Contents

1 Introduction

The clustering problem is the 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 machine learning field, and it arises in 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 summing 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 method can be used for data compression, data correction and pattern recognition.

There are several categories of clustering methods, each has a direct impact of 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 the 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 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 a NP-hard problem, and hence cannot be solved efficiently. Most algorithms seek to minimize some mathematical criteria, 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) K-means algorithm: This algorithm is probably the most famous within the clustering scope, and dates back to MacQueen (1967). The k-means algorithm partitions the

This algorithm can be described as an optimization algorithm (see precise details below) which minimizes the following

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. ~~The~~ objective function ~~that the k-means algorithm minimizes is~~

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

The simplicity of the algorithm both in the theoretical and implementation aspects made it very popular.

(ii) 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. The objective function that the FKM algorithm minimizes is

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

The parameter 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)

(iii) Expectation-Maximization (EM) algorithm: The EM algorithm 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

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

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.

An interesting paper of Teboulle [8] shows that these center-based clustering algorithms can be recovered from the proposed continuous optimization framework. The smoothing methodologies for the clustering problem are based on nonlinear means and on approximation of appropriate asymptotic functions.

Most of the existing clustering methods are sensitive to the starting point, namely choosing different starting point result in significant changes in the final clustering. There are plethora of heuristic initialization method. One such initialization method is choosing random data points as staring centers for clustering, 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 point from the points chosen so far.

to chose k

We begin this with

The starting point of our work is a formulation of the clustering problem which consists of minimizing the sum of finite collection of min-functions, which is a nonsmooth and nonconvex optimization problem, in its most the general case. The clustering problem is given by

This

where $x = (x^1, x^2, \ldots, x^k) \in \mathbb{R}^{nk}$

$$\min_{x \in \mathbb{R}^{nk}} \left\{ F(x) := \sum_{i=1}^{m} \min_{1 \le l \le k} d(x^l, a^i) \right\}, \tag{1.1}$$

This paragraph should be here

with $d(\cdot, \cdot)$ being a distance-like function.

The lack of smoothness in this formulation can be overcome, yet the nonconvex nature of the clustering problem shall 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 the common approach is usually achieved by relaxation of the original problem. Motivated by papers of Attouch et al. [1, 2] that established convergence of the sequences generated by the proximal Gauss-Seidel scheme in the general nonconvex and nonsmooth settings, and similar result for the proximal-forward-backward 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. [5], where the objective function is a function of finite blocks of variables.

We focus on two cases of distance-like functions. The first is the squared Euclidean norm, which is the standard proximity measure used by k-means. For this case, we derive an equivalent smooth optimization problem for the clustering problem presented in (1.1) and prove our convergence result for the suggested algorithm via the methodology which is discussed in [5]. The second distance-like function that we study is the Euclidean norm. In this case we present two approximations, in order to overcome the lack of smoothness in the problem, and then proceed with the same methodology as in the first case. 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.

Outline of the thesis. This work is organized as follows. In the following section we transform the initial formulation of the clustering problem into a smooth one. In addition, we introduce the KL theory and the general methodology that will be used in our analysis of the proposed algorithms. In section 3 and 4 we suggest two algorithms, KPLAM and ε-KPALM, respectively. KPALM addresses the clustering problem with squared Euclidean norm distance-like function and ε-KPALM the standard Euclidean norm distance function. Employing the methodology of Section 2 we establish our convergence results. In section 5 we prove the convergence of k-means algorithm to a critical point and under certain assumption extend the convergence to a local minimum. Finally, in Section 6 we compare the performance of these algorithm according to some common criteria.

In this thesis you don't write it like that! See the format!

In this thesis we take a different route and consider the problem in its original nonconvex form. Very recently this complicated route become more relevant and interesting thanks to few papers which pave the way for dealing with nonconvex problems using sophisticated mathematical tools as will be explained later.

2 Problem Reformulation and Mathematical Tools

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:

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

where Δ denotes the well-known simplex defined by

$$\Delta = \left\{ u \in \mathbb{R}^k : \sum_{l=1}^{k} u_l = 1,\ u \ge 0 \right\}.$$

Using this fact in Problem (1.1) and introducing new variables $w^i \in \mathbb{R}^k$, $i = 1, 2, \ldots, m$, gives a smooth reformulation of the clustering problem

$$\min_{x\in\mathbb{R}^{nk}} \sum_{i=1}^{m} \min_{w^i\in\Delta} \langle w^i, d^i(x)\rangle, \tag{2.1}$$

where

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

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

$$\min_{x\in\mathbb{R}^{nk}, w\in\mathbb{R}^{km}} \left\{ \sum_{i=1}^{m} \left(\langle w^i, d^i(x)\rangle + \delta_\Delta(w^i) \right) \right\}, \tag{2.2}$$

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 (2.2) with the functions defined above gives a compact equivalent form of the original clustering problem

$$\min\left\{ \Psi(z) := H(w, x) + G(w) \mid z := (w, x) \in \mathbb{R}^{km} \times \mathbb{R}^{nk} \right\}. \tag{2.3}$$

2.2 Introduction to the PALM Theory

In this subsection we give a brief review of the main developments established in [5]. These developments which include the proximal alternating linearized minimization (PALM) algorithm and general procedure for proving global convergence of generic algorithm play a central rule in this work. First, let us recall several definitions which are needed for the upcoming discussion.

Definition 1 (Subdifferentials). *Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.*

(i) For a given $x \in dom\ \sigma := \{x \in \mathbb{R}^d : \sigma(x) < \infty\}$, 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 PALM's analysis. Let $\eta \in (0, +\infty]$. Denote the following class of functions

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

Definition 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 $\overline{u} \in dom\ \partial\sigma := \{u \in \mathbb{R}^d : \partial\sigma \neq \emptyset\}$ 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.

Definition 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*

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

As mentioned in [5], the PALM algorithm is nothing but alternating the classical proximal gradient (aka Proximal Forward-Backward) over

where $f : \mathbb{R}^n \to (-\infty, +\infty]$ and $g : \mathbb{R}^n \to (-\infty, +\infty]$ are proper and lower semicontinuous 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 $H(x, \cdot) \in C^{1,1}_{L_2(x)}$. PALM is alternating the steps of the proximal forward-backward (PFB) scheme. PFB scheme tackles the problem of minimizing the sum of a smooth function h with a nonsmooth one σ, and can be viewed as the proximal regularization of h linearized at a given point x^k, that is,

$$x^{k+1} \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \langle x - x^k, \nabla h(x^k) \rangle + \frac{t}{2} \|x - x^k\|^2 + \sigma(x) \right\}, \ (t > 0).$$

This

Adopting this scheme on the two blocks (x, y) leads towards the following approximations

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

and

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

can be summarized as follows

Thus, PALM is alternating between the following two subproblems

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

Assuming Ψ is KL function and the generated sequence by PALM, $\{z^k := (x^k, y^k)\}_{k \in \mathbb{N}}$, is bounded, Bolte et al. [5] proved that the sequence obeys properties (C1) and (C2), and it converges to a critical point of Ψ.

thus

a gradient-like descent sequence

3 Clustering: The Squared Euclidean Norm Case

3.1 Clustering with PALM

In this section we tackle the clustering problem, given in (2.3), 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 subsection. 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.

Attouch et al. [1,2] establishe convergence ... (continue as in page 4) ... contribution of [5] is the general methodology to prove convergence of generic algorithm in the setting of noncomvex and nonsmooth optimization problems

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

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

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

Theorem 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 σ.*

The class of semi-algebraic function 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.

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

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

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

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

The purpose of the proposed methodology is

~~Building upon [1, 2], the following three requirements are sufficient~~ to assure the convergence of the whole sequence $\{z^k\}_{k\in\mathbb{N}}$ to a critical point of Ψ.

(C1) *Sufficient decrease property:* There exists a positive constant ρ_1, such that

$$\rho_1 \|z^{k+1} - z^k\|^2 \leq \Psi(z^k) - \Psi(z^{k+1}), \quad \forall\, k = 0, 1, \ldots.$$

(C2) *A subgradient lower bound for iterates gap:* Assuming that the sequence generated by the algorithm $\mathcal{A}$ is bounded. There exists a positive constant ρ_2, such that

$$\|w^{k+1}\| \leq \rho_2 \|z^{k+1} - z^k\|, \quad w^k \in \partial\Psi(z^k) \quad \forall\, k = 0, 1, \ldots.$$

Please replace this with what I sent you!

(C3) *KL property:* The function Ψ is a KL function.

~~Due to Theorem 1, property (C3) follows whenever Ψ is a semi-algebraic function.~~ Finally, we present the proximal alternating linearized minimization (PALM) algorithm which solves the nonconvex and nonsmooth minimization problem of the following form

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