Tel-Aviv University
Raymond and Beverly Sackler Faculty of Exact Sciences

A Novel Class of Globally Convergent Algorithms For Clustering Problems

Thesis submitted in partial fulfillment of graduation requirements
for a degree of M.Sc. at Tel-Aviv University

Tel-Aviv University
School of Mathematical Sciences
Department of Statistics and Operations Research

By
Sergey Voldman

The research work in this thesis has been conducted
under the supervision of Prof. Marc Teboulle
and Prof. Shoham Sabach

February 2016

Abstract

The clustering problem is one of the fundamental problems in unsupervised machine learning, and arises in a wide scope of applications. The clustering problem is a nonconvex and nonsmooth optimization problem. We propose two clustering center-based algorithms, each tackles a different distance function. We prove the global convergence of these algorithm to a critical point via a new methodology which is based on the powerful Kurdyka-Łojasiewicz property. As an illustration of the theoretical results, we present numerical tests which demonstrate the effectiveness of the proposed algorithms.

Acknowledgements

I would like to thank my advisor, Prof. Marc Teboulle, for introducing me to the interesting world of continuous optimization during my graduate studies. I am deeply grateful to Prof. Teboulle for motivating and encouraging me to research the clustering problem and for his helpful ideas how to tackle this interesting and important problem.

I would like to acknowledge my co-advisor, Prof. Shoham Sabach, for his insightful comments and constructive criticisms at different stages of my research and for his dedication and patience in helping me to complete this thesis.

None of this would have been possible without the love and support of my family, to whom this thesis is dedicated to. My family has been a constant source of love, concern, support and strength.

Finally, I appreciate the financial support from Tel-Aviv university which enabled me to focus on my research, and achieve this thesis.

February 2016
Sergey Voldman

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter presents the importance of the clustering problem and describes the usefulness of clustering in many applications. Moreover, we describe the existing types and approaches of clustering, and review the literature on the most popular clustering center-based algorithms. We discuss the contributions of this thesis, mainly in developing new algorithms and the use of a novel methodology to prove their convergence results.

1.1 Background and Motivation

The clustering problem is a task of grouping objects which are similar. It consists of partitioning a dataset into subsets, called clusters, such that the data points in each cluster are similar with respect to a specific criteria.

The clustering problem is a fundamental problem in the machine learning field, and it arises in a wide scope of applications, such as data mining, pattern recognition, information retrieval and many others. For example, in image segmentation, one is interested in partitioning the pixels of an image into objects, where each pixel can be described via its location in the image and its color given in RGB format. Another example is learning the probability density of some data, where the data is assumed to be drawn from a mixtures of distributions. Each partition of the data is represented by a unimodal probability density model, and a summation of all the cluster models gives a multimodal density for the entire dataset. Vector quantization is yet another example, where large sets of points are represented by their centroid point. This approach can be used for data compression, data correction and pattern recognition.

There are several categories of clustering methods, each has a direct impact on the final clustering structure.

(i) Hierarchical versus partitioning clustering. In partitioning clustering the dataset is divided into clusters, whereas in hierarchical clustering each cluster may have sub-clusters, thus forming a tree which leaves are single points of the dataset.

(ii) Hard versus soft and fuzzy clustering. In hard clustering each data point is assigned to single cluster, versus a soft clustering where each point may be assigned to more than one cluster, hence clusters may overlap. In fuzzy clustering for each point there is a distribution that describes the probability of a point to be part of a certain cluster.

(iii) Complete versus partial clustering. In complete clustering all points in the dataset are assigned to clusters, whereas in partial clustering some points may be intentionally skipped and are not being assigned to a cluster.

Finding the optimal partition of a fixed number of clusters for some given dataset is known to be an NP-hard problem (see [11]), and hence cannot be solved efficiently. Most algorithms seek to minimize certain objective function, and usually achieve local rather than global minimum solution. In this work we focus on partitioning clustering, where the number of clusters in known in advance. Most partitioning clustering methods iteratively update the cluster centers, and hence they are often referred as center-based clustering methods.

We introduce few notations for the upcoming discussion. Let $\mathcal{A} = \{a^1, a^2, \ldots, a^m\}$ be a given set of points in $\mathbb{R}^n$, and let $1 < k < m$ be a fixed given number of clusters. The clustering problem consists of partitioning the dataset $\mathcal{A}$ into k subsets $\{C^1, C^2, \ldots, C^k\}$, called clusters. For each $l = 1, 2, \ldots, k$, the cluster C^l is represented by its center $x^l \in \mathbb{R}^n$. We describe several well-known center-based clustering algorithms.

(i) The k-means algorithm. This algorithm is probably the most famous within the clustering scope, and dates back to Steinhaus (1956), MacQueen (1967) and Lloyd (1982) (see [18, 15, 14]). The k-means algorithm partitions the data into k sets. The solution is then a set of k centers, each of which is located at the centroid of the data for which it is the closest center. The k-means algorithm performs hard clustering, and each point is labeled according to its closest center. This algorithm can be discribed as an optimization algorithm (see precise details

below) which minimizes the following objective function

$$f_{KM}(x^1, x^2, \cdots, x^k) = \sum_{i=1}^{m} \min_{1 \le l \le k} \left\| a^i - x^l \right\|^2 .$$

The simplicity of the algorithm both in the theoretical and implementation aspects made it very popular. There has been enormous improved techniques designed targeting a variety of applications (see [12] for a review).

(ii) The fuzzy k-means (FKM) algorithm. The FKM algorithm is a soft clustering method. For each data point the result of the FKM algorithm is a distribution of membership over the clusters (see [5] for the original paper by Bezdek on FKM). The objective function that the FKM algorithm minimizes is

$$f_{FKM}(w^1, w^2, \cdots, w^m, x^1, x^2, \cdots, x^k) = \sum_{i=1}^{m} \sum_{l=1}^{k} (w_l^i)^\beta \left\| a^i - x^l \right\|^2 .$$

The variable w_l^i denotes the probability that data point a^i is assigned to cluster x^l, thus it is under the constraints $\sum_{l=1}^{k} w_l^i = 1$ for all $1 \le i \le m$ and $w_l^i \ge 0$. The parameter $\beta > 1$ governs the "fuzzy partition". Setting $\beta = 1$ results in the standard k-means algorithm (see Section 2.1 for more details). Applying the Gauss-Seidel algorithm with the objective f_{FKM}, namely alternating between the $w^1, w^2, \cdots, w^m$ variables and the $x^1, x^2, \cdots, x^k$ variables, yields the FKM algorithm (see [10, p. 528]).

(iii) The Expectation-Maximization (EM) algorithm. The EM algorithm (Dempster et al. [9]) is used extensively in statistical estimation problems for learning mixtures of distributions. It is a soft clustering algorithm. The objective function that EM maximizes is

$$f_{EM}(x^1, x^2, \cdots, x^k) = \sum_{i=1}^{m} \log \left(\sum_{l=1}^{k} p\left(a^i | x^l\right) p\left(x^l\right) \right),$$

where $p\left(a^i | x^l\right)$ is the probability of a^i given that it is generated by the Gaussian distribution with center x^l and $p\left(x^l\right)$ is the prior probability of center x^l. The algorithm is guaranteed to converge to a local maximum of the likelihood function f_{EM} (see [21]).

An interesting paper of Teboulle [19] shows that these center-based clustering algorithms can be recovered from a certain proposed continuous optimization framework

which will be used in this work (see more details below). The k-means objective function, f_{KM}, can be extended to a more general form,

$$F(x) = \sum_{i=1}^{m} v_i \min_{1\le l\le k} d\left(x^l, a^i\right), \tag{1.1.1}$$

where $d(\cdot,\cdot)$ is some distance-like function and v_i positive weights which satisfy $\sum_{i=1}^{m} v_i = 1$. The smoothing methods in [19] are based on replacing the nonsmooth term $\min_{1\le l\le k} d\left(x^l, a^i\right) = -\sigma_\Delta(-d(x))$ with an approximation of an asymptotic convex nonlinear mean function defined by

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

where $\pi \in \Delta_+ := int\,(\Delta)$ and $h \in \mathcal{H}$ which is the class of function defined by

$$\mathcal{H} = \left\{h \in C^3(int(\mathrm{dom} h)) : h' > 0,\ h'' > 0,\ \text{and } t \mapsto -\frac{h'(t)}{h''(t)} \text{ is convex}\right\}.$$

Therefore, the approximating smooth objective function of the original nonsmooth clustering problem function, defined in (1.1.1), takes the following form

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

where $s > 0$ serves as a smoothing parameter. The suggested center-based algorithm for soft clustering, the Smooth k-means (SKM) algorithm, generates a sequence whose each limit point is a stationary point of F_s.

Most of the existing clustering methods are sensitive to the starting point, namely choosing different starting point result in significant differences in the final clustering. There are plethora of heuristic initialization methods. One such initialization method is to choose k random data points as staring centers, assuming uniform distribution or some other prior distribution on the data. Another popular method is k-means++, where the first center is chosen at random from the dataset, and for each $2 \le l \le k$, the center x^l is the furthest data point from the data points chosen so far.

We begin this work with a formulation of the clustering problem which consists of minimizing the sum of finite collection of min-functions. This is a nonsmooth and nonconvex optimization problem, in its most general case. The clustering problem is given by

$$\min_{x\in\mathbb{R}^{nk}} \left\{F(x) := \sum_{i=1}^{m} \min_{1\le l\le k} d(x^l, a^i)\right\}, \tag{1.1.2}$$

where $x = \left(x^1, x^2, \ldots, x^k\right) \in \mathbb{R}^{nk}$ with $d(\cdot,\cdot)$ being a distance-like function.

We focus on two cases of distance-like functions. The first is the squared Euclidean norm, which is the standard proximity measure used in the k-means algorithm. For this case, we use an equivalent smooth optimization problem for the clustering problem presented in (1.1.2) and prove convergence result for the suggested algorithm via the methodology which was recently developed in [7] and will be discussed in great details below. The second distance-like function that we study is the Euclidean norm. In this case we present an approximation model, in order to overcome the lack of smoothness in the problem. Then we propose an algorithm to solve the approximated model which combines ideas which were used in the squared Euclidean case with a classical smoothing ideas which was used in [3]. We present numeric experiments, that show the superiority of the Euclidean norm distance function for datasets in which the data points are spread relatively sparsely form their centers.

The lack of smoothness in this model can be overcome, yet the nonconvex nature of the clustering problem will accompany the discussions throughout this work. Significant amount of studies have been made on convex models, even though in many cases the original optimization problem is nonconvex. To overcome the lack of convexity, one of the common approaches is usually achieved by considering a convex relaxation of the original problem. In this thesis we take a different route and consider the problem in its original nonconvex form. Very recently this complicated route became more relevant and interesting thanks to few papers (see [1, 2, 7] and the references therein) which pave the way for dealing with nonconvex problems using sophisticated mathematical tools as will be explained later in Section 2.2.

1.2 Outline and Contributions of The Thesis

Our main objectives and contributions in this thesis are as follows.

- To develop algorithms that address the clustering problem for two different distance-like functions and present numerical tests which demonstrate the effectiveness of the proposed algorithms.
- To demonstrate the usefulness of KL property and the general methodology developed in [7] to tackle the clustering problem.
- To prove the convergence of k-means to a critical point.

We outline now the contents of this thesis.

- In Chapter 2 we transform the initial discrete formulation of the clustering problem (see (1.1.1)) into a smooth model. In addition, we recall the KL theory and the general methodology, which was developed in [7], that will be used in our analysis of the proposed algorithms.

- In Chapter 3 we tackle the clustering problem with the squared Euclidean norm distance-like function, which is the most common distance used in many other clustering algorithms. The proposed clustering algorithm is based on the alternating minimization method, and it is similar to the k-means algorithm. In this case the objective function is smooth and we can apply the general methodology, and prove convergence of the generated sequence to a critical point of the corresponding objective function.

- In Chapter 4 we tackle the clustering problem with the Euclidean norm distance-like function. Providing an approximation to the original objective function which overcome the lack of smoothness and then proceed with the general methodology and prove again convergence of the generated sequence to a critical point of the approximated smooth objective function.

- In Chepter 5 we show that the k-means algorithm can be recovered from the proposed model of the clustering problem presented in Chapter 3. In addition, we prove that k-means algorithm convergence to a critical point, and under additional assumption, we extend the convergence to a local minimum.

- In Chapter 6 we compare the performance of the proposed algorithms, and some existing center-based clustering algorithms, according to some common criteria.

1.3 Notation and Terminology

The following notations will be used throughout this thesis

Table 1.1: Table of Notations

$\mathcal{A}$	dataset for clustering of size m
k	the number of clusters
x^l	center of cluster l, for each $l = 1, 2, \ldots, k$; $x = \left(x^1, x^2, \cdots, x^k\right)$
$\langle \cdot, \cdot \rangle$	the standard dot product in Euclidean space, that is $\langle u, v \rangle = \sum_{i=1}^{d} u_l v_l$
$\lVert \cdot \rVert$	Euclidean norm $\lVert x \rVert = \sqrt{\sum_{l=1}^{d} {x_l}^2}$
Δ	the simplex i.e., $\Delta = \left\{ u \in \mathbb{R}^d : \sum_{l=1}^{d} u_l = 1,\ u \geq 0 \right\}$
$\delta_S(\cdot)$	delta function of set $S \subset \mathbb{R}^d$, which is defined to be 0 in S and ∞ otherwise
$\mathrm{dom}\sigma$	domain of function σ, which is all vectors v such that $\sigma(v) < \infty$
$\partial\sigma$	subdifferential of function σ (see Definition 2.2.1)
$\mathrm{crit}\sigma$	set of all critical points of function σ, that is all vectors v such that $0 \in \partial\sigma(v)$
$\mathrm{dist}(u, S)$	distance function, for any point $u \in \mathbb{R}^d$ and set $S \in \mathbb{R}^d$, $\mathrm{dist}(u, S) := \inf \{\lVert u - v \rVert : v \in S\}$
$H(w, x)$	sum of distances of x^l from each data point in $\mathcal{A}$, adjusted by the non-negative weights w^i, $H(w, x) = \sum_{i=1}^{m} \sum_{l=1}^{k} w_l^i d\left(x^l, a^i\right)$
$G(w)$	sum of delta functions which constrain each w^i to be in the simplex, that is $G(w) = \sum_{i=1}^{m} \delta_\Delta(w^i)$
Ψ	the objective function in the clustering problem, defined by $\Psi(w, x) = H(w, x) + G(w)$

Chapter 2

Problem Reformulation and Mathematical Tools

Transforming the initial discrete clustering problem given in (1.1.2) into a smooth form is the first main objective of this chapter. The second objective is to present a sufficient mathematical background which leads to the general methodology, developed in [7], that enables to analyze algorithms in the nonconvex and nonsmooth setting.

2.1 Reformulation of the Clustering Problem

We begin with a reformulation of the clustering problem which will be the basis for our developments in this work. The reformulation is based on the following fact:

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

where Δ denotes the well-known simplex defined by

$$\Delta = \left\{ w \in \mathbb{R}^k : \sum_{l=1}^{k} w_l = 1,\ w \geq 0 \right\}.$$

Using this fact in Problem (1.1.2) and introducing new variables $w^i \in \mathbb{R}^k$, $i = 1, 2, \ldots, m$, gives a smooth reformulation of the clustering problem

$$\min_{x\in\mathbb{R}^{nk}} \sum_{i=1}^{m} \min_{w^i\in\Delta} \langle w^i, d^i(x)\rangle, \tag{2.1.1}$$

where

$$d^i(x) = (d(x^1, a^i), d(x^2, a^i), \ldots, d(x^k, a^i)) \in \mathbb{R}^k, \quad i = 1, 2, \ldots, m. \tag{2.1.2}$$

Replacing further the constraint $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$, which is defined to be 0 in Δ and ∞ otherwise, to the objective function, results in a equivalent formulation

$$\min_{x \in \mathbb{R}^{nk}, w \in \mathbb{R}^{km}} \left\{ \sum_{i=1}^{m} \left(\langle w^i, d^i(x) \rangle + \delta_\Delta(w^i) \right) \right\}, \tag{2.1.3}$$

where $w = (w^1, w^2, \ldots, w^m) \in \mathbb{R}^{km}$. Finally, for the simplicity of the yet to come expositions, we define the following functions

$$H(w, x) := \sum_{i=1}^{m} H^i(w, x) = \sum_{i=1}^{m} \langle w^i, d^i(x) \rangle \quad \text{and} \quad G(w) = \sum_{i=1}^{m} G^i(w^i) := \sum_{i=1}^{m} \delta_\Delta(w^i).$$

Replacing the terms in Problem (2.1.3) with the functions defined above gives a compact equivalent form of the original clustering problem

$$\min \left\{ \Psi(z) := H(w, x) + G(w) \mid z := (w, x) \in \mathbb{R}^{km} \times \mathbb{R}^{nk} \right\}. \tag{2.1.4}$$

2.2 Convergence Methodology

In this subsection we give a brief review of the main developments established in [7]. These developments include on one hand the proximal alternating linearized minimization (PALM) algorithm and on the other hand, a general procedure for proving global convergence of generic algorithm which will play a central role in this work. First, let us recall several definitions which are needed for the upcoming discussion.

Definition 2.2.1 (Subdifferentials)**.** *Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.*

(i) For a given $x \in dom\ \sigma := \left\{ x \in \mathbb{R}^d : \ \sigma(x) < \infty \right\}$, the Fréchet subdifferential of σ at x, written $\widehat{\partial}\sigma(x)$, is the set of all vectors $u \in \mathbb{R}^d$ which satisfy

$$\liminf_{y \neq x \ y \to x} \frac{\sigma(y) - \sigma(x) - \langle u, y - x \rangle}{\|y - x\|} \geq 0.$$

When $x \notin dom\sigma$, we set $\widehat{\partial}\sigma(x) = \emptyset$.

(ii) The limiting-subdifferential, or subdifferential in short, of σ at $x \in \mathbb{R}^n$, written $\partial\sigma(x)$, is defined through the following closure process

$$\partial\sigma(x) := \left\{ u \in \mathbb{R}^d : \exists x^k \to x,\ \sigma(x^k) \to \sigma(x) \text{ and } u^k \in \widehat{\partial}\sigma(x^k) \text{ as } k \to \infty \right\}.$$

In the nonsmooth context, as in the smooth case, the well-known Fermat's rule remains unchanged, that is, if $x \in \mathbb{R}^d$ is a local minimizer of σ then $0 \in \partial\sigma(x)$. Points whose subdifferential contains 0 are called *critical points*, and the set of all critical points of σ is denoted by critσ.

Now we present the Kurdyka-Łojasiewicz property, which plays a central role in the general methodology which was developed in [7]. Let $\eta \in (0, +\infty]$. Denote the following class of concave functions

$$\Phi_\eta = \left\{ \varphi \in C\left([0,\eta), \mathbb{R}_+\right) :\ \varphi \in C^1\left((0,\eta)\right),\ \varphi' > 0,\ \varphi(0) = 0 \right\}.$$

Definition 2.2.2 (Kurdyka-Łojasiewicz property)**.** *Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and lower semicontinuous.*

(i) The function σ is said to have the Kurdyka-Łojasiewicz (KL) property at $\overline{u} \in$ dom $\partial\sigma := \left\{ u \in \mathbb{R}^d :\ \partial\sigma \neq \emptyset \right\}$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\overline{u}$ and a function $\varphi \in \Phi_\eta$, such that for all

$$u \in U \cap \left\{ x \in \mathbb{R}^d :\ \sigma(\overline{u}) < \sigma(x) < \sigma(\overline{u}) + \eta \right\},$$

the following inequality holds

$$\varphi'(\sigma(u) - \sigma(\overline{u}))dist(0, \partial\sigma(u)) \geq 1,$$

where $dist(x, S) := \inf\{\|y - x\| :\ y \in S\}$ denotes the distance from $x \in \mathbb{R}^d$ to $S \subset \mathbb{R}^d$.

(ii) If σ satisfy the KL property at each point of dom σ then σ is called a KL function.

As it can be seen from the definition above, verifying that a given function satisfies the KL property is quite involved. This can be overcome by using an important result of Bolte et al. (see [6]). Before presenting this result we will recall the definition of semi-algebraic function.

Definition 2.2.3 (Semi-algebraic sets and functions)**.** *(i) A subset $S \subset \mathbb{R}^d$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^d \to \mathbb{R}$ such that*

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{ u \in \mathbb{R}^d :\ g_{ij} = 0 \text{ and } h_{ij}(u) < 0 \right\}$$

(ii) A function $h : \mathbb{R}^d \to (-\infty, +\infty]$ is called semi-algebraic if its graph

$$\left\{(u,t) \in \mathbb{R}^{d+1} : \ h(u) = t\right\},$$

is a semi-algebraic subset of $\mathbb{R}^{d+1}$.

Theorem 2.2.1. *Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. If σ is semi-algebraic then if satisfies the KL property at any point of $dom\sigma$.*

The class of semi-algebraic functions is very broad, it includes real polynomial functions, indicator functions of semi-algebraic sets, finite sums and products of semi-algebraic functions, composition of semi-algebraic functions, and many more.

Attouch et al. [1, 2] established convergence of sequences generated by the proximal Gauss-Seidel scheme in the general nonconvex and nonsmooth setting, and by the proximal-forward-backward (aka Proximal Gradient, more on this method see [13, 20, 8]) algorithm applied to the nonconvex and nonsmooth minimization of the sum of a nonsmooth function with a smooth one. This approach assumes that the objective function to be minimized satisfies the Kurdyka-Łojasiewicz (KL) property. The convergence results were further extended in the recent work by Bolte et al. [7], to the PALM algorithm which is a novel algorithm that combines the two basic and old ideas of Alternating Minimization and Proximal Gradient (see more details below). Additional contribution of [7] is the general methodology to prove convergence of a generic algorithm in the setting of nonconvex and nonsmooth optimization problems.

Equipped with these definitions, we present the general methodology that will be used several times throughout this work. Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function which is bounded from below and consider the problem

$$(P) \quad \min\left\{\sigma(z) : z \in \mathbb{R}^d\right\}.$$

Suppose that we are given a generic algorithm $\mathcal{A}$ which generates a sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ via the following scheme:

$$z^0 \in \mathbb{R}^d,\ z^{k+1} \in \mathcal{A}\left(z^k\right), \quad k = 0, 1, \ldots.$$

The purpose of the proposed methodology is to assure the convergence of the whole sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ to a critical point of σ. The set of all limit points is denoted by $\omega\left(z^0\right)$, and defined by

$$\left\{\overline{z} \in \mathbb{R}^d : \exists \text{ an increasing sequence of integers } \{k_l\}_{l\in\mathbb{N}} \text{ such that } z^{k_l} \to \overline{z} \text{ as } l \to \infty\right\}.$$

Definition 2.2.4. *Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. A sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ is called a gradient-like descent sequence for σ if for all $k \in \mathbb{N}$ the following two conditions hold:*

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

$$\rho_1 \left\|z^{k+1} - z^k\right\|^2 \leq \sigma\left(z^k\right) - \sigma\left(z^{k+1}\right).$$

(C2) A subgradient lower bound for the iterates gap:

- *$\left\{z^k\right\}_{k\in\mathbb{N}}$ is bounded.*
- *There exists a positive scalar ρ_2 such that*

$$\left\|w^{k+1}\right\| \leq \rho_2 \left\|z^{k+1} - z^k\right\|, \; w^{k+1} \in \partial\sigma\left(z^{k+1}\right).$$

The two conditions (C1) and (C2) defining a gradient-like descent sequence for a given σ are typical for any descent type algorithm, and provide the basic tools to prove that the limit of any convergent subsequence of $\left\{z^k\right\}_{k\in\mathbb{N}}$ is a critical point of σ. More precisely, from [7, Lemma 5, p. 476] we have

Lemma 2.2.1. *If $\left\{z^k\right\}_{k\in\mathbb{N}}$ is a gradient-like descent sequence for a given function σ, which is lsc and proper on $\mathbb{R}^d$, then $\omega\left(z^0\right)$ is a nonempty, compact and connected set, and we have*

$$\lim_{k\to\infty} dist\left(z^k, \omega\left(z^0\right)\right) = 0.$$

This result can thus be applied to any algorithm that produces a gradient-like descent to establish convergence in accumulation points. The main goal is to establish global convergence, i.e., that the whole sequence converges to a critical point of σ. This can be achieved by imposing an additional assumption on the class of functions σ, it must satisfy the Kurdyka-Lojasiewiez property.

As proven in [7], relying on a key uniformization of the KL property it is possible to establish global convergence of any gradient-like descent sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$, independently of the algorithm used. Verifying the KL property of a given function might often be a difficult task. However, thanks to a Theorem 2.2.1, any proper and lsc function σ which is semi-algebraic satisfies the KL property at any point in domσ. We summarize the general methodology and convergence results of [7] in the following abstract convergence result.

Theorem 2.2.2. *Let $\sigma : \mathbb{R}^d \to (-\infty, \infty]$ be a proper, lower semicontinuous and semi-algebraic function with $\inf\sigma > -\infty$, and assume that $\left\{z^k\right\}_{k\in\mathbb{N}}$ is a gradient-like descent sequence for σ. If $\omega\left(z^0\right) \subset crit\sigma$ then the sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ convergences to a critical point z^* of σ.*

Remark 2.2.1. *Under the premises of this theorem, it is also possible to derive a rate of convergence result for the sequence* $\left\{z^k\right\}_{k\in\mathbb{N}}$ *of the form* $\left\|z^k - z^*\right\| \leq Ck^{-\gamma}$, *for some positive constant* C *and where* $\gamma > 0$ *is a so-called KL exponent.*

Finally, we present the proximal alternating linearized minimization (PALM) algorithm which solves the nonconvex and nonsmooth minimization problem of the following form

$$(M) \quad \text{minimize } \sigma(x,y) := f(x) + g(y) + H(x,y) \text{ over all } (x,y) \in \mathbb{R}^n \times \mathbb{R}^m,$$

where $f : \mathbb{R}^n \to (-\infty, +\infty]$ and $g : \mathbb{R}^n \to (-\infty, +\infty]$ are proper and lower semi-continuous functions while $H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a C^1 function. In addition, partial gradients of H are Lipschitz continuous, namely, $H(\cdot, y) \in C^{1,1}_{L_1(y)}$ and $H(x, \cdot) \in C^{1,1}_{L_2(x)}$.

As mentioned in [7] the PALM algorithm is nothing but alternating the classical proximal gradient over the two blocks (x, y). This leads towards the following approximations

$$\widehat{\sigma}\left(x, y^k\right) = \left\langle x - x^k, \nabla_x H\left(x^k, y^k\right)\right\rangle + \frac{c_k}{2}\left\|x - x^k\right\|^2 + f(x), \ (c_k > 0),$$

and

$$\widetilde{\sigma}\left(x^{k+1}, y\right) = \left\langle y - y^k, \nabla_y H\left(x^{k+1}, y^k\right)\right\rangle + \frac{d_k}{2}\left\|y - y^k\right\|^2 + g(y), \ (d_k > 0).$$

Thus, PALM can be summarized as follows

$$x^{k+1} \in \operatorname{argmin}\left\{\widehat{\sigma}(x, y^k) : \ x \in \mathbb{R}^n\right\} \quad \text{and} \quad y^{k+1} \in \operatorname{argmin}\left\{\widetilde{\sigma}\left(x^{k+1}, y\right) : \ y \in \mathbb{R}^m\right\}.$$

Assuming Ψ is KL function and the generated sequence by PALM, $\left\{\left(x^k, y^k\right)\right\}_{k\in\mathbb{N}}$, is bounded, Bolte et al. [7] proved that the sequence is a gradient-like descent sequence, and thus it converges to a critical point of σ.

Chapter 3

Clustering: The Squared Euclidean Norm Case

We develop an algorithm for the clustering problem given in (2.1.4) with squared Euclidean distance-like function. We show that the acquired model is a sum of smooth and nonsmooth function, therefore it allows us to apply the convergence methodology described in Section 2.2 and prove that the generated sequence converges to a critical point of the objective function Ψ of the clustering problem.

3.1 Clustering with PALM

In this section we tackle the clustering problem, given in (2.1.4), for which the proximity function $d(\cdot,\cdot)$ is taken to be the classical distance function defined by $d(u,v) = \|u-v\|^2$. We devise a PALM-like algorithm, based on the discussion in the previous section. Since the clustering problem has a specific structure, we are ought to exploit it in the following manner.

(1) The function $w \mapsto H(w,x)$, for fixed x, is linear and therefore there is no need to linearize it as suggested in the framework which was discussed in Section 2.2.

(2) The function $x \mapsto H(w,x)$, for fixed w, is quadratic and convex. Hence, there is no need to add a proximal term as suggested in the framework which was discussed in Section 2.2.

As in the PALM algorithm, our algorithm is based on the old approach of alternating minimization, with the following adaptations which are motivated by the

observations mentioned above. More precisely, with respect to w we suggest to regularize the first subproblem with proximal term as follows

$$w^i(t+1) = \operatorname*{argmin}_{w^i \in \Delta} \left\{ \langle w^i, d^i(x(t)) \rangle + \frac{\alpha_i(t)}{2} \|w^i - w^i(t)\|^2 \right\}, \quad i = 1, 2, \ldots, m, \tag{3.1.1}$$

where $\alpha_i(t) > 0$ for all $i = 1, 2, \ldots, m$. On the other hand, with respect to x we perform exact minimization, that is,

$$x(t+1) = \operatorname{argmin} \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\}. \tag{3.1.2}$$

It is easy to check that all subproblems, with respect to w^i, $i = 1, 2, \ldots, m$, and x, can be rewritten explicitly (where we use P_Δ for the orthogonal projection onto the set Δ). Thus, we can present now the KPALM algorithm.

KPALM

(1) Initialization: $(w(0), x(0)) \in \Delta^m \times \mathbb{R}^{nk}$.

(2) General step ($t = 0, 1, \ldots$):

(2.1) Cluster assignment: choose certain $\alpha_i(t) > 0$, $i = 1, 2, \ldots, m$, and compute

$$w^i(t+1) = P_\Delta \left(w^i(t) - \frac{d^i(x(t))}{\alpha_i(t)} \right). \tag{3.1.3}$$

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

$$x^l(t+1) = \frac{\sum_{i=1}^m w_l^i(t+1) a^i}{\sum_{i=1}^m w_l^i(t+1)}. \tag{3.1.4}$$

We begin our analysis of the KPALM algorithm with the following boundedness property of the generated sequence. For simplicity, from now on, we denote $z(t) := (w(t), x(t))$, $t \in \mathbb{N}$.

Proposition 3.1.1 (Boundedness of KPALM sequence)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM. Then, the following statements hold true.*

(i) For all $l = 1, 2, \ldots, k$, the sequence $\left\{x^l(t)\right\}_{t\in\mathbb{N}}$ is contained in $Conv(\mathcal{A})$, the convex hull of $\mathcal{A}$, and therefore bounded by $M = \max_{1\le i\le m} \|a^i\|$.

(ii) The sequence $\{z(t)\}_{t\in\mathbb{N}}$ is bounded in $\mathbb{R}^{km} \times \mathbb{R}^{nk}$.

Proof. (i) Let $1 \leq l \leq k$. We set $\lambda_i = w_l^i(t) / \sum_{j=1}^{m} w_l^j(t), i = 1, 2, \ldots, m$, then $\lambda_i \geq 0$ and $\sum_{i=1}^{m} \lambda_i = 1$. From (3.1.4) we have

$$x^l(t) = \frac{\sum_{i=1}^{m} w_l^i(t) a^i}{\sum_{i=1}^{m} w_l^i(t)} = \sum_{i=1}^{m} \left(\frac{w_l^i(t)}{\sum_{j=1}^{m} w_l^j(t)} \right) a^i = \sum_{i=1}^{m} \lambda_i a^i \in Conv(\mathcal{A}), \quad (3.1.5)$$

which proves that $x^l(t)$ is in the convex hull of $\mathcal{A}$, for all $l = 1, 2, \ldots, k$ and $t \in \mathbb{N}$. Taking the norm of $x^l(t)$ and using (3.1.5) yields that

$$\|x^l(t)\| = \left\| \sum_{i=1}^{m} \lambda_i a^i \right\| \leq \sum_{i=1}^{m} \lambda_i \|a^i\| \leq \sum_{i=1}^{m} \lambda_i \max_{1 \leq i \leq m} \|a^i\| = M.$$

(ii) The sequence $\{w(t)\}_{t \in \mathbb{N}}$ is bounded, since $w^i(t) \in \Delta$ for all $i = 1, 2, \ldots, m$ and $t \in \mathbb{N}$. Combined with the previous item, the result follows.

□

The following assumption will be crucial for the coming analysis.

Assumption 3.1.1. *(i) The chosen sequences of parameters $\{\alpha_i(t)\}_{t \in \mathbb{N}}$, $i = 1, 2, \ldots, m$, are bounded, that is, there exist $\underline{\alpha_i} > 0$ and $\overline{\alpha_i} < \infty$ for all $i = 1, 2, \ldots, m$, such that*

$$\underline{\alpha_i} \leq \alpha_i(t) \leq \overline{\alpha_i}, \quad \forall\, t \in \mathbb{N}. \quad (3.1.6)$$

(ii) For all $t \in \mathbb{N}$ there exists $\underline{\beta} > 0$ such that

$$2 \min_{1 \leq l \leq k} \sum_{i=1}^{m} w_l^i(t) := \beta(w(t)) \geq \underline{\beta}. \quad (3.1.7)$$

It should be noted that Assumption 3.1.1(i) is very mild since the parameters $\alpha_i(t)$, $1 \leq i \leq m$ and $t \in \mathbb{N}$, can be chosen arbitrarily by the user and therefore it can be controlled such that the boundedness property holds true. Assumption 3.1.1(ii) is essential since if it is not true then $w_l^i(t) = 0$ for all $1 \leq i \leq m$, which means that the center x^l does not play any role in the solution process which is, of course, meaningless situation.

Lemma 3.1.1 (Strong convexity of $H(w, x)$ in x)**.** *The function $x \mapsto H(w, x)$ is strongly convex with parameter $\beta(w)$ which defined in (3.1.7), whenever $\beta(w) > 0$.*

Proof. Since the function $x \mapsto H(w,x) = \sum_{l=1}^{k}\sum_{i=1}^{m} w_l^i \|x^l - a^i\|^2$ is C^2, it is strongly convex if and only if the smallest eigenvalue of the corresponding Hessian matrix is positive. Indeed, the Hessian is given by

$$\nabla_{x^j}\nabla_{x^l} H(w,x) = \begin{cases} 0 & \text{if } j \neq l, \quad 1 \leq j,l \leq k, \\ 2\sum_{i=1}^{m} w_l^i & \text{if } j = l, \quad 1 \leq j,l \leq k. \end{cases}$$

Since the Hessian is a diagonal matrix, the smallest eigenvalue is $\beta(w) = 2\min_{1\leq l\leq k}\sum_{i=1}^{m} w_l^i$, and the result follows. □

Now we are ready to prove global convergence of the sequence $\{z(t)\}_{t\in\mathbb{N}}$ generated by KPALM to a critical point of Ψ given in (2.1.4). We will follow here the general procedure which was discussed is Section 2.2. Therefore we need to prove that $\{z(t)\}_{t\in\mathbb{N}}$ is a gradient-like descent sequence (see Definition 2.2.4), that is, the conditions (C1) and (C2) hold. We begin by proving condition (C1).

Proposition 3.1.2 (Sufficient decrease property)**.** *Suppose that Assumption 3.1.1 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM. Then there exists $\rho_1 > 0$ such that*

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

Proof. From step (3.1.3), see also (3.1.1), we derive, for each $i = 1, 2, \ldots, m$, the following inequality

$$\begin{aligned} H^i(w(t+1), x(t)) + \frac{\alpha_i(t)}{2}\|w^i(t+1) - w^i(t)\|^2 &= \\ &= \langle w^i(t+1), d^i(x(t))\rangle + \frac{\alpha_i(t)}{2}\|w^i(t+1) - w^i(t)\|^2 \\ &\leq \langle w^i(t), d^i(x(t))\rangle + \frac{\alpha_i(t)}{2}\|w^i(t) - w^i(t)\|^2 \\ &= \langle w^i(t), d^i(x(t))\rangle \\ &= H^i(w(t), x(t)). \end{aligned}$$

Hence, we obtain for all $t \in \mathbb{N}$, that

$$\frac{\alpha_i(t)}{2}\|w^i(t+1) - w^i(t)\|^2 \leq H^i(w(t), x(t)) - H^i(w(t+1), x(t)). \tag{3.1.8}$$

Denote $\underline{\alpha} = \min_{1\le i\le m} \underline{\alpha_i}$. Summing inequality (3.1.8) over $i = 1, 2, \ldots, m$ yields

$$\begin{aligned}
\frac{\underline{\alpha}}{2}\|w(t+1) - w(t)\|^2 &= \frac{\underline{\alpha}}{2}\sum_{i=1}^{m} \|w^i(t+1) - w^i(t)\|^2 \\
&\le \sum_{i=1}^{m} \frac{\alpha_i(t)}{2}\|w^i(t+1) - w^i(t)\|^2 \\
&\le \sum_{i=1}^{m} \left[H^i(w(t), x(t)) - H^i(w(t+1), x(t))\right] \\
&= H(w(t), x(t)) - H(w(t+1), x(t)), \qquad (3.1.9)
\end{aligned}$$

where the first inequality follows from Assumption 3.1.1(i) and the definition of $\underline{\alpha}$.

From Assumption 3.1.1(ii) we have that $\beta(w(t)) \ge \underline{\beta}$, for all $t \in \mathbb{N}$, and from Lemma 3.1.1 it follows that the function $x \mapsto H(w(t), x)$ is strongly convex with parameter $\beta(w(t))$. H using the strong convexity yields that

$$\begin{aligned}
H(w(t+1), x(t)) - H(w(t+1), x(t+1)) &\ge \langle \nabla_x H(w(t+1), x(t+1)), x(t) - x(t+1)\rangle \\
&\quad + \frac{\beta(w(t))}{2}\|x(t) - x(t+1)\|^2 \\
&= \frac{\beta(w(t))}{2}\|x(t+1) - x(t)\|^2 \\
&\ge \frac{\underline{\beta}}{2}\|x(t+1) - x(t)\|^2, \qquad (3.1.10)
\end{aligned}$$

where the equality follows from (3.1.2), since $\nabla_x H(w(t+1), x(t+1)) = 0$. Set $\rho_1 = \frac{1}{2}\min\left\{\underline{\alpha}, \underline{\beta}\right\}$, by combining (3.1.9) and (3.1.10), we get

$$\begin{aligned}
\rho_1\|z(t+1) - z(t)\|^2 &= \rho_1\left(\|w(t+1) - w(t)\|^2 + \|x(t+1) - x(t)\|^2\right) \\
&\le [H(w(t), x(t)) - H(w(t+1), x(t))] \\
&\qquad + [H(w(t+1), x(t)) - H(w(t+1), x(t+1))] \\
&= H(z(t)) - H(z(t+1)) \\
&= \Psi(z(t)) - \Psi(z(t+1)),
\end{aligned}$$

where the last equality follows from the fact that $G(w(t)) = 0$, since $w(t) \in \Delta^m$ for all $t \in \mathbb{N}$, and therefore $H(z(t)) = \Psi(z(t))$, $t \in \mathbb{N}$. This proves the desired result. □

Now, we are focusing on proving condition (C2) and we will need the following technical result.

Lemma 3.1.2. *Denote* $S = \left\{x \in \mathbb{R}^k : x^l \in Conv(\mathcal{A}),\ l = 1, 2, \cdots, k\right\}$. *Then, the function* d^i *defined in (2.1.2) is Lipschitz continuous in* S *with constant* $4M$, *that is*

$$\left\|d^i(x) - d^i(y)\right\| \le 4M\left\|x - y\right\|, \quad \forall\, x, y \in S,$$

where $M = \max_{1\le i\le m} \|a^i\|$.

Proof. Take $x, y \in S$, since $d(u,v) = \|u-v\|^2$, we get that

$$\begin{aligned}
\left\|d^i(x) - d^i(y)\right\| &= \left[\sum_{l=1}^{k} \left|\left\|x^l - a^i\right\|^2 - \left\|y^l - a^i\right\|^2\right|^2\right]^{\frac{1}{2}} \\
&= \left[\sum_{l=1}^{k} \left|\left\|x^l\right\|^2 - 2\left\langle x^l, a^i\right\rangle + \left\|a^i\right\|^2 - \left\|y^l\right\|^2 + 2\left\langle y^l, a^i\right\rangle - \left\|a^i\right\|^2\right|^2\right]^{\frac{1}{2}} \\
&\le \left[\sum_{l=1}^{k} \left(\left|\left\|x^l\right\|^2 - \left\|y^l\right\|^2\right| + \left|2\left\langle y^l - x^l, a^i\right\rangle\right|\right)^2\right]^{\frac{1}{2}} \\
&\le \left[\sum_{l=1}^{k} \left(\left|\left\|x^l\right\| - \left\|y^l\right\|\right| \cdot \left|\left\|x^l\right\| + \left\|y^l\right\|\right| + 2\left\|y^l - x^l\right\| \cdot \left\|a^i\right\|\right)^2\right]^{\frac{1}{2}} \\
&\le \left[\sum_{l=1}^{k} \left(\left\|x^l - y^l\right\| \cdot 2M + 2\left\|x^l - y^l\right\| \cdot M\right)^2\right]^{\frac{1}{2}} \\
&= \left[\sum_{l=1}^{k} (4M)^2 \left\|x^l - y^l\right\|^2\right]^{\frac{1}{2}} \\
&= 4M\left\|x - y\right\|,
\end{aligned}$$

where the last inequality follows from the fact that $x^l \in Conv(\mathcal{A})$ and hence $\left\|x^l\right\| \le M$ for all $l = 1, 2, \ldots, k$. This proves the desired result. □

Now, using this result we can show that $\{z(t)\}_{t\in\mathbb{N}}$ satisfies condition (C2).

Proposition 3.1.3 (Subgradient lower bound for the iterates gap)**.** *Let* $\{z(t)\}_{t\in\mathbb{N}}$ *be the sequence generated by KPALM. Then, there exists* $\rho_2 > 0$ *and* $\gamma(t+1) \in \partial\Psi(z(t+1))$ *such that*

$$\|\gamma(t+1)\| \le \rho_2 \|z(t+1) - z(t)\|, \quad \forall\, t \in \mathbb{N}.$$

Proof. By the definition of Ψ (see (2.1.4)) we get

$$\partial\Psi = \nabla H + \partial G = \left(\left(\nabla_{w^i} H^i + \partial_{w^i}\delta_\Delta\right)_{i=1,2,\ldots,m}, \nabla_x H\right).$$

Evaluating the last relation at $z(t+1)$ yields

$$\begin{aligned}
\partial\Psi(z(t+1)) &= \Big(\big(\nabla_{w^i} H^i(w(t+1), x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\big)_{i=1,2,\ldots,m}, \\
&\qquad \nabla_x H(w(t+1), x(t+1))\Big) \\
&= \Big(\big(d^i(x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\big)_{i=1,2,\ldots,m}, \\
&\qquad \nabla_x H(w(t+1), x(t+1))\Big) \\
&= \Big(\big(d^i(x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\big)_{i=1,2,\ldots,m}, \mathbf{0}\Big), \qquad (3.1.11)
\end{aligned}$$

where the last equality follows from (3.1.2), that is, the optimality condition of $x(t+1)$.

The optimality condition of $w^i(t+1)$ which derived from (3.1.1), yields that for all $i = 1, 2, \ldots, m$ there exists $u^i(t+1) \in \partial\delta_\Delta(w^i(t+1))$ such that

$$d^i(x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1) = \mathbf{0}. \qquad (3.1.12)$$

Setting $\gamma(t+1) := \Big((d^i(x(t+1)) + u^i(t+1))_{i=1,2,\ldots,m}, \mathbf{0}\Big)$, it follows from (3.1.11) that $\gamma(t+1) \in \partial\Psi(z(t+1))$. Using (3.1.12) we obtain

$$\gamma(t+1) = \Big(\big(d^i(x(t+1)) - d^i(x(t)) - \alpha_i(t)(w^i(t+1) - w^i(t))\big)_{i=1,2,\ldots,m}, \mathbf{0}\Big).$$

Hence, by defining $\overline{\alpha} = \max\limits_{1\le i\le m} \overline{\alpha_i}$, we obtain

$$\begin{aligned}
\|\gamma(t+1)\| &\le \sum_{i=1}^{m} \|d^i(x(t+1)) - d^i(x(t)) - \alpha_i(t)\left(w^i(t+1) - w^i(t)\right)\| \\
&\le \sum_{i=1}^{m} \|d^i(x(t+1)) - d^i(x(t))\| + \sum_{i=1}^{m} \alpha_i(t)\|w^i(t+1) - w^i(t)\| \\
&\le \sum_{i=1}^{m} 4M\|x(t+1) - x(t)\| + \overline{\alpha}\sqrt{m}\|w(t+1) - w(t)\| \\
&\le \left(4Mm + \overline{\alpha}\sqrt{m}\right)\|z(t+1) - z(t)\|,
\end{aligned}$$

where the third inequality follows from Lemma 3.1.2. Define $\rho_2 = 4Mm + \overline{\alpha}\sqrt{m}$, and the result follows. □

Chapter 4

Clustering: The Euclidean Norm Case

We develop an algorithm for the clustering problem given in (2.1.4) with Euclidean distance-like function. Due to the lack of smoothness in the obtained model, can not apply the general methodology of Section 2.2 directly. Therefore, we approximate the H part of the objective function with a smooth function H_ε and replace the original objective function of the clustering problem, Ψ, with its approximation Ψ_ε. The proposed algorithm ε-KPALM, performs alternation between cluster assignment and center update steps. The cluster assignment step is equivalent to the same step as in KPALM, whereas for the center update step it performs a certain gradient step with respect to H_ε. We prove that the sequence which is generated by ε-KPALM globally converges to critical point.

4.1 A Smoothed Clustering Problem

In the previous section we have formulated the clustering problem in the following equivalent form

$$\min \left\{ \Psi(z) := H(w,x) + G(w) \mid z := (w,x) \in \mathbb{R}^{km} \times \mathbb{R}^{nk} \right\},$$

where, in this setting, the involved functions are

$$H(w,x) = \sum_{i=1}^{m} \langle w^i, d^i(x) \rangle = \sum_{i=1}^{m} \sum_{l=1}^{k} w_l^i \|x^l - a^i\| \quad \text{and} \quad G(w) = \sum_{i=1}^{m} \delta_\Delta(w^i).$$

In order to be able to use the theory mentioned in Section 2.2, we have used, in Section 3.1, the fact that the coupled function $H(w,x)$ is smooth, which is not the

case now. Therefore, for any $\varepsilon > 0$, it leads us to the following smoothed form of the clustering problem

$$\min \left\{ \Psi_\varepsilon(z) := H_\varepsilon(w,x) + G(w) \mid z := (w,x) \in \mathbb{R}^{km} \times \mathbb{R}^{nk} \right\}, \tag{4.1.1}$$

where

$$H_\varepsilon(w,x) = \sum_{l=1}^{k} H_\varepsilon^l(w,x) = \sum_{l=1}^{k} \sum_{i=1}^{m} w_l^i \left(\|x^l - a^i\|^2 + \varepsilon^2 \right)^{1/2}, \tag{4.1.2}$$

and for all $i = 1, 2, \ldots, m$,

$$d_\varepsilon^i(x) = \left(\left(\|x^1 - a^i\|^2 + \varepsilon^2 \right)^{1/2}, \left(\|x^2 - a^i\|^2 + \varepsilon^2 \right)^{1/2}, \ldots, \left(\|x^k - a^i\|^2 + \varepsilon^2 \right)^{1/2} \right) \in \mathbb{R}^k. \tag{4.1.3}$$

Note that $\Psi_\varepsilon(z)$ is a perturbed form of $\Psi(z)$ for a small $\varepsilon > 0$, and obviously $\Psi_0(z) = \Psi(z)$. The following lemma shows that the smoothed function $H_\varepsilon(w,x)$ indeed approximates $H(w,x)$.

Lemma 4.1.1 (Closeness of smooth)**.** *For any $(w,x) \in \Delta^m \times \mathbb{R}^{nk}$ and $\varepsilon > 0$ the following relations hold true*

$$H(w,x) \leq H_\varepsilon(w,x) \leq H(w,x) + m\varepsilon.$$

Proof. It is clear that for all $\lambda \geq 0$ we have

$$\forall\, \lambda \geq 0, \quad \lambda \leq \sqrt{\lambda^2 + \varepsilon^2} \leq \lambda + \varepsilon.$$

Applying this inequality with $\lambda = \|x^l - a^i\|$, yields

$$\|x^l - a^i\| \leq \left(\|x^l - a^i\|^2 + \varepsilon^2 \right)^{1/2} \leq \|x^l - a^i\| + \varepsilon,$$

for all $l = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, m$. By multiplying each inequality by w_l^i and summing over $l = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, m$ we obtain

$$H(w,x) \leq H_\varepsilon(w,x) \leq H(w,x) + \sum_{i=1}^{m} \sum_{l=1}^{k} w_l^i \varepsilon.$$

Since for all $i = 1, 2, \ldots, m,\ w^i \in \Delta$, the result follows. □

Now we would like to develop an algorithm which is based on the methodology of PALM (see Section 2.2) to solve Problem (4.1.1). It is easy to see that with respect to w, the objective function Ψ_ε keeps on the same structure as Ψ and therefore we apply the same step as in KPALM. More precisely, for all $i = 1, 2, \ldots, m$, we have

$$\begin{aligned} w^i(t+1) &= \arg\min_{w^i \in \Delta} \left\{ \left\langle w^i, d_\varepsilon^i(x(t)) \right\rangle + \frac{\alpha_i(t)}{2} \|w^i - w^i(t)\|^2 \right\} \\ &= P_\Delta \left(w^i(t) - \frac{d_\varepsilon^i(x(t))}{\alpha_i(t)} \right), \quad \forall\, t \in \mathbb{N}, \end{aligned}$$

where $\alpha_i(t)$, $i = 1, 2, \ldots, m$, is arbitrarily chosen. On the other hand, with respect to x we tackle the subproblem differently than in KPALM. Here we follow exactly the idea of PALM, that is, linearizing the function $x \to H(w, \cdot)$, for fixed w, and adding a regularizing term

$$x^l(t+1) = \operatorname*{argmin}_{x^l} \left\{ \left\langle x^l - x^l(t), \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \right\rangle + \frac{L^l_\varepsilon(w(t+1), x(t))}{2} \|x^l - x^l(t)\|^2 \right\},$$

where

$$L^l_\varepsilon(w(t+1), x(t)) := \sum_{i=1}^{m} \frac{w^i_l(t+1)}{(\|x^l(t) - a^i\|^2 + \varepsilon^2)^{1/2}}, \quad \forall\, l = 1, 2, \ldots, k. \tag{4.1.4}$$

The motivation to use this specific regularizing parameter (see (4.1.4)) will be discussed later.

Now we present our algorithm for solving Problem (4.1.1), we call it ε-KPALM. The algorithm alternates between cluster assignment step, similar to KPALM, and centers update step that is based on certain gradient step.

ε-KPALM

(1) Initialization: $(w(0), x(0)) \in \Delta^m \times \mathbb{R}^{nk}$.

(2) General step ($t = 0, 1, \ldots$):

(2.1) Cluster assignment: choose certain $\alpha_i(t) > 0$, $i = 1, 2, \ldots, m$, and compute

$$w^i(t+1) = P_\Delta \left(w^i(t) - \frac{d^i_\varepsilon(x(t))}{\alpha_i(t)} \right). \tag{4.1.5}$$

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

$$x^l(t+1) = x^l(t) - \frac{1}{L^l_\varepsilon(w(t+1), x(t))} \nabla_{x^l} H_\varepsilon(w(t+1), x(t)). \tag{4.1.6}$$

Remark 4.1.1. *Similarly to the KPALM algorithm, the sequence generated by ε-KPALM is also bounded, since here we also have that*

$$\begin{aligned} x^l(t+1) &= x^l(t) - \frac{1}{L^l_\varepsilon(w(t+1), x(t))} \nabla_{x^l} H(w(t+1), x(t)) \\ &= x^l(t) - \frac{1}{L^l_\varepsilon(w(t+1), x(t))} \sum_{i=1}^{m} w^i_l(t+1) \cdot \frac{x^l(t) - a^i}{(\|x^l(t) - a^i\|^2 + \varepsilon^2)^{1/2}} \\ &= \frac{1}{L^l_\varepsilon(w(t+1), x(t))} \sum_{i=1}^{m} \left(\frac{w^i_l(t+1)}{(\|x^l(t) - a^i\|^2 + \varepsilon^2)^{1/2}} \right) a^i \in Conv(\mathcal{A}). \end{aligned}$$

Before we will be able to prove the two properties (see Section 2.2) needed for global convergence of the sequence $\{z(t)\}_{t\in\mathbb{N}}$ generated by ε-KPALM, we will need several auxiliary results. For the simplicity of the expositions we define the function $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ by

$$f_\varepsilon(x) = \sum_{i=1}^{m} v_i \left(\|x - a^i\|^2 + \varepsilon^2\right)^{1/2},$$

for fixed non-negative numbers (not all zero) $v_1, v_2, \ldots, v_m \in \mathbb{R}$ and $a^i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$. We also need the following auxiliary function $h_\varepsilon : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$h_\varepsilon(x, y) = \sum_{i=1}^{m} \frac{v_i \left(\|x - a^i\|^2 + \varepsilon^2\right)}{\left(\|y - a^i\|^2 + \varepsilon^2\right)^{1/2}}.$$

Finally we introduce the following modulus, $L_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ defined by

$$L_\varepsilon(x) = \sum_{i=1}^{m} \frac{v_i}{\left(\|x - a^i\|^2 + \varepsilon^2\right)^{1/2}}.$$

Lemma 4.1.2 (Properties of the auxiliary function h_ε)**.** *The following properties of h_ε hold.*

(i) For any $y \in \mathbb{R}^n$,

$$h_\varepsilon(y, y) = f_\varepsilon(y).$$

(ii) For any $x, y \in \mathbb{R}^n$,

$$h_\varepsilon(x, y) \geq 2f_\varepsilon(x) - f_\varepsilon(y).$$

(iii) For any $x, y \in \mathbb{R}^n$,

$$f_\varepsilon(x) \leq f_\varepsilon(y) + \langle \nabla f_\varepsilon(y), x - y\rangle + \frac{L_\varepsilon(y)}{2}\|x - y\|^2.$$

Proof. (i) Follows by substituting $x = y$ in $h_\varepsilon(x, y)$.

(ii) For any two numbers $a \in \mathbb{R}$ and $b > 0$ the inequality

$$\frac{a^2}{b} \geq 2a - b,$$

holds true. Thus, for every $i = 1, 2, \ldots, m$, we have that

$$\frac{\|x - a^i\|^2 + \varepsilon^2}{\left(\|y - a^i\|^2 + \varepsilon^2\right)^{1/2}} \geq 2\left(\|x - a^i\|^2 + \varepsilon^2\right)^{1/2} - \left(\|y - a^i\|^2 + \varepsilon^2\right)^{1/2}.$$

Multiplying the last inequality by v_i and summing over $i = 1, 2, \ldots, m$, the results follows.

(iii) The function $x \mapsto h_\varepsilon(x,y)$ is quadratic with associated matrix $L_\varepsilon(y)\mathbf{I}$. Therefore, its second-order taylor expansion around y leads to the following identity

$$h_\varepsilon(x,y) = h_\varepsilon(y,y) + \langle \nabla_x h_\varepsilon(y,y), x-y\rangle + L_\varepsilon(y)\|x-y\|^2.$$

Using the first two items and the fact that $\nabla_x h_\varepsilon(y,y) = 2\nabla f_\varepsilon(y)$ yields the desired result.

□

Now we get back to the ε-KPALM algorithm and prove few technical results about the involved functions which are based on the properties obtained above.

Proposition 4.1.1 (Bounds for L_ε^l)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM. Then, the following two statements hold true.*

(i) For all $t \in \mathbb{N}$ and $l = 1,2,\ldots,k$ we have

$$L_\varepsilon^l(w(t+1),x(t)) \geq \frac{\underline{\beta}}{\left({d_{\mathcal{A}}}^2+\varepsilon^2\right)^{1/2}},$$

where $d_{\mathcal{A}} = diam(Conv(\mathcal{A}))$ is the diameter of $Conv(\mathcal{A})$ and $\underline{\beta}$ is given in (3.1.7).

(ii) For all $t \in \mathbb{N}$ and $l = 1,2,\ldots,k$ we have

$$L_\varepsilon^l(w(t+1),x(t)) \leq \frac{m}{\varepsilon}.$$

Proof. (i) From Assumption 3.1.1(ii) and the fact that $x^l(t) \in Conv(\mathcal{A})$ for all $1 \leq l \leq k$, it follows that

$$L_\varepsilon^l(w(t{+}1),x(t)) = \sum_{i=1}^m \frac{w_l^i(t+1)}{(\|x^l(t)-a^i\|^2+\varepsilon^2)^{1/2}} \geq \frac{\sum_{i=1}^m w_l^i(t+1)}{\left({d_{\mathcal{A}}}^2+\varepsilon^2\right)^{1/2}} \geq \frac{\underline{\beta}}{\left({d_{\mathcal{A}}}^2+\varepsilon^2\right)^{1/2}},$$

where the first inequality follows from the fact that $\left\|x^l(t)-a^i\right\| \leq d_{\mathcal{A}}$, for all $1 \leq l \leq k$.

(ii) Since $w(t+1) \in \Delta^m$ we have

$$L_\varepsilon^l(w(t+1),x(t)) = \sum_{i=1}^m \frac{w_l^i(t+1)}{(\|x^l(t)-a^i\|^2+\varepsilon^2)^{1/2}} \leq \sum_{i=1}^m \frac{1}{\varepsilon} = \frac{m}{\varepsilon},$$

as asserted.

□

Now we prove the following result.

Proposition 4.1.2. *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM. Then, for all $t \in \mathbb{N}$, we have*

$$H_\varepsilon(w(t+1), x(t+1)) \leq H_\varepsilon(w(t+1), x(t)) + \langle \nabla_x H_\varepsilon(w(t+1), x(t)), x(t+1) - x(t)\rangle + \sum_{l=1}^{k} \frac{L_\varepsilon^l(w(t+1), x(t))}{2} \|x^l(t+1) - x^l(t)\|^2.$$

Proof. By definition (see (4.1.2)) we have, for $i = 1, 2, \ldots, m$, that

$$H_\varepsilon^l(w(t+1), x(t)) = f_\varepsilon(x^l(t)),$$

where $v_i = w_l^i(t+1)$, $i = 1, 2, \ldots, m$. Therefore, by applying Lemma 4.1.2(iii) with $x = x^l(t+1)$ and $y = x^l(t)$, we get

$$H_\varepsilon^l(w(t+1), x(t+1)) \leq H_\varepsilon^l(w(t+1), x(t)) + \left\langle \nabla_{x^l} H_\varepsilon^l(w(t+1), x(t)), x(t+1) - x(t)\right\rangle + \frac{L_\varepsilon^l(w(t+1), x(t))}{2} \|x^l(t+1) - x^l(t)\|^2.$$

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

$$H_\varepsilon(w(t+1), x(t+1)) \leq H_\varepsilon(w(t+1), x(t)) + \sum_{l=1}^{k} \frac{L_\varepsilon^l(w(t+1), x(t))}{2} \|x^l(t+1) - x^l(t)\|^2 + \sum_{l=1}^{k} \left\langle \nabla_{x^l} H_\varepsilon(w(t+1), x(t)), x^l(t+1) - x^l(t)\right\rangle.$$

Replacing the last term with the following compact form

$$\sum_{l=1}^{k} \left\langle \nabla_{x^l} H_\varepsilon(w(t+1), x(t)), x^l(t+1) - x^l(t)\right\rangle = \langle \nabla_x H_\varepsilon(w(t+1), x(t)), x(t+1) - x(t)\rangle,$$

and the result follows. □

Now we are finally ready to prove the two properties needed for guaranteeing that the sequence $\{z(t)\}_{t\in\mathbb{N}}$ which is generated by ε-KPALM converges to a critical point of Ψ_ε.

Proposition 4.1.3 (Sufficient decrease property)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM. Then, there exists $\rho_1 > 0$ such that*

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

Proof. As we already mentioned, the step with respect to w of KPALM and ε-KPALM are similar in nature and therefore following the same arguments given at the beginning of the proof of Proposition 3.1.2 we have that

$$\frac{\underline{\alpha}}{2}\|w(t+1)-w(t)\|^2 \leq H_\varepsilon(w(t),x(t)) - H_\varepsilon(w(t+1),x(t)), \tag{4.1.7}$$

where $\underline{\alpha} = \min\limits_{1\leq i\leq m} \alpha_i$. Applying Proposition 4.1.2 and using (4.1.6) we get for all $t \in \mathbb{N}$ that

$$\begin{aligned} H_\varepsilon(w(t+1),x(t)) - H_\varepsilon(w(t+1),x(t+1)) \geq \\ &\geq \sum_{l=1}^{k} \frac{L_\varepsilon^l(w(t+1),x(t))}{2}\|x^l(t+1)-x^l(t)\|^2 \\ &\geq \frac{\underline{\beta}}{\left({d_{\mathcal{A}}}^2+\varepsilon^2\right)^{1/2}} \sum_{l=1}^{k} \|x^l(t+1)-x^l(t)\|^2 \\ &\geq \frac{\underline{\beta}}{\left({d_{\mathcal{A}}}^2+\varepsilon^2\right)^{1/2}} \|x(t+1)-x(t)\|^2, \end{aligned} \tag{4.1.8}$$

where the second inequality follows from Proposition 4.1.1(i). Set $\rho_1 = \frac{1}{2}\min\left\{\underline{\alpha}, \underline{\beta}/\left(d_{\mathcal{A}}^2+\varepsilon^2\right)^{1/2}\right\}$. Summing (4.1.7) and (4.1.8) yields

$$\begin{aligned} \rho_1\|z(t+1)-z(t)\|^2 &= \rho_1\left(\|w(t+1)-w(t)\|^2 + \|x(t+1)-x(t)\|^2\right) \\ &\leq [H_\varepsilon(w(t),x(t)) - H_\varepsilon(w(t+1),x(t))] \\ &\qquad + [H_\varepsilon(w(t+1),x(t)) - H_\varepsilon(w(t+1),x(t+1))] \\ &= H_\varepsilon(z(t)) - H_\varepsilon(z(t+1)) \\ &= \Psi_\varepsilon(z(t)) - \Psi_\varepsilon(z(t+1)), \end{aligned}$$

where the last equality follows from the fact that $G(w(t)) = 0$, since $w(t) \in \Delta^m$ for all $t \in \mathbb{N}$. This proves the desired result. □

The next two lemmas will be useful in proving the subgradient lower bounds for the iterates gap property of the sequence generated by ε-KPALM.

Lemma 4.1.3. *For all $y, z \in \mathbb{R}^n$ the following statement holds true*

$$\|\nabla f_\varepsilon(y) - \nabla f_\varepsilon(z)\| \leq \frac{2L_\varepsilon(z)L_\varepsilon(y)}{L_\varepsilon(z)+L_\varepsilon(y)}\|z-y\|.$$

Proof. Let $z \in \mathbb{R}^n$ be a fixed vector. Define the following function

$$\widetilde{f}_\varepsilon(y) = f_\varepsilon(y) - \langle \nabla f_\varepsilon(z), y\rangle,$$

hence,

$$f_\varepsilon(y) = \widetilde{f}_\varepsilon(y) + \langle \nabla f_\varepsilon(z), y \rangle . \tag{4.1.9}$$

Substituting (4.1.9) into Lemma 4.1.2(iii) yields

$$\widetilde{f}_\varepsilon(x) \leq \widetilde{f}_\varepsilon(y) + \left\langle \nabla \widetilde{f}_\varepsilon(y), x - y \right\rangle + \frac{L_\varepsilon(y)}{2} \|x - y\|^2. \tag{4.1.10}$$

It is clear that the optimal point of $\widetilde{f}_\varepsilon$ is z since $\nabla \widetilde{f}_\varepsilon(z) = 0$, therefore using (4.1.10) with
$x = y - (1/L_\varepsilon(y))\, \nabla \widetilde{f}_\varepsilon(y)$ yields

$$\begin{aligned}
\widetilde{f}_\varepsilon(z) &\leq \widetilde{f}_\varepsilon \left(y - \frac{1}{L_\varepsilon(y)} \nabla \widetilde{f}_\varepsilon(y) \right) \\
&\leq \widetilde{f}_\varepsilon(y) + \left\langle \nabla \widetilde{f}_\varepsilon(y), -\frac{1}{L_\varepsilon(y)} \nabla \widetilde{f}_\varepsilon(y) \right\rangle + \frac{L_\varepsilon(y)}{2} \left\| \frac{1}{L_\varepsilon(y)} \nabla \widetilde{f}_\varepsilon(y) \right\|^2 \\
&= \widetilde{f}_\varepsilon(y) - \frac{1}{2L_\varepsilon(y)} \left\| \nabla \widetilde{f}_\varepsilon(y) \right\|^2 .
\end{aligned}$$

Thus, using the definition of $\widetilde{f}_\varepsilon$ and the fact that $\nabla \widetilde{f}_\varepsilon(y) = \nabla f_\varepsilon(y) - \nabla f_\varepsilon(z)$, yields that

$$f_\varepsilon(z) \leq f_\varepsilon(y) + \langle \nabla f_\varepsilon(z), z - y \rangle - \frac{1}{2L_\varepsilon(y)} \|\nabla f_\varepsilon(y) - \nabla f_\varepsilon(z)\|^2.$$

Now, following the same arguments we can show that

$$f_\varepsilon(y) \leq f_\varepsilon(z) + \langle \nabla f_\varepsilon(y), y - z \rangle - \frac{1}{2L_\varepsilon(z)} \|\nabla f_\varepsilon(z) - \nabla f_\varepsilon(y)\|^2.$$

Combining the last two inequalities yields that

$$\left(\frac{1}{2L_\varepsilon(z)} + \frac{1}{2L_\varepsilon(y)} \right) \|\nabla f_\varepsilon(y) - \nabla f_\varepsilon(z)\|^2 \leq \langle \nabla f_\varepsilon(z) - \nabla f_\varepsilon(y), z - y \rangle ,$$

that is,

$$\|\nabla f_\varepsilon(y) - \nabla f_\varepsilon(z)\| \leq \frac{2L_\varepsilon(z) L_\varepsilon(y)}{L_\varepsilon(z) + L_\varepsilon(y)} \|z - y\|,$$

for all $z, y \in \mathbb{R}^n$. This proves the desired result. □

Lemma 4.1.4. *For any $x, y \in \mathbb{R}^{nk}$ such that $x^l, y^l \in Conv(\mathcal{A})$ for all $1 \leq l \leq k$ the following inequality holds*

$$\|d^i_\varepsilon(x) - d^i_\varepsilon(y)\| \leq \frac{d_{\mathcal{A}}}{\varepsilon} \|x - y\|, \quad \forall\, i = 1, 2, \ldots, m,$$

with $d_{\mathcal{A}} = diam(Conv(\mathcal{A}))$ and $d^i_\varepsilon(\cdot)$ is defined in (4.1.3).

Proof. Define $\psi(t) = \sqrt{t+\varepsilon^2}$, for $t \geq 0$. Using the Lagrange mean value theorem over $a > b \geq 0$ yields

$$\frac{\psi(a)-\psi(b)}{a-b} = \psi'(c) = \frac{1}{2\sqrt{c+\varepsilon^2}} \leq \frac{1}{2\varepsilon},$$

where $c \in (b,a)$. Therefore, for all $i = 1,2,\ldots,m$ and $l = 1,2,\ldots,k$ we have

$$\begin{aligned}
\left|\left(\|x^l-a^i\|^2+\varepsilon^2\right)^{1/2} - \left(\|y^l-a^i\|^2+\varepsilon^2\right)^{1/2}\right| &\leq \frac{1}{2\varepsilon}\left|\|x^l-a^i\|^2+\varepsilon^2-\left(\|y^l-a^i\|^2+\varepsilon^2\right)\right| \\
&= \frac{1}{2\varepsilon}\left|\|x^l-a^i\|^2-\|y^l-a^i\|^2\right| \\
&= \frac{1}{2\varepsilon}\left|\|x^l-a^i\|+\|y^l-a^i\|\right|\cdot\left|\|x^l-a^i\|-\|y^l-a^i\|\right| \\
&\leq \frac{1}{\varepsilon}\, d_{\mathcal{A}}\|x^l-y^l\|,
\end{aligned}$$

where the last inequality follows from $\left\|x^l-a^i\right\|, \left\|y^l-a^i\right\| \leq d_{\mathcal{A}}$ and the reverse triangle inequality. Therefore,

$$\begin{aligned}
\|d^i_\varepsilon(x)-d^i_\varepsilon(y)\| &= \left[\sum_{l=1}^{k}\left|\left(\|x-a^i\|^2+\varepsilon^2\right)^{1/2}-\left(\|y-a^i\|^2+\varepsilon^2\right)^{1/2}\right|^2\right]^{\frac{1}{2}} \\
&\leq \left[\sum_{l=1}^{k}\left(\frac{1}{\varepsilon}\, d_{\mathcal{A}}\|x^l-y^l\|\right)^2\right]^{\frac{1}{2}} \\
&= \frac{d_{\mathcal{A}}}{\varepsilon}\|x-y\|,
\end{aligned}$$

as asserted. □

Proposition 4.1.4 (Subgradient lower bound for the iterates gap)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM. Then, there exists $\rho_2 > 0$ and $\gamma(t+1) \in \partial\Psi_\varepsilon(z(t+1))$ such that*

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

Proof. Repeating the steps of the proof in the case of KPALM (see Proposition 3.1.3) yields that

$$\begin{aligned}
\gamma(t+1) := &\left(\left(d^i_\varepsilon(x(t+1))+u^i(t+1)\right)_{i=1,\ldots,m}, \nabla_x H_\varepsilon(w(t+1),x(t+1))\right) \\
&\in \partial\Psi_\varepsilon(z(t+1)),
\end{aligned} \tag{4.1.11}$$

where $u^i(t+1) \in \partial\delta_\Delta(w^i(t+1))$, $i = 1,2,\ldots,m$. Now, writing the optimality condition of step (4.1.5), yields that

$$d^i_\varepsilon(x(t)) + \alpha_i(t)\left(w^i(t+1)-w^i(t)\right) + u^i(t+1) = \mathbf{0}. \tag{4.1.12}$$

Plugging (4.1.12) into (4.1.11), and taking the norm yields

$$\begin{aligned}
\|\gamma(t+1)\| &\leq \sum_{i=1}^{m} \|d_\varepsilon^i(x(t+1)) - d_\varepsilon^i(x(t)) - \alpha_i(t)\left(w^i(t+1) - w^i(t)\right)\| \\
&\qquad + \|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| \\
&\leq \sum_{i=1}^{m} \|d_\varepsilon^i(x(t+1)) - d_\varepsilon^i(x(t))\| + \sum_{i=1}^{m} \alpha_i(t)\|w^i(t+1) - w^i(t)\| \\
&\qquad + \|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| \\
&\leq \frac{m d_{\mathcal{A}}}{\varepsilon}\|x(t+1) - x(t)\| + \overline{\alpha}\sqrt{m}\|w(t+1) - w(t)\| \\
&\qquad + \|\nabla_x H_\varepsilon(w(t+1), x(t+1))\|,
\end{aligned}$$

where the last inequality follows from Lemma 4.1.4 and the fact that $\overline{\alpha} = \max_{1\leq i\leq m} \overline{\alpha_i}$. Next we will show that $\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| \leq c\|x(t+1) - x(t)\|$, for some constant $c > 0$. Indeed, for all $l = 1, 2, \ldots, k$, we have

$$\begin{aligned}
\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) &= \nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\
&\qquad + \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\
&= \nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\
&\qquad + L_\varepsilon^l(w(t+1), x(t))\left(x^l(t) - x^l(t+1)\right), \qquad (4.1.13)
\end{aligned}$$

where the last equality follows from (4.1.6). Therefore,

$$\begin{aligned}
\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| &\leq \sum_{l=1}^{k} \|\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1))\| \\
&\leq \sum_{l=1}^{k} L_\varepsilon^l(w(t+1), x(t))\|x^l(t+1) - x^l(t)\| \\
&\qquad + \sum_{l=1}^{k} \|\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t))\| \\
&\leq \frac{m}{\varepsilon}\sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\| \\
&\qquad + \sum_{l=1}^{k} \gamma^l(t)\|x^l(t+1) - x^l(t)\|, \qquad (4.1.14)
\end{aligned}$$

where the last inequality follows from Proposition 4.1.1(ii) and Lemma 4.1.3 using

$$\gamma^l(t) = \frac{2L_\varepsilon^l(w(t+1), x(t))L_\varepsilon^l(w(t+1), x(t+1))}{L_\varepsilon^l(w(t+1), x(t)) + L_\varepsilon^l(w(t+1), x(t+1))}, \quad l = 1, 2, \ldots, k.$$

From Proposition 4.1.1(ii) we obtain that

$$\gamma^l(t) = \frac{2}{\frac{1}{L_\varepsilon^l(w(t+1),x(t))} + \frac{1}{L_\varepsilon^l(w(t+1),x(t+1))}} \leq \frac{2}{\frac{\varepsilon}{m} + \frac{\varepsilon}{m}} = \frac{m}{\varepsilon}.$$

Hence, from (4.1.14), we have

$$\begin{aligned} \|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| &\leq \frac{2m}{\varepsilon} \sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\| \\ &\leq \frac{2m\sqrt{k}}{\varepsilon} \|x(t+1) - x(t)\|. \end{aligned} \tag{4.1.15}$$

Therefore, setting $\rho_2 = \frac{md_{\mathcal{A}}}{\varepsilon} + \overline{\alpha}\sqrt{m} + \frac{2m\sqrt{k}}{\varepsilon}$, yields the result. □

4.2 Different Approach Towards Solving the Smoothed Hε

In this section we describe a different approach towards solving the smoothed clustering problem described in (4.1.1). Using the Arithmetic-Geometric inequality we derive the following simple observation

$$\frac{1}{2} \min_{s \geq 0} \left\{ s\lambda + \frac{1}{\lambda} \right\} \geq \min_{s \geq 0} \left\{ \sqrt{s\lambda \cdot \frac{1}{s}} \right\} = \sqrt{\lambda}, \quad \forall\, \lambda \geq 0,$$

and the unique minimizer is given by $s^* = 1/\sqrt{\lambda}$. Using this fact we can write

$$\sqrt{\|u\|^2 + \varepsilon^2} = \frac{1}{2} \min_{v \geq 0} \left\{ v \left(\|u\|^2 + \varepsilon^2 \right) + \frac{1}{v} \right\}, \tag{4.2.16}$$

with $v^* = 1/\sqrt{\|u\|^2 + \varepsilon^2}$. Thus, instead of solving problem (4.1.1) with $H_\varepsilon(\cdot,\cdot)$, defined in (4.1.2), we replace it with the following function

$$B_\varepsilon(v, w, x) = \frac{1}{2} \sum_{i=1}^{m} \sum_{l=1}^{k} \left\{ v_l^i w_l^i \left(\left\| x^l - a^i \right\|^2 + \varepsilon^2 \right) + \frac{w_l^i}{v_l^i} \right\}, \tag{4.2.17}$$

where $v = (v^1, v^2, \ldots, v^m)$, and then problem (4.1.1) can be written equivalently as

$$\min_{x,v,w} \left\{ B_\varepsilon(v, w, x) + G(w) : \; v \geq 0 \right\}.$$

For all $i = 1, 2, \ldots, m$ we define $b_\varepsilon^i : \mathbb{R}^{mk} \times \mathbb{R}^{nk} \to \mathbb{R}^k$ by

$$b_\varepsilon^i(v, x) = \left(\frac{1}{2} v_l^i \left(\left\| x^l - a^i \right\|^2 + \varepsilon^2 \right) + \frac{1}{2v_l^i} \right)_{l=1,2,\ldots,k} \in \mathbb{R}^k,$$

and we have that

$$B_\varepsilon(v,w,x) = \sum_{i=1}^{m} \left\langle w^i, b_\varepsilon^i(v,x) \right\rangle . \tag{4.2.18}$$

Now the situation is similar to that of Section 3.1, namely

(1) The function $w \mapsto B_\varepsilon(v,w,x)$, for fixed v and x, is linear;

(2) The function $x \mapsto B_\varepsilon(v,w,x)$, for fixed v and w, is quadratic and convex.

Hence we can tackle these two steps as in KPALM.

Equipped with these observation we proceed to a PALM-like algorithm, which is based on three steps alternating minimization. More precisely, with respect to v we perform exact minimization

$$v(t+1) = \operatorname{argmin} \{B_\varepsilon(v, w(t), x(t)) : \ v \geq 0\}$$

It should be noted that this problem can be written equivalently by

$$v(t+1) = \operatorname{argmin} \left\{B_\varepsilon(v, w(t), x(t)) : \ v \in I^{mk}\right\},$$

where $I := [1/\kappa, 1/\varepsilon]$ and $\kappa = \sqrt{d_{\mathcal{A}}^2 + \varepsilon^2}$. With respect to w, as in KPALM case, for each $i = 1, 2, \ldots, m$, we need to solve the subproblem given by

$$w^i(t+1) = \arg\min_{w^i \in \Delta} \left\{ \langle w^i, b_\varepsilon^i(v(t+1), x(t)) \rangle + \frac{\alpha_i(t)}{2} \|w^i - w^i(t)\|^2 \right\}, \tag{4.2.19}$$

where $\alpha_i(t) > 0$, $i = 1, 2, \ldots, m$. With respect to x, again as in KPALM case, we perform exact minimization

$$x(t+1) = \operatorname{argmin} \left\{B_\varepsilon(v(t+1), w(t+1), x) \mid x \in \mathbb{R}^{nk}\right\}. \tag{4.2.20}$$

It is easy to check that explicit solutions to all three subproblems are given by

$$v_l^i(t+1) = \frac{1}{\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2}}, \quad i = 1, 2, \ldots, m, \ l = 1, 2, \ldots, k, \tag{4.2.21}$$

$$w^i(t+1) = P_\Delta \left(w^i(t) - \frac{b_\varepsilon^i(v(t+1), x(t))}{\alpha_i(t)} \right), \quad i = 1, 2, \ldots, m, \tag{4.2.22}$$

and

$$x^l(t+1) = \frac{\sum_{i=1}^m w_l^i(t+1) v_l^i(t+1) a^i}{\sum_{i=1}^m w_l^i(t+1) v_l^i(t+1)}, \quad l = 1, 2, \ldots, k. \tag{4.2.23}$$

From the subproblem for v and the observation given in (4.2.16), we derive the following three relations

$$B_\varepsilon(v(t+1), w, x(t)) = H_\varepsilon(w, x(t)), \quad \forall\, t \in \mathbb{N}, \ \forall\, w \in \Delta^m, \tag{4.2.24}$$

$$b_\varepsilon^i(v(t+1), x(t)) = d_\varepsilon^i(x(t)), \quad \forall\, t \in \mathbb{N},\ i = 1, 2, \ldots, m, \tag{4.2.25}$$

where d_ε^i is defined in (4.1.3), and

$$B_\varepsilon(v, w, x) \geq H_\varepsilon(w, x), \quad \forall\, (v, w, x) \in I^{mk} \times \Delta^m \times \mathbb{R}^{nk}. \tag{4.2.26}$$

Substituting (4.2.25) into (4.2.22) yields

$$w^i(t+1) = P_\Delta\left(w^i(t) - \frac{d_\varepsilon^i(x(t))}{\alpha_i(t)}\right), \quad i = 1, 2, \ldots, m.$$

Moreover, substituting (4.2.21) into (4.2.23) yields

$$x^l(t+1) = \frac{1}{L_\varepsilon^l(w(t+1), x(t))} \sum_{i=1}^{m} \left(\frac{w_l^i(t+1)}{\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2}}\right) a^i, \quad l = 1, 2, \ldots, k,$$

where L_ε^l is defined in (4.1.4). Thus, we recover the ε-KPALM algorithm, which means these two different approaches lead to the same iterative algorithm. However, with the current approach we can swiftly prove the sufficient decrease and the subgradient lower bound for the iterates gap properties, which are needed to obtain global convergence of $\{z(t)\}_{t\in\mathbb{N}}$ that generated by ε-KPALM.

Proposition 4.2.1 (Sufficient decrease property)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM. Then, there exists $\rho_1 > 0$ such that*

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

Proof. From (4.2.19) we have

$$\left\langle w^i(t+1), b_\varepsilon^i(v(t+1), x(t))\right\rangle + \frac{\alpha_i(t)}{2}\left\|w^i(t+1) - w^i(t)\right\|^2 \leq \left\langle w^i(t), b_\varepsilon^i(v(t+1), x(t))\right\rangle.$$

Summing the last inequality over $i = 1, 2, \ldots, m$ and applying (4.2.18) yields

$$B_\varepsilon(v(t+1), w(t+1), x(t)) + \sum_{i=1}^{m} \frac{\alpha_i(t)}{2}\left\|w^i(t+1) - w^i(t)\right\|^2 \leq B_\varepsilon(v(t+1), w(t), x(t)).$$

Using Assumption 3.1.1(i) we derive

$$\begin{aligned}\frac{\alpha}{2}\|w(t+1) - w(t)\|^2 &\leq B_\varepsilon(v(t+1), w(t), x(t)) - B_\varepsilon(v(t+1), w(t+1), x(t)) \\ &\leq H_\varepsilon(w(t), x(t)) - H_\varepsilon(w(t+1), x(t)),\end{aligned} \tag{4.2.27}$$

where the last inequality follows from (4.2.24) and (4.2.26).

Since the function $x \mapsto B_\varepsilon(v, w, x)$ is C^2, and

$$\nabla_{x^j}\nabla_{x^l}B_\varepsilon(v,w,x) = \begin{cases} 0 & \text{if } j \neq l, \quad 1 \leq j,l \leq k, \\ \sum_{i=1}^{m} w_l^i v_l^i & \text{if } j = l, \quad 1 \leq j,l \leq k, \end{cases}$$

it follows that, the function $x \mapsto B_\varepsilon(v(t+1), w(t), x)$ is strongly convex with parameter $\underline{\beta}/2\kappa$, for all $t \in \mathbb{N}$. Indeed,

$$\nabla^2_{x^l}B_\varepsilon(v(t+1), w(t), x) = \sum_{i=1}^{m} w_l^i(t)v_l^i(t+1) \geq \frac{1}{\kappa}\sum_{i=1}^{m} w_l^i(t) \geq \frac{\beta(w^i(t))}{2\kappa} > \frac{\underline{\beta}}{2\kappa} > 0,$$

where the first inequality follows from the fact that $v_l^i(t) \in I$ for all $t \in \mathbb{N}$, $\beta(\cdot)$ is defined in (3.1.7), and the second inequality is due to Assumption 3.1.1(ii). Using the strong convexity property we deduce the sufficient decrease in x, as follows,

$$\begin{aligned} \frac{\underline{\beta}}{4\kappa}\|x(t+1) - x(t)\|^2 &= \langle \nabla_x B_\varepsilon(z(t+1)), x(t) - x(t+1)\rangle + \frac{\underline{\beta}}{4\kappa}\|x(t+1) - x(t)\|^2 \\ &\leq B_\varepsilon(v(t+1), w(t+1), x(t)) - B_\varepsilon(z(t+1)) \\ &\leq H_\varepsilon(w(t+1), x(t)) - H_\varepsilon(w(t+1), x(t+1)), \end{aligned} \tag{4.2.28}$$

where the first equality follows from (4.2.20), the second inequality follows from the strong convexity, and the last inequality is due to (4.2.24) and (4.2.26). Set $\rho_1 = \min\{\underline{\alpha}/2, \underline{\beta}/4\kappa\}$. Summing (4.2.27) and (4.2.28), we get

$$\begin{aligned} \rho_1\|z(t+1) - z(t)\|^2 &= \rho_1\left(\|w(t+1) - w(t)\|^2 + \|x(t+1) - x(t)\|^2\right) \\ &\leq [H_\varepsilon(z(t)) - H_\varepsilon(w(t+1), x(t))] \\ &\qquad + [H_\varepsilon(w(t+1), x(t)) - H_\varepsilon(z(t+1))] \\ &= H_\varepsilon(z(t)) - H_\varepsilon(z(t+1)) \\ &= \Psi_\varepsilon(z(t)) - \Psi_\varepsilon(z(t+1)), \end{aligned}$$

where the last equality follows from the fact that $G(w(t)) = 0$, since $w(t) \in \Delta^m$ for all $t \in \mathbb{N}$. This proves the desired result. □

Proposition 4.2.2 (Subgradient lower bound for the iterates gap)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM. Then, there exists $\rho_2 > 0$ and $\gamma(t+1) \in \partial\Psi_\varepsilon(z(t+1))$ such that*

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

Proof. By the definition of Ψ_ε (see (4.1.1)) we get

$$\Psi_\varepsilon(w,x) = H_\varepsilon(w,x) + \sum_{i=1}^{m}\delta_\Delta(w^i) \tag{4.2.29}$$

Differentiating (4.2.29) with respect to x and evaluating it in $z(t+1)$ yields

$$\partial_x \Psi_\varepsilon(z(t+1)) = \nabla_x H_\varepsilon(z(t+1)). \tag{4.2.30}$$

Similarly, differentiating (4.2.29) with respect to w^i and evaluating it in $z(t+1)$ yields

$$\partial_{w^i} \Psi_\varepsilon(z(t+1)) = d^i_\varepsilon(x(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)). \tag{4.2.31}$$

The optimality condition of $w^i(t+1)$ which derived from (4.2.19), yields that for all $i = 1, 2, \ldots, m$ there exists $u^i(t+1) \in \partial\delta_\Delta(w^i(t+1))$ such that

$$\begin{aligned} \mathbf{0} &= b^i_\varepsilon(v(t+1), x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1) \\ &= d^i_\varepsilon(x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1), \end{aligned} \tag{4.2.32}$$

where the last equality follows from (4.2.25). Substituting (4.2.32) into (4.2.31) and combining with (4.2.30) we deduce that

$$\begin{aligned} \gamma(t+1) := &\Big(\left(d^i(x(t+1)) - d^i(x(t)) - \alpha_i(t)(w^i(t+1) - w^i(t))\right)_{i=1,2,\ldots,m}, \nabla_x H_\varepsilon(z(t+1))\Big) \\ &\in \partial\Psi_\varepsilon(z(t+1)). \end{aligned}$$

Therefore,

$$\begin{aligned} \|\gamma(t+1)\| &\le \sum_{i=1}^m \left\|d^i(x(t+1)) - d^i(x(t)) - \alpha_i(t)(w^i(t+1) - w^i(t))\right\| + \|\nabla_x H_\varepsilon(z(t+1))\| \\ &\le \sum_{i=1}^m \left\|d^i(x(t+1)) - d^i(x(t))\right\| + \overline{\alpha} \sum_{i=1}^m \left\|w^i(t+1) - w^i(t))\right\| \\ &\qquad + \frac{2m\sqrt{k}}{\varepsilon} \|x(t+1) - x(t)\| \\ &\le \sum_{i=1}^m \frac{d_{\mathcal{A}}}{\varepsilon} \|x(t+1) - x(t)\| + \overline{\alpha}\sqrt{m} \|w(t+1) - w(t)\| \\ &\qquad + \frac{2m\sqrt{k}}{\varepsilon} \|x(t+1) - x(t)\| \\ &\le \left(\frac{md_{\mathcal{A}}}{\varepsilon} + \overline{\alpha}\sqrt{m} + \frac{2m\sqrt{k}}{\varepsilon}\right) \|z(t+1) - z(t)\|, \end{aligned}$$

where the second inequality was established in Proposition 4.1.4 (see (4.1.15)) and the third inequality follows from Lemma 4.1.4. Define $\rho_2 = \frac{md_{\mathcal{A}}}{\varepsilon} + \overline{\alpha}\sqrt{m} + \frac{2m\sqrt{k}}{\varepsilon}$, and the result follows. □

Chapter 5

Returning to k-means

This chapter proves that the k-means algorithm can be recovered from the model developed in Chapter 2 and it is a special case of KPALM algorithm. We prove the convergence of k-means to critical point, again with the general methodology of Section 2.2. Assuming the uniqueness of labeling in the output of k-means algorithm we improve the convergence result to local minimum.

5.1 Similarity of KPALM to k-means

The famous k-means algorithm has close relation to KPALM algorithm. k-means alternates between cluster assignment and centers update steps as well. In detail, we can write its steps in the following manner

k-means

(1) Initialization: $x(0) \in \mathbb{R}^{nk}$.

(2) General step ($t = 0, 1, \ldots$):

(2.1) Cluster assignment: for $i = 1, 2, \ldots, m$ compute

$$w^i(t+1) = \arg\min_{w^i \in \Delta} \left\{ \langle w^i, d^i(x(t)) \rangle \right\}. \tag{5.1.1}$$

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

$$x^l(t+1) = \frac{\sum_{i=1}^m w_l^i(t+1) a^i}{\sum_{i=1}^m w_l^i(t+1)}. \tag{5.1.2}$$

It is easy to see that if we take $\alpha_i(t) = 0$ for all $1 \leq i \leq m$ and $t \in \mathbb{N}$, then KPALM becomes k-means. We aim to use the theory described in Section 2.2 once again and show that the sequence generated by k-means converges to a critical point of $\Psi(\cdot)$, as defined is (2.1.4). The sufficient decrease proof of Section 3.1 collapses in this case, since it is based on Assumption 3.1.1(i), that is, $\alpha_i(t) > \underline{\alpha_i} > 0$, for all $t \in \mathbb{N}$ and $i = 1, 2, \ldots, m$. However, the proof of the subgradient lower bound for the iterates gap property follows through as is. In the following discussion we present the means to treat the case that $\alpha_i(t) = 0$, and prove the sufficient decrease property.

Lemma 5.1.1. *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by k-means. Then, there exists $c > 0$ such that*

$$\|w^i(t+1) - w^i(t)\| \leq c\|x(t+1) - x(t)\|, \quad \forall\, i = 1, 2, \ldots\ m,\ t \in \mathbb{N}.$$

Proof. At each iteration k-means partitions the set $\mathcal{A}$ into k clusters, and the center of each cluster is its mean. Since the number of these partitions if finite, there exists a finite set $\mathcal{C} = \left\{x^1, x^2, \ldots, x^N\right\} \subset \mathbb{R}^{nk}$ such that for all $t \in \mathbb{N}$, $x(t) \in \mathcal{C}$. We denote

$$r = \min_{1\leq j<l\leq N} \|x^j - x^l\|,$$

and set $c = \sqrt{2}/r$. At each iteration, the point a^i can move from one cluster to another, hence

$$\|w^i(t+1) - w^i(t)\| \leq \sqrt{2}.$$

Therefore, combining these arguments yields

$$\frac{\|w^i(t+1) - w^i(t)\|}{\|x(t+1) - x(t)\|} \leq \frac{\sqrt{2}}{r}.$$

In case that $x(t+1) = x(t)$, this implies that none of the clusters has changed, hence we proved the statement in both cases. □

Equipped with the last lemma we briefly prove the sufficient decrease property of k-means.

Proposition 5.1.1 (Sufficient decrease property for k-means sequence)**.** *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by k-means. Then, there exists $\rho_1 > 0$ such that*

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

Proof. The function $x \mapsto H(w(t), x)$ remains strongly convex with parameter $\beta(w(t))$ (see (3.1.10)), hence we have a sufficient decrease in the x variable, namely,

$$\frac{\beta}{2}\|x(t+1) - x(t)\|^2 \leq H(w(t), x(t)) - H(w(t+1), x(t+1)). \tag{5.1.3}$$

Setting $\rho_1 = \underline{\beta}/2(1+mc^2)$, we can write

$$\begin{aligned}
\rho_1\|z(t+1)-z(t)\|^2 &= \rho_1 \sum_{i=1}^{m} \|w^i(t+1)-w^i(t)\|^2 + \rho_1\|x(t+1)-x(t)\|^2 \\
&\leq \rho_1(1+mc^2)\|x(t+1)-x(t)\|^2 \\
&\leq H(w(t),x(t)) - H(w(t+1),x(t+1)) \\
&= \Psi(z(t)) - \Psi(z(t+1))
\end{aligned}$$

where the first inequality follows from Lemma 5.1.1, the second follows from (5.1.3), and the last equality follows from the fact that $G(w(t)) = 0$, for all $t \in \mathbb{N}$. □

5.2 k-means Local Minima Convergence Proof

In this section we present a simple and direct proof that k-means converges to local minima. We start with rewriting the k-means algorithm, in its most familiar form

k-means

(1) Initialization: $x(0) \in \mathbb{R}^{nk}$.

(2) General step ($t = 0, 1, \ldots$):

(2.1) Cluster assignment: for $l = 1, 2, \ldots, k$ compute

$$C^l(t+1) = \left\{a \in \mathcal{A} : \|a - x^l(t)\| \leq \|a - x^j(t)\|, \quad \forall 1 \leq j \leq k\right\}. \tag{5.2.4}$$

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

$$x^l(t+1) = mean(C^l(t+1)) := \frac{1}{|C^l(t+1)|} \sum_{a \in C^l(t+1)} a. \tag{5.2.5}$$

(2.3) Stopping criteria: halt if

$$\forall 1 \leq l \leq k \quad C^l(t+1) = C^l(t) \tag{5.2.6}$$

As in KPALM, k-means needs Assumption 3.1.1(ii) for step (5.2.5) to be well defined. In order to prove the convergence of k-means to local minimum, we will need to following assumption.

Assumption 5.2.1. *Let $t \in \mathbb{N}$ be the final iteration of k-means run, then we assume that each $a \in \mathcal{A}$ belongs exclusively to single cluster $C^l(t)$.*

For any $x \in \mathbb{R}^{nk}$ we denote the super-partition of $\mathcal{A}$ with respect to x by $\overline{C^l}(x) = \{a \in \mathcal{A} \mid \|a - x^l\| \leq \|a - x^j\|, \quad \forall j \neq l\}$, for all $1 \leq l \leq k$, and the sub-partition of $\mathcal{A}$ by $\underline{C^l}(x) = \{a \in \mathcal{A} \mid \|a - x^l\| < \|a - x^j\|, \quad \forall j \neq l\}$. Moreover, denote $R_{lj}(t) = \min\limits_{a \in C^l(t)} \{\|a - x^j(t)\| - \|a - x^l(t)\|\}$ for all $1 \leq l, j \leq k$, and $r(t) = \min\limits_{l \neq j} R_{lj}$.
Due to Assumption 5.2.1 we have that $\overline{C^l}(x(t)) = \underline{C^l}(x(t)) = C^l(t+1)$, for all $1 \leq l \leq k$, $t \in \mathbb{N}$, we also have that $r(t) > 0$ for all $t \in \mathbb{N}$.

Proposition 5.2.1. *Let $(C(t), x(t))$ be the clusters and centers k-means returns. Denote by $U = B\left(x^1(t), \frac{r(t)}{2}\right) \times B\left(x^2(t), \frac{r(t)}{2}\right) \times \cdots \times B\left(x^l(t), \frac{r(t)}{2}\right)$ an open neighbourhood of $x(t)$, then for any $x \in U$ we have $C^l(t) = \underline{C^l}(x)$ for all $1 \leq l \leq k$.*

Proof. Pick some $a \in C^l(t)$, then $x^l(t-1)$ is the closest center among the centers of $x(t-1)$. Since k-means halts at step t, then from (5.2.6) we have $x(t) = x(t-1)$, thus $x^l(t)$ is the closest center to a among the centers of $x(t)$. Further we have

$$r(t) \leq \|x^j(t) - a\| - \|x^l(t) - a\| \quad \forall j \neq l. \tag{5.2.7}$$

Next, we show that $a \in \underline{C^l}(x)$, indeed

$$\begin{aligned}
\|a - x^l\| - \|a - x^j\| &\leq \|a - x^l(t)\| + \|x^l(t) - x^l\| - \big(\|a - x^j(t)\| - \|x^j(t) - x^j\|\big) \\
&= \|a - x^l\| - \|a - x^j(t)\| + \|x^l(t) - x^l\| + \|x^j(t) - x^j\| \\
&< \|a - x^l\| - \|a - x^j(t)\| + r(t) \\
&\leq -r(t) + r(t) = 0,
\end{aligned}$$

where the second inequality holds since $x^l \in B\left(x^l(t), \frac{r(t)}{2}\right)$ and $x^j \in B\left(x^j(t), \frac{r(t)}{2}\right)$, and the third inequality follows from (5.2.7), and we get that $C^l(t) \subseteq \underline{C^l}(x)$. By definition of $\underline{C^l}(x)$ we have that for any $l \neq j$, $\underline{C^l}(x) \cap \underline{C^j}(x) = \emptyset$, and for all $1 \leq l \leq k$, $\underline{C^l}(x) \subseteq \mathcal{A}$. Now, since $C(t)$ is a partition of $\mathcal{A}$, then $C^l(t) = \underline{C^l}(x)$ for all $1 \leq l \leq k$. □

Proposition 5.2.2 (k-means converges to local minimum)**.** *Let $(C(t), x(t))$ be the clusters and centers k-means returns, then $x(t)$ is local minimum of F in $U = B\left(x^1(t), \frac{r(t)}{2}\right) \times B\left(x^2(t), \frac{r(t)}{2}\right) \times \cdots \times B\left(x^l(t), \frac{r(t)}{2}\right) \subset \mathbb{R}^{nk}$.*

Proof. The minimum of F in U is

$$\min_{x \in U} F(x) = \min_{x \in U} \sum_{l=1}^{k} \sum_{a \in C^l(x)} \|a - x^l\|^2 = \min_{x \in U} \sum_{l=1}^{k} \sum_{a \in C^l(t)} \|a - x^l\|^2,$$

where the last equality follows from Proposition 5.2.1.

The function $x \mapsto \sum_{l=1}^{k} \sum_{a \in C^l(t)} \|a - x^l\|^2$ is strictly convex, separable in x^l for all $1 \leq l \leq k$, and reaches its minimum at $\frac{1}{|C^l(t)|} \sum_{a \in C^l(t)} a = mean(C^l(t)) = x^l(t)$, and the result follows. $\square$

Chapter 6

Numeric Results

In this section we show the numeric results and compare the algorithms presented in this work with other algorithms that are commonly used to address the clustering problem.

The initialization points used within the implementation of the compared algorithms are as follows. k-means staring point is constructed by randomly choosing k different points from the dataset. The same technique is employed in the cases of KPALM and ε-KPALM, for the $x(0)$ variable. Whereas for the $w(0)$ variable, it is chosen at random from Δ^m. k-means++ takes also part in our comparison, and it is basically the same as k-means, with the exception of its starting point that is constructed in the following manner. The first center $x^1(0)$ is chosen randomly from the dataset $\mathcal{A}$. Suppose that $1 \leq l < k$ centers have already been chosen, set $x^{l+1}(0)$ to be the point in the dataset that is the furthest from its closest center.

Since it is impractical to compare the function values achieved with the algorithms which solve the squared Euclidean clustering problem with that of the algorithms which solve the Euclidean clustering problem, we used some criteria devised to compare clustering partitions. Criteria such as *variation of information (VI)*, *Mirkin metric*, and *Van Dongen metric*, are few examples for metrics that measure the difference between two clustering partitions (see [16]). With these metrics we compared the similarity of the partition achieved with each algorithm to the desired partition of each dataset. The goal is to decrease the value of the metrics.

6.1 Iris Dataset

We used the famous Iris dataset to test the performance of the KPALM algorithm. It is important to note that choosing the parameter α is left to the user, and as

presented below, has a significant effect on the convergence rate and the quality of the achieved clustering, namely the value of the objective function over the generated series. All the plots in this section are made by averaging over 100 trials, each trial with random starting point.

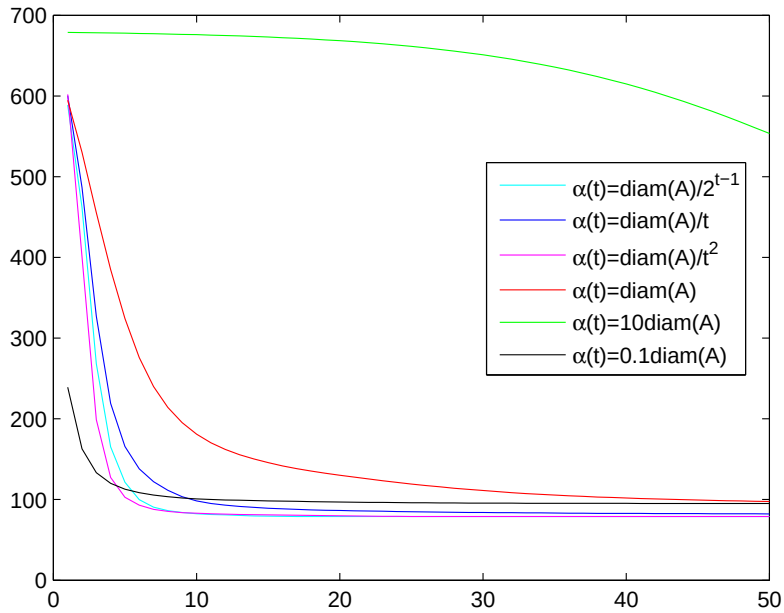


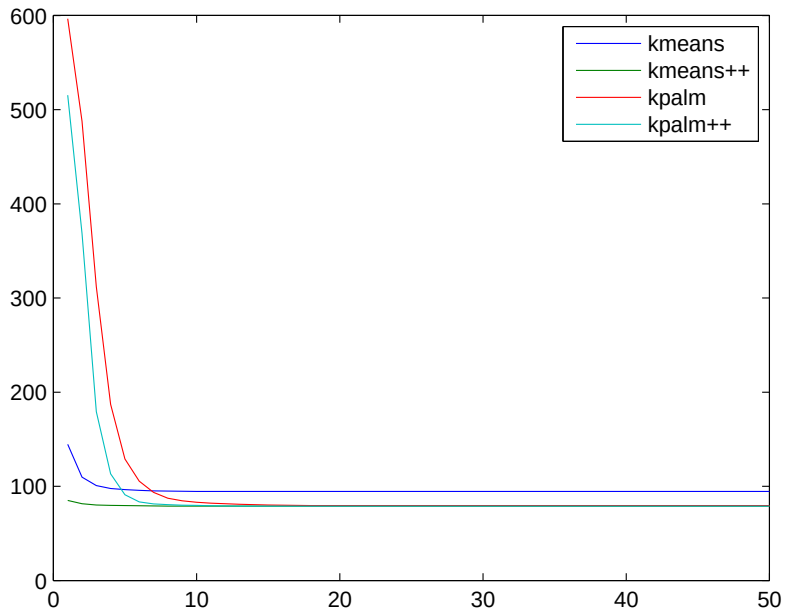
Figure 6.1: Comparison of the objective values for different values of α.

Figure 6.1 shows that dynamic values of the parameter α which decreases fast, such as $\alpha_i(t) = \frac{diam(\mathcal{A})}{2^{t-1}}$, achieve smaller function values. In Figure 6.2 we made a comparison between KPALM with dynamic rule for choosing the parameter α, that is $\alpha_i(t) = \frac{diam(\mathcal{A})}{2^{t-1}}$, with k-means and k-means++. It demonstrates that KPALM can reach lower objective function values then k-means, and these are similar to the values achieved with KMEAN++. In addition, the KPALM++ are the objective function values achieved with KPALM when the x variable is initialized as in k-means++. Unlike k-means, the objective function values KPALM converge to are more stable and less sensitive to its starting point.

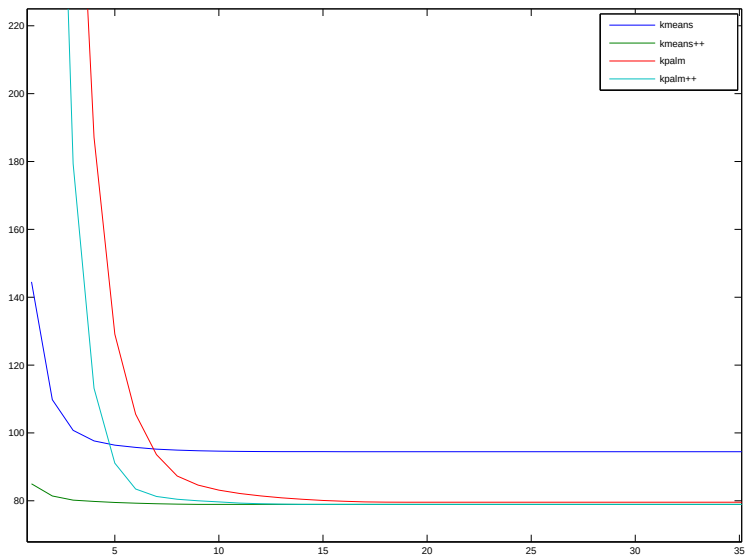
Figure 6.3 shows the number of iteration needed to reach precision of 1e-3 between consecutive objective function values. Similarly to Figure 6.1, in Figure 6.4 we can see a comparison of the objective values of Ψ_ε for different function values. The value of ε is set to be 1e-5.

6.2 Synthetic Dataset

In this section we show that ε-KPALM is less sensitive to outliners in the data verses algorithms that suit the squared Euclidean norm (e.g., k-means, k-means++ and KPALM). We generated two synthetic datasets, each contains 300 points in the plane, by sampling three two-dimensional Gaussian, 100 samples each. In Figure 6.5(6.5a) the clusters are denser than in Figure 6.5(6.5b). Then we run the clustering algorithms and compared their clustering results, namely, how many points were clustered correctly. From Figure 6.6(6.6a) it is evident that k-means is superior to other algorithms in the dense case and ε-KPALM is quite sensitive. Whereas, in the sparse case in Figure 6.6(6.6b), ε-KPALM is superior, and less sensitive to outliners. In Figure 6.7 we compare the distance of clusterings achieved with different algorithms to the desired clustering, where kpalm1, kpalm2 and kpalm3 match using $\alpha(t) = diam(\mathcal{A})/2^{t-1}$, $\alpha(t) = diam(\mathcal{A})/t^2$ and $\alpha(t) = diam(\mathcal{A})$ respectively, and similarly for ε-kpalm$i$, $i \in \{1, 2, 3\}$. In Figure 6.7(6.7a) we witness that for dense dataset, the resulting clusterings of squared Euclidean algorithms, namely, k-means, k-means++ and KPALM, are superior to the clustering ε-KPALM, where KPALM with $\alpha(t) = diam(\mathcal{A})/2^{t-1}$ gives the best result, that is, the clustering in this setting is the closest to the desired clustering. Whereas in the sparse dataset, the clustering achieved with ε-KPALM with $\alpha(t) = diam(\mathcal{A})/t^2$ is the closest to the desired clustering, as reflected from Figure 6.7(6.7b).

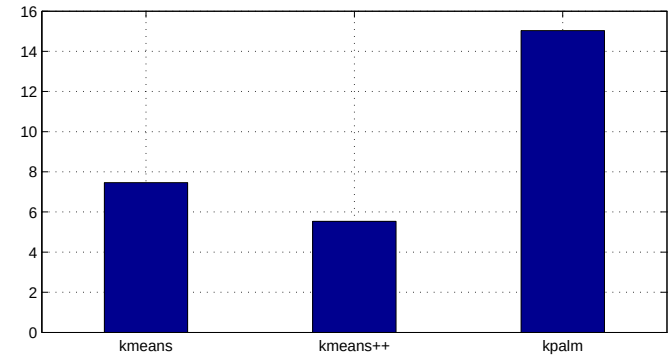


(a) Comparison of objective function values.

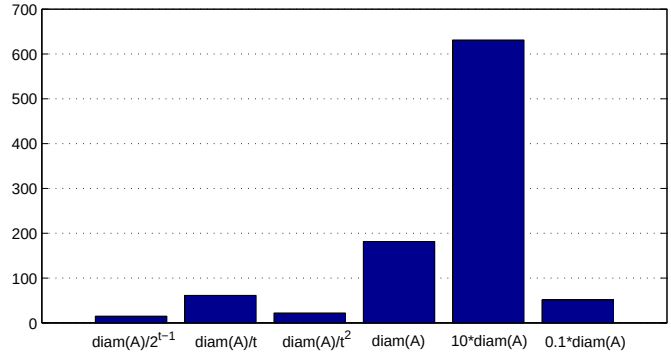


(b) Zoom of Figure 6.2a.

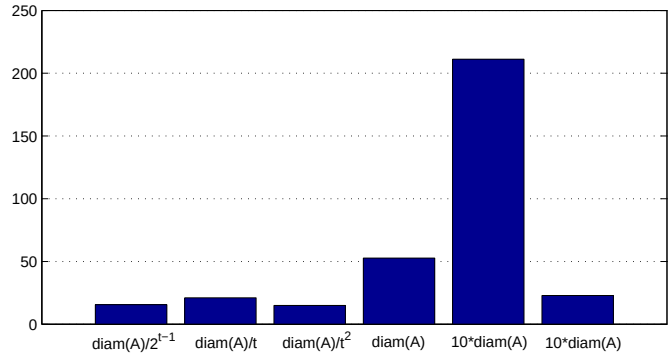
Figure 6.2: Comparison of objective function values for k-means, k-means++, KPALM and KPALM++.



(a) Number of iterations of k-means, k-means++ and KPALM with $\alpha(t) = diam(\mathcal{A})/2^{t-1}$.



(b) Number of iterations of KPALM with different updates of $\alpha(t)$.



(c) Number of iterations of ε-KPALM with different updates of $\alpha(t)$.

Figure 6.3: Comparison of number of iterations needed to reach 1e-3 precision of Ψ.

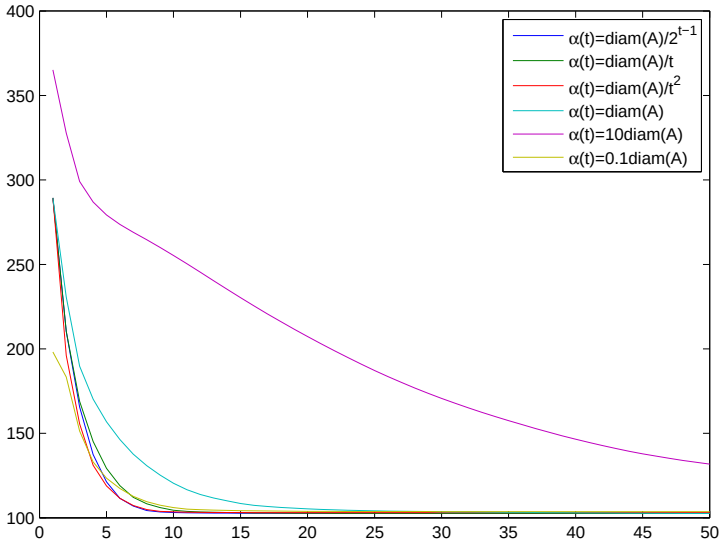
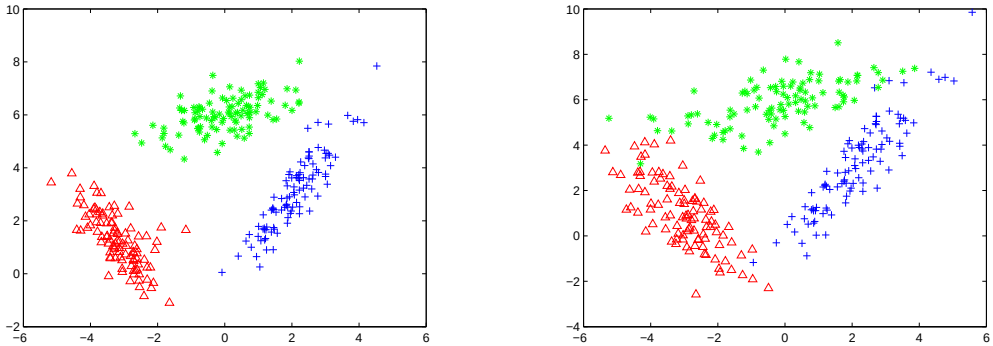


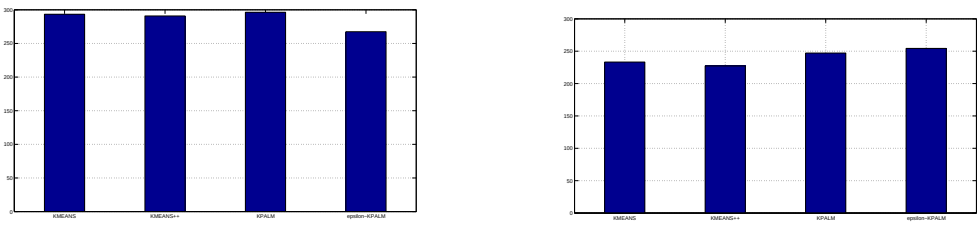
Figure 6.4: Comparison of the objective values for different values of α.



(a) Dense Gaussians.

(b) Sparse Gaussians.

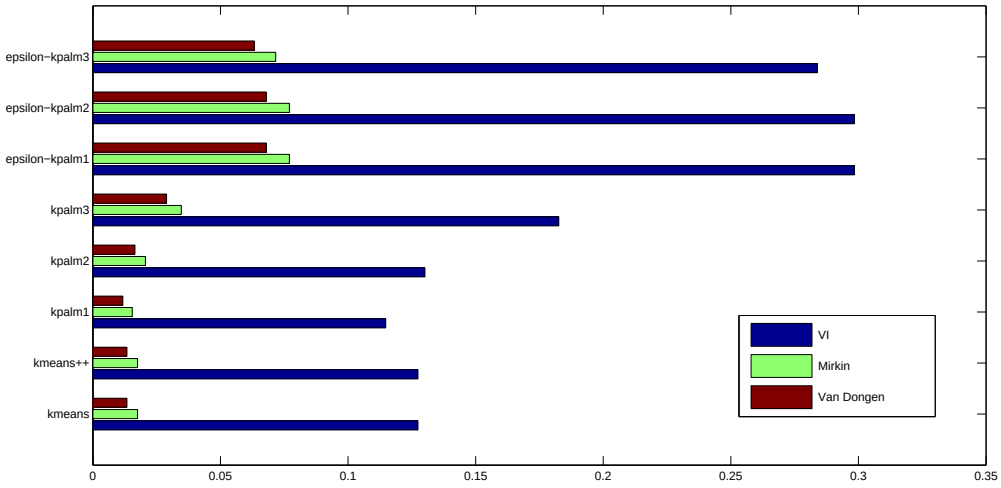
Figure 6.5: Two datasets, each 300 points.



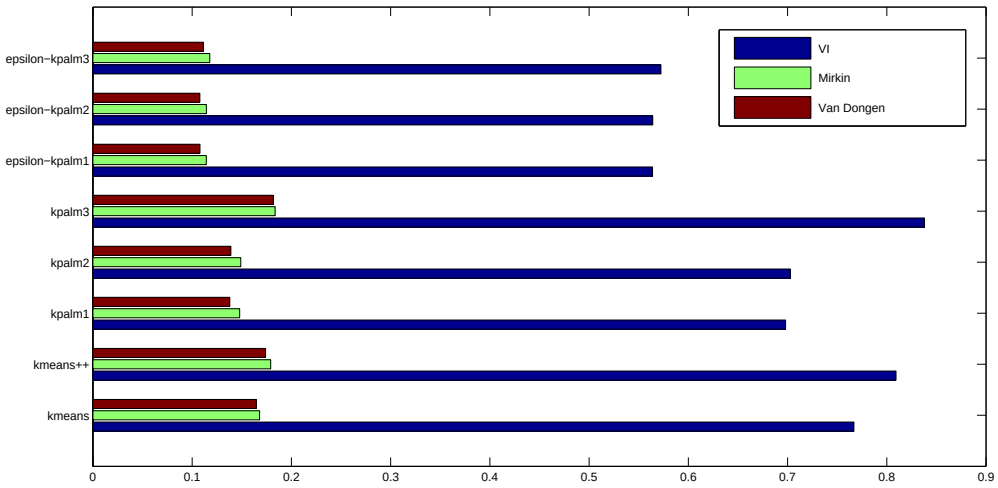
(a) Dense Gaussians clustering.

(b) Sparse Gaussians clustering.

Figure 6.6: Results of clustering algorithms for dense and sparse datasets.



(a) Dense Gaussians metrics comparison.



(b) Sparse Gaussians metrics comparison.

Figure 6.7: Comparison of metrics between clusterings for dense and sparse datasets.

Bibliography

[1] H. Attouch, J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, Vol. 116, pp. 5-16 (2009).

[2] H. Attouch, J. Bolte, B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, Vol. 137, 91-129 (2013).

[3] A. Beck, S. Sabach. Weiszfeld's Method: Old and New Results. *Journal of Optimization Theory and Applications*, Vol. 164, pp. 1-40 (2015).

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

[5] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York (1981).

[6] J. Bolte, A. Daniilidis, A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal of Optimization*, Vol 17, pp. 1205-1223 (2006).

[7] J. Bolte, S. Sabach, M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, Vol. 146, pp. 459-494 (2014).

[8] G. Chen, R. Rockafellar. Convergence rates in forward-backward splitting. *SIAM Journal on Optimization*, Vol. 7, pp. 421-444 (1997).

[9] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Vol. 39, pp. 1-38 (1977).

[10] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., second edition (2001).

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

[12] A.K. Jain, M.N. Murty, P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, Vol. 31 pp. 264-323 (1999).

[13] P. Lions, B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis* Vol. 16, pp. 964-979 (1979).

[14] S.P. Lloyd. Least squares quantization in PCM. Bell Telephone Laboratories Paper, Murray Hill, NJ (1957). Also in, *IEEE Transactions on Information Theory*, Vol. 28 pp. 127-135 (1982).

[15] J.MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Math., Stat. and Probability*, pp. 281-296 (1967).

[16] M. Meila. Comparing Clusterings - An Axiomatic View. In *ICML '05: Proceedings of the 22nd international conference on Machine Learning ACM*, New York, pp. 577-584 (2005).

[17] S.Z. Selim, M.A. Ismail. K-Means-Type Algorithms. A Generalized Convergence Theorem and Characterization of Local Optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, pp. 81-87 (1984).

[18] H. Steinhaus. Sur la division des corps materiels en parties. *Bull. Acad. Polon. Sci.*, C1. III, Vol. IV, pp. 801-804 (1956).

[19] M. Teboulle. A Unified Continuous Optimization Framework for Center-Based Clustering Methods. *The Journal of Machine Learning Research*, Vol. 8, pp. 65-102 (2007).

[20] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, Vol. 29, pp. 119-138 (1991).

[21] C.F.J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, Vol. 11 pp. 95-103 (1983).