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 $x^l \in \mathbb{R}^n$, and we ~~want~~ are intersted to determine k cluster centers $\{x^1, x^2, \ldots, x^k\}$ such that the sum of certain proximity measures from each point $a^i, i = 1, 2, \ldots, m$, to a nearest cluster center x^l is minimized. W ~~we~~ define the vector of all centers by $x = (x^1, x^2, \ldots, x^k) \in \mathbb{R}^{nk}$.

The clustering problem is given by

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

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

2 Problem Reformulation and Notations

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

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

where Δ ~~is~~ denotes the well-known simplex defined by

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

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

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

write it as formula in a separate line

where $\boxed{d^i(x) = (d(x^1, a^i), d(x^2, a^i), \ldots, d(x^k, a^i)) \in \mathbb{R}^k, i = 1, 2, \ldots, m.}$ Replacing further the constraint $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$, which defined to be ~~0~~ in Δ and ∞ otherwise, to the objective function, results in a equivalent formulation

OK

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

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

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

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

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

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, given in (2.3), 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.

(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 PALM.

(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 PALM.

the

As in PALM algorithm, our algorithm is based on alternating minimization, with the following adaptations which are motivated by the observations mentioned above. More precisely, with respect to w we suggest to regularize the subproblem with proximal term as follows:

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

On the other hand, with respect to x we perform exact minimization

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

It is easy to check that all subproblems, with respect to w^i, $i = 1, 2, \ldots, m$, and x, can be simplified as follows:

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

where P_Δ is the orthogonal projection onto the set Δ, and

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

Therefore we can record now the suggested KPALM algorithm.

KPALM

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

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

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

a

$$w^i(t+1) = P_\Delta\left(w^i(t) - \frac{d^i(x(t))}{\alpha_i(t)}\right). \tag{3.5}$$

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

u

$$x^l(t+1) = \frac{\sum_{i=1}^m w_l^i(t+1)a^i}{\sum_{i=1}^m w_l^i(t+1)}. \tag{3.6}$$

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

Lemma 3.0.1 (Boundedness of KPALM sequence). *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM. Then, the following statements hold true.*

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

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

Proof. (i) 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 $\sum_{i=1}^m \lambda_i = 1$. From (3.4) we have

$$x^l(t) = \frac{\sum_{i=1}^m w_l^i(t)a^i}{\sum_{i=1}^m w_l^i(t)} = \sum_{i=1}^m \left(\frac{w_l^i(t)}{\sum_{j=1}^m w_l^j(t)}\right) a^i = \sum_{i=1}^m \lambda_i a^i \in Conv(\mathcal{A}). \tag{3.7}$$

Hence $x^l(t)$ is in the convex hull of $\mathcal{A}$, for all $l = 1, 2, \ldots, k$ and $t \in \mathbb{N}$. Taking the norm of $x^l(t)$ and using (3.7) yields that

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

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

□

The following assumption will be crucial for the ~~comping~~ analysis.

coming

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

$$\underline{\alpha_i} \leq \alpha_i(t) \leq \overline{\alpha_i}, \quad \forall\, t \in \mathbb{N}. \tag{3.8}$$

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

$$\min_{1\leq l\leq k} \sum_{i=1}^{m} w_l^i(t) \geq \underline{\beta}. \tag{3.9}$$

$\alpha_i(t)$

It should be noted that Assumption 1(i) is very mild since the parameters ~~α_i~~, $1 \leq i \leq m$ and $t \in \mathbb{N}$, can be chosen arbitrarily by the user and therefore it can be controlled such that the boundedness property holds true. Assumption 1(ii) is essential since if it is not true then $w_l^i(t) = 0$ for all $1 \leq i \leq m$, which means that the center x^l does not involved in the objective function.

Lemma 3.0.2 (Strong convexity of $H(w, x)$ in x). *The function $x \mapsto H(w, x)$ is strongly convex with parameter $\beta(w) := 2 \min\limits_{1\leq l\leq k} \left\{ \sum\limits_{i=1}^{m} w_l^i \right\}$, 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. Indeed, the Hessian is given by

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

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

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

Proposition 3.1 (Sufficient decrease property). *Suppose that Assumption 1 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM. Then, there exists $\rho_1 > 0$ such that*

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

Proof. From the step (3.5) we derive the following inequality

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

Hence, we obtain

$$\frac{\alpha_i(t)}{2}\|w^i(t+1)-w^i(t)\|^2 \leq H_i(w(t),x(t)) - H_i(w(t+1),x(t)). \tag{3.10}$$

Denote $\underline{\alpha} = \min\limits_{1\leq i\leq m} \underline{\alpha_i}$. Summing inequality (3.10) over $i = 1, 2, \ldots, m$ yields

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

where the first inequality follows from Assumption 1(i).

From Assumption 1(ii) we have that $\beta(w(t)) = 2\min\limits_{1\leq l\leq k}\left\{\sum\limits_{i=1}^{m} w_l^i(t)\right\} \geq \underline{\beta}$, 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

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

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

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

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

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

Lemma 3.1.1. *Let* $\{z(t)\}_{t\in\mathbb{N}}$ *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

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

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

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

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

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

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

this proves the desired result. □

Proposition 3.2 (Subgradient lower bound for the iterates gap). *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM. Then there exists $\rho_2 > 0$ and $\gamma(t+1) \in \partial\Psi(z(t+1))$ such that*

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

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

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

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

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

where the last equality follows from (3.6), that is, the optimality condition of $x(t+1)$.

The optimality condition of $w^i(t+1)$ which 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

$$d^i(x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1) = \mathbf{0}. \tag{3.11}$$

Setting $\gamma(t+1) := \left((d^i(x(t+1)) + u^i(t+1))_{i=1,2,\ldots,m}, \mathbf{0}\right) \in \partial\Psi(z(t+1))$. Using (3.11) we obtain

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

Hence,

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

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

4 Clustering via Alternation with Weiszfeld Step

4.1 Algorithm to the Smoothed Clustering Problem

In the previous section we showed that Problem (2.1) 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

$$H(w,x) = \sum_{i=1}^{m} \langle w^i, d^i(x)\rangle = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i\|x^l - a^i\|,$$

and

$$G(w) = \sum_{i=1}^{m} \delta_\Delta(w^i).$$

In ~~However, in~~ order to be able to use the theory mentioned in Section (3.1), we need the coupled function $H(w,x)$ to be smooth, which is not the case now. Therefore, for any $\varepsilon > 0$, it leads us to the following smoothed form of the clustering problem

$$\min\left\{\Psi_\varepsilon(z) := H_\varepsilon(w,x) + G(w) \mid z := (w,x) \in \mathbb{R}^{km} \times \mathbb{R}^{nk}\right\}, \tag{4.1}$$

~~$\sum_{i=1}^{m} H_\varepsilon^i(w,x)$~~

where

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

and

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

for all $i = 1, 2, \ldots, m$. Note that $\Psi_\varepsilon(z)$ is a perturbed form of $\Psi(z)$ for some small $\varepsilon > 0$, and $\Psi_0(z) = \Psi(z)$.

Now we would like to develop an algorithm which is based on the methodology of PALM to solve Problem (4.1). It is easy to see that with respect to w, the objective Ψ_ε keeps on the same structure as Ψ and therefore we apply the same step as in KPALM. More precisely, for all $i = 1, 2, \ldots, m$, we have

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

where $\alpha_i(t)$, $i = 1, 2, \ldots, m$, is arbitrarily chosen. On the other hand, with respect to x we tackle the subproblem differently than in KPALM ~~then PALM.~~ Here we follow exactly the idea of PALM, that is :

$$x^l(t+1) = \arg\min_{x^l} \left\{ \left\langle x^l - x^l(t), \nabla_{x^l} \underset{\text{OK}}{\cancel{H}_\varepsilon}(w(t+1), x(t)) \right\rangle + \frac{L_\varepsilon^l(w(t+1), x(t))}{2} \|x^l - x^l(t)\|^2 \right\},$$

where

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

Now we present our algorithm for solving Problem (4.1), we call it ε-KPALM. The algorithm alternates between cluster assignment step, similar to that as in KPALM, and centers update step that is based on certain gradient step.

ε-KPALM

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

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

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

$$w^i(t+1) = P_\Delta \left(w^i(t) - \frac{d_\varepsilon^i(x(t))}{\alpha_i(t)} \right). \tag{4.2}$$

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

$$x^l(t+1) = x^l(t) - \frac{1}{L_\varepsilon^l(w(t+1), x(t))} \nabla_{x^l} H_\varepsilon(w(t+1), x(t)). \tag{4.3}$$

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

$$x^l(t+1) = x^l(t) - \frac{1}{L_\varepsilon^l(w(t+1), x(t))} \nabla_{x^l} H(w(t+1), x(t))$$
$$= x^l(t) - \frac{1}{L_\varepsilon^l(w(t+1), x(t))} \sum_{i=1}^m w_l^i(t+1) \cdot \frac{x^l(t) - a^i}{(\|x^l(t) - a^i\|^2 + \varepsilon^2)^{1/2}}$$
$$= \frac{1}{L_\varepsilon^l(w(t+1), x(t))} \sum_{i=1}^m \frac{w_l^i(t+1) a^i}{(\|x^l(t) - a^i\|^2 + \varepsilon^2)^{1/2}} \in Conv(\mathcal{A}).$$

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

delete

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

for fixed positive numbers $w_1, w_2, \ldots, w_m \in \mathbb{R}$ and $b^i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$. We also need the following auxiliary function $h_\varepsilon : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

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

Finally need introduce the following two operators, $T_\varepsilon : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$T_\varepsilon(x) = \frac{1}{\sum\limits_{i=1}^{m} \frac{w_i}{\left(\|x-b^i\|^2+\varepsilon^2\right)^{1/2}}} \sum_{i=1}^{m} \frac{w_i b^i}{\left(\|x - b^i\|^2 + \varepsilon^2\right)^{1/2}},$$

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(iii) The function $x \mapsto h_\varepsilon(x, y)$ is strongly convex and its unique minimizer is determined by the optimality equation

$$\nabla_x h_\varepsilon(x, y) = \sum_{i=1}^{m} \frac{2w_i\left(x - b^i\right)}{\left(\|y - b^i\|^2 + \varepsilon^2\right)^{1/2}} = 0.$$

Simple algebraic manipulation leads to the relation

$$x = T_\varepsilon(y),$$

and the desired results follows.

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

□

~~The following proofs are based on the properties of the auxiliary function h_ε, and they are similar to the proofs in [BS2015], hence we will just state them here. Lemma 4.0.5 does not appear in that paper, and its proof is given here~~.

~~**Lemma 4.0.2** (Monotonicity property of T_ε, similar to (BS2015, Lemma 3.2, page 7)). *For every $y \in \mathbb{R}^n$ we have*~~

$$\cancel{f_\varepsilon(T_\varepsilon(y)) \le f_\varepsilon(y)}.$$

~~**Lemma 4.0.3** (Decent lemma for function f_ε, similar to (BS2015, Lemma 5.1, page 10)). *For every $y \in \mathbb{R}^n$ we have*~~

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

~~**Lemma 4.0.4** (Similar to (BS2015, Lemma 5.2, page 12)). *For every $x, y \in \mathbb{R}^n$ we have*~~

$$\cancel{f_\varepsilon(T_\varepsilon(y)) - f_\varepsilon(x) \le \frac{L_\varepsilon(y)}{2}\left(\|y - x\|^2 - \|T_\varepsilon(y) - x\|^2\right)}.$$

Lemma 4.0.5. *For all* ~~y^0, y~~ $\in \mathbb{R}^n$ *the following statement holds true*

y, z

z

$$\|\nabla f_\varepsilon(y) - \nabla f_\varepsilon(\cancel{y^0})\| \le \frac{2L_\varepsilon(\cancel{y^0})L_\varepsilon(y)}{L_\varepsilon(\cancel{y^0}) + L_\varepsilon(y)}\|\cancel{y^0} - y\|.$$

z

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

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

and

$$\widetilde{h_\varepsilon}(x, y) = h_\varepsilon(x, y) - \langle \nabla f_\varepsilon(z), x\rangle.$$

It is clear that $x \mapsto \widetilde{h_\varepsilon}(x, y)$ is also a quadratic function with associated matrix $L_\varepsilon(y)\mathbf{I}$. Therefore, from 4.0.1(i) we can write

$$\begin{aligned}\widetilde{h_\varepsilon}(x, y) &= \widetilde{h_\varepsilon}(y, y) + \left\langle \nabla_x \widetilde{h_\varepsilon}(y, y), x - y\right\rangle + L_\varepsilon(y)\|x - y\|^2 \\ &= \widetilde{f_\varepsilon}(y) + \langle 2\nabla f_\varepsilon(y) - \nabla f_\varepsilon(z), x - y\rangle + L_\varepsilon(y)\|x - y\|^2.\end{aligned} \tag{4.4}$$

Lemma

Now we can prove that the function f_ε has Lipschitz continuou gradient.

Lemma

On the other hand, from 4.0.1(ii) we have that

$$\begin{aligned}\widetilde{h}_\varepsilon(x,y) &= h_\varepsilon(x,y) - \langle \nabla f_\varepsilon(z), x\rangle \geq 2f_\varepsilon(x) - f_\varepsilon(y) - \langle \nabla f_\varepsilon(z), x\rangle \\ &= 2\widetilde{f}_\varepsilon(x) - \widetilde{f}_\varepsilon(y) + \langle \nabla f_\varepsilon(z), x-y\rangle\,,\end{aligned} \tag{4.5}$$

where the last equality follows from the definition of $\widetilde{f}_\varepsilon$. Combining (4.4) and (4.5) yields

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

Dividing the last inequality by 2 leads to

$$\widetilde{f}_\varepsilon(x) \leq \widetilde{f}_\varepsilon(y) + \left\langle \nabla \widetilde{f}_\varepsilon(y), x-y\right\rangle + \frac{L_\varepsilon(y)}{2}\|x-y\|^2. \tag{4.6}$$

It is clear that the optimal point of $\widetilde{f}_\varepsilon$ is z since $\nabla \widetilde{f}_\varepsilon(z) = 0$, therefore using (4.6) with $x = y - \cancel{\frac{1}{L_\varepsilon(y)}} (1/L_\varepsilon(y)) \nabla \widetilde{f}_\varepsilon(y)$ yields

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

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

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

Now, following the same arguments we can show that

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

Combining the ~~and combining~~ last two inequalities yields that

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

that is,

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

for all $z, y \in \mathbb{R}^n$. □

Now we are finally ready to prove the two properties that needed ~~by PALM, and deduce that~~ for guaranting that the sequence that is generated by ε-KPALM converge s to critical point of Ψ_ε.

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

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

Now we get back to ε-KPALM algorithm and prove few technical results about the ~~These main properties are given in the following technical result~~. involved functions which are based on the auxiliary results obtained above.

Proposition 1. Let $\{z(t)\}_{t\in N}$ be a sequence generated by ε-KPALM. Then, the following two statments hold true.

(1) For all $t\in N$ and $l = 1, 2, \ldots, K$ we have

$$L_\varepsilon^l\left(w(t+1), x(t)\right) \geqslant \frac{\beta}{(d_{il}^2 + \varepsilon^2)^{1/2}}\,, \longrightarrow \text{see Page 18}$$

As we already the steps of KPALM and ε-KPALM with respect to w are similar and therefore following the same arguments given at the begining of the proof of Proposition 3.1 we have that

Proof. ~~Similar steps to the ones in the proof of sufficient decrease property of KPALM lead to~~

$$\frac{\underline{\alpha}(\cancel{t})}{2}\|w(t+1)-w(t)\|^2 \le H_\varepsilon(w(t),x(t)) - H_\varepsilon(w(t+1),x(t)), \qquad (4.7)$$

where $\underline{\alpha}(\cancel{t}) = \min_{1\le i\le m}\{\overset{\alpha_i}{\cancel{\alpha_i(t)}}\}$.

~~Applying Lemma 4.0.4 with respect to $H_\varepsilon^{w_l(t+1)}(\cdot)$ yields~~

~~$H_\varepsilon^{w_l(t+1)}(x^l(t+1)) - H_\varepsilon^{w_l(t+1)}(x^l) \le \frac{L_\varepsilon^{w_l(t+1)}(x^l(t))}{2}\left(\|x^l(t)-x^l\|^2 - \|x^l(t+1)-x^l\|^2\right), \quad \forall x^l \in \mathbb{R}^n,$~~

~~for all $l = 1,2,\ldots,k$. Setting $x^l = x^l(t)$ and rearranging yields~~

~~$\frac{L_\varepsilon^{w_l(t+1)}(x^l(t))}{2}\|x^l(t+1)-x^l(t)\|^2 \le H_\varepsilon^{w_l(t+1)}(x^l(t)) - H_\varepsilon^{w_l(t+1)}(x^l(t+1)), \quad \forall 1 \le l \le k.$~~ (4.8)

~~Denote $\underline{L}(t) = \min_{1\le l\le k}\left\{L_\varepsilon^{w_l(t+1)}(x^l(t))\right\}$. Summing (4.8) over $l = 1,2,\ldots,k$ leads to~~

~~$\frac{\underline{L}(t)}{2}\|x(t+1)-x(t)\|^2 = \frac{\underline{L}(t)}{2}\sum_{l=1}^k \|x^l(t+1)-x^l(t)\|^2$~~

~~$\le \sum_{l=1}^k \frac{L_\varepsilon^{w_l(t+1)}(x^l(t))}{2}\|x^l(t+1)-x^l(t)\|^2$~~ (4.9)

~~$\le \sum_{l=1}^k \left(H_\varepsilon^{w_l(t+1)}(x^l(t)) - H_\varepsilon^{w_l(t+1)}(x^l(t+1))\right)$~~

~~$= H_\varepsilon(w(t+1),x(t)) - H_\varepsilon(w(t+1),x(t+1)).$~~

$$\rho_1 = \frac{1}{2}\min\left\{\underline{\alpha}, \beta/(d_{\mathcal{A}}^2+\varepsilon^2)^{1/2}\right\}$$

Set $\cancel{\rho_1} = \frac{1}{2}\min_{t\in\mathbb{N}}\{\underline{\alpha}(\cancel{t}), \underline{L}(\cancel{t})\}$, ~~and note that since $x^l(t) \in Conv(\mathcal{A})$ for all $1 \le l \le k$, then~~

~~$L_\varepsilon^{w_l(t+1)}(x^l(t)) = \sum_{i=1}^m \frac{w_l^i(t+1)}{(\|x^l(t)-a^i\|^2+\varepsilon^2)^{1/2}} \ge \frac{\sum_{i=1}^m w_l^i(t+1)}{(d_{\mathcal{A}}{}^2+\varepsilon^2)^{1/2}},$~~

~~where $d_{\mathcal{A}} = diam(Conv(\mathcal{A}))$, hence together with Assumption 1 assures that $\rho_1 > 0$.~~ Combining (4.7) and (4.9) yields

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

where the last equality follows from the fact that $G(w(t))=0$, $t\in\mathbb{N}$.

This ~~which~~ proves the desired result. □

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

Applying Proposition 2 we get for all $t\in\mathbb{N}$ that

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

where the last inequality follows from Proposition 1(i).

Lemma 4.1.1. *For any $x, y \in \mathbb{R}^{nk}$ such that $x^l, y^l \in Conv(\mathcal{A})$ for all $1 \le l \le k$ the following inequality holds*

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

with $d_{\mathcal{A}} = diam(Conv(\mathcal{A}))$.

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

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

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

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

Hence,

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

as asserted. □

Lemma 4.1.2 (Upper bound of the sequence $\left\{\overline{L}(x(t))\right\}_{t\in\mathbb{N}}$)**.** *Let $\{z(t)\}_{t\in\mathbb{N}} = \{(w(t), x(t))\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM, then for any $t \in \mathbb{N}$ we have*

$$\overline{L}(x(t)) = \max_{1\le l\le k}\left\{L_\varepsilon^{w_l(t+1)}(x^l(t)) + \frac{2L_\varepsilon^{w_l(t+1)}(x^l(t))L_\varepsilon^{w_l(t+1)}(x^l(t+1))}{L_\varepsilon^{w_l(t+1)}(x^l(t)) + L_\varepsilon^{w_l(t+1)}(x^l(t+1))}\right\} \le \frac{2m}{\varepsilon}.$$

Proof. For any $w_l \in [0, 1]^m$ and $x^l \in \mathbb{R}^n$ we have

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

Therefore,

$$\overline{L}(x(t)) = \max_{1\le l\le k}\left\{L_\varepsilon^{w_l(t+1)}(x^l(t)) + \frac{2}{\frac{1}{L_\varepsilon^{w_l(t+1)}(x^l(t))} + \frac{1}{L_\varepsilon^{w_l(t+1)}(x^l(t+1))}}\right\} \le \frac{m}{\varepsilon} + \frac{2}{\frac{2\varepsilon}{m}} = \frac{2m}{\varepsilon},$$

this proves the desired result. □

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 ε-KPALM, then there exists $\rho_2 > 0$ and $\gamma(t+1) \in \partial\Psi_\varepsilon(z(t+1))$ such that*

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

Proof. Repeating the steps of the proof in the case of KPALM yields that

$$\gamma(t+1) := \left(\left(d_\varepsilon^i(x(t+1)) + u^i(t+1)\right)_{i=1,\ldots,m}, \nabla_x H_\varepsilon(w(t+1),x(t+1))\right) \in \partial\Psi_\varepsilon(z(t+1)), \tag{4.10}$$

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

$$d_\varepsilon^i(x(t)) + \alpha_i(t)\left(w^i(t+1)-w^i(t)\right) + u^i(t+1) = \mathbf{0}. \tag{4.11}$$

Plugging (4.11) into (4.10), and taking norm yields

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

You can say that with $\sqrt{m}$!

where the last inequality follows from Lemma 4.1.1 and the fact that $\overline{\alpha}\cancel{(t)} = \max_{1\le i\le m}\overline{\alpha_i}\cancel{(t)}$.
Next we bound $\|\nabla_x H_\varepsilon(w(t+1),x(t+1))\| \le c\|x(t+1)-x(t)\|$, for some constant $c>0$. Indeed, for all $l=1,2,\ldots,k$, we have

$$\|\nabla_x H_\varepsilon(w(t+1),x(t+1))\| \le \sum_{l=1}^k \|\overset{\nabla_{x^l}H_\varepsilon(w(t+1),x(t+1))}{\cancel{\nabla H_\varepsilon^{w_l(t+1)}(x^l(t+1))}}\|$$

$$\le \sum_{l=1}^k \|\cancel{\nabla H_\varepsilon^{w_l(t+1)}(x^l(t))}\| + \sum_{l=1}^k \|\overset{\nabla_{x^l}H_\varepsilon(w(t+1),x(t+1))}{\cancel{\nabla H_\varepsilon^{w_l(t+1)}(x^l(t+1))}} - \overset{\nabla_{x^l}H_\varepsilon(w(t+1),x(t))}{\cancel{\nabla H_\varepsilon^{w_l(t+1)}(x^l(t))}}\|. \tag{4.12}$$

[illegible]

$L_\varepsilon^l(\ )\,\|x^l(t+1)-x^l(t)\|$

~~From (4.3) we have~~

see Page 18 (*)

$$\cancel{\nabla H_\varepsilon^{w_l(t+1)}(x^l(t)) = L_\varepsilon^{w_l(t+1)}(x^l(t))\left(x^l(t+1)-x^l(t)\right), \quad \forall\, 1\le l\le k,}$$

~~applying Lemma 4.0.5 with respect to $H_\varepsilon^{w_l(t+1)}(\cdot)$ and plugging into (4.12) yields~~

$$\cancel{\|\nabla_x H(w(t+1),x(t+1))\| \le}$$

$$\cancel{\le \sum_{l=1}^k \left(L_\varepsilon^{w_l(t+1)}(x^l(t)) + \frac{2L_\varepsilon^{w_l(t+1)}(x^l(t))L_\varepsilon^{w_l(t+1)}(x^l(t+1))}{L_\varepsilon^{w_l(t+1)}(x^l(t)) + L_\varepsilon^{w_l(t+1)}(x^l(t+1))}\right)\|x^l(t+1)-x^l(t)\|.}$$

Therefore, ~~denote $\overline{L}(x(t)) = \max_{1\le l\le k}\left\{L_\varepsilon^{w_l(t+1)}(x^l(t)) + \frac{2L_\varepsilon^{w_l(t+1)}(x^l(t))L_\varepsilon^{w_l(t+1)}(x^l(t+1))}{L_\varepsilon^{w_l(t+1)}(x^l(t))+L_\varepsilon^{w_l(t+1)}(x^l(t+1))}\right\}$, and~~ setting $\rho_2 = \sqrt{m}\left(\frac{d_A}{\varepsilon} + \overline{\alpha}\cancel{(t)}\right) + \overset{\frac{2m\sqrt{k}}{\varepsilon}}{\cancel{k\overline{L}(x(t))}}$, ~~note that Lemma 4.1.2 together with Assumption 1(i) imply that ρ_2 is bounded from above, and~~ yields the result follows. □

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

where the last equality follows from (4.3). Therefore

The following lemma shows that the smoothed function indeed $H_\varepsilon(w,x)$ approximates $H(w,x)$.

Lemma 4.2.1 (Closeness of smooth). *For any* $(w,x) \in \Delta^m \times \mathbb{R}^{nk}$ *and* $\varepsilon > 0$ *the following inequalities hold true*

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

Proof. Applying the inequality

$$(a+b)^\lambda \leq a^\lambda + b^\lambda, \quad \forall\, a,b \geq 0,\ \lambda \in (0,1],$$

with $a = \|x^l - a^i\|^2$, $b = \varepsilon^2$ and $\lambda = \frac{1}{2}$, yields

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

Together with the fact that

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

yields the following inequality

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

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

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

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

5 Returning to KMENAS

5.1 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

$$v^i(t+1) = \arg\min_{v^i \in \Delta} \left\{ \langle v^i, d^i(y(t)) \rangle \right\}. \tag{5.1}$$

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

$$y^l(t+1) = \frac{\sum_{i=1}^m v_l^i(t+1) a^i}{\sum_{i=1}^m v_l^i(t+1)}. \tag{5.2}$$

The KMEANS algorithm obviously resemble KPALM algorithm. Denote $\overline{\alpha}(t) = \max_{1 \leq i \leq m} \alpha_i(t)$. 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.

5.2 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

$$C^l(t+1) = \left\{ a \in \mathcal{A} \mid \|a - x^l(t)\| \leq \|a - x^j(t)\|, \quad \forall 1 \leq l \leq k \right\}. \tag{5.3}$$

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

$$x^l(t+1) = mean(C^l(t)) := \frac{1}{|C^l(t)|} \sum_{a \in C^l(t)} a. \tag{5.4}$$

(2.3) Stopping criteria: Halt if

$$\forall 1 \leq l \leq k \quad C^l(t+1) = C^l(t) \tag{5.5}$$

As in KPALM, KMEANS needs Assumption 1(ii) for step (5.4) 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 $R_{lj}(t) = \min\limits_{a \in C^l(t)} \{\|a - x^j(t)\| - \|a - x^l(t)\|\}$ for all $1 \leq l, j \leq k$, and $r(t) = \min\limits_{l \neq j} R_{lj}$.

Due to Assumption 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 5.1. *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 (5.5) we have $x(t) = x(t-1)$, thus $x^l(t)$ is the closest center to a among the centers of $x(t)$. Further we have

$$r(t) \leq \|x^j(t) - a\| - \|x^l(t) - a\| \quad \forall j \neq l. \tag{5.6}$$

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 (5.6), and we get that $C^l(t) \subseteq \underline{C^l}(x)$. By definition of $\underline{C^l}(x)$ we have that for any $l \neq j$, $\underline{C^l}(x) \cap \underline{C^j}(x) = \emptyset$, and for all $1 \leq l \leq k$, $\underline{C^l}(x) \subseteq \mathcal{A}$. Now, since $C(t)$ is a partition of $\mathcal{A}$, then $C^l(t) = \underline{C^l}(x)$ for all $1 \leq l \leq k$. □

Proposition 5.2 (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 5.1.

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

where $d_{\mathcal{A}}$ is the diameter of $\mathcal{A}$ and β is given in (3.9).

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

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

Proof.

Now we prove the following result.

Proposition. Let $\{z(t)\}_{t \in \mathbb{N}}$ be a sequence generated by ε-KPALM. Then, ~~the following assertions hold true.~~ for

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

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

~~(2) For all $t \in \mathbb{N}$ we have~~ Proof. Please prove it!

(*)
$$\le \frac{m}{\varepsilon} \sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\| + \sum_{l=1}^{k} \gamma^l(t) \|x^l(t+1) - x^l(t)\|, \qquad (1)$$

where the last inequality follows from Proposition 1(ii) and Lemma 4.0.5 where

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

From Proposition 1(ii) we obtain that

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

Hence, from (1), we have

$$\|\nabla_x H_{\varepsilon}(w(t+1), x(t+1))\| \le \frac{2m}{\varepsilon} \sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\| \le \frac{2m\sqrt{k}}{\varepsilon} \|x(t+1) - x(t)\|.$$