1 The Clustering Problem

Let $\mathcal{A} = \left\{a^1, a^2, \ldots, a^m\right\}$ 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 data $\mathcal{A}$ into k subsets $\left\{A^1, A^2, \ldots, A^k\right\}$, called clusters. For each $l = 1, 2, \ldots, k$, the cluster A_l is represented by its center x^l, and we want to determine k cluster centers $\left\{x^1, x^2 \ldots, x^k\right\}$ such that the sum of proximity measures from each point $a^i, i = 1, 2, \ldots, m$, to a nearest cluster center x^l is minimized.

The clustering problem formulation is given by

$$\min_{x^1,\ldots,x^k \in \mathbb{R}^n} \sum_{i=1}^{m} \min_{1 \le l \le k} d(x^l, a^i), \tag{1.1}$$

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

2 Problem Reformulation and Notations

We introduce some notations that will be used throughout this document.

$A = (a^1, \ldots, a^m) \in (\mathbb{R}^n)^m$, where $a^i \in \mathbb{R}^n, i = 1, 2, \ldots, m$

$W = (w^1, \ldots, w^m) \in \left(\mathbb{R}^k\right)^m$, where $w^i \in \mathbb{R}^k, i = 1, 2, \ldots, m$

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

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

$\Delta = \left\{u \in \mathbb{R}^k \mid \sum\limits_{l=1}^{k} u_l = 1, u_l \ge 0, l = 1, 2, \ldots, k\right\}$

For some $S \subseteq \mathbb{R}^n$, the indicator function of the set$s$ is defined and denoted as follows

$\delta_S(p) = \begin{cases} 0 & \text{if } p \in S \\ \infty & \text{if } p \notin S \end{cases}.$

Using the fact that $\min\limits_{1 \le l \le k} u_l = \min\left\{\langle u, v\rangle \mid v \in \Delta\right\}$, and applying it over (1.1), gives a smooth reformulation of the clustering problem

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

Further replacing the constrain over $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$ to the objective function, results in a equivalent formulation

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

Finally, introducing several more useful definitions, for each $i = 1, 2, \ldots, m$

$$H_i(W,X) = \langle w^i, d^i(X)\rangle \quad G(w^i) = \delta_\Delta(w^i)$$
$$H(W,X) = \sum_{i=1}^{m} H_i(W,X) \quad G(W) = \sum_{i=1}^{m} G(w^i)$$

Replacing the terms in (2.1) with the functions above gives a compact form that is equivalent to the original clustering problem

$$\min\left\{\Psi(Z) := H(W,X) + G(W) \mid Z := (W,X) \in (\mathbb{R}^k)^m \times (\mathbb{R}^n)^k\right\} \tag{2.3}$$

3 Clustering via PALM Approach

3.1 Introduction to PALM Theory

Presentation of PALM's requirements and of the algorithm steps ...

3.2 Clustering with PALM for Squared Euclidean Norm Distance-Like Function

In this section we tackle the clustering problem with distance-like function $d(u,v) = \|u-v\|^2$. We devise a PALM-like algorithm, based on the discussion about PALM in the previous subsection. Since the clustering problem has a specific structure, we are ought to exploit it in the following manner. First we notice that the map $W \mapsto H(W,X)$ is linear in W, so there is no need to linearize it. In addition, the map $X \mapsto H(W,X) = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i\|x^l - a^i\|^2 = \sum_{l=1}^{k}\sum_{i=1}^{m} w_l^i\|x^l - a^i\|^2$ is convex and quadratic in X, hence we do not need to add a proximal term as in PALM algorithm.

Now we propose the PALM-like algorithm for clustering, we name it KPALM.

(1) Initialization: Set $t = 0$, and pick random vectors $(W(0), X(0)) \in \Delta^m \times (\mathbb{R}^n)^k$

(2) For each $t = 0, 1, \ldots$ generate a sequence $\{(W(t), X(t))\}_{t\in\mathbb{N}}$ as follows:

(2.1) Cluster Assignment: Take $\nu \in (0,1]$, compute $\beta(t) = \min_{1\le l\le k}\left\{\sum_{i=1}^{m} w_l^i(t)\right\}$, set $\alpha(t) = \nu\beta(t)$ and for each $i = 1, 2, \ldots, m$ compute

$$w^i(t+1) = \arg\min_{w^i\in\Delta}\left\{\langle w^i, d^i(X(t))\rangle + \frac{\alpha(t)}{2}\|w^i - w^i(t)\|^2\right\} \tag{3.1}$$

(2.2) Centers Update: For each $l = 1, 2, \ldots, k$ compute $x^l \in \mathbb{R}^n$ via

$$X(t+1) = \arg\min\left\{H(W(t+1), X) \mid X \in (\mathbb{R}^n)^k\right\} \tag{3.2}$$

At each step $t \in \mathbb{N}$, the KPALM algorithm alternates between cluster assignment and centers update. The explicit formulas for step t are given below

$$w^i(t+1) = \Pi_\Delta \left(w^i(t) - \frac{d^i(X(t))}{\alpha(t)} \right), \quad i = 1, 2, \ldots, m, \tag{3.3}$$

$$x^l(t+1) = \frac{\sum_{i=1}^m w_l^i(t+1) a^i}{\sum_{i=1}^m w_l^i(t+1)}, \quad l = 1, 2, \ldots, k. \tag{3.4}$$

Remark 1. *(i) $\alpha(t)$ is the step-size, and it must be positive. If for some step $t \in \mathbb{N}$ $\alpha(t) = 0$ then there exists $1 \le l' \le k$ such that $\beta(t) = \sum_{i=1}^m w_{l'}^i = 0$, since for all $1 \le l \le k$, and for all $1 \le i \le m$ such that $w_l^i \ge 0$ then for all $1 \le i \le m$, $\quad w_{l'}^i = 0$. Thus, none of the points in $\mathcal{A}$ belong to cluster l', in that case the algorithm can halt. Hence from now on we assume that for all $t \in \mathbb{N}$, $\beta(t) = \min_{1 \le l \le k} \left\{ \sum_{i=1}^m w_l^i(t) \right\} > 0$, and it follows that $\alpha(t) > 0$.*

(ii) Since for all $t \in \mathbb{N}$ $\quad W(t) \in \Delta^m$ then $\Psi(Z(t)) = H(W(t), X(t)) + G(W(t)) = H(W(t), X(t))$.

(iii) Note that the function $X \mapsto H(W, X)$ is separable in x^l for all $l = 1, 2, \ldots, k$. Thus, regardless of the specific distance-like function $d(\cdot, \cdot)$, the centers update step is $x^l(t+1) = \arg\min_{x^l \in \mathbb{R}^k} \left\{ \sum_{i=1}^m w_l^i d(x^l, a^i) \right\}, l = 1, 2, \ldots, k$, and in the case of squared Euclidean norm the result is as in (3.4).

Lemma 3.0.1 (Boundedness of KPALM sequence)**.**
Let $\{Z(t)\}_{t \in \mathbb{N}} = \{W(t), X(t)\}_{t \in \mathbb{N}}$ be the sequence generated by KPALM. Then

(i) $x^l(t) \in Conv(\mathcal{A})$ for all $l = 1, 2, \ldots, k$ and $t \in \mathbb{N}$, where $Conv(\mathcal{A})$ is the convex hull of $\mathcal{A}$.

(ii) For all $l = 1, 2, \ldots, k$ and $t \in \mathbb{N}$, $\|x^l(t)\|$ is bounded by $M = \max_{1 \le i \le m} \|a^i\|$.

(iii) $\{Z(t)\}_{t \in \mathbb{N}}$ is a bounded sequence in $(\mathbb{R}^k)^m \times (\mathbb{R}^n)^k$.

Proof. (i) Set $\alpha_i = \frac{w_l^i(t)}{\sum_{j=1}^m w_l^j(t)}, i = 1, 2, \ldots, m$, then $\alpha_i \ge 0$ and $\sum_{i=1}^m \alpha_i = 1$. From (3.4) we have

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

Hence $x^l(t)$ is in the convex hull of $\mathcal{A}$, for all $l = 1, 2, \ldots, k$, and $t \in \mathbb{N}$.

(ii) Taking the norm of $x^l(t)$ yields

$$\|x^l(t)\| = \left\| \sum_{i=1}^m \left(\frac{w_l^i(t)}{\sum_{j=1}^m w_l^j(t)} \right) a^i \right\| \le \sum_{i=1}^m \left(\frac{w_l^i(t)}{\sum_{j=1}^m w_l^j(t)} \right) \|a^i\| \le \sum_{i=1}^m \alpha_i \max_{1 \le i \le m} \|a^i\| = M.$$

(iii) $w^i(t)$ 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.

□

Lemma 3.0.2 (Strong convexity of $H(W, X)$ in X)**.** *At step t, the function $X \mapsto H(W(t), X)$ is strongly convex iff $\beta(t) > 0$.*

Proof. Since the function $X \mapsto H(W(t), X) = \sum_{l=1}^{k} \sum_{i=1}^{m} w_l^i \|x^l - a^i\|^2$ is C^2, it is strongly convex iff the smallest eigenvalue of the Hessian matrix is positive. Thus

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

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

Now we are ready to prove the decrease property of KPALM algorithm.

Proposition 3.1 (Sufficient decrease property)**.**
Let $\{Z(t)\}_{t \in \mathbb{N}} = \{W(t), X(t)\}_{t \in \mathbb{N}}$ be the sequence generated by KPALM, then there exists $\rho_1 > 0$ such that $\rho_1 \|Z(t+1) - Z(t)\|^2 \leq \Psi(Z(t)) - \Psi(Z(t+1))$ for all $t \in \mathbb{N}$.

Proof. From (3.1) we derive the following inequality

$$\begin{aligned} H_i(W(t+1), X(t)) + \frac{\alpha(t)}{2} \|w^i(t+1) - w^i(t)\|^2 &= \langle w^i(t+1), d^i(X(t)) \rangle + \frac{\alpha(t)}{2} \|w^i(t+1) - w^i(t)\|^2 \\ &\leq \langle w^i(t), d^i(X(t)) \rangle + \frac{\alpha(t)}{2} \|w^i(t) - w^i(t)\|^2 \\ &= \langle w^i(t), d^i(X(t)) \rangle \\ &= H_i(W(t), X(t)) \end{aligned}$$

Hence, we obtain

$$\frac{\alpha(t)}{2} \|w^i(t+1) - w^i(t)\|^2 \leq H_i(W(t), X(t)) - H_i(W(t+1), X(t)).$$

Summing this inequality over $i = 1, 2, \ldots, m$ yields

$$\begin{aligned} \frac{\alpha(t)}{2} \|W(t+1) - W(t)\|^2 &= \frac{\alpha(t)}{2} \sum_{i=1}^{m} \|w^i(t+1) - w^i(t)\|^2 \\ &\leq \sum_{i=1}^{m} H_i(W(t), X(t)) - \sum_{i=1}^{m} H_i(W(t+1), X(t)) \\ &= H(W(t), X(t)) - H(W(t+1), X(t)) \end{aligned}$$

From lemma 3.0.2 we have that the function $X \mapsto H(W(t), X)$ is strongly convex with parameter $2\beta(t) > 0$, hence it follows that

$$\begin{aligned} H(W(t+1), X(t)) - H(W(t+1), X(t+1)) \geq \\ &\geq \nabla_X H(W(t+1), X(t+1))^T (X(t) - X(t+1)) + \frac{2\beta(t)}{2} \|X(t) - X(t+1)\|^2 \\ &= \beta(t) \|X(t) - X(t+1)\|^2 \end{aligned}$$

where the last equality follows from (3.2), since $\nabla_X H(W(t+1), X(t+1)) = 0$.
Set $\rho_1 = \min\{\alpha(t), \beta(t)\} = \alpha(t)$, combined with the previous inequalities, we have

$$\begin{aligned} \rho_1 \|Z(t+1) - Z(t)\|^2 &= \rho_1 \left(\|W(t+1) - W(t)\|^2 + \|X(t+1) - X(t)\|^2 \right) \\ &\leq [H(W(t), X(t)) - H(W(t+1), X(t))] \\ &\quad + [H(W(t+1), X(t)) - H(W(t+1), X(t+1))] \\ &= H(Z(t)) - H(Z(t+1)) = \Psi(Z(t)) - \Psi(Z(t+1)), \end{aligned}$$

where the last equality follows from remark 1(ii). □

Next, we aim to prove the subgradient lower bound for iterates gap property. The following lemma will be handy in our proof.

Lemma 3.1.1.
Let $\{Z(t)\}_{t\in\mathbb{N}} = \{W(t), X(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM, then $\|d^i(X(t+1) - d^i(X(t))\| \leq 4M\|X(t+1) - X(t)\|$ for all $i = 1, 2, \ldots, m$ and $t \in \mathbb{N}$, where $M = \max_{1\leq i\leq m} \|a^i\|$.

Proof.

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

□

Proposition 3.2 (Subgradient lower bound for iterates gap property)**.**
Let $\{Z(t)\}_{t\in\mathbb{N}} = \{W(t), X(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM, then there exists $\rho_2 > 0$ and $\gamma(t+1) \in \partial\Psi(Z(t+1))$ such that $\|\gamma(t+1)\| \leq \rho_2\|Z(t+1) - Z(t)\|^2$, for all $t \in \mathbb{N}$.

Proof. $\Psi = H + G$, then

$$
\begin{aligned}
\partial\Psi = \nabla H + \partial G &= (\nabla_W H, \nabla_X H) + \left((\partial_{w^i}\delta_\Delta)_{i=1,\dots,m}, \mathbf{0} \right) \\
&= \left((\nabla_{w^i} H_i + \partial_{w^i}\delta_\Delta)_{i=1,\dots,m}, \nabla_X H \right).
\end{aligned}
$$

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

$$
\begin{aligned}
&\partial\Psi(Z(t+1)) = \\
&\quad = \left(\left(\nabla_{w^i} H_i(W(t+1), X(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\right)_{i=1,\dots,m}, \nabla_X H(W(t+1), X(t+1)) \right) \\
&\quad = \left(\left(d^i(X(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\right)_{i=1,\dots,m}, \nabla_X H(W(t+1), X(t+1)) \right) \\
&\quad = \left(\left(d^i(X(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\right)_{i=1,\dots,m}, \mathbf{0} \right),
\end{aligned}
$$

where the last equality follows from (3.2), that is the optimality condition of $X(t+1)$.
Taking the norm of the last equation yields

$$
\|\partial\Psi(Z(t+1))\| \leq \sum_{i=1}^{m} \|d^i(X(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\|. \tag{3.5}
$$

The optimality condition of $w^i(t+1)$ that is derived from (3.1), 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

$$d^i(X(t)) + \alpha(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1) = 0. \tag{3.6}$$

Setting $\gamma(t+1) := \left(\left(d^i(X(t+1)) + u^i(t+1)\right)_{i=1,\ldots,m}, \mathbf{0}\right) \in \partial\Psi(Z(t+1))$, and plugging (3.6) into (3.5) we have

$$\begin{aligned}
\|\gamma(t+1)\| &\leq \sum_{i=1}^{m} \|d^i(X(t+1)) - d^i(X(t)) - \alpha(t)\left(w^i(t+1) - w^i(t)\right)\| \\
&\leq \sum_{i=1}^{m} \|d^i(X(t+1)) - d^i(X(t))\| + \sum_{i=1}^{m} \alpha(t)\|w^i(t+1) - w^i(t)\| \\
&\leq \sum_{i=1}^{m} 4M\|X(t+1)) - X(t)\| + m\alpha(t)\|Z(t+1) - Z(t)\| \\
&\leq m\left(4M + \alpha(t)\right)\|Z(t+1) - Z(t)\|
\end{aligned}$$

where the third inequality follows from lemma 3.1.1.
Define $\rho_2 = m\left(4M + \alpha(t)\right)$ and the result follows. □

3.3 Similarity to KMEANS

The famous KMEANS algorithm has close proximity to KPALM algorithm. KMEANS alternates between cluster assignments and center updates as well. In detail, we can write its steps in the following manner

(1) Initialization: Set $t = 0$, and pick random centers $Y(0) \in (\mathbb{R}^n)^k$

(2) For each $t = 0, 1, \ldots$ generate a sequence $\{(V(t), Y(t))\}_{t\in\mathbb{N}}$ as follows:

(2.1) Cluster Assignment: For $i = 1, 2, \ldots, m$ compute

$$v^i(t+1) = \arg\min_{v^i \in \Delta} \left\{\langle v^i, d^i(Y(t))\rangle\right\} \tag{3.7}$$

(2.2) Center Update: For $l = 1, 2, \ldots, k$ compute

$$y^l(t+1) = \frac{\sum_{i=1}^{m} v_l^i(t+1)a^i}{\sum_{i=1}^{m} v_l^i(t+1)} \tag{3.8}$$

The KMEANS algorithm obviously resemble KPALM algorithm. Assuming same starting point $X(0) = Y(0)$ and by taking $\nu \to 0$, we have

$$V(t) = \lim_{\nu\to 0} W(t), \quad Y(t) = \lim_{\nu\to 0} X(t),$$

meaning, both algorithms converge to the same result.