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 $\left\{C^1, C^2, \ldots, C^k\right\}$, called clusters. For each $l = 1, 2, \ldots, k$, the cluster C^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 is given by

$$\min_{x^1, x^2, \ldots, x^k \in \mathbb{R}^n} F(x^1, x^2, \ldots, x^k) := \sum_{i=1}^{m} \min_{1 \leq l \leq k} d(x^l, a^i), \tag{1.1}$$

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.

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

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

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

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

$\Delta = \left\{u \in \mathbb{R}^k \mid \sum_{l=1}^{k} u_l = 1,\ u_l \geq 0,\ l = 1, 2, \ldots, k\right\}$.

Let $S \subseteq \mathbb{R}^n$. The indicator function of 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 $\min\limits_{1 \leq l \leq k} u_l = \min\left\{\langle u, v \rangle \mid v \in \Delta\right\}$, and applying it over (1.1), gives a smooth reformulation of the clustering problem

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

Replacing further the constraint $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$ to the objective function, results in a equivalent formulation

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

Finally, introducing several more useful notations is needed. For each $i = 1, 2, \ldots, m$, we denote

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

Replacing the terms in (2.1) with the functions defined above gives a compact form of the original clustering problem

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

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 Function

In this section we tackle the clustering problem with the classical distance function defined by $d(u,v) = \|u-v\|^2$. 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 function $w \mapsto H(w,x)$ is linear in w, so there is no need to linearize it. In addition, the function $x \mapsto H(w,x) = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i \|x^l - a^i\|^2 = \sum_{l=1}^{k}\sum_{i=1}^{m} w_l^i \|x^l - a^i\|^2$ is convex and quadratic in x, hence we do not need to add a proximal term as in PALM algorithm.

Now we propose a PALM-like algorithm for clustering, which we call KPALM.

(1) Initialization: Set $t = 0$, and pick random vectors $(w(0), x(0)) \in \Delta^m \times \mathbb{R}^{nk}$.

(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 any $\alpha_i(t) > 0$ and for each $i = 1, 2, \ldots, m$ compute

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

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

$$x(t+1) = \arg\min \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\}. \tag{3.2}$$

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

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

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

where P_Δ is the orthogonal projection onto the set Δ.

Assumption 1. *We assume that none of the clusters $C^l, l = 1, 2, \ldots, k$, get empty during this process, hence for all $1 \le l \le k$ and $t \in \mathbb{N}$ we have that $\sum\limits_{i=1}^m w_l^i(t) > 0$.*

Remark 1. *(i) Since for all $t \in \mathbb{N}$ we have that $w(t) \in \Delta^m$ then $G(w(t)) = 0$ and therefore $\Psi(z(t)) = H(w(t), x(t))$.*

(ii) For any choice of distance-like function $d(\cdot,\cdot)$, the function $x \mapsto H(w,x)$ is separable in x^l for all $l = 1, 2, \ldots, k$. Thus, regardless the choice of distance-like function $d(\cdot,\cdot)$, the centers update step can be done in parallel over all centers, that is, $x^l(t+1) = \arg\min\limits_{x^l \in \mathbb{R}^k}\left\{\sum\limits_{i=1}^m w_l^i d(x^l, a^i)\right\}$, $l = 1, 2, \ldots, k$, and in the case of the squared Euclidean norm the result is given 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, the following statements hold true.*

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

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

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

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

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

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 again from (3.4) that

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

(iii) 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.

$\square$

Lemma 3.0.2 (Strong convexity of $H(w,x)$ in x)**.** *The function $x \mapsto H(w,x)$ is strongly convex with parameter $\beta(w) = 2 \min_{1\le l \le k} \left\{ \sum_{i=1}^{m} w_l^i \right\}$, whenever $\beta(w) > 0$.*

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 if and only if the smallest eigenvalue of the corresponding Hessian matrix is positive. Thus

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

Since the Hessian is a diagonal matrix, the smallest eigenvalue is $\min_{1\le l\le k} 2\sum_{i=1}^{m} w_l^i = \beta(w)$, 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*

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

Proof. From (3.1) we derive the following inequality

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

Hence, we obtain

$$\frac{\alpha_i(t)}{2}\|w^i(t+1) - w^i(t)\|^2 \le H_i(w(t),x(t)) - H_i(w(t+1),x(t)). \tag{3.5}$$

Denote $\alpha(t) = \min_{1\le i\le m}\{\alpha_i(t)\}$. Summing inequality (3.5) over $i = 1,2,\ldots,m$ yields

$$\begin{aligned} \frac{\alpha(t)}{2}\|w(t+1) - w(t)\|^2 &= \frac{\alpha(t)}{2}\sum_{i=1}^{m}\|w^i(t+1) - w^i(t)\|^2 \\ &\le \sum_{i=1}^{m}\frac{\alpha_i(t)}{2}\|w^i(t+1) - w^i(t)\|^2 \\ &\le \sum_{i=1}^{m} H_i(w(t),x(t)) - \sum_{i=1}^{m} H_i(w(t+1),x(t)) \\ &= H(w(t),x(t)) - H(w(t+1),x(t)). \end{aligned}$$

From Assumption 1 we have that $\beta(w(t)) = 2 \min\limits_{1\leq l\leq k} \left\{ \sum\limits_{i=1}^{m} w_l^i(t) \right\} > 0$, and from Lemma 3.0.2 it follows that the function $x \mapsto H(w(t), x)$ is strongly convex with parameter $\beta(w(t))$, hence it follows that

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

where the last equality follows from (3.2), since $\nabla_x H(w(t+1), x(t+1)) = 0$. Set $\rho_1 = \frac{1}{2} \min\{\alpha(t), \beta(w(t))\}$, combined with the previous inequalities, we have

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

where the last equality follows from Remark 1(i). □

Next, we aim to prove the subgradient lower bound for iterates gap property. The following lemma will be essential 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*

$$\|d^i(x(t+1) - d^i(x(t))\| \leq 4M \|x(t+1) - x(t)\|, \quad \forall i = 1, 2, \ldots, m,\ t \in \mathbb{N},$$

where $M = \max\limits_{1\leq i\leq m} \|a^i\|$.

Proof. Since $d(u, v) = \|u - v\|^2$, we get that

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

$$= \left[\sum_{l=1}^{k} (4M)^2 \|x^l(t+1) - x^l(t)\|^2 \right]^{\frac{1}{2}} = 4M\|x(t+1) - x(t)\|,$$

this proves the desired result. $\square$

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*

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

Proof. By the definition of Ψ (see (2.3)) we get

$$\partial\Psi = \nabla H + \partial G = \Big((\nabla_{w^i} H_i + \partial_{w^i}\delta_\Delta)_{i=1,\dots,m}, \nabla_x H \Big).$$

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

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

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

$$\|\partial\Psi(z(t+1))\| \leq \sum_{i=1}^{m} \|d^i(x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\|. \tag{3.6}$$

The optimality condition of $w^i(t+1)$ that is derived from (3.1), yields that for all $i = 1, 2, \dots, m$ there exists $u^i(t+1) \in \partial\delta_\Delta(w^i(t+1))$ such that

$$d^i(x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1) = \mathbf{0}. \tag{3.7}$$

Setting $\gamma(t+1) := \Big(\big(d^i(x(t+1)) + u^i(t+1)\big)_{i=1,\dots,m}, \mathbf{0} \Big) \in \partial\Psi(z(t+1))$, and plugging (3.7) into (3.6) we have

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

where the third inequality follows from Lemma 3.1.1, and $\overline{\alpha}(t) = \max_{1\leq i\leq m} \alpha_i(t)$. Define $\rho_2 = m\left(4M + \overline{\alpha}(t)\right)$ and the result follows. $\square$

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 $y(0) \in \mathbb{R}^{nk}$.

(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

$$v^i(t+1) = \arg\min_{v^i\in\Delta} \left\{\langle v^i, d^i(y(t))\rangle\right\}. \tag{3.8}$$

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

$$y^l(t+1) = \frac{\sum_{i=1}^m v_l^i(t+1)a^i}{\sum_{i=1}^m v_l^i(t+1)}. \tag{3.9}$$

The KMEANS algorithm obviously resemble KPALM algorithm. Denote $\overline{\alpha}(t) = \max\limits_{1\leq i\leq m} \alpha_i(t)$. Assuming same starting point $x(0) = y(0)$ and by taking $\overline{\alpha}(t) \to 0$, we have

$$v(t) = \lim_{\overline{\alpha}(t)\to 0} w(t), \quad y(t) = \lim_{\overline{\alpha}(t)\to 0} x(t),$$

meaning, both algorithms converge to the same result.

3.4 KMEANS Convergence Proof

We start with rewriting the KMEANS algorithms, in its most familiar form

(1) Initialization: Set $t = 0$, and pick random centers $x(0) \in \mathbb{R}^{nk}$.

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

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

$$C^l(t+1) = \left\{a \in \mathcal{A} \mid \|a - x^l(t)\| \leq \|a - x^j(t)\|, \quad \forall 1 \leq l \leq k\right\}. \tag{3.10}$$

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

$$x^l(t+1) = mean(C^l(t)) := \frac{1}{|C^l(t)|}\sum_{a\in C^l(t)} a. \tag{3.11}$$

(2.3) Stopping criteria: Halt if

$$\forall 1 \leq l \leq k \quad C^l(t+1) = C^l(t) \tag{3.12}$$

As in KPALM, KMEANS needs Assumption 1 for step (3.11) to be well defined. In addition, to prove the convergence of KMEANS to local minimum, we will need to following assumption.

Assumption 2. *For any step $t \in \mathbb{N}$, 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 $R_{lj}(t) = \min\limits_{a \in C^l(t)} \left\{\|a - x^j(t)\| - \|a - x^l(t)\|\right\}$ for all $1 \leq l, j \leq k$, and $r(t) = \min\limits_{l \neq j} R_{lj}$.

Due to Assumption 2 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 3.3. *Let $(C(t), x(t))$ be the clusters and centers KMEANS returns. Denote an open neighborhood of $x(t)$ by $U = B\left(x^1(t), \frac{r(t)}{2}\right) \times B\left(x^2(t), \frac{r(t)}{2}\right) \times \cdots \times B\left(x^l(t), \frac{r(t)}{2}\right)$, then for any $x \in U$ we have $\underline{C^l}(x) = C^l(t)$ 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 KMEANS halts at step t, then from (3.12) 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

$$r(t) \leq \|x^j(t) - a\| - \|x^l(t) - a\| \quad \forall j \neq l. \tag{3.13}$$

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

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

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 (3.13), and we get that $C^l(t) \subseteq \underline{C^l}(x)$. Now, since $C(t)$ is a partition of $\mathcal{A}$ then $\underline{C^l}(x) = C^l(t)$ for all $1 \leq l \leq k$. $\square$

Proposition 3.4 (KMEANS converges to local minimum)**.** *Let $(C(t), x(t))$ be the clusters and centers KMEANS 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 3.3.

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

4 Clustering via Alternation with Weiszfeld Step

In this section we tackle the clustering problem with distance-like function being the Euclidean norm in $\mathbb{R}^n$, namely

$$\min_{x^1,x^2,\ldots,x^k\in\mathbb{R}^n}\left\{\sum_{i=1}^{m}\min_{1\le l\le k} d(x^l,a^i)\right\}, \tag{4.1}$$

where $d(u,v)=\|u-v\|$.
In the previous sections we showed that (4.1) has 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 $H(w,x)=\sum_{i=1}^{m}\left\langle w^i,d^i(x)\right\rangle=\sum_{i=1}^{m}\sum_{l=1}^{k}w_l^i d(x^l,a^i)=\sum_{l=1}^{k}\sum_{i=1}^{m}w_l^i d(x^l,a^i)=\sum_{l=1}^{k}\sum_{i=1}^{m}w_l^i\|x^l-a^i\|$, and $G(w)=\sum_{i=1}^{m}\delta_\Delta(w^i)$.

4.1 Weiszfeld Algorithm Scheme

Short overview of the problem and the algorithm ...

4.2 Clustering with Weiszfeld Step

We introduce some useful notations that will be useful for this section. For $1\le l\le k$ denote $L_l(w,x)=\sum_{i=1}^{m}\frac{w_l^i}{\|x_l-a^i\|}$, and $H_l(w,x)=\sum_{i=1}^{m}w_l^i\|x^l-a^i\|$.
Weiszfeld single iteration is defined for each $1\le l\le k$ via

$$T_l(w,x)=\frac{\sum_{i=1}^{m}\frac{w_l^i a^i}{\|x^l-a^i\|}}{\sum_{i=1}^{m}\frac{w_l^i}{\|x^l-a^i\|}}. \tag{4.2}$$

Next we present our algorithm for solving problem (4.1). The algorithm alternates between clusters assignment step, which is exactly as in KPALM, and centers update step that is based on a single Weizsfeld iteration.

(1) Initialization: Set $t=0$, and pick random vectors $(w(0),x(0))\in\Delta^m\times\mathbb{R}^{nk}$.

(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 any $\alpha_i(t)>0$ and for each $i=1,2,\ldots,m$ compute

$$w^i(t+1)=\arg\min_{w^i\in\Delta}\left\{\langle w^i,d^i(x(t))\rangle+\frac{\alpha_i(t)}{2}\|w^i-w^i(t)\|^2\right\}. \tag{4.3}$$

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

$$x^l(t+1)=T_l(w(t+1),x(t)). \tag{4.4}$$

Assumption 3. *For any step $t \in \mathbb{N}$ and for all $1 \le l \le k$, we assume that $x^l(t) \notin \mathcal{A}$.*

Remark 2. *(i) Due to Assumption 3 it follows that the centers update step in (4.4) is well defined.*

(ii) It is easy to verify that for all $1 \le l \le k$ the following equations hold true:

$$\nabla_{x^l} H_l(w, x) = \sum_{i=1}^{m} w_l^i \frac{x^l - a^i}{\|x^l - a^i\|}, \quad \forall x^l \notin \mathcal{A}, \tag{4.5}$$

and that

$$T_l(w, x) = x^l - \frac{1}{L_l(w, x)} \nabla_{x^l} H_l(w, x), \quad \forall x^l \notin \mathcal{A}. \tag{4.6}$$

As in KPALM case, we aim to prove the the sufficient decrease property and subgradient lower bounds for iterates gap property. Note that

$$x^l(t+1) = T_l(w(t+1), x(t)) = \frac{\sum\limits_{i=1}^{m} \frac{w_l^i(t+1)a^i}{\|x^l(t)-a^i\|}}{\sum\limits_{i=1}^{m} \frac{w_l^i(t+1)}{\|x^l(t)-a^i\|}} = \sum_{i=1}^{m} \frac{\frac{w_l^i(t+1)}{\|x^l(t)-a^i\|}}{\sum\limits_{j=1}^{m} \frac{w_l^j(t+1)}{\|x^l(t)-a^j\|}} a^i \in Conv(\mathcal{A}),$$

hence the sequence generated by (Alg-Name) is bounded as well.

Proposition 4.1 (Sufficient decrease property)**.** *Let $\{z(t)\}_{t\in\mathbb{N}} = \{(w(t), x(t))\}_{t\in\mathbb{N}}$ be the sequence generated by (Alg-Name), then there exists $\rho_1 > 0$ such that*

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

Proof. In the proof of sufficient decrease property of KPALM we showed that

$$\frac{\alpha(t)}{2} \|w(t+1) - w(t)\|^2 \le H(w(t), x(t)) - H(w(t+1), x(t)), \tag{4.7}$$

where $\alpha(t) = \min\limits_{1\le i\le m} \{\alpha_i(t)\}$. This part was proven independently of the distance function $d(\cdot, \cdot)$, and since in (Alg-Name) the clusters assignment step is identical to KPALM, the claim is correct in the case of (Alg-Name) as well.

Applying Lemma 4.2 from [reference to Weizsfeld paper] yields

$$\begin{aligned} &H_l(w(t+1), x(t+1)) - H_l(w(t+1), x) \le \\ &\quad \le \frac{L_l(w(t+1), x(t))}{2} \left(\|x^l(t) - x^l\|^2 - \|x^l(t+1) - x^l\|^2 \right), \quad \forall x \in \mathbb{R}^{nk}, 1 \le l \le k. \end{aligned}$$

Setting $x = x(t)$ and rearranging yields

$$\frac{L_l(w(t+1), x(t))}{2} \|x^l(t+1) - x(t)\|^2 \le H_l(w(t+1), x(t)) - H_l(w(t+1), x(t+1)), \quad \forall 1 \le l \le k. \tag{4.8}$$

Denote $L(t) = \min\limits_{1 \le l \le k} \{L_l(w(t+1), x(t))\}$. Summing (4.8) over $l = 1, 2, \ldots, k$ leads to

$$\begin{aligned}\frac{L(t)}{2}\|x(t+1-x(t)\|^2 &= \frac{L(t)}{2}\sum_{l=1}^{k}\|x^l(t+1-x^l(t)\|^2 \\ &\le \sum_{l=1}^{k}\frac{L_l(t)}{2}\|x^l(t+1-x^l(t)\|^2 \\ &\le \sum_{l=1}^{k}\left(H_l(w(t+1),x(t)) - H_l(w(t+1),x(t+1))\right) \\ &= H(w(t+1),x(t)) - H(w(t+1),x(t+1)).\end{aligned} \tag{4.9}$$

Set $\rho_1 = \frac{1}{2}\min\{\alpha(t), L(t)\}$, and again Assumption 1 assures that $\rho_1 > 0$. Combined with (4.7) and (4.9) we have that

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

□