ (see (3.1)) we obtain that

$$F_s(x(t+1), w(t+1)) \leq F_s(x(t), w(t+1)) - \frac{\bar{\beta}}{2}\|x(t+1) - x(t)\|^2. \tag{4.4}$$

The result now follows by combining (4.3) and (4.4). $\square$

Before proving that SAMBA satisfies condition (C2'), we need the following technical result.

Lemma 4.3. *Let $\{z(t)\}_{t\in\mathbb{N}}$ be a bounded sequence, which is generated by SAMBA. For each $t \in \mathbb{N}$, there exist $R > 0$ such that*

$$\|\nabla_x H_s(z(t+1))\| \leq R\|x(t+1) - x(t)\|.$$

Proof. Fix $l = 1, 2, \ldots, k$. Since $\left\{x^l(t)\right\}_{t\in\mathbb{N}}$ is bounded, and $l_s(\cdot)$ is a continuous function (see Definition 2.1(ii)), it follows that $\left\{l_s\left(x^l(t)\right)\right\}_{t\in\mathbb{N}}$ is bounded from above for all $t \in \mathbb{N}$, which together with the boundedness of $\{w(t)\}_{t\in\mathbb{N}}$, implies that $L_s^l(t) := \sum_{i=1}^m w_l^i(t+1) l_s(x^l(t))$ is bounded from above for all $t \in \mathbb{N}$. Thus, there exists $M > 0$ such that $L_s(t) \leq M$. Now, from Lemma 2.1, we obtain that the convex function $H_s^l\left(\cdot, w^l(t+1)\right)$, $t \in \mathbb{N}$, satisfies the classical Descent Lemma [9], and hence $\nabla_x H_s^l\left(\cdot, w^l(t+1)\right)$ is M-Lipschitz continuous, i.e.,

$$\left\|\nabla_x H_s^l\left(x^l(t+1), w^l(t+1)\right) - \nabla_x H_s^l\left(x^l(t), w^l(t+1)\right)\right\| \leq M\left\|x^l(t+1) - x^l(t)\right\|. \tag{4.5}$$

From step (3.5) we have that

$$x^l(t+1) = x^l(t) - \frac{1}{L_s^l(t)}\nabla_x H_s^l\left(p^l(t)\right),$$

where, for simplicity, in this proof we denote $p^l(t) := \left(x^l(t), w(t+1)\right)$. Therefore, using (4.5) we obtain

$$\begin{aligned}
\left\|\nabla_x H_s^l\left(z^l(t+1)\right)\right\| &= \left\|\nabla_x H_s^l\left(z^l(t+1)\right) - \nabla_x H_s^l\left(p^l(t)\right) + \right. \\
&\quad \left. L_s^l(t)\left(x^l(t) - x^l(t+1)\right)\right\| \\
&\leq \left\|\nabla_x H_s^l\left(z^l(t+1)\right) - \nabla_x H_s^l\left(p^l(t)\right)\right\| \\
&\quad + L_s^l(t)\left\|x^l(t+1) - x^l(t)\right\| \\
&\leq M\left\|x^l(t+1) - x^l(t)\right\| + M\left\|x^l(t+1) - x^l(t)\right\| \\
&= 2M\left\|x^l(t+1) - x^l(t)\right\|.
\end{aligned}$$

Thus

$$\begin{aligned}
\|\nabla_x H_s(z(t+1))\| &\leq \sum_{l=1}^k \left\|\nabla_x H_s^l\left(z^l(t+1)\right)\right\| \\
&\leq 2M\sum_{l=1}^k \left\|x^l(t+1) - x^l(t)\right\| \\
&\leq 2M\sqrt{k}\,\|x(t+1) - x(t)\|,
\end{aligned}$$

where the first and the last inequalities follow from Lemma 4.2. This proves the stated result. □

We will now prove that the generated sequence $\{(x(t), w(t))\}_{t\in\mathbb{N}}$ also satisfies condition (C2'), where the bound is measured again in terms of the sequence $\{x(t)\}_{t\in\mathbb{N}}$, as required in Definition 4.2.

Proposition 4.2 (Subgradient lower bound)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be a bounded sequence which is generated by SAMBA. For each $t \in \mathbb{N}$ there exist some $u(t+1) \in \partial F_s(z(t+1))$ and $\rho_2 > 0$ such that*

$$\|u(t+1)\| \leq \rho_2 \|x(t+1) - x(t)\|, \quad \forall\, t \in \mathbb{N}.$$

Proof. Let $t \in \mathbb{N}$. By the definition of $F_s(\cdot,\cdot)$ (see (3.1)) we get

$$\partial F_s(z(t+1)) = \nabla H_s(z(t+1)) + \left(\mathbf{0}, \left[\partial \delta_\Delta\left(w^i(t+1)\right)\right]_{i=1}^m\right),$$

where $\left[v^i\right]_{i=1}^m := \left(v^1, v^2, \ldots, v^m\right)$. Therefore, by using the optimality condition of step (3.3), we have that (with $p(t) := \left(x^l(t), w(t+1)\right)$)

$$\mathbf{0} \in \nabla_{w^i} H_s(p(t)) + \partial \delta_\Delta\left(w^i(t+1)\right).$$

Thus, with

$$u(t+1) := \left(\nabla_x H_s(z(t+1)), \nabla_w H_s(z(t+1)) - \nabla_w H_s(p(t))\right),$$

we have that $u(t+1) \in \partial F_s(z(t+1))$. By the definition of $H_s(\cdot,\cdot)$, we have for all $i = 1, 2, \ldots, m$ that

$$\nabla_{w^i} H_s(z(t+1)) - \nabla_{w^i} H_s(p(t)) = \rho_s^i(x(t+1)) - \rho_s^i(x(t)).$$

Now, from the definition of $\rho_s^i(\cdot)$, $i = 1, 2, \ldots, m$, we obtain

$$\begin{aligned}
\|\nabla_w H_s(z(t+1)) - \nabla_w H_s(p(t))\| &\leq \sum_{i=1}^m \|\nabla_{w^i} H_s(z(t+1)) - \nabla_{w^i} H_s(p(t))\| \\
&= \sum_{i=1}^m \left\|\rho_s^i(x(t+1)) - \rho_s^i(x(t))\right\| \\
&\leq \sum_{i=1}^m \sum_{l=1}^k \left|d_s^i\left(x^l(t+1)\right) - d_s^i\left(x^l(t)\right)\right|, \qquad (4.6)
\end{aligned}$$

where the first and last inequalities use the left-hand side inequality of (4.2), as established in Lemma 4.2.

Now, since $\left\{x^l(t)\right\}_{t\in\mathbb{N}}$, $l = 1, 2, \ldots, k$, is bounded, invoking the classical Lipschitzian property (cf. [27, Theorem 10.4]), for the convex function $d_s^i(\cdot)$, $i = 1, 2, \ldots, m$, there exists $M > 0$ such that

$$\left|d_s^i\left(x^l(t+1)\right) - d_s^i\left(x^l(t)\right)\right| \leq M\left\|x^l(t+1) - x^l(t)\right\|. \qquad (4.7)$$

Therefore, by combining (4.6) and (4.7) we obtain that

$$\begin{aligned}
\|\nabla_w H_s(z(t+1)) - \nabla_w H_s(p(t))\| &\leq M \sum_{i=1}^m \sum_{l=1}^k \left\|x^l(t+1) - x^l(t)\right\| \\
&= mM \sum_{l=1}^k \left\|x^l(t+1) - x^l(t)\right\| \\
&\leq mM\sqrt{k}\,\|x(t+1) - x(t)\|, \qquad (4.8)
\end{aligned}$$

where the last inequality follows from Lemma 4.2 (see the right-hand side inequality of (4.2)). Now, from Lemma 4.3 we obtain that there exists $R > 0$ such that

$$\|\nabla_x H_s(z(t+1))\| \leq R\|x(t+1) - x(t)\|. \tag{4.9}$$

By combining (4.8) and (4.9) we derive

$$\begin{aligned}\|u(t+1)\| &\leq \|\nabla_x H_s(z(t+1))\| + \|\nabla_w H_s(z(t+1)) - \nabla_w H_s(p(t))\| \\ &\leq R\|x(t+1) - x(t)\| + mM\sqrt{k}\|x(t+1) - x(t)\| \\ &= \left(R + mM\sqrt{k}\right)\|x(t+1) - x(t)\|,\end{aligned}$$

which proves the desired result with $\rho_2 = R + mM\sqrt{k}$. □

We have proven, so far, that SAMBA generates a sequence which satisfies conditions (C1') and (C2'). Therefore, all we need now, is to show that condition (C3) holds true. Theorem 4.1 can then be applied directly to obtain the partial global convergence. This is recorded in the following result.

Theorem 4.2. *Let $\{z(t)\}_{t\in\mathbb{N}}$ be a bounded sequence, which is generated by SAMBA and assume that each $d_s^i(\cdot)$, $i = 1, 2, \ldots, m$, is smooth approximation and semi-algebraic. Then, the sequence $\{x(t)\}_{t\in\mathbb{N}}$ has finite length, i.e., $\sum_{t=1}^{\infty}\|x(t+1) - x(t)\| < \infty$, and it converges to some x^*. In addition, for any limit point w^* of $\{w(t)\}_{t\in\mathbb{N}}$, $z^* = (x^*, w^*)$ is a critical point of $F_s(\cdot,\cdot)$.*

Proof. In Propositions 4.1 and 4.2, we have proven that $\{z(t)\}_{t\in\mathbb{N}}$ satisfies conditions (C1') and (C2') of Definition 4.2, respectively. We will now show that condition (C3) also holds true in this case. As assumed, the sequence $\{z(t)\}_{t\in\mathbb{N}}$ is bounded and, therefore, there exists a subsequence $\{z(t_q)\}_{q\in\mathbb{N}}$ which converges to some $\bar{z} = (\bar{x}, \bar{w})$. From the continuity of $H_s(\cdot,\cdot)$ and the facts that $w(t) \in \Delta$, $t \in \mathbb{N}$, and $\bar{w} \in \Delta$ we obtain that

$$\limsup_{t\to\infty} F_s(x(t), w(t)) = \limsup_{t\to\infty} H_s(x(t), w(t)) = H_s(\bar{x}, \bar{w}) = F_s(\bar{x}, \bar{w}).$$

It is easy to check that in our setting, $F_s(\cdot,\cdot)$ is semi-algebraic, and since $\{z(t)\}_{t\in\mathbb{N}}$ is a gradient-like descent sequence for minimizing $F_s(\cdot,\cdot)$ according to Definition 4.2, we get the desired result from Theorem 4.1, as stated. □

5. Smoothing Approaches and SAMBA for Clustering

In this section, we apply our results to the nonsmooth and nonconvex clustering problem which was described in the introduction, namely

$$\min_{x\in\mathbb{R}^{nk}} \sum_{i=1}^{m} \min_{1\leq l\leq k} \left\|x^l - a^i\right\|. \tag{CP-N}$$

Hence, here we have $d^i\left(x^l\right) = \left\|x^l - a^i\right\|$, $i = 1, 2, \ldots, m$ and $l = 1, 2, \ldots, k$, which clearly fits to our general optimization model (NS).

To apply SAMBA and its convergence analysis developed in Section 4, all we need is to identify for each $i = 1, 2, \ldots, m$, an adequate smoothing function $d_s^i(\cdot)$, namely, to show that $d^i \in \mathcal{S}[\mathbb{R}^n, l_s]$. To that end, we will use two very well-known and natural smoothing approaches for the Euclidean norm (see, e.g., [7, 6] and the references therein):

(a) A direct smoothing approach.
(b) Smoothing via Moreau's Envelope.

For simplicity, we drop all the indices. Let $a \in \mathbb{R}^n$. For any $u \in \mathbb{R}^n$ let $d(u) := \|u - a\|$ and for any $s > 0$, let $d_s(\cdot)$ be a convex smoothing function of $d(\cdot)$. Below, we consider two convex smooth functions in $\mathcal{S}[\mathbb{R}^n, l_s]$.

- **Direct Smoothing**

  Consider the following smooth approximation function:

$$d_s(u) := \left(\|u - a\|^2 + s^2\right)^{1/2}. \tag{5.1}$$

  Clearly, for every $u \in \mathbb{R}^n$, we have

$$d(u) \leq d_s(u) \leq d(u) + s, \text{ and } \lim_{s \to 0^+} d_s(u) = d(u).$$

  Moreover, note that $d_s(u) = \psi(u - a)$ with $\psi(z) := \sqrt{\|z\|^2 + s^2}$, and a straightforward computation shows that:

$$\nabla^2 \psi(z) = \frac{\mathbf{I}}{\psi(z)} - \frac{zz^T}{\psi^3(z)} \preceq \frac{\mathbf{I}}{\psi(z)} \preceq s^{-1}\mathbf{I},$$

  where $\mathbf{I}$ is the identity $n \times n$ matrix. Therefore, $\left\|\nabla^2 \psi(z)\right\| \leq s^{-1}$, and hence the function $d_s(\cdot)$ belongs to the class of continuously differentiable functions which have Lipschitz continuous gradient with Lipschitz constant s^{-1}. Therefore, the smoothing function $d_s(\cdot)$ satisfies the premises of Definition 2.1, here with a constant function $l_s(\cdot) = 1/s$.

- **Smoothing via Moreau's Envelope**

  A classical way of smoothing a convex function is obtained via the so-called Moreau envelope [26]; see also [6] for a more recent study and the references therein. We first recall some fundamental results regarding Moreau's envelope and proximal mapping, which will be essential to our discussion. Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper, lower semicontinuous and convex function. For $t > 0$, the *Moreau Envelope* is defined by

$$M_{t\varphi}(u) := \min_{v \in \mathbb{R}^n} \left\{\varphi(v) + \frac{1}{2t}\|v - u\|^2\right\}, \tag{5.2}$$

  and the *proximal mapping* is defined by

$$\text{prox}_{t\varphi}(u) = \text{argmin}_{v \in \mathbb{R}^n} \left\{\varphi(v) + \frac{1}{2t}\|v - u\|^2\right\}. \tag{5.3}$$

It is well-known [26] that the Moreau's envelope $M_{t\varphi}(\cdot)$ is convex and continuously differentiable on $\mathbb{R}^n$, with a $(1/t)$-Lipschitz continuous gradient given by

$$\nabla M_{t\varphi}(u) = \frac{1}{t}\left(u - \operatorname{prox}_{t\varphi}(u)\right). \tag{5.4}$$

Using the Moreau envelope, we thus consider here the following smooth approximation function:

$$d_s(u) = M_{sd}(u),$$

where here $d(x) = \|x - a\|$. Therefore, a simple calculation shows that the Moreau envelope (see (5.2)) and the proximal mapping (see (5.3)) are respectively given by:

$$d_s(u) = M_{sd}(u) = \begin{cases} \|u-a\| - \frac{s}{2}, & \|u-a\| > s, \\ \frac{1}{2s}\|u-a\|^2, & \|u-a\| \le s, \end{cases} \tag{5.5}$$

and

$$\begin{aligned} \operatorname{prox}_{sd}(u) &= \left(1 - \frac{s}{\max\{\|u-a\|, s\}}\right)(u-a) + a \\ &= u - \frac{s(u-a)}{\max\{\|u-a\|, s\}}. \end{aligned} \tag{5.6}$$

Clearly, we thus have

$$d(u) - \frac{s}{2} \le d_s(u) = M_{sd}(u) \le d(u), \text{ and } \lim_{s\to 0^+} d_s(u) = d(u).$$

Moreover, thanks the properties of the Moreau Envelope just alluded above, it follows that $d_s(\cdot)$ is convex and has $1/s$-Lipschitz continuous gradient, and hence $d_s(\cdot)$ satisfies the premises of Definition 2.1(ii) with $l_s(\cdot) = 1/s$.

Both the direct and the Moreau smoothing approaches provide us with a smooth and convex function $d_s(\cdot) \in \mathcal{S}[\mathbb{R}^n, l_s]$, with the same $l_s(\cdot) = 1/s$. Thus, we can activate SAMBA for the clustering problem in both cases. However, recalling the step-size of SAMBA (cf. (3.4)), in both cases, for fixed $w \in \Delta^m$ and all $l = 1, 2, \ldots, k$, we obtain that

$$L_s^l(u, w) = \sum_{i=1}^{m} \frac{w_l^i}{s},$$

which for small values of the smoothing parameter s, can be a large number. To overcome this potential difficulty, we will show below that in both cases, we can improve the situation by using a *dynamic* and smaller $L_s^l(u, w)$, which takes into consideration the data-set of the clustering problem, resulting in a larger and better step-size in SAMBA for the clustering problem. Moreover, thanks to the analysis developed below, we will also show that the resulting update for the cluster center in SAMBA admits a delightful

convex combination formula of the data-set, similar to the usual k-means scheme based on the squared Euclidean norm.

5.1. **Dynamic Step-Size for the Smooth Approximation Function.** The next result establishes a *descent-like lemma*, however, where the usual Lipschitz constant is replaced by a function.

Lemma 5.1. *For any $s > 0$, let $d_s(\cdot)$ be the smooth approximation function obtained via direct (see (5.1)) or Moreau (see (5.5)) smoothing techniques. Then, for all $u, v \in \mathbb{R}^n$, we have*

$$d_s(v) \leq d_s(u) + \langle \nabla d_s(u), v - u \rangle + \frac{l_s(u)}{2} \|v - u\|^2,$$

where

(a) *for direct smoothing:*

$$l_s(u) = \frac{1}{\left(\|u - a\|^2 + s^2\right)^{1/2}}, \tag{5.7}$$

(b) *for Moreau smoothing:*

$$l_s(u) = \frac{1}{\max\{\|u - a\|, s\}}.$$

Proof. We first prove the Case (a), that is, for the direct smoothing. Define the following auxiliary function $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$h(v, u) := \frac{\|v - a\|^2 + s^2}{\left(\|u - a\|^2 + s^2\right)^{1/2}} = \frac{d_s^2(v)}{d_s(u)}.$$

Then, clearly $h(u, u) = d_s(u)$, and it easily seen that $\nabla_v h(u, u) = 2\nabla d_s(u)$. Invoking, for all $\alpha \in \mathbb{R}$ and $\beta > 0$, the basic fact:

$$\frac{\alpha^2}{\beta} \geq 2\alpha - \beta, \tag{5.8}$$

we obtain

$$h(v, u) = \frac{d_s^2(v)}{d_s(u)} \geq 2d_s(v) - d_s(u).$$

In addition, the function $v \mapsto h(v, u)$ is quadratic with the associated matrix $l_s(u)\mathbf{I}$. Therefore, its second-order Taylor expansion around u leads to the following identity

$$h(v, u) = h(u, u) + \langle \nabla_v h(u, u), v - u \rangle + l_s(u) \|v - u\|^2,$$

and using the facts proven above, the desired result follows.

Now we prove the Case (b), i.e., for the Moreau Smoothing. In a similar vein, define the following auxiliary function $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$h(v, u) = \begin{cases} \frac{\|v-a\|^2}{\|u-a\|} - \frac{s}{2}, & \|u - a\| > s, v \in \mathbb{R}^n, \\ \frac{1}{2s} \|v - a\|^2, & \|u - a\| \leq s, v \in \mathbb{R}^n. \end{cases}$$

Then, clearly $h(u,u) = M_{sd}(u) \equiv d_s(u)$, and it is easily seen that

$$\nabla_v h(u,u) = \begin{cases} \frac{u-a}{d(u)}, & \|u-a\| > s, \\ \frac{u-a}{s}, & \|u-a\| \leq s. \end{cases}$$

When, $\|u-a\| \leq s$ and $v \in \mathbb{R}^n$, since $d_s(u) = (1/2s)\|u-a\|^2$, we immediately obtain

$$d_s(v) = d_s(u) + \langle \nabla d_s(u), v-u \rangle + \frac{1}{s}\|v-u\|^2.$$

For the other case, namely when $\|u-a\| > s$ and $v \in \mathbb{R}^n$, invoking again (5.8), it follows that

$$\begin{aligned} h(v,u) &\geq 2\|v-a\| - \|u-a\| - \frac{s}{2} \\ &= 2\left(\|v-a\| - \frac{s}{2}\right) - \left(\|u-a\| - \frac{s}{2}\right) = 2d_s(v) - d_s(u). \end{aligned}$$

Moreover, since the function $v \mapsto h(v,u)$ is quadratic, with the associated matrix $(1/d(u))\mathbf{I}$, then like in the Case (a), its second-order Taylor expansion around u leads to an identity, and combining all the established facts, the desired result follows. □

We can now develop SAMBA where the smoothing functions $d_s^i(\cdot)$, $i = 1,2,\ldots,m$, are defined either trough the direct smoothing or the Moreau smoothing. SAMBA for the clustering problem reads as follows:

SAMBA for Clustering

(1) Input: Data set $\mathcal{A}$, $s > 0$ and the function $d_s(\cdot)$ given in (5.1) or (5.5).
(2) Initialization: Start with $x(0) \in \mathbb{R}^{nk}$.
(3) General step ($t = 0,1,\ldots$):
  (3.1) Cluster assignment: for all $i = 1,2,\ldots,m$ compute

$$w^i(t+1) = \operatorname{argmin}\left\{\left\langle w^i, \rho_s^i(x(t))\right\rangle : w^i \in \Delta\right\}. \tag{5.9}$$

  (3.2) Center update: for each $l = 1,2,\ldots,k$ compute

$$x^l(t+1) = \sum_{i=1}^{m} \lambda_i^l(t)\, a^i,$$

where λ_i^l, $i = 1,2,\ldots,m$, is computed through $l_s(\cdot)$ which is given via either one of the formulas (5.10) or (5.11) below.

As can be seen, SAMBA for clustering, which solves problem (CP-N) shares the pattern of most center-based clustering algorithms, namely, this algorithm alternates between cluster assignment (the subproblem with respect to w) and the centers update (the subproblem with respect to x). In fact, we will now show that the x-step is indeed like in the *k-means* scheme, which means a convex combination of the data points in $\mathcal{A}$. This will, in

turn, prove that SAMBA for clustering generates a bounded sequence, as recorded in the following result, a property which is needed to apply the global convergence result (see Theorem 4.2) in this case.

Lemma 5.2. *The sequence* $\{(x(t), w(t))\}_{t\in\mathbb{N}}$, *generated by SAMBA for clustering, is bounded. In particular, the* x*-step reads, for each* $l = 1, 2, \ldots, k$, *as follows*

$$x^l(t+1) = \sum_{i=1}^{m} \lambda_i^l(t)\, a^i, \text{ with } \lambda_i^l(t) = \frac{w_l^i(t+1)}{L_s^l(t)} l_s\left(x^l(t)\right), \quad i = 1, 2, \ldots, m,$$

with $L_s^l(t) := \sum_{i=1}^m w_l^i(t+1)\, l_s\left(x^l(t)\right)$, *and where for* $i = 1, 2, \ldots, m$ *and each* $l = 1, 2, \ldots, k$,

(a) *for direct smoothing:*

$$l_s\left(x^l(t)\right) := \frac{1}{\left(\|x^l(t) - a^i\|^2 + s^2\right)^{1/2}} \equiv \frac{1}{d_s\left(x^l(t)\right)}, \tag{5.10}$$

(b) *for Moreau smoothing:*

$$l_s\left(x^l(t)\right) := \frac{1}{\max\{\|x^l(t) - a^i\|, s\}}. \tag{5.11}$$

Proof. It is straightforward from the algorithm that $w^i(t) \in \Delta$ for all $t \in \mathbb{N}$ and $i = 1, 2, \ldots, m$. Therefore, we obviously have that $\{w(t)\}_{t\in\mathbb{N}}$ is bounded. In addition, we will now show that $x(t)$, $t \in \mathbb{N}$, can be written as a weighted arithmetic mean of the data points in $\mathcal{A}$, which completes the boundedness assertion of the sequence $\{(x(t), w(t))\}_{t\in\mathbb{N}}$. Indeed, for any $l = 1, 2, \ldots, k$, we have by definition of the x-step in SAMBA:

$$x^l(t+1) = x^l(t) - \frac{1}{L_s^l(t)} \nabla H_s^l\left(x^l(t), w(t+1)\right)$$

where we recall that by definition, for $w \in \Delta^m$ and all $l = 1, 2, \ldots, k$, we have that (cf (2.2)):

$$H_s^l(u, w) := \sum_{i=1}^{m} w_l^i d_s^i(u) \quad \text{and} \quad L_s^l(u, w) = \sum_{i=1}^{m} w_l^i l_s(u). \tag{5.12}$$

We will now show that in both Cases (a) and (b) the cluster centers admit a convex combination representation of the data-set, with the corresponding convex weights. First, we note that for every $i = 1, 2, \ldots, m$ we have for Case (a) the following:

$$\nabla d_s^i(u) = \frac{u - a^i}{d_s(u)} = \left(u - a^i\right) l_s(u) \quad \text{with} \quad l_s(u) := \frac{1}{d_s(u)},$$

and for Case (b), from (5.6), we have the following

$$\nabla d_s^i(u) = \frac{u - \operatorname{prox}_{sd^i}(u)}{s} = \frac{u - a^i}{\max\{\|u - a^i\|, s\}} = \left(u - a^i\right) l_s(u),$$

with

$$l_s(u) := \frac{1}{\max\{\|u - a^i, s\}\|}.$$

Thus, for both cases we obtain:

$$\nabla H_s^l\left(x^l(t), w(t+1)\right) = \sum_{i=1}^{m} w_l^i(t+1)\left(x^l(t) - a^i\right) l_s\left(x^l(t)\right),$$

and

$$L_s^l(t) \equiv L_s^l\left(x^l(t), w(t+1)\right) = \sum_{i=1}^{m} w_l^i(t+1)\, l_s\left(x^l(t)\right),$$

where $l_s\left(x^l(t)\right)$ are respectively given by (5.10) and (5.11). Therefore, we obtain for both cases:

$$\begin{aligned} x^l(t+1) &= x^l(t) - \frac{1}{L_s^l(t)} \sum_{i=1}^{m} w_l^i(t+1)\left(x^l(t) - a^i\right) l_s\left(x^l(t)\right) \\ &= \frac{1}{L_s^l(t)} \sum_{i=1}^{m} w_l^i(t+1)\, l_s\left(x^l(t)\right) a^i \in \mathrm{Conv}(\mathcal{A}), \end{aligned}$$

and this complete the proof of the desired result with the corresponding convex weights λ_l^i for both Cases (a) and (b). □

Equipped the above results, we immediately obtain the following convergence result of SAMBA for clustering.

Theorem 5.1. *Let $\{z(t)\}_{t\in\mathbb{N}}$ be a sequence which is generated by SAMBA for clustering with either one of the smooth function* (5.1) *or* (5.5)*. Then, the sequence $\{x(t)\}_{t\in\mathbb{N}}$ has finite length, i.e., $\sum_{t=1}^{\infty} \|x(t+1) - x(t)\| < \infty$, and it converges to some x^*. In addition, for any limit point w^* of $\{w(t)\}_{t\in\mathbb{N}}$, $z^* = (x^*, w^*)$ is a critical point of $F_s(\cdot,\cdot)$.*

6. Numerical Results

In this section we report numerical experiments comparing the performance of the two variants of SAMBA for to the clustering problem, as developed in the previous section. For simplicity, from now on, SAMBA with direct smoothing will be denoted below by SAMBA - D while SAMBA with Moreau's smoothing is denoted by SAMBA - M.

We consider four data-sets taken from the UC Irvine Machine Learning Repository [15], where in Table 1 we summarize the parameters of every data-set, namely the number of data points m, their corresponding dimension n and the number of clusters k.

Since our data-sets are in higher dimensions than 2D or 3D, visualization of the clustering results are not possible. Therefore, we will follow some classical mathematical indices for evaluating the clustering performances of the two algorithms (see [33] for details and more information about these

Dataset Name	m	n	k
Glass (Gl)	214	9	6
Iris (Ir)	150	4	3
Salary (Sa)	30162	14	2
Seeds (Se)	210	7	3

TABLE 1. Parameters of the UCI Machine Learning Repository databases.

aspects). We computed the Rand, Jaccard (Jacc), Purity-Efficiency (PE) and Variation of Information (VI) indices, which are so-called *external* measures. In addition, we also computed the Davies-Bouldin (BD) index, that is so-called *internal* measure. It should be noted that the indices Rand, Jacc and PE should be maximized, i.e., clustering result with higher value for these indices is better. Whereas, the VI and the DB indices should be minimized.

Experiments Information. For each data-set, we executed SAMBA - D and SAMBA - M, 100 times with different initial centers $x(0)$. In each execution we have computed all the five indices, where in Tables (2) (for the maximizing indices) and (3) (for the minimizing indices) we present the average results over the 100 tests. Since these two algorithms also depend on the value of the smoothing parameter s, we have repeated all these experiments with four different values of this parameter, which are $10, 1, 0.1$ and 0.01. It should be noted that we normalized each feature in each data-set by subtracting its mean and dividing by its standard deviation.

To conclude, clearly, for larger values of the smoothing parameter ($s = 10$ and $s = 1$), the variant of SAMBA with direct smoothing approach (SAMBA - D) is better in the majority of cases. However, for small smoothing parameter ($s = 0.1$), the SAMBA with Moreau's smoothing achieves slightly better results. In the case of the smallest smoothing parameter ($s = 0.01$), both algorithms perform the same way.

7. Appendix: Proof of Theorem 4.1

In order to prove Theorem 4.1, we will need first to prove the following result which shows that conditions (C1') and (C2') are enough to guarantee that a gradient-like descent sequence is subsequently converges to a point in $\operatorname{crit}\Psi$, that is, a critical point of Ψ. For simplicity, we denote the set of all limit points of $\{z^t\}_{t\in\mathbb{N}}$ by $\omega\left(z^0\right)$.

The next result establishes the promised subsequential convergence.

Lemma 7.1 (Subsequence Convergence)**.** *Let $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a lower semicontinuous function and let $\left\{z^t = \left(x^t, w^t\right)\right\}_{t\in\mathbb{N}}$ be a bounded*

		Rand		Jacc		PE	
		S - D	S - M	S - D	S - M	S - D	S - M
$s = 10$	Gl	0.6714	**0.6718**	0.2479	**0.2502**	0.5671	**0.5714**
	Ir	**0.8087**	0.8075	**0.5731**	0.5718	**1.0415**	1.0404
	Sa	**0.5584**	0.5573	**0.4614**	0.4605	**0.9099**	0.9086
	Se	**0.9015**	0.9014	**0.7407**	0.7405	**1.2035**	1.2033
$s = 1$	Gl	0.6814	**0.6818**	**0.2236**	0.2212	**0.5238**	0.5194
	Ir	**0.8017**	0.8016	**0.5681**	0.568	1.0395	1.0395
	Sa	0.5525	**0.5528**	**0.4624**	0.4618	**0.9095**	0.9086
	Se	**0.905**	0.8995	**0.7486**	0.7359	**1.2108**	1.1991
$s = 0.1$	Gl	0.6811	**0.6813**	0.2202	**0.2203**	0.5204	**0.5206**
	Ir	0.7999	**0.8**	0.5705	**0.5706**	1.0449	1.0449
	Sa	0.5551	0.5551	0.4657	0.4657	0.9139	0.9139
	Se	0.8886	0.8886	0.7121	0.7121	1.1759	1.1759
$s = 0.01$	Gl	0.6783	0.6783	0.2206	0.2206	0.5199	0.5199
	Ir	0.8034	0.8034	0.5734	0.5734	1.0453	1.0453
	Sa	0.552	0.552	0.4599	0.4599	0.9058	0.9058
	Se	0.889	0.889	0.7142	0.7142	1.1779	1.1779

TABLE 2. Values of the five indices for clustering produced by SAMBA - D and SAMBA - M. Each run is halted after 50 iterations, which is enough in most cases to achieve stable values of x and w. The best results in each data-set and each index are marked by boldface.

gradient-like descent sequence for minimizing Ψ*. Then,* $\omega\left(z^0\right)$ *is a nonempty and compact subset of* $\operatorname{crit}\Psi$*, and we have*

$$\lim_{t\to\infty} \operatorname{dist}\left(z^t, \omega\left(z^0\right)\right) = 0. \tag{7.1}$$

In addition, the objective function Ψ *is finite and constant on* $\omega\left(z^0\right)$*.*

Proof. Since $\{z^t\}_{t\in\mathbb{N}}$ is bounded, there is $z^* \in \mathbb{R}^d$ and a subsequence $\left\{z^{t_q}\right\}_{q\in\mathbb{N}}$ such that $z^{t_q} \to z^*$ as $q \to \infty$ and hence $\omega\left(z^0\right)$ is nonempty. Moreover, the set $\omega\left(z^0\right)$ is compact, since it can be viewed as an intersection of compact sets. Now, from condition (C3) and the lower semicontinuity of Ψ, we obtain

$$\lim_{q\to\infty} \Psi\left(z^{t_q}\right) = \Psi\left(z^*\right). \tag{7.2}$$

On the other hand, from conditions (C1') and (C2'), we know that there is $u^t \in \partial\Psi\left(z^t\right)$, $t \in \mathbb{N}$, such that $u^t \to 0$ as $t \to \infty$. The closedness property[1] of $\partial\Psi$ implies thus that $0 \in \partial\Psi\left(z^*\right)$. This proves that z^* is a critical point of Ψ, and hence (7.1) is valid.

[1]Let $\left\{\left(q^t, p^t\right)\right\}_{t\in\mathbb{N}}$ be a sequence in $\operatorname{graph}(\partial\Psi)$ that converges to (q,p) as $t \to \infty$. By the very definition of $\partial\Psi(q)$, if $\Psi\left(q^t\right)$ converges to $\Psi(q)$ as $t \to \infty$, then $(q,p) \in \operatorname{graph}(\partial\Psi)$.

		VI		DB	
		S - D	S - M	S - D	S - M
$s = 10$	Gl	2.023	**2.0108**	1.1907	**1.172**
	Ir	**0.7552**	0.756	0.8129	**0.8125**
	Sa	**0.9958**	0.996	**2.5411**	2.544
	Se	**0.5862**	0.5868	**0.9226**	0.9227
$s = 1$	Gl	**2.1696**	2.1842	**1.3145**	1.3376
	Ir	0.7428	**0.7427**	0.8001	**0.8**
	Sa	**1.0012**	1.0016	2.5371	**2.5369**
	Se	**0.5735**	0.5968	**0.9271**	0.9279
$s = 0.1$	Gl	2.2014	2.2014	**1.3628**	1.3634
	Ir	0.7247	**0.7245**	0.7918	0.7918
	Sa	0.9984	0.9984	2.5207	2.5207
	Se	0.6429	**0.6428**	0.9323	0.9323
$s = 0.01$	Gl	**2.1981**	2.1983	1.3468	1.3468
	Ir	0.7311	0.7311	0.8004	0.8004
	Sa	1.0055	1.0055	2.5647	2.5647
	Se	0.6373	0.6373	0.9365	0.9365

TABLE 3. Values of the five indices for clustering produced by SAMBA - D and SAMBA - M. Each run is halted after 50 iterations, which is enough in most cases to achieve stable values of x and w. The best results in each data-set and each index are marked by boldface.

To complete the proof, let $\lim_{t\to\infty} \Psi\left(z^t\right) = \alpha \in \mathbb{R}$. Then $\left\{\Psi\left(z^{t_q}\right)\right\}_{q\in\mathbb{N}}$ converges to α and from (7.2) we have that $\Psi\left(z^*\right) = \alpha$. Hence the restriction of Ψ to $\omega\left(z^0\right)$ equals α. □

To achieve our main goal, i.e., to establish global convergence of the *whole* sequence, we need an additional assumption on the class of functions Ψ: it must satisfy the so-called nonsmooth Kurdyka-Łojasiewicz (KL) property [11] (see [23, 24] for smooth cases). We refer the reader to [12] for an in depth study of the class of KL functions, as well as references therein. We now provide the formal definition of the KL property and two important results.

Denote $[\alpha < \Psi < \beta] := \left\{z \in \mathbb{R}^d : \ \alpha < \Psi\left(z\right) < \beta\right\}$. Let $\eta > 0$, and set

$$\Phi_\eta = \left\{\varphi \in C^0[0,\eta) \cap C^1(0,\eta) : \ \varphi\left(0\right) = 0, \varphi \text{ concave and } \varphi' > 0\right\}.$$

Definition 7.1 (The nonsmooth KL property)**.** A proper and lower semi-continuous function $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$ has the Kurdyka-Łojasiewicz (KL)

property locally at $\overline{z} \in \operatorname{dom} \Psi$, if there exist $\eta > 0$, $\varphi \in \Phi_\eta$, and a neighborhood $U(\overline{z})$ such that

$$\varphi'(\Psi(z) - \Psi(\overline{z})) \operatorname{dist}(0, \partial\Psi(z)) \geq 1,$$

for all $z \in U(\overline{z}) \cap [\Psi(\overline{z}) < \Psi(z) < \Psi(\overline{z}) + \eta]$.

Verifying the KL property of a given function might often be a difficult task. However, thanks to a fundamental result established in [11], it holds for the broad class of *semi-algebraic* functions.

Theorem 7.1. *Let $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. If Ψ is semi-algebraic, then it satisfies the KL property at any point of* $\operatorname{dom} \Psi$.

Our last ingredient is a key uniformization of the KL property proven in [13, Lemma 6, p. 478], which we record below.

Lemma 7.2 (Uniformized KL Property)**.** *Let Ω be a compact set and let $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. Assume that Ψ is constant on Ω and satisfies the KL property at each point of Ω. Then, there exist $\varepsilon > 0$, $\eta > 0$ and $\varphi \in \Phi_\eta$ such that for all $\overline{z}$ in Ω one has,*

$$\varphi'(\Psi(z) - \Psi(\overline{z})) \operatorname{dist}(0, \partial\Psi(z)) \geq 1, \tag{7.3}$$

and all $z \in \left\{x \in \mathbb{R}^d : \operatorname{dist}(x, \Omega) < \varepsilon\right\} \cap [\Psi(\overline{z}) < \Psi(z) < \Psi(\overline{z}) + \eta]$.

We can now prove the following abstract convergence result, which is a weaker variant of the global convergence result obtained in [13] (see also [14] for another variant of the abstract convergence result, which was designed for analyzing Lagrangian-based methods).

Theorem 7.2 (Global Convergence)**.** *Let $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a lower semicontinuous function and let $\left\{z^t = \left(x^t, w^t\right)\right\}_{t\in\mathbb{N}}$ be a bounded gradient-like descent sequence for minimizing Ψ. If Ψ is semi-algebraic, then the sequence $\{x^t\}_{t\in\mathbb{N}}$ has finite length, i.e., $\sum_{t=1}^{\infty} \left\|x^{t+1} - x^t\right\| < \infty$, and it converges to some x^*. In addition, for any limit point w^* of $\{w^t\}_{t\in\mathbb{N}}$, $z^* = (x^*, w^*)$ is a critical point of Ψ.*

Proof. Since $\{z^t\}_{t\in\mathbb{N}}$ is bounded, there exists a subsequence $\left\{z^{t_q}\right\}_{q\in\mathbb{N}}$ such that $z^{t_q} \to \overline{z}$ as $q \to \infty$. In a similar way as in Lemma 7.1, we get that

$$\lim_{t\to\infty} \Psi\left(z^t\right) = \Psi(\overline{z}). \tag{7.4}$$

If there exists an integer $\bar{t}$ for which $\Psi\left(z^{\bar{t}}\right) = \Psi(\overline{z})$, then condition (C1') would imply that $z^{\bar{t}+1} = z^{\bar{t}}$. A trivial induction then shows that the sequence $\{z^t\}_{t\in\mathbb{N}}$ is stationary and the announced results are obvious. Since $\left\{\Psi\left(z^t\right)\right\}_{t\in\mathbb{N}}$ is a nonincreasing sequence, it is clear from (7.4) that $\Psi(\overline{z}) < \Psi\left(z^t\right)$ for all $t > 0$. Again from (7.4) for any $\eta > 0$ there exists a nonnegative integer t_0 such that $\Psi\left(z^t\right) < \Psi(\overline{z}) + \eta$ for all $t > t_0$. From Lemma 7.1, we know that $\lim_{t\to\infty} \operatorname{dist}\left(z^t, \omega\left(z^0\right)\right) = 0$. This means that for any $\varepsilon > 0$,

there exists a positive integer t_1 such that $\operatorname{dist}\left(z^t, \omega\left(z^0\right)\right) < \varepsilon$ for all $t > t_1$.

From Lemma 7.1, we know that $\omega\left(z^0\right)$ is nonempty and compact, the function Ψ is finite and constant on $\omega\left(z^0\right)$. Hence, we can apply the Uniformization Lemma (see Lemma 7.2) with $\Omega = \omega\left(z^0\right)$. Therefore, for any $t > \bar{t} := \max\{t_0, t_1\}$, we have

$$\varphi'\left(\Psi\left(z^t\right) - \Psi\left(\overline{z}\right)\right) \operatorname{dist}\left(0, \partial\Psi\left(z^t\right)\right) \geq 1. \tag{7.5}$$

This makes sense, since we know that $\Psi\left(z^t\right) > \Psi\left(\overline{z}\right)$ for any $t > \bar{t}$. From condition (C2'), we get that

$$\varphi'\left(\Psi\left(z^t\right) - \Psi\left(\overline{z}\right)\right) \geq \frac{1}{\rho_2}\left\|x^t - x^{t-1}\right\|^{-1}. \tag{7.6}$$

For convenience, we define for all $p, q \in \mathbb{N}$ and $\overline{z}$ the following quantity

$$\Delta_{p,q} := \varphi\left(\Psi\left(z^p\right) - \Psi\left(\overline{z}\right)\right) - \varphi\left(\Psi\left(z^q\right) - \Psi\left(\overline{z}\right)\right).$$

From the concavity of φ we get that

$$\Delta_{t,t+1} \geq \varphi'\left(\Psi\left(z^t\right) - \Psi\left(\overline{z}\right)\right)\left(\Psi\left(z^t\right) - \Psi\left(z^{t+1}\right)\right). \tag{7.7}$$

Combining condition (C1') with (7.6) and (7.7) yields, for any $t > \bar{t}$, that

$$\Delta_{t,t+1} \geq \frac{\left\|x^{t+1} - x^t\right\|^2}{\rho\left\|x^t - x^{t-1}\right\|}, \quad \text{where } \rho := \rho_2/\rho_1. \tag{7.8}$$

Using the fact that $2\sqrt{\alpha\beta} \leq \alpha + \beta$ for all $\alpha, \beta \geq 0$, we infer from the later inequality that

$$2\left\|x^{t+1} - x^t\right\| \leq \left\|x^t - x^{t-1}\right\| + \rho\Delta_{t,t+1}. \tag{7.9}$$

Summing up (7.9) for $i = \bar{t}+1, \ldots, t$ yields

$$\begin{aligned}
2\sum_{i=\bar{t}+1}^{t}\left\|x^{i+1} - x^i\right\| &\leq \sum_{i=\bar{t}+1}^{t}\left\|x^i - x^{i-1}\right\| + \rho\sum_{i=\bar{t}+1}^{t}\Delta_{i,i+1} \\
&\leq \sum_{i=\bar{t}+1}^{t}\left\|x^{i+1} - x^i\right\| + \left\|x^{\bar{t}+1} - x^{\bar{t}}\right\| + \rho\sum_{i=\bar{t}+1}^{k}\Delta_{i,i+1} \\
&= \sum_{i=\bar{t}+1}^{t}\left\|x^{i+1} - x^i\right\| + \left\|x^{\bar{t}+1} - x^{\bar{t}}\right\| + \rho\Delta_{\bar{t}+1,t+1},
\end{aligned}$$

where the last inequality follows from the fact that $\Delta_{p,q} + \Delta_{q,r} = \Delta_{p,r}$ for all $p, q, r \in \mathbb{N}$. Since $\varphi \geq 0$, recalling the definition of $\Delta_{\bar{t}+1,t+1}$, we thus have for any $k > l$ that

$$\sum_{i=\bar{t}+1}^{t}\left\|x^{i+1} - x^i\right\| \leq \left\|x^{\bar{t}+1} - x^{\bar{t}}\right\| + \rho\varphi\left(\Psi\left(z^{\bar{t}+1}\right) - \Psi\left(\overline{z}\right)\right),$$

which implies that $\sum_{t=1}^{\infty} \left\|x^{t+1} - x^t\right\| < \infty$, i.e., $\{x^t\}_{t\in\mathbb{N}}$ is a Cauchy sequence and hence converges to some x^*. Let w^* be a limit point of $\{t^k\}_{k\in\mathbb{N}}$, then from Lemma 7.1 we obtain that $z^* = (x^*, w^*)$ is a critical point of Ψ. □

Acknowledgment. The research of Marc Teboulle was partially supported by the Israel Science Foundation, under ISF Grant 1844-16 and the German-Israel Foundation, under Grant GIF-GI1253304.6-2014.

References

[1] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1-2, Ser. B):5–16, 2009.
[2] A. Auslender. *Optimisation: Méthodes numériques*. Masson, Paris, 1976.
[3] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.
[4] A. Beck and S. Sabach. Weiszfeld's method: old and new results. *J. Optim. Theory Appl.*, 164(1):1–40, 2015.
[5] A. Beck, S. Sabach, and M. Teboulle. An alternating semiproximal method for nonconvex regularized structured total least squares problems. *SIAM J. Matrix Anal. Appl.*, 37(3):1129–1150, 2016.
[6] A. Beck and M. Teboulle. Smoothing and first order methods: a unified framework. *SIAM J. Optim.*, 22(2):557–580, 2012.
[7] A. Ben-Tal and M. Teboulle. A smoothing technique for nondifferentiable optimization problems. In *Optimization. Lecture Notes in Mathematics, vol. 1405.*, pages 1–11. Springer Berlin Heidelberg, Berlin, 1989.
[8] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Computer Science and Applied Mathematics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1982.
[9] D. P. Bertsekas. *Nonlinear Programming*. Belmont MA: Athena Scientific, second edition, 1999.
[10] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York-London, 1981.
[11] J. Bolte, A. Daniilidis, and A. Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4):1205–1223, 2006.
[12] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362(6):3319–3363, 2010.
[13] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2, Ser. A):459–494, 2014.
[14] J. Bolte, S. Sabach, and M. Teboulle. Nonconvex Lagrangian-Based Optimization: Monitoring Schemes and Global Convergence *Math. Oper. Res.*, 2018.
[15] D. Dheeru and E. K. Taniskidou. UCI machine learning repository, 2017.
[16] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley-Interscience, New York, second edition, 2001.
[17] E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21(3):768–769, 1965.
[18] R. Hesse, D. R. Luke, S. Sabach, and M. Tam. The proximal heterogeneous block implicit-explicit method and application to blind ptychographic imaging. *SIAM J. Imaging Sci.*, 8(1):426–457, 2015.
[19] C. Iyigun and A. Ben-Israel. Probabilistic D-clustering. *J. Classification*, 25(1):5–26, 2008.

[20] C. Iyigun and A. Ben-Israel. A generalized Weiszfeld method for the multi-facility location problem. *Oper. Res. Lett.*, 38(3):207–214, 2010.
[21] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.
[22] J. Kogan, C. Nicholas, and M. Teboulle, editors. *Grouping Multidimensional Data: Recent Advances in Clustering*. Springer Berlin Heidelberg, Berlin, 2006.
[23] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)*, 48(3):769–783, 1998.
[24] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles (Paris, 1962)*, pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.
[25] D. S. Modha and W. S. Spangler. Feature weighting in k-means clustering. *Mach. Learn.*, 52(3):217–237, 2003.
[26] J. J. Moreau. Proximitéet dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.
[27] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.
[28] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.
[29] H. Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci. Cl. III.*, 4:801–804, 1956.
[30] M. Teboulle. A unified continuous optimization framework for center-based clustering methods. *J. Mach. Learn. Res.*, 8:65–102, 2007.
[31] 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*, pages 127–160. Springer Berlin Heidelberg, Berlin, 2006.
[32] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal*, 43:355–386, 1937.
[33] D. Xu and Y. Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

(S. Sabach) Faculty of Industrial Engineering, The Technion, Haifa, 32000, Israel
E-mail address: ssabach@ie.technion.ac.il

(M. Teboulle) School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978
E-mail address: teboulle@post.tau.ac.il

(S. Voldman) Faculty of Industrial Engineering, The Technion, Haifa, 32000, Israel
E-mail address: sergeyv@campus.technion.ac.il.