1   The Clustering Problem

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 data $\mathcal{A}$ 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, and we want to determine k cluster centers $\{x^1, x^2, \ldots, x^k\}$ 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 $\min_{1 \leq l \leq k} u_l = \min\{\langle u, v \rangle \mid v \in \Delta\}$, 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. *For all $1 \leq l \leq k$ and $t \in \mathbb{N}$ we assume that $\sum_{i=1}^m w_l^i(t) > 0$, or equivalently that for all $t \in \mathbb{N}$ $\min_{1 \leq l \leq k} \left\{ \sum_{i=1}^m w_l^i(t) \right\} > 0$.*

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, $x^l(t+1) = \arg \min_{x^l \in \mathbb{R}^k} \left\{ \sum_{i=1}^m w_l^i(t) d(x^l, a^i) \right\}$, $l = 1, 2, \ldots, k$, 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 $M = \max_{1 \leq i \leq m} \|a^i\|$.

(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 $\lambda_i \geq 0$ and $\sum_{i=1}^m \lambda_i = 1$. 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

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

(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 $\beta(w) = 2 \min\limits_{1 \le l \le k} \left\{ \sum\limits_{i=1}^{m} w_l^i \right\}$, whenever $\beta(w) > 0$.*

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

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

Since the Hessian is a diagonal matrix, the smallest eigenvalue is $\min\limits_{1 \le l \le k} 2 \sum\limits_{i=1}^{m} w_l^i = \beta(w)$, and the result follows. $\qquad\square$

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 $\alpha(t) = \min\limits_{1 \le i \le m} \{\alpha_i(t)\}$. Summing inequality (3.5) over $i = 1, 2, \dots, m$ yields

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

4

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

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

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

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

where the last equality follows from Remark 1(i). $\qquad\square$

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

$$\|d^i(x(t+1)) - d^i(x(t))\| = \left[\sum_{l=1}^{k} \left| \|x^l(t+1) - a^i\|^2 - \|x^l(t) - a^i\|^2 \right|^2 \right]^{\frac{1}{2}}$$

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

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

$$\leq \left[\sum_{l=1}^{k} \left(\left| \|x^l(t+1)\| - \|x^l(t)\| \right| \cdot \left| \|x^l(t+1)\| + \|x^l(t)\| \right| + 2\|x^l(t) - x^l(t+1)\| \cdot \|a^i\| \right)^2 \right]^{\frac{1}{2}}$$

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

5

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

this proves the desired result. $\qquad\qquad\square$

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*

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

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

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

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

$$\partial\Psi(z(t+1)) =$$
$$= \left((\nabla_{w^i} H_i(w(t+1), x(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)))_{i=1,\dots,m}, \nabla_x H(w(t+1), x(t+1)) \right)$$
$$= \left((d^i(x(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)))_{i=1,\dots,m}, \nabla_x H(w(t+1), x(t+1)) \right)$$
$$= \left((d^i(x(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)))_{i=1,\dots,m}, \mathbf{0} \right),$$

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

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

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

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

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

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 $\overline{\alpha}(t) = \max_{1 \le i \le m} \alpha_i(t)$. 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

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

   (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

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

As in KPALM, KMEANS needs Assumption 1 for step (3.11) to be well defined. In order 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 $R_{lj}(t) = \min_{a \in C^l(t)} \{\|a - x^j(t)\| - \|a - x^l(t)\|\}$ for all $1 \leq l, j \leq k$, and $r(t) = \min_{l \neq j} R_{lj}$.

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 neighbourhood 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 $1 \leq l \leq k$. Let $(C(t), x(t))$ be the clusters and centers KMEANS 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 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

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

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 $C^l(t) \subseteq \underline{C^l}(x)$. 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 $1 \leq l \leq k$. $\qquad\square$

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

8

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

$$\min_{x^1,x^2,\ldots,x^k\in\mathbb{R}^n}\left\{\sum_{i=1}^{m}\min_{1\leq l\leq k}\|x^l-a^i\|\right\}. \tag{4.1}$$

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

4.1   The Smoothed Fermat-Weber Problem

Solving the smoothed Fermat-Weber plays a significant role in the algorithm that addresses the clustering problem with Euclidean norm distance-like function. The Fermat-Weber problem is formulated as follows

$$\min_{x\in\mathbb{R}^n}\left\{f(x):=\sum_{i=1}^{m}w_i\|x-a^i\|\right\}, \tag{4.2}$$

where $w_i>0$, $i=1,2,\ldots,m$, are given positive weights and $\mathcal{A}=\left\{a^1,a^2,\ldots,a^m\right\}\subset\mathbb{R}^n$ are given vectors. As shown in [BS2015] this problem can be solved via the consecutive appliance of the operator $T:\mathbb{R}^n\to\mathbb{R}^n$ defined by

$$T(x)=\frac{\displaystyle\sum_{i=1}^{m}\frac{w_ia^i}{\|x-a^i\|}}{\displaystyle\sum_{i=1}^{m}\frac{w_i}{\|x-a^i\|}}.$$

It is easily noticed that $f(x)$ is not differentiable over $\mathcal{A}$. For our purposes we are interested in the smoothed Fermat-Weber problem, that can be formulated in the following manner

$$\min_{x\in\mathbb{R}^n}\left\{f_\epsilon(x):=\sum_{i=1}^{m}w_i\left(\|x-a^i\|^2+\epsilon^2\right)^{1/2}\right\}, \tag{4.3}$$

with $\epsilon>0$ being some small perturbation constant. Next we introduce the operator $T_\epsilon:\mathbb{R}^n\to\mathbb{R}^n$ defined by

$$T_\epsilon(x)=\frac{\displaystyle\sum_{i=1}^{m}\frac{w_ia^i}{(\|x-a^i\|^2+\epsilon^2)^{1/2}}}{\displaystyle\sum_{i=1}^{m}\frac{w_i}{(\|x-a^i\|^2+\epsilon^2)^{1/2}}}.$$

This version of the smooth T operator together with its properties that are to be discussed below are the cornerstone to prove the properties needed by PALM, and in turn to show the convergence of the sequence generated by the algorithm proposed to tackle the smooth version of the clustering problem presented later on. In order to prove some properties of T_ϵ, which are the same as the

properties of T described in [BS2015], we also will need an auxiliary function $h_\epsilon : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

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

Another useful function $L_\epsilon : \mathbb{R}^n \to \mathbb{R}$ that serves somewhat like Lipschitz function for the gradient of f_ϵ is defined by

$$L_\epsilon(x) = \sum_{i=1}^m \frac{w_i}{(\|x - a^i\|^2 + \epsilon^2)^{1/2}}.$$

It is easy to verify the following equality

$$T_\epsilon(x) = x - \frac{\nabla f_\epsilon(x)}{L_\epsilon(x)} \quad \forall x \in \mathbb{R}^n. \tag{4.4}$$

Lemma 4.0.1 (Properties of the auxiliary function h_ϵ). *The following properties of h_ϵ hold.*

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

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

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

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

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

$$T_\epsilon(y) = \arg\min_{x \in \mathbb{R}^n} h_\epsilon(x, y).$$

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

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

4.2   Clustering with Weiszfeld Step

We introduce some useful notations that will be useful for this section. For $1 \leq l \leq k$ 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 $1 \leq l \leq k$ via

$$T_l(w, x) = \frac{\sum_{i=1}^m \frac{w_l^i a^i}{\|x^l - a^i\|}}{\sum_{i=1}^m \frac{w_l^i}{\|x^l - a^i\|}}. \tag{4.5}$$

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 Weiszfeld 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

$$w^i(t+1) = \arg \min_{w^i \in \Delta} \left\{ \langle w^i, d^i(x(t)) \rangle + \frac{\alpha_i(t)}{2} \|w^i - w^i(t)\|^2 \right\}. \tag{4.6}$$

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

$$x^l(t+1) = T_l(w(t+1), x(t)). \tag{4.7}$$

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

Remark 2. *(i) Due to Assumption 3 it follows that the centers update step in (4.7) is well defined.*

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

$$\nabla_{x^l} H_l(w, x) = \sum_{i=1}^{m} w_l^i \frac{x^l - a^i}{\|x^l - a^i\|}, \quad \forall x^l \notin \mathcal{A}, \tag{4.8}$$

and that

$$T_l(w, x) = x^l - \frac{1}{L_l(w, x)} \nabla_{x^l} H_l(w, x), \quad \forall x^l \notin \mathcal{A}. \tag{4.9}$$

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

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

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*

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

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

$$\frac{\alpha(t)}{2} \|w(t+1) - w(t)\|^2 \leq H(w(t), x(t)) - H(w(t+1), x(t)), \tag{4.10}$$

where $\alpha(t) = \min_{1 \leq i \leq m} \{\alpha_i(t)\}$. 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 Weiszfeld paper] yields

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

11

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

$$\frac{L_l(w(t+1), x(t))}{2} \|x^l(t+1) - x^l(t)\|^2 \leq H_l(w(t+1), x(t)) - H_l(w(t+1), x(t+1)), \quad \forall 1 \leq l \leq k. \quad (4.11)$$

Denote $L(t) = \min_{1 \leq l \leq k} \{L_l(w(t+1), x(t))\}$. Summing (4.11) over $l = 1, 2, \ldots, k$ leads to

$$
\begin{aligned}
\frac{L(t)}{2} \|x(t+1) - x(t)\|^2 &= \frac{L(t)}{2} \sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\|^2 \\
&\leq \sum_{l=1}^{k} \frac{L_l(t)}{2} \|x^l(t+1) - x^l(t)\|^2 \\
&\leq \sum_{l=1}^{k} \left(H_l(w(t+1), x(t)) - H_l(w(t+1), x(t+1)) \right) \\
&= H(w(t+1), x(t)) - H(w(t+1), x(t+1)).
\end{aligned}
\quad (4.12)
$$

Set $\rho_1 = \frac{1}{2} \min \{\alpha(t), L(t)\}$, and then Assumption 1 assures that $\rho_1 > 0$. Combining (4.10) and (4.12) yields

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

which proves the desired result. $\qquad\square$

Next we prove the subgradient lower bounds for iterates gap property of the sequence generated by (Alg-Name), we start with proving two useful lemmas.

Lemma 4.1.1. *Let $\{z(t)\}_{t \in \mathbb{N}} = \{(w(t), x(t))\}_{t \in \mathbb{N}}$ be the sequence generated by (Alg-Name), then*

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

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

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

$\qquad\square$

The following lemma is based on the results within [reference to Weiszfeld paper]. We remind the notations in that context. For non-negative weights $w_i \geq 0, i = 1, 2, \ldots, m$, and m points $\mathcal{A} = \{a^1, a^2, \ldots, a^m\} \subset \mathbb{R}^n$, the functions $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \times \mathbb{R}^n \setminus \mathcal{A} \to \mathbb{R}$ and $L : \mathbb{R}^n \to \mathbb{R}$ are respectively defined by

$$f(x) = \sum_{i=1}^m w_i \|x - a^i\|, \quad h(x,y) := \sum_{i=1}^m w_i \frac{\|x - a^i\|^2}{\|y - a^i\|}, \quad L(x) := \sum_{i=1}^m \frac{w_i}{\|x - a^i\|}.$$

Lemma 4.1.2. *For all $y^0, y \in \mathbb{R}^n \setminus \mathcal{A}$ the following statement holds true*

$$\|\nabla f(y) - \nabla f(y^0)\| \leq \frac{2L(y^0)L(y)}{L(y^0) + L(y)} \|y^0 - y\|.$$

Proof. Let $y^0 \in \mathbb{R}^n \setminus \mathcal{A}$ be a fixed vector. Define the following function

$$\widetilde{f}(y) = f(y) - \langle \nabla f(y^0), y \rangle,$$

and similarly to the paper [reference to Weiszfeld paper] we define

$$\widetilde{h}(x,y) = h(x,y) - \langle \nabla f(y^0), x \rangle.$$

It is clear that $x \mapsto \widetilde{h}(x,y)$ is still quadratic function with associated matrix $L(y)\mathbf{I}$. Therefore, we can write

$$\begin{aligned}
\widetilde{h}(x,y) &= \widetilde{h}(y,y) + \langle \nabla_x \widetilde{h}(y,y), x - y \rangle + L(y)\|x - y\|^2 \\
&= \widetilde{f}(y) + \langle 2\nabla f(y) - \nabla f(y^0), x - y \rangle + L(y)\|x - y\|^2.
\end{aligned} \tag{4.13}$$

On the other hand, from [BS2015, Lemma 3.1(ii), Page 7] we have that

$$\begin{aligned}
\widetilde{h}(x,y) &= h(x,y) - \langle \nabla f(y^0), x \rangle \geq 2f(x) - f(y) - \langle \nabla f(y^0), x \rangle \\
&= 2\widetilde{f}(x) - \widetilde{f}(y) + \langle \nabla f(y^0), y - x \rangle,
\end{aligned} \tag{4.14}$$

where the last equality follows from the definition of $\widetilde{f}$. Combining (4.13) and (4.14) yields

$$\begin{aligned}
2\widetilde{f}(x) &\leq 2\widetilde{f}(y) + 2\langle \nabla f(y) - \nabla f(y^0), x - y \rangle + L(y)\|x - y\|^2 \\
&= 2\widetilde{f}(y) + 2\langle \nabla \widetilde{f}(y), x - y \rangle + L(y)\|x - y\|^2.
\end{aligned}$$

Dividing the last inequality by 2 leads to

$$\widetilde{f}(x) \leq \widetilde{f}(y) + \langle \nabla \widetilde{f}(y), x - y \rangle + \frac{L(y)}{2}\|x - y\|^2. \tag{4.15}$$

It is clear that the optimal point of $\widetilde{f}$ is y^0 since $\nabla \widetilde{f}(y^0) = 0$, therefore from (4.15) we obtain

$$\begin{aligned}
\widetilde{f}(y^0) &\leq \widetilde{f}\left(y - \frac{1}{L(y)}\nabla \widetilde{f}(y)\right) \leq \widetilde{f}(y) + \left\langle \nabla \widetilde{f}(y), -\frac{1}{L(y)}\nabla \widetilde{f}(y) \right\rangle + \frac{L(y)}{2}\left\|\frac{1}{L(y)}\nabla \widetilde{f}(y)\right\|^2 \\
&= \widetilde{f}(y) - \frac{1}{2L(y)}\left\|\nabla \widetilde{f}(y)\right\|^2.
\end{aligned}$$

Thus, using the definition of $\widetilde{f}$ and the face that $\nabla \widetilde{f}(y) = \nabla f(y) - \nabla f(y^0)$, yields that

$$f(y^0) \leq f(y) + \langle \nabla f(y^0), y^0 - y \rangle - \frac{1}{2L(y)}\|\nabla f(y) - \nabla f(y^0)\|^2.$$

13

Now, following the same arguments we can show that

$$f(y) \leq f(y^0) + \langle \nabla f(y), y - y^0 \rangle - \frac{1}{2L(y^0)} \|\nabla f(y^0) - \nabla f(y)\|^2$$

and combining last two inequalities yields that

$$\left(\frac{1}{2L(y^0)} + \frac{1}{2L(y)} \right) \|\nabla f(y) - \nabla f(y^0)\|^2 \leq \langle \nabla f(y^0) - \nabla f(y), y^0 - y \rangle,$$

that is

$$\|\nabla f(y) - \nabla f(y^0)\| \leq \frac{2L(y^0)L(y)}{L(y^0) + L(y)} \|y^0 - y\|,$$

for all $y^0, y \in \mathbb{R}^n \setminus \mathcal{A}$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Proposition 4.2 (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 (Alg-Name), 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. Repeating the steps of the proof in the case of KPALM yields that

$$\gamma(t+1) := \left(\left(d^i(x(t+1)) + u^i(t+1) \right)_{i=1,\ldots,m}, \nabla_x H(w(t+1), x(t+1)) \right) \in \partial\Psi(z(t+1)), \quad (4.16)$$

where for all $1 \leq i \leq m$, $u^i(t+1) \in \partial\delta_\Delta(w^i(t+1))$ such that

$$d^i(x(t)) + \alpha_i(t) \left(w^i(t+1) - w^i(t) \right) + u^i(t+1) = \mathbf{0}. \qquad (4.17)$$

Plugging (4.17) into (4.16), and taking norm yields

$$\|\gamma(t+1)\| \leq \sum_{i=1}^m \|d^i(x(t+1)) - d^i(x(t)) - \alpha_i(t) \left(w^i(t+1) - w^i(t) \right)\| + \|\nabla_x H(w(t+1), x(t+1))\|$$

$$\leq \sum_{i=1}^m \|d^i(x(t+1)) - d^i(x(t))\| + \sum_{i=1}^m \alpha_i(t) \|w^i(t+1) - w^i(t)\| + \|\nabla_x H(w(t+1), x(t+1))\|$$

$$\leq m\|x(t+1) - x(t)\| + m\overline{\alpha}(t)\|w(t+1) - w(t)\| + \|\nabla_x H(w(t+1), x(t+1))\|,$$

where the last inequality follows from Lemma 4.1.1 and the fact that $\overline{\alpha}(t) = \max_{1 \leq i \leq m} \alpha_i(t)$.

Next we will derive the following bound $\|\nabla_x H(w(t+1), x(t+1))\| \leq c\|x(t+1) - x(t)\|$, for some constant $c > 0$. Indeed, we have

$$\|\nabla_x H(w(t+1), x(t+1))\| \leq \sum_{l=1}^k \|\nabla_{x^l} H_l(w(t+1), x(t+1))\|$$

$$\leq \sum_{l=1}^k \|\nabla_{x^l} H_l(w(t+1), x(t))\| + \sum_{l=1}^k \|\nabla_{x^l} H_l(w(t+1), x(t+1)) - \nabla_{x^l} H_l(w(t+1), x(t))\|. \qquad (4.18)$$

From (4.9) we have

$$\nabla_{x^l} H_l(w(t+1), x(t)) = L_l(w(t+1), x(t)) \left(x^l(t+1) - x^l(t) \right), \quad \forall 1 \leq l \leq k,$$

14

applying Lemma 4.1.2 with respect to $H_l(w(t+1), \cdot)$ and plugging into (4.18) yields

$$\|\nabla_x H(w(t+1), x(t+1))\| \leq$$

$$\leq \sum_{l=1}^{k} \left(L_l(w(t+1), x(t)) + \frac{2L_l(w(t+1), x(t))L_l(w(t+1), x(t+1))}{L_l(w(t+1), x(t)) + L_l(w(t+1), x(t+1))} \right) \|x^l(t+1) - x^l(t)\|.$$

Therefore, denote $\overline{L}(t) = \max_{1 \leq l \leq k} \left\{ L_l(w(t+1), x(t)) + \frac{2L_l(w(t+1), x(t))L_l(w(t+1), x(t+1))}{L_l(w(t+1), x(t)) + L_l(w(t+1), x(t+1))} \right\}$, and set $\rho_2 = m + m\overline{\alpha}(t) + k\overline{L}(t)$, and the result follows. $\qquad\square$

4.3 Refactoring the Method

In this section we use the modification of the Weiszfeld method as presented in [reference to Weiszfeld paper]. We will show that this approach allows us to drop Assumption 3 and treat the cases when $x^l(t) \in \mathcal{A}$ for some $1 \leq l \leq k$ and $t \in \mathbb{N}$.

At first we return to the notations that are used in [reference to Weiszfeld paper]. We remind that in order to solve a single update of cluster center we need to solve the following minimization problem

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \sum_{i=1}^{m} w_i \|x - a_i\| \right\}. \tag{4.19}$$

The operator that is invloved in solving this problem $\tilde{T} : \mathbb{R}^n \to \mathbb{R}^n$ is defined as follows

$$\tilde{T}(x) = \begin{cases} T(x), & x \notin \mathcal{A}, \\ a_j, & x = a_j \ (1 \leq j \leq m) \text{ and } \|R_j\| \leq w_j, \\ S(a_j), & x = a_j \ (1 \leq j \leq m) \text{ and } \|R_j\| > w_j, \end{cases}$$

where the operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$T(x) = \frac{\sum_{i=1}^{m} \frac{w_i a_i}{\|x - a_i\|}}{\sum_{i=1}^{m} \frac{w_i}{\|x - a_i\|}},$$

the values R_j are computed via

$$R_j = \sum_{i=1, i \neq j} w_i \frac{a_j - a_i}{\|a_i - a_j\|}, \quad j = 1, 2, \ldots, m,$$

and the operator $S : \mathcal{A} \to \mathbb{R}^n$ is defined as follows

$$S(a_j) = a_j + t_j d_j,$$

with the direction $d_j = -\frac{R_j}{\|R_j\|}$ and the stepsize $t_j = \frac{\|R_j\| - w_j}{L(a_j)}$. The function $L : \mathbb{R}^n \to \mathbb{R}$ that is used in the definition of the stepsize t_j is also extended to $\mathcal{A}$ in the following manner

$$L(x) = \begin{cases} \sum_{i=1}^{m} \frac{w_i}{\|x - a_i\|}, & x \notin \mathcal{A}, \\ \sum_{\substack{i=1 \\ i \neq j}}^{m} \frac{w_i}{\|a_j - a_i\|}, & x = a_j \ (1 \leq j \leq m). \end{cases}$$

15

Based on these notations, the algorithm that is proposed in [reference to Weiszfeld paper] to solve the problem (4.19) is called the Modified Weiszfeld Method, and it is a simple iterative application of operator $\tilde{T}$, as specified below.

(1) Initialization: Pick random vector $x_0 \in \mathbb{R}^n$.

(2) General Step: For each $k = 0, 1, \ldots$ generate a sequence $x_{k+1} = \tilde{T}(x_k)$.

Lemma 4.2.1. *Let $\{x_k\}_{t \in \mathbb{N}}$ be the sequence generated by the Modified Weiszfeld Method, and assume that for some $k \in \mathbb{N}$ either $x_k \in \mathcal{A}$ or $x_{k+1} \in \mathcal{A}$, then the following statements hold true.*

(i) *There exists $M_1 > 0$ such that $M_1 \|x_{k+1} - x_k\|^2 \leq f(x_k) - f(x_{k+1})$.*

(ii) *There exist $M_2 > 0$ and $v_{k+1} \in \partial f(x_{k+1})$ such that $\|v_{k+1}\| \leq M_2 \|x_{k+1} - x_k\|$.*

Proof. (i) First assume that $x_k = a_j \in \mathcal{A}$. If $\|R_j\| \leq w_j$ then $x_{k+1} = a_j$ and the result follows. If $\|R_j\| > w_j$ then $x_{k+1} = S(a_j)$ then from [BS 2015, Lemma 7.1, page 20] we have that

$$f(x_k) - f(x_{k+1}) \geq \frac{L(a_j)}{2} t_j{}^2 = \frac{L(a_j)}{2} \|x_{k+1} - x_k\|^2.$$

Now, suppose $x_k \notin \mathcal{A}$, and $x_{k+1} = a_j \in \mathcal{A}$. Plugging $x = x_k$ in [BS 2015, Lemma 5.2, page 12] yields

$$f(x_{k+1}) - f(x_k) \leq \frac{L(x_k)}{2} \left(-\|x_{k+1} - x_k\|^2 \right).$$

Thus, setting $M_1 = L(x_k)$ gives the desired result for all cases.

(ii) Again, first we inspect the case $x_k = a_j \in \mathcal{A}$. If $\|R_j\| \leq w_j$ then according to [BS 2015] $x = a_j$ is the optimal solution to problem (4.19), hence $\mathbf{0} \in \partial f(a_j)$ and the results follows. If $\|R_j\| > w_j$ then $x_{k+1} = S(a_j)$, hence

$$0 < t_j = \|x_{k+1} - x_k\|. \tag{4.20}$$

Moreover, evaluating the subgradient of f leads to

$$\partial f(x) = \begin{cases} \{\nabla f(x)\} = \left\{ \sum\limits_{i=1}^{m} w_i \frac{x - a_i}{\|x - a_i\|} \right\}, & x \notin \mathcal{A}, \\[2em] \left\{ \sum\limits_{\substack{i=1 \\ i \neq j}}^{m} w_i \frac{x - a_i}{\|a_j - a_i\|} + w_j z_j \mid \|z_j\| \leq 1 \right\}, & x = a_j \in \mathcal{A} \end{cases}.$$

Thus we have that

$$\|\partial f(x)\| \leq \sum_{i=1}^{m} w_i \leq m. \tag{4.21}$$

From (4.20) and (4.21) it follows that there exists $M_2 > 0$ such that $\|\partial f(x)\| \leq M_2 \|x_{k+1} - x_k\|$. Now, assume that $x_k \notin \mathcal{A}$, and $x_{k+1} = a_j \in \mathcal{A}$. Thus $\|x_{k+1} - x_k\| > 0$ and combining the fact that $\|\partial f(x)\| \leq m$ then again we can assure that there exists $M_2 > 0$ such that $\|\partial f(x)\| \leq M_2 \|x_{k+1} - x_k\|$.

$\square$

The last lemma allows us to drop Assumption 3 and bypass points in $\mathcal{A}$ by simply replacing the operator T in center update step in (Alg-Name) (4.7) with the operator $\tilde{T}$, and still reserve both, the sufficient decrease and the subgradient lower bound for iterates gap properties, that are essential for the conversion of the algorithm to some limit point.

5 Clustering via ADMM Approach

Introducing some new variable into the problem leads to the following clustering problem notation

$$\min_{x\in\mathbb{R}^{nk}} \min_{w\in\mathbb{R}^{km}} \left\{ \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i d(x^l, a^i) \mid w^i \in \Delta, i = 1, 2, \ldots, m \right\}$$

$$= \min_{x\in\mathbb{R}^{nk}, w\in\mathbb{R}^{km}, z\in\mathbb{R}^{km}} \left\{ \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i z_l^i \;\middle|\; \begin{array}{ll} w^i \in \Delta, & i = 1, 2, \ldots, m, \\ z_l^i = d(x^l, a^i), & i = 1, 2, \ldots, m, \quad l = 1, 2, \ldots, k \end{array} \right\}.$$

The augmented Lagrangian that is associated with this problem is

$$L_\rho(w, x, z, y) = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i z_l^i + \sum_{i=1}^{m}\sum_{l=1}^{k} y_l^i(z_l^i - d(x^l, a^i)) + \frac{\rho}{2}\sum_{i=1}^{m}\sum_{l=1}^{k}\left(z_l^i - d(x^l, a^i)\right)^2. \tag{5.1}$$

Thus the ADMM formulas for (4.2) are as follows

$$w(t+1) = \arg\min_{w\in\Delta^m} L_\rho(w, x(t), z(t), y(t)),$$

$$\Rightarrow \quad w^i(t+1) = \arg\min_{w^i\in\Delta} \sum_{l=1}^{k} w_l^i z_l^i(t) = \arg\min_{w^i\in\Delta} \left\langle w^i, z^i(t) \right\rangle, \quad 1 \le i \le m,$$

$$x(t+1) = \arg\min_{x\in\mathbb{R}^{nk}} L_\rho(w(t+1), x, z(t), y(t)),$$

$$\Rightarrow \quad x^l(t+1) = \arg\min_{x^l\in\mathbb{R}^{nk}} -\sum_{i=1}^{m} y_l^i(t) d(x^l, a^i) + \frac{\rho}{2}\sum_{i=1}^{m}\left(z_l^i - d(x^l, a^i)\right)^2, \quad 1 \le l \le k,$$

$$z(t+1) = \arg\min_{z\in\mathbb{R}^{km}} L_\rho(w(t+1), x(t+1), z, y(t)),$$

$$\Rightarrow \quad z^i(t+1) = \arg\min_{z^i\in\mathbb{R}^{km}} \left\langle w^i(t+1), z^i \right\rangle + \left\langle y^i(t), z^i \right\rangle + \frac{\rho}{2}\left\| z^i - \left(d(x^l(t+1), a^i)\right)_{l=1,\ldots,k} \right\|^2$$

$$= \left(d(x^l(t+1), a^i)\right)_{l=1,\ldots,k} - \frac{1}{\rho}\left(w^i(t+1) + y^i(t)\right), \quad 1 \le i \le m,$$

$$y_l^i(t+1) = y_l^i(t) + \rho(z_l^i(t+1) - d(x^l(t+1), a^i)), \quad 1 \le i \le m, 1 \le l \le k.$$