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

we center $x^l \in \mathbb{R}^n$, and are interested ~~want~~ to determine k cluster centers $\{x^1, x^2, \ldots, x^k\}$ such that

delete 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. 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

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

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

2 Problem Reformulation and Notations

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

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

where Δ denotes the well-known simplex defined by

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

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

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

where

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

Replacing further the constraint $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$, which defined to be 0 in Δ and ∞ otherwise, to the objective function, results in a equivalent formulation

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

where $w = (w^1, w^2, \ldots, w^m) \in \mathbb{R}^{km}$. Finally, for the simplicity of the yet to come expositions, we define the following functions

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

H^i

Clustering: The Squared Euclidean Norm Case

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 the PALM Theory

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

You can fill this part!

3.2 Clustering with PALM ~~for Squared Euclidean Norm Distance Function~~

delete

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.

As in the 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 first subproblem with proximal term as follows~~/~~

delete

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

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

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

written explicitly

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

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

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

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

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.

$\square$

The following assumption will be crucial for the coming analysis.

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

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

$:= \beta(w(t))$

It should be noted that Assumption 1(i) is very mild since the parameters $\alpha_i(t)$, $1 \leq i \leq m$ 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 $\beta(w)$* $\cancel{:= 2 \min_{1\leq l\leq k} \left\{ \sum_{i=1}^{m} w_l^i \right\}}$ which defines in (3.9), *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. Indeed, the Hessian is given by

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

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

Now we are ready to prove the ~~decrease~~ descent 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), see also (3.1), we derive , for each $i = 1, 2, \ldots, m$, the following inequality

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

H^i H^i

Hence, we obtain

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

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

H^i

(נוסח 1)

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

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

From Assumption 1(ii) we have that $\beta(w(t)) = 2\min_{1\le l\le k}\left\{\sum_{i=1}^{m} w_l^i(t)\right\} \ge \underline{\beta}$, 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

(נוסח 2)

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

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~~ by combining (נוסח 1) and (נוסח 2), we ~~have~~ get

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

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

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*

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

where $M = \max_{1\le i\le m}\|a^i\|$.

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

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

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

$$
\begin{aligned}
&\partial\Psi(z(t+1)) = \\
&\quad = \left((\nabla_{w^i} H_i(w(t+1), x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1)))_{i=1,2,\ldots,m}, \nabla_x H(w(t+1), x(t+1)) \right) \\
&\quad = \left((d^i(x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1)))_{i=1,2,\ldots,m}, \nabla_x H(w(t+1), x(t+1)) \right) \\
\text{(3.2)} &\quad = \left((d^i(x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1)))_{i=1,2,\ldots,m}, \mathbf{0} \right), \qquad \text{(3.new)}
\end{aligned}
$$

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

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

and from (3.new) it follows that $\gamma(t+1) \in \partial\Psi(z(t+1))$

by defining $\overline{\alpha} = \max_{1 \le i \le m} \overline{\alpha_i}$, we obtain

Hence,

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

delete

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

4 Clustering ~~via Alternation with Weiszfeld Step~~

in the Euclidean Norm Case

4.1 ~~Algorithm to the~~ Smoothed Clustering Problem

A

have formulated

In the previous section we ~~showed that~~ Problem (2.1) ~~has~~ the following equivalent form

[redacted] in

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

where , in this setting, the involved functions are

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

and

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

delete

have used

fact that the [redacted]

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

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

where

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

(4 zoom)

and

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

$\odot$

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

$$= \sum_{l=1}^{k} H_\varepsilon^l(w,x)$$

as Ψ and therefore we apply the same step as in KPALM. More precisely, for all $i = 1, 2, \ldots, m$, we have

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

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. Here we follow exactly the idea of PALM, that is,

Linearizing the function and adding regularizing term

$$x^l(t+1) = \operatorname*{argmin}_{x^l} \left\{ \left\langle x^l - x^l(t), \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \right\rangle + \frac{L_\varepsilon^l(w(t+1), x(t))}{2} \|x^l - x^l(t)\|^2 \right\},$$

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.

delete

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

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

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

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

Please change: $b^i \longrightarrow a^i$ all over the section!

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 function $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ by

$w_i \xrightarrow{\text{ch.}} v_i$

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

for fixed positive numbers $\overset{v_1}{\cancel{w_1}}, \overset{v_2}{\cancel{w_2}}, \ldots, \overset{v_m}{\cancel{w_m}} \in \mathbb{R}$ 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

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

Finally we introduce the following operator, $L_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ defined by

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

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 $x, y \in \mathbb{R}^n$,

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

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

(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 $\overset{v_i}{\cancel{w_i}}$ and summing over $i = 1, 2, \ldots, m$, the results follows.

(iii) 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.~~ following identity

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

Using the first two items and the fact that $\nabla_x h_\varepsilon(y,y) = 2\nabla f_\varepsilon(y)$ yields the desired result.

□

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

Lemma 4.0.2. *For all $y, z \in \mathbb{R}^n$ the following statement holds true*

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

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 Lemma 4.0.1(i) we can write

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

On the other hand, from Lemma 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 $x = y - (1/L_\varepsilon(y))\nabla \widetilde{f}_\varepsilon(y)$ 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

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

This proves the desired result

Combining the 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 $z, y \in \mathbb{R}^n$. $\square$

Now we get back to ε-KPALM algorithm and prove few technical results about the involved functions which are based on the auxiliary results obtained above.

Proposition 4.1 (Bounds for L_ε^l). *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM. Then, the following two statements hold true.*

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

$$L_\varepsilon^l(w(t+1), x(t)) \geq \frac{\underline{\beta}}{\left(d_{\mathcal{A}}{}^2 + \varepsilon^2\right)^{1/2}},$$

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

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

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

Fact

Proof. (i) From Assumption 1(ii) and the ~~face~~ that $x^l(t) \in Conv(\mathcal{A})$ for all $1 \leq l \leq k$, it follows that

$$L_\varepsilon^l(w(t+1), x(t)) = \sum_{i=1}^m \frac{w_l^i(t+1)}{\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2}} \geq \frac{\sum_{i=1}^m w_l^i(t+1)}{\left(d_{\mathcal{A}}{}^2 + \varepsilon^2\right)^{1/2}} \geq \frac{\underline{\beta}}{\left(d_{\mathcal{A}}{}^2 + \varepsilon^2\right)^{1/2}},$$

as asserted.

(ii) Since $w(t+1) \in \Delta^m$ we have

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

as asserted. $\square$

Now we prove the following result.

By definition (see (4.30N)) we have, [redacted] for $l=1,2,\dots,k$, that

$$H_\varepsilon^l(\overset{w(t+1)}{[redacted]}, \overset{x}{[redacted]}) = \cancel{\sum_{i=1}^{m} w_l^i(t)\left(\|x^l(t)-a^i\|^2+\varepsilon^2\right)^{1/2}} = f_\varepsilon(\overset{x^l}{[redacted]}),$$

where $v_i = w_l^i(t+1)$, $i=1,2,\dots,m$. Therefore, by applying Lemma 4.0.1(iii) [redacted] with $x = x^l(t+1)$ and $y = x^l(t)$, we get

Proposition 4.2. *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM. Then, for all $t \in \mathbb{N}$ we have*

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

Proof. ~~Applying Lemma 4.0.1(iii) over~~ $\cancel{x = x^l(t+1)}$ ~~and~~ $\cancel{y = x^l(t)}$~~, with~~ $\cancel{(w_1, w_2, \dots, w_m) = (w_l^1(t+1), w_l^2(t+1), \dots, w_l^m(t+1))}$ ~~and~~ $\cancel{(b^1, b^2, \dots, b^m) = (a^1, a^2, \dots, a^m)}$ ~~yields~~

$$h_\varepsilon\cancel{(x^l(t+1), x^l(t)) = h_\varepsilon(x^l(t), x^l(t)) + \left\langle \nabla_x h_\varepsilon(x^l(t), x^l(t)), x^l(t+1) - x^l(t) \right\rangle + L_\varepsilon^l(x^l(t))\|x^l(t+1) - x^l(t)\|^2.} \tag{4.7}$$

~~From Lemma 4.0.1(i) we have~~

$$h_\varepsilon\cancel{(x^l(t), x^l(t)) = f_\varepsilon(x^l(t)) = \sum_{i=1}^{m} w_l^i(t+1)\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2}.}$$

~~Using the fact that~~ $\cancel{\nabla_x h_\varepsilon(y,y) = 2 f_\varepsilon(y)}$~~, for all~~ $\cancel{y \in \mathbb{R}^n}$~~, yields~~

$$\cancel{\nabla_x h_\varepsilon(x^l(t), x^l(t)) = 2\nabla f_\varepsilon(x^l(t)) = 2\sum_{i=1}^{m} \frac{w_l^i(t+1)(x^l(t) - a^i)}{\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2}} = 2\nabla_{x^l} H_\varepsilon(w(t+1), x(t)).}$$

~~Plugging last two identities into (4.7) yields~~

$$h_\varepsilon\cancel{(x^l(t+1), x^l(t)) = \sum_{i=1}^{m} w_l^i(t+1)\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2} + 2\left\langle \nabla_{x^l} H_\varepsilon(w(t+1), x(t)), x^l(t+1) - x^l(t) \right\rangle + L_\varepsilon^l(x^l(t))\|x^l(t+1) - x^l(t)\|^2.} \tag{4.8}$$

~~From Lemma 4.0.1(ii) we obtain~~

$$\cancel{h_\varepsilon(x^l(t+1), x^l(t)) \ge 2 f_\varepsilon(x^l(t+1)) - f_\varepsilon(x^l(t)).} \tag{4.9}$$

~~Combining (4.8) with (4.9), dividing by 2 and rearranging the terms yields~~

$$\cancel{\sum_{i=1}^{m} w_l^i(t+1)\left(\|x^l(t+1) - a^i\|^2 + \varepsilon^2\right)^{1/2} \le \sum_{i=1}^{m} w_l^i(t+1)\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2}} + \left\langle \cancel{\nabla_{x^l} H_\varepsilon(w(t+1), x(t)), x^l(t+1) - x^l(t)} \right\rangle + \frac{L_\varepsilon^l(x^l(t))}{2}\cancel{\|x^l(t+1) - x^l(t)\|^2}.$$

Summing the last inequality over $l = 1, 2, \dots, k$, yields

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

$$H_\varepsilon(w(t+1), x(t+1)) \le H_\varepsilon(x(t)) + \sum_{l=1}^{k} \frac{\cancel{L^l(x(t))}}{2}\|x^l(t+1) - x^l(t)\|^2 + \sum_{l=1}^{k} \left\langle \nabla_{x^l} H_\varepsilon(w(t+1), x(t)), x^l(t+1) - x^l(t) \right\rangle.$$

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

Replacing the last term with the following compact form

$$\sum_{l=1}^{k} \left\langle \nabla_{x^l} H_\varepsilon(w(t+1), x(t)), x^l(t+1) - x^l(t) \right\rangle = \left\langle \nabla_x H_\varepsilon(w(t+1), x(t)), x(t+1) - x(t) \right\rangle,$$

and the result follows. □

Now we are finally ready to prove the two properties that needed for guaranteeing that the sequence that is generated by ε-KPALM converges to critical point of Ψ_ε.

Proposition 4.3 (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*

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

(?) *Proof.* As we already mentioned the steps of KPALM and ε-KPALM with respect to w are similar and therefore following the same arguments given at the beginning of the proof of Proposition 3.2 we have that

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

where $\underline{\alpha} = \min\limits_{1\leq i\leq m} \alpha_i$.

Applying Proposition 4.2 we get for all $t \in \mathbb{N}$ that

$$\begin{aligned} H_\varepsilon(w(t+1), x(t)) - H_\varepsilon(w(t+1), x(t+1)) &\geq \sum_{l=1}^{k} \frac{L_\varepsilon^l(w(t+1), x(t))}{2}\|x^l(t+1) - x^l(t)\|^2 \\ &\geq \frac{\underline{\beta}}{\left({d_\mathcal{A}}^2 + \varepsilon^2\right)^{1/2}} \sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\|^2 \\ &\geq \frac{\underline{\beta}}{\left({d_\mathcal{A}}^2 + \varepsilon^2\right)^{1/2}} \|x(t+1) - x(t)\|^2, \end{aligned} \tag{4.11}$$

where the ~~last~~ second inequality follows from Proposition 4.1(i).

Set $\rho_1 = \frac{1}{2}\min\left\{\underline{\alpha}, \underline{\beta}/\left(d_\mathcal{A}^2 + \varepsilon^2\right)^{1/2}\right\}$. Summing (4.10) and (4.11) yields

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

where the last equality follows from the fact that $G(w(t)) = 0$, for all $t \in \mathbb{N}$. This proves the desired result. □

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

Lemma 4.3.1. *For any $x, y \in \mathbb{R}^{nk}$ such that $x^l, y^l \in Conv(\mathcal{A})$ for all $1 \leq l \leq k$ the following inequality holds*

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

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

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

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

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

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

Hence,

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

as asserted. $\square$

Proposition 4.4 (Subgradient lower bound for the iterates gap). *Let $\{z(t)\}_{t \in \mathbb{N}}$ 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*

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

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

$$\gamma(t+1) := \left(\left(d_\varepsilon^i(x(t+1)) + u^i(t+1) \right)_{i=1,\ldots,m}, \nabla_x H_\varepsilon(w(t+1), x(t+1)) \right) \in \partial \Psi_\varepsilon(z(t+1)), \quad (4.12)$$

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

$$d_\varepsilon^i(x(t)) + \alpha_i(t) \left(w^i(t+1) - w^i(t) \right) + u^i(t+1) = \mathbf{0}. \quad (4.13)$$

Plugging (4.13) into (4.12), and taking norm yields

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

where the last inequality follows from Lemma 4.3.1 and the fact that $\overline{\alpha} = \max_{1 \le i \le m} \overline{\alpha_i}$.
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

$$\begin{aligned} \nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) &= \nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\ &\quad + \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\ &= \nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\ &\quad + L_\varepsilon^l(w(t+1), x(t)) \left(x^l(t) - x^l(t+1)\right), \end{aligned} \tag{4.14}$$

where the last equality follows from (4.3). Therefore,

$$\begin{aligned} \|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| &\le \sum_{l=1}^{k} \|\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1))\| \qquad L_\varepsilon^l(w(t+1), x(t)) \\ &\le \sum_{l=1}^{k} \cancel{L_\varepsilon^l} \|x^l(t+1) - x^l(t)\| \\ &\quad + \sum_{l=1}^{k} \|\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t))\| \\ &\le \frac{m}{\varepsilon} \sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\| + \sum_{l=1}^{k} \gamma^l(t) \|x^l(t+1) - x^l(t)\|, \end{aligned} \tag{4.15}$$

where the last inequality follows from Proposition 4.1(ii) and Lemma 4.0.2 where

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

From Proposition 4.1(ii) we obtain that

(4.15)

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

Hence, from ~~4~~.15, we have

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

Therefore, setting $\rho_2 = \sqrt{m}\left(\frac{d_A}{\varepsilon} + \overline{\alpha}\right) + \frac{2m\sqrt{k}}{\varepsilon}$, yields and the result. $\square$

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

Lemma 4.4.1 (Closeness of smooth). *For any $(w, x) \in \Delta^m \times \mathbb{R}^{nk}$ and $\varepsilon > 0$ the following inequalities hold true*

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

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. $\square$

5 Returning to KMEANS

relation

5.1 Similarity to KMEANS

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

> **KMEANS**
>
> (1) Initialization: $x(0) \in \mathbb{R}^{nk}$.
>
> (2) General step ($t = 0, 1, \ldots$):
>
> (2.1) Cluster assignment: for $i = 1, 2, \ldots, m$ compute
>
> $$w^i(t+1) = \arg\min_{w^i \in \Delta} \left\{ \langle w^i, d^i(x(t)) \rangle \right\}. \tag{5.1}$$
>
> (2.2) Centers update: for $l = 1, 2, \ldots, k$ compute
>
> $$x^l(t+1) = \frac{\sum_{i=1}^m w_l^i(t+1) a^i}{\sum_{i=1}^m w_l^i(t+1)}. \tag{5.2}$$
>
> ~~(2.3) Stopping criteria: halt if~~
>
> $$\cancel{\forall 1 \le l \le k \quad C^l(t+1) = C^l(t)} \tag{5.3}$$

delete

easy to see that

It is ~~easily noted that~~ if we take $\alpha_i(t) = 0$ for all $1 \le i \le m$ and $t \in \mathbb{N}$, then KPALM becomes KMEANS.

Lemma 5.0.2. *Let $\{z(t)\}_{t \in \mathbb{N}}$ be the sequence generated by KMEANS. Then, there exists $c > 0$ such that*

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

Proof. At each iteration KMEANS partitions the set $\mathcal{A}$ into k clusters, and the center of each cluster is its mean. Since the number of these partitions if finite, then there exists a finite set $\mathcal{C} = \left\{x^1, x^2, \ldots, x^N\right\} \subset \mathbb{R}^{nk}$ such that for all $t \in \mathbb{N}$, $x(t) \in \mathcal{C}$. We denote

$$r = \min_{1 \le j < l \le N} \|x^j - x^l\|,$$

and set $c = \sqrt{2}/r$. At each iteration, the point a^i can move from one cluster to another, hence

$$\|w^i(t+1) - w^i(t)\| \le \sqrt{2}.$$

Therefore, combining these arguments yields

$$\frac{\|w^i(t+1 - w^i(t)\|}{\|x(t+1) - x(t)\|} \le \frac{\sqrt{2}}{r}.$$

I don't understand this proof! What is r? Shouldn't be depend on t? This is true that we need an inequality as you wrote in the lemma but I am not sure if your proof can be fixed

In case that $x(t+1) = x(t)$, this implies that none of the clusters has changed, hence we proved the statement in both cases. $\square$

The next part is written in an informal manner.

I believe that with Lemma 5.0.2 we can show the usual sufficient decrease property of PALM in the following manner. As we discussed the following inequality still holds in the case of $\alpha = 0$

$$\frac{\beta}{2}\|x(t+1) - x(t)\|^2 \leq H(w(t), x(t)) - H(w(t+1), x(t+1)). \tag{5.4}$$

Setting $\rho_1 = \underline{\beta}/2(1 + mc^2)$, where c is as in Lemma 5.0.2, and we can write

$$\begin{aligned}
\rho_1\|z(t+1) - z(t)\|^2 &= \rho_1 \sum_{i=1}^{m} \|w^i(t+1) - w^i(t)\|^2 + \rho_1\|x(t+1) - x(t)\|^2 \\
&\leq \rho_1(1 + mc^2)\|x(t+1) - x(t)\|^2 \\
&\leq H(w(t), x(t)) - H(w(t+1), x(t+1)) \\
&= \Psi(z(t)) - \Psi(z(t+1))
\end{aligned}$$

where the first inequality follows from Lemma 5.0.2, the second follows from (5.4), and the last equality follows from the fact that $G(w(t)) = 0$, for all $t \in \mathbb{N}$. I think this proves the sufficient decrease part.

I reviewed the subgradient lower bound for iterates gap proof, and I think it follows through as is, where in all places that refer to α we set 0.

You need to write everything formally as you did before!