Introduction

1 ~~The Clustering Problem~~

Should be filled with a lot of "bla-bla"!

Let $\mathcal{A} = \{a^1, a^2, \dots, a^m\}$ 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 $\{C^1, C^2, \dots, C^k\}$, called clusters. For each $l = 1, 2, \dots, k$, the cluster C^l is represented by its center $x^l \in \mathbb{R}^n$, and we are interested to determine k cluster centers $\{x^1, x^2, \dots, x^k\}$ such that the sum of certain proximity measures from each point a^i, $i = 1, 2, \dots, m$, to a nearest cluster center x^l is minimized. We define the vector of all centers by $x = (x^1, x^2, \dots, x^k) \in \mathbb{R}^{nk}$. set data

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~~ Mathematical Tools

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:

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

where Δ denotes the well-known simplex defined by

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

Using this fact in Problem (1.1) and introducing new variables $w^i \in \mathbb{R}^k$, $i = 1, 2, \dots, m$, 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

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

Replacing further the constraint $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$, which is 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 $w = (w^1, w^2, \dots, w^m) \in \mathbb{R}^{km}$. Finally, for the simplicity of the yet to come expositions, we define the following functions

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

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

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

Should be moved to Section 2

3 Clustering: The Squared Euclidean Norm Case

3.1 Introduction to the PALM Theory

whole this subsection should be rewritten. You say many times PALM but don't present it!!!

In this subsection we give a brief review of the problem structure that PALM theory treats, the requirements needed and the results it assures. The PALM algorithm solves the nonconvex and nonsmooth minimization problem of the following form

$$(M) \quad \text{minimize } \Psi(x,y) := f(x) + g(y) + H(x,y) \text{ over all } (x,y) \in \mathbb{R}^n \times \mathbb{R}^m,$$

where $f : \mathbb{R}^n \to (-\infty, +\infty]$ and $g : \mathbb{R}^n \to (-\infty, +\infty]$ are proper and lower semicontinuous functions while $H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a C^1 function.

Suppose that we are given a generic algorithm $\mathcal{A}$ which solves problem (M) and generates a sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ via the following:

$$z^0 \in \mathbb{R}^d,\ z^{k+1} \in \mathcal{A}\left(z^k\right), \quad k = 0, 1, \ldots.$$

delete

There are three basic requirements necessary ~~for PALM~~ to assure the convergence of the whole sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ to a critical point of Ψ.

(1) Sufficient decrease property: There exists a positive constant ρ_1, such that

$$\rho_1\|z^{k+1} - z^k\|^2 \le \Psi(z^k) - \Psi(z^{k+1}), \quad \forall\, k = 0, 1, \ldots.$$

(2) A subgradient lower bound for iterates gap: Assuming that the sequence generated by the algorithm $\mathcal{A}$ is bounded. There exists a positive constant ρ_2, such that

$$\|w^{k+1}\| \le \rho_2\|z^{k+1} - z^k\|, \quad w^k \in \partial\Psi\left(z^k\right) \quad \forall\, k = 0, 1, \ldots.$$

(3) The function Ψ is a KL function.

With regard to the last item, see the definition of KL property in [BST2014]. The objective functions to be discussed in the current work are all KL functions, since it is either a weighed sum of squared Euclidean norms, or a weighed sum of Euclidean norms.

for which the proximity function $d(\cdot,\cdot)$ is chosen to be

3.2 Clustering with PALM

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 in the previous subsection. Since the clustering problem has a specific structure, we are ought to exploit it in the following manner.

the framework which
was discussed in Section 3.1

(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 old
approach of

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

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

where $\alpha_i(t) > 0$ for all $i = 1, 2, \ldots, m$. On the other hand, with respect to x we perform exact minimization

, that is,

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

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

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

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

delete

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

(where we use P_Δ for the orthogonal projection onto the set Δ).
Thus, we can present now the KPALM algorithm.

~~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

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

delete

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

Proposition 3.1 (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 $\mathbb{R}^{km} \times \mathbb{R}^{nk}$.

(3.6)

Let $1 \le l \le k$, we set

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

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

which proves that ~~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 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

$$2 \min_{1\leq l\leq k} \sum_{i=1}^{m} w_l^i(t) := \beta(w(t)) \geq \underline{\beta}. \tag{3.9}$$

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.~~ play any rule in the solution process which is, of course, meaningless situation.

Lemma 3.1.1 (Strong convexity of $H(w, x)$ in x). *The function $x \mapsto H(w, x)$ is strongly convex with parameter $\beta(w)$ which defined in (3.9), whenever $\beta(w) > 0$.*

Proof. Since the function $x \mapsto H(w, x) = \sum_{l=1}^{k}\sum_{i=1}^{m} w_l^i \|x^l - a^i\|^2$ 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. □

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

Proposition 3.2 (Sufficient decrease property). Let *~~Suppose that Assumption 1 holds true and let~~ $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM. ~~Then,~~* Suppose that Assumption 1 holds true, 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 step (3.5), see also (3.1), we derive, for each $i = 1, 2, \ldots, m$, the following inequality

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

Now we are ready to prove global convergence of the sequence $\{z(t)\}_{t\in\mathbb{N}}$ generated by KPALM to a critical point of Ψ given in (2.3). We will follow here the general procedure which was discussed in Section 3.1. Therefore, we need to prove that $\{z(t)\}$ satisfies the conditions (C1) and (C2). We begin by proving condition (C1).

for all $t \in \mathbb{N}$, that

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

$$\begin{aligned}
\frac{\underline{\alpha}}{2}\|w(t+1) - w(t)\|^2 &= \frac{\underline{\alpha}}{2}\sum_{i=1}^{m} \|w^i(t+1) - w^i(t)\|^2 \\
&\leq \sum_{i=1}^{m} \frac{\alpha_i(t)}{2}\|w^i(t+1) - w^i(t)\|^2 \\
&\leq \sum_{i=1}^{m} \left[H^i(w(t), x(t)) - H^i(w(t+1), x(t))\right] \\
&= H(w(t), x(t)) - H(w(t+1), x(t)),
\end{aligned} \tag{3.11}$$

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

and the definition of $\underline{\alpha}$.

From Assumption 1(ii) we have that $\beta(w(t)) \geq \underline{\beta}$, and from Lemma 3.1.1 it follows that the function $x \mapsto H(w(t), x)$ is strongly convex with parameter $\beta(w(t))$, ~~hence it follows that~~

for all $t \in \mathbb{N}$,

. Hence, using the strong convexity, yields that

$$\begin{aligned}
&H(w(t+1), x(t)) - H(w(t+1), x(t+1)) \geq \\
&\quad \geq \langle \nabla_x H(w(t+1), x(t+1)), x(t) - x(t+1)\rangle + \frac{\beta(w(t))}{2}\|x(t) - x(t+1)\|^2 \\
&\quad = \frac{\beta(w(t))}{2}\|x(t+1) - x(t)\|^2 \\
&\quad \geq \frac{\underline{\beta}}{2}\|x(t+1) - x(t)\|^2,
\end{aligned} \tag{3.12}$$

where the equality follows from (3.2), since $\nabla_x H(w(t+1), x(t+1)) = 0$. Set $\rho_1 = \frac{1}{2}\min\{\underline{\alpha}, \underline{\beta}\}$, by combining (3.11) and (3.12), we get

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

where the last equality follows from the fact that $G(w(t)) = 0$, since $w(t) \in \Delta^m$ 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.~~

are focusing on proving condition (C2). To this end we will need the following technical result.

Lemma 3.2.1. *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 $M = \max_{1 \leq i \leq m} \|a^i\|$.

. Then [redacted] we have a Lipschitz property of $d^i(\cdot)$ for all $t \in \mathbb{N}$, given by

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

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

(see Proposition 3.1(i)), This

where the last inequality follows from the fact that $x^l(t) \in Conv(\mathcal{A})$ and hence $\|x^l(t)\| \leq M$ for all $t \in \mathbb{N}$ and $l = 1, 2, \ldots, k$, ~~this~~ proves the desired result. □

Proposition 3.3 (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*

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

Now, using this result we can show that $\{z(t)\}_{t\in\mathbb{N}}$ satisfies condition (C2):

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

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

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

$$
\begin{aligned}
&\partial\Psi(z(t+1)) = \\
&\quad = \left(\left(\nabla_{w^i} H^i(w(t+1), x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\right)_{i=1,2,\ldots,m}, \nabla_x H(w(t+1), x(t+1)) \right) \\
&\quad = \left(\left(d^i(x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\right)_{i=1,2,\ldots,m}, \nabla_x H(w(t+1), x(t+1)) \right) \\
&\quad = \left(\left(d^i(x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1))\right)_{i=1,2,\ldots,m}, \mathbf{0} \right),
\end{aligned}
\tag{3.13}
$$

where the last equality follows from (3.2), 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

$$d^i(x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1) = \mathbf{0}. \tag{3.14}$$

, it follows from (3.13)

Setting $\gamma(t+1) := \left(\left(d^i(x(t+1)) + u^i(t+1)\right)_{i=1,2,\ldots,m}, \mathbf{0}\right)$ ~~and from (3.13) it follows~~ that $\gamma(t+1) \in \partial\Psi(z(t+1))$. Using (3.14) we obtain

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

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

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

where the third inequality follows from Lemma 3.2.1. Define $\rho_2 = 4Mm + \overline{\alpha}\sqrt{m}$, and the result follows. $\square$

4 Clustering: The Euclidean Norm Case

4.1 A Smoothed Clustering Problem

In the previous section we have formulated the clustering problem in 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, in this setting, the involved functions are

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

, in Section 3,

In order to be able to use the theory mentioned in Section ~~3~~.1 2.4, we have used the fact that the coupled function $H(w,x)$ 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

varepsilon

$$H_\varepsilon(w,x) = \sum_{l=1}^{k} H_\varepsilon^l(w,x) = \sum_{l=1}^{k}\sum_{i=1}^{m} w_l^i\left(\|x^l - a^i\|^2 + \varepsilon^2\right)^{1/2}, \tag{4.2}$$

and for all $i = 1, 2, \ldots, m$,

$$d_\varepsilon^i(x) = \left(\left(\|x^1 - a^i\|^2 + \varepsilon^2\right)^{1/2}, \left(\|x^2 - a^i\|^2 + \varepsilon^2\right)^{1/2}, \ldots, \left(\|x^k - a^i\|^2 + \varepsilon^2\right)^{1/2}\right) \in \mathbb{R}^k. \tag{4.3}$$

Note that $\Psi_\varepsilon(z)$ is a perturbed form of $\Psi(z)$ for a small $\varepsilon > 0$, and obviously $\Psi_0(z) = \Psi(z)$. The following lemma shows that the smoothed function $H_\varepsilon(w,x)$ indeed approximates $H(w,x)$.

relations hold

Lemma 4.0.1 (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~~ It is clear that for all $\lambda \geq 0$ we have

$$\forall\, \lambda \geq 0, \quad \lambda \leq \sqrt{\lambda^2 + \varepsilon^2} \leq \lambda + \varepsilon.$$

Applying this inequality with $\lambda = \|x^l - a^i\|$, yields

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

for all $l = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, m$. By 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. □

(see Section 2.1)

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 function 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

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

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

a

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

where

$$L_\varepsilon^l(w(t+1), x(t)) := \sum_{i=1}^{m} \frac{w_l^i(t+1)}{\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2}}, \quad \forall\, l = 1, 2, \ldots, k. \tag{4.4}$$

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

$x \to H(w, \cdot)$, for fixed w,

The motivation to use this specific regularizing parameter (see (4.4)) will be discussed later.

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

$$w^i(t+1) = P_\Delta \left(w^i(t) - \frac{d^i_\varepsilon(x(t))}{\alpha_i(t)} \right). \tag{4.5}$$

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

$$x^l(t+1) = x^l(t) - \frac{1}{L^l_\varepsilon(w(t+1), x(t))} \nabla_{x^l} H_\varepsilon(w(t+1), x(t)). \tag{4.6}$$

please modify it as remark

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

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

(see section 2.1)

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

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

for fixed non-negative numbers (not all zero) $v_1, v_2, \ldots, v_m \in \mathbb{R}$ and $a^i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$. We also need the following auxiliary function $h_\varepsilon : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

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

mudulos

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

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

Lemma 4.0.2 (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 $x, y \in \mathbb{R}^n$,

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

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

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

Multiplying the last inequality by v_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 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 get back to the ε-KPALM algorithm and prove few technical results about the involved functions which are based on the ~~auxiliary results~~ properties 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 $d_{\mathcal{A}} = diam(Conv(\mathcal{A}))$ 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

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

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

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

where the first inequality follows from the fact that $\|x^l(t) - a^i\| \le d_\mathcal{A}$, for all $1 \le l \le k$.

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

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

as asserted. □

Now we prove the following result.

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*

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

Proof. By definition (see (4.2)) we have, for $i = 1, 2, \ldots, m$, that

$$H_\varepsilon^l(w(t+1), x(t)) = f_\varepsilon(x^l(t)),$$

where $v_i = w_l^i(t+1)$, $i = 1, 2, \ldots, m$. Therefore, by applying Lemma 4.0.2(iii) with $x = x^l(t+1)$ and $y = x^l(t)$, we get

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

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

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

Replacing the last term with the following compact form

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

and the result follows. □

$\{z(t)\}_{t\in\mathbb{N}}$

Now we are finally ready to prove the two properties needed for guaranteeing that the sequence which is generated by ε-KPALM converges to a 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*

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

delete

Proof. As we already mentioned, the ~~steps~~ with respect to w of KPALM and ε-KPALM are similar in nature and therefore following the same arguments given at the beginning of the proof of Proposition 3.2 we have that

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

where $\underline{\alpha} = \min\limits_{1\leq i\leq m} \alpha_i$. Applying Proposition 4.2 and using ~~with~~ (4.6) we get for all $t \in \mathbb{N}$ that

$$\begin{aligned} H_\varepsilon(w(t+1), x(t)) - H_\varepsilon(w(t+1), x(t+1)) &\geq \sum_{l=1}^{k} \frac{L_\varepsilon^l(w(t+1), x(t))}{2}\|x^l(t+1) - x^l(t)\|^2 \\ &\geq \frac{\underline{\beta}}{\left(d_A{}^2 + \varepsilon^2\right)^{1/2}} \sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\|^2 \\ &\geq \frac{\underline{\beta}}{\left(d_A{}^2 + \varepsilon^2\right)^{1/2}}\|x(t+1) - x(t)\|^2, \end{aligned} \tag{4.8}$$

where the second inequality follows from Proposition 4.1(i). Set $\rho_1 = \frac{1}{2}\min\left\{\underline{\alpha}, \underline{\beta}/\left(d_A^2 + \varepsilon^2\right)^{1/2}\right\}$. Summing (4.7) and (4.8) yields

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

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

The next two lemmas 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 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 function

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

hence,

$$f_\varepsilon(y) = \widetilde{f}_\varepsilon(y) + \langle \nabla f_\varepsilon(z), y \rangle. \tag{4.9}$$

Substituting (4.9) into Lemma 4.0.2(iii) yields

$$\widetilde{f}_\varepsilon(x) \leq \widetilde{f}_\varepsilon(y) + \left\langle \nabla \widetilde{f}_\varepsilon(y), x - y \right\rangle + \frac{L_\varepsilon(y)}{2}\|x-y\|^2. \tag{4.10}$$

It is clear that the optimal point of $\widetilde{f}_\varepsilon$ is z since $\nabla \widetilde{f}_\varepsilon(z) = 0$, therefore using (4.10) 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

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

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 $z, y \in \mathbb{R}^n$. This proves the desired result. □

Lemma 4.3.2. *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 $d_{\mathcal{A}} = diam(Conv(\mathcal{A}))$ and $d_\varepsilon^i(\cdot)$ is defined in (4.3).

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

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

where the last inequality follows from $\|x^l - a^i\|, \|y^l - a^i\| \le d_A$ and the reverse triangle inequality. Therefore,

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

as asserted. □

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*

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

(see Proposition 3.3)

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

$$\gamma(t+1) := \left(\left(d_\varepsilon^i(x(t+1)) + u^i(t+1)\right)_{i=1,\ldots,m}, \nabla_x H_\varepsilon(w(t+1), x(t+1))\right) \in \partial\Psi_\varepsilon(z(t+1)), \quad (4.11)$$

where ~~for all $1 \le i \le m$,~~ $u^i(t+1) \in \partial\delta_\Delta(w^i(t+1))$ ~~such that~~ $i=1,2,\ldots,m$,

Now, writing the optimality condition of step (4.5), yields that

$$d_\varepsilon^i(x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1) = \mathbf{0}. \quad (4.12)$$

Plugging (4.12) into (4.11), and taking the norm yields

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

where the last inequality follows from Lemma 4.3.2 and the fact that $\overline{\alpha} = \max_{1\le i\le m} \overline{\alpha_i}$.
Next we will show that $\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| \le c\|x(t+1) - x(t)\|$, for some constant $c > 0$. Indeed, for all $l = 1, 2, \ldots, k$, we have

$$\begin{aligned}
\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) &= \nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\
&\quad + \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\
&= \nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t)) \\
&\quad + L_\varepsilon^l(w(t+1), x(t))\left(x^l(t) - x^l(t+1)\right), \quad (4.13)
\end{aligned}$$

where the last equality follows from (4.6). Therefore,

$$
\begin{aligned}
\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| &\le \sum_{l=1}^{k} \|\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1))\| \\
&\le \sum_{l=1}^{k} L_\varepsilon^l(w(t+1), x(t))\|x^l(t+1) - x^l(t)\| \\
&\quad + \sum_{l=1}^{k} \|\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t))\| \\
&\le \frac{m}{\varepsilon}\sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\| + \sum_{l=1}^{k} \gamma^l(t)\|x^l(t+1) - x^l(t)\|, \qquad (4.14)
\end{aligned}
$$

where the last inequality follows from Proposition 4.1(ii) and Lemma 4.3.1 using

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

From Proposition 4.1(ii) we obtain that

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

Hence, from (4.14), we have

$$
\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| \le \frac{2m}{\varepsilon}\sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\| \le \frac{2m\sqrt{k}}{\varepsilon}\|x(t+1) - x(t)\|. \qquad (4.15)
$$

Therefore, setting $\rho_2 = \frac{md_A}{\varepsilon} + \overline{\alpha}\sqrt{m} + \frac{2m\sqrt{k}}{\varepsilon}$, yields the result. □

and the unique minimizer is given by $s^* = 1/\sqrt{\lambda}$.

H_ε

4.2 Different Approach Towards Solving the Smoothed $H\varepsilon$

In this section we describe a different approach towards solving the smoothed clustering problem described in (4.1). Using the Arithmetic-Geometric inequality we derive the following simple observation

Delete

$$\frac{1}{2}\min_{s\geq 0}\left\{s\lambda+\frac{1}{\lambda}\right\}\geq \min_{s\geq 0}\left\{\sqrt{s\lambda\cdot\frac{1}{s}}\right\}=\sqrt{\lambda},$$

, $\lambda \geq 0$,

Using this ~~inequality~~ fact we can write

$$\sqrt{\|u\|^2+\varepsilon^2}=\frac{1}{2}\min_{v\geq 0}\left\{v\left(\|u\|^2+\varepsilon^2\right)+\frac{1}{v}\right\}, \tag{4.16}$$

with $v^* = 1/\sqrt{\|u\|^2+\varepsilon^2}$. Thus, instead of solving ~~H_ε directly, defined in (4.2)~~ problem (4.1) with $H_\varepsilon(\cdot,\cdot)$ defined in (4.2), we replace it with the following function

$$B_\varepsilon(v,w,x)=\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{k}\left\{v_l^i w_l^i\left(\left\|x^l-a^i\right\|^2+\varepsilon^2\right)+\frac{w_l^i}{v_l^i}\right\}, \tag{4.17}$$

where $v = (v^1, v^2, \ldots, v^m)$ ~~$\in I^{km}$, where $I := [1/\kappa, 1/\varepsilon]$ and $\kappa = \sqrt{d_A^2+\varepsilon^2}$~~. For all $i = 1, 2, \ldots, m$ we define $b_\varepsilon^i : \mathbb{R}^{mk}\times\mathbb{R}^{nk}\to\mathbb{R}^k$ by

Delete

$$b_\varepsilon^i(v,x)=\left(\frac{1}{2}v_l^i\left(\left\|x^l-a^i\right\|^2+\varepsilon^2\right)+\frac{1}{2v_l^i}\right)_{l=1,2,\ldots,k}\in\mathbb{R}^k,$$

and we have that

$$B_\varepsilon(v,w,x)=\sum_{i=1}^{m}\left\langle w^i, b_\varepsilon^i(v,x)\right\rangle. \tag{4.18}$$

Now the situation is similar to that of ~~the squared Euclidean norm~~ Section 3, namely

(1) The function $w \mapsto B_\varepsilon(v,w,x)$, for fixed v and x, is linear; ~~and therefore there is no need to linearize it as suggested in PALM.~~

(2) The function $x \mapsto B_\varepsilon(v,w,x)$, for fixed v and w, is quadratic and convex; ~~Hence, there is no need to add a proximal term as suggested in PALM.~~

Hence we can tackle these two ~~we can~~ ~~tackle the~~ steps as in KPALM.

Equipped with these observation we proceed to a PALM-like algorithm, which is based on three steps alternating minimization. More precisely, with respect to v we perform exact minimization

$$v(t+1)=\operatorname{argmin}\{B_\varepsilon(v,w(t),x(t))\mid v \geq 0\}. \tag{4.19}$$

It should be noted that this problem can be written equivalently by $v(t+1) = \operatorname{argmin}\{B_\varepsilon(v, w(t), x(t)) : v \in I^{mk}\}$ where $I := [1/\kappa, 1/\varepsilon]$ and $\kappa = \sqrt{d_A^2+\varepsilon^2}$

With respect to w, as in KPALM case, we ~~suggest to regularize~~ need to solve the subproblem ~~with proximal term as follows~~ given by

$$w^i(t+1)=\arg\min_{w^i\in\Delta}\left\{\langle w^i, b_\varepsilon^i(v(t+1),x(t))\rangle+\frac{\alpha_i(t)}{2}\|w^i-w^i(t)\|^2\right\},\quad i=1,2,\ldots,m, \tag{4.20}$$

where $\alpha_i(t) > 0$, ~~for all~~ $i = 1, 2, \ldots, m$. ~~And with~~ With respect to x, again as in KPALM case, we perform exact minimization

$$x(t+1)=\operatorname{argmin}\left\{B_\varepsilon(v(t+1),w(t+1),x)\mid x\in\mathbb{R}^{nk}\right\}. \tag{4.21}$$

, and then problem (4.1) can be written equivalently as

$$\min_{x,v,w}\left\{B_\varepsilon(v,w,x)+G(w)\ :\ v\geq 0\right\}.$$

delete

It is easy to check that ~~the~~ explicit solutions to all three subproblems are given by

$$v_l^i(t+1) = \frac{1}{\left(\|x^l(t) - a^i\|^2 + \varepsilon^2\right)^{1/2}}, \quad i = 1, 2, \ldots, m, \; l = 1, 2, \ldots, k, \tag{4.22}$$

$$w^i(t+1) = P_\Delta\left(w^i(t) - \frac{b_\varepsilon^i(v(t+1), x(t))}{\alpha_i(t)}\right), \quad i = 1, 2, \ldots, m, \tag{4.23}$$

and

$$x^l(t+1) = \frac{\sum_{i=1}^m w_l^i(t+1) v_l^i(t+1) a^i}{\sum_{i=1}^m w_l^i(t+1) v_l^i(t+1)}, \quad l = 1, 2, \ldots, k. \tag{4.24}$$

From the subproblem for v and the ~~initial~~ observation given in ~~(see~~ (4.16)~~)~~, we derive the following three relations

$$B_\varepsilon(v(t+1), w, x(t)) = H_\varepsilon(w, x(t)), \quad \forall\, t \in \mathbb{N},\ \forall\, w \in \Delta^m, \tag{4.25}$$

$$b_\varepsilon^i(v(t+1), x(t)) = d_\varepsilon^i(x(t)), \quad \forall\, t \in \mathbb{N},\ i = 1, 2, \ldots, m, \tag{4.26}$$

where d_ε^i is defined in (4.3), and

$$B_\varepsilon(v, w, x) \geq H_\varepsilon(w, x), \ \forall\, (v, w, x) \in I^{mk} \times \Delta^m \times \mathbb{R}^{nk}. \tag{4.27}$$

Substituting (4.26) into ~~w update step (see~~ (4.23)~~)~~ yields

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

Moreover, substituting ~~v update step (see~~ (4.22)~~)~~ into ~~x update step (see~~ (4.24)~~)~~ yields

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

Thus, we recover the algorithm approach

where L_ε^l is defined in (4.4). ~~These are exactly the update steps w and x in~~ ε-KPALM, which means ~~the two algorithms are the same~~. However, with the current ~~version~~ we can swiftly prove the sufficient decrease and the subgradient lower bound for the iterates gap properties, ~~similar to the KPALM case~~. which are needed to obtain global converges of $\{z(t)\}_{t\in\mathbb{N}}$ that generated by ε-KPALM

these two different approaches lead to the same iterative algorithm,

Proposition 4.5 (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*

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

Proof. From (4.20) we have

$$\left\langle w^i(t+1), b_\varepsilon^i(v(t+1), x(t))\right\rangle + \frac{\alpha_i(t)}{2}\left\|w^i(t+1) - w^i(t)\right\|^2 \leq \left\langle w^i(t), b_\varepsilon^i(v(t+1), x(t))\right\rangle$$

Summing the last inequality over $i = 1, 2, \ldots, m$ and applying (4.18) yields

$$B_\varepsilon(v(t+1), w(t+1), x(t)) + \sum_{i=1}^m \frac{\alpha_i(t)}{2}\left\|w^i(t+1) - w^i(t)\right\|^2 \leq B_\varepsilon(v(t+1), w(t), x(t))$$

Using Assumption 1(i) we derive

$$\frac{\underline{\alpha}}{2}\|w(t+1)-w(t)\|^2 \leq B_\varepsilon(v(t+1),w(t),x(t)) - B_\varepsilon(v(t+1),w(t+1),x(t))$$
$$\leq H_\varepsilon(w(t),x(t)) - H_\varepsilon(w(t+1),x(t)), \tag{4.28}$$

where the last inequality follows from (4.25) and (4.27).

Since the function $x \mapsto B_\varepsilon(v,w,x)$ is C^2, and

it follows that

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

~~hence,~~ the function $x \mapsto B_\varepsilon(v(t+1),w(t),x)$ is strongly convex with parameter $\underline{\beta}/2\kappa$, for all $t \in \mathbb{N}$, ~~indeed,~~ Indeed

$$\nabla^2_{x^l}B_\varepsilon(v(t+1),w(t),x) = \sum_{i=1}^{m} w_l^i(t)v_l^i(t+1) \cancel{\geq} \frac{1}{\kappa}\sum_{i=1}^{m} w_l^i(t) \cancel{\geq} \frac{\underline{\beta}(w^i(t))}{2\kappa} \cancel{\geq} \frac{\underline{\beta}}{2\kappa} > 0,$$

$\geq$

$>$

where the first inequality follows from the fact that $v_l^i(t) \in I$ for all $t \in \mathbb{N}$, $\underline{\beta}(\cdot)$ is defined in (3.9), and the second inequality is due to Assumption 1(ii). Using the strong convexity property we deduce the sufficient decrease in x, ~~indeed,~~ as follows

$$\frac{\underline{\beta}}{4\kappa}\|x(t+1)-x(t)\|^2 = \langle \nabla_x B_\varepsilon(z(t+1)), x(t)-x(t+1)\rangle + \frac{\underline{\beta}}{4\kappa}\|x(t+1)-x(t)\|^2$$
$$\leq B_\varepsilon(v(t+1),w(t+1),x(t)) - B_\varepsilon(z(t+1))$$
$$\leq H_\varepsilon(w(t+1),x(t)) - H_\varepsilon(w(t+1),x(t+1)), \tag{4.29}$$

where the first equality follows from (4.21), the second inequality follows from the strong convexity, and the last inequality is due to (4.25) and (4.27). Set $\rho_1 = \min\{\underline{\alpha}/2, \underline{\beta}/4\kappa\}$. Summing (4.28) and (4.29), we get

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

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

Proposition 4.6 (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. By the definition of Ψ_ε (see (4.1)) we get

$$\Psi_\varepsilon(w,x) = H_\varepsilon(w,x) + \sum_{i=1}^{m}\delta_\Delta(w^i) \tag{4.30}$$

with respect to

it

Differentiating (4.30) ~~w.r.t.~~ x and evaluating in $z(t+1)$ yields

$$\partial_x \Psi_\varepsilon(z(t+1)) = \nabla_x H_\varepsilon(z(t+1)). \tag{4.31}$$

Similarly, d ~~D~~ifferentiating (4.30) ~~w.r.t.~~ w^i and evaluating it in $z(t+1)$ yields

$$\partial_{w^i} \Psi_\varepsilon(z(t+1)) = d_\varepsilon^i(x(t+1)) + \partial_{w^i}\delta_\Delta(w^i(t+1)). \tag{4.32}$$

The optimality condition of $w^i(t+1)$ which derived from (4.20), 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

$$\begin{aligned}\mathbf{0} &= b_\varepsilon^i(v(t+1), x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1)\\ &= d_\varepsilon^i(x(t)) + \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) + u^i(t+1),\end{aligned} \tag{4.33}$$

where the last equality follows from (4.26). Substituting (4.33) into (4.32) and combining with (4.31) we deduce that

$$\gamma(t+1) := \left(\left(d^i(x(t+1)) - d^i(x(t)) - \alpha_i(t)(w^i(t+1) - w^i(t))\right)_{i=1,2,\ldots,m}, \nabla_x H_\varepsilon(z(t+1))\right) \in \partial\Psi_\varepsilon(z(t+1)).$$

Therefore,

$$\begin{aligned}\|\gamma(t+1)\| &\leq \sum_{i=1}^m \left\|d^i(x(t+1)) - d^i(x(t)) - \alpha_i(t)(w^i(t+1) - w^i(t))\right\| + \|\nabla_x H_\varepsilon(z(t+1))\|\\ &\leq \sum_{i=1}^m \left\|d^i(x(t+1)) - d^i(x(t))\right\| + \overline{\alpha}\sum_{i=1}^m \left\|w^i(t+1) - w^i(t)\right\| + \frac{2m\sqrt{k}}{\varepsilon}\|x(t+1) - x(t)\|\\ &\leq \sum_{i=1}^m \frac{d_A}{\varepsilon}\|x(t+1) - x(t)\| + \overline{\alpha}\sqrt{m}\,\|w(t+1) - w(t)\| + \frac{2m\sqrt{k}}{\varepsilon}\|x(t+1) - x(t)\|\\ &\leq \left(\frac{md_A}{\varepsilon} + \overline{\alpha}\sqrt{m} + \frac{2m\sqrt{k}}{\varepsilon}\right)\|z(t+1) - z(t)\|,\end{aligned}$$

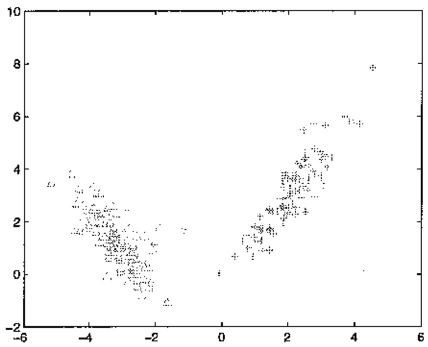
where the second inequality was established in Proposition 4.4 (see (4.15)) and the third inequality follows from Lemma 4.3.2. Define $\rho_2 = \frac{md_A}{\varepsilon} + \overline{\alpha}\sqrt{m} + \frac{2m\sqrt{k}}{\varepsilon}$, and the result follows. □

4.3 Synthetic Dataset

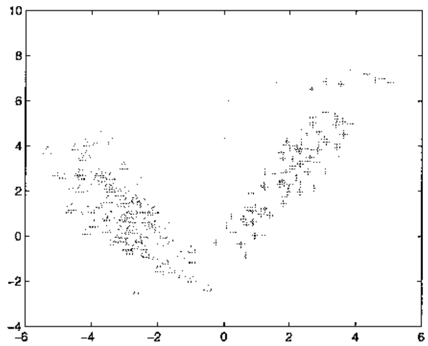
In this section we show that ε-KPALM is less sensitive to outliners in the data verses algorithms that suit the squared Euclidean norm (e.g. KMEANS, KMEANS++ and KPALM). We generated two synthetic datasets, each contains 300 points in the plane, by sampling three two-dimensional Gaussian, 100 samples each. In Figure 1(1a) the clusters are denser than in Figure 1(1b).

Then we run the clustering algorithms and compared their clustering results, namely, how many points were clustered correctly. From Figure 2(2a) it is evident that KMEANS is superior to other algorithms in the dense case and ϵ-KPALM is quite sensitive. Whereas, in the sparse case in Figure 2(2b), ε-KPALM is superior, and less sensitive to outliners.

Varepsilon

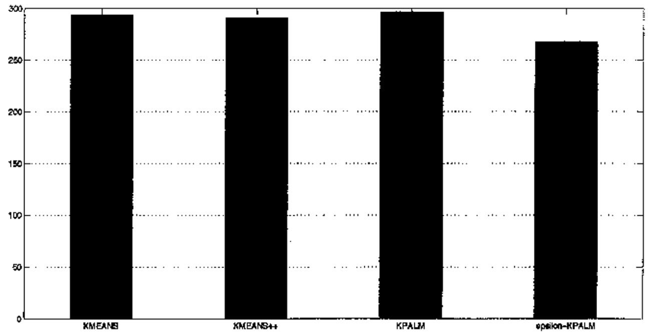


(a) Dense Gaussians.

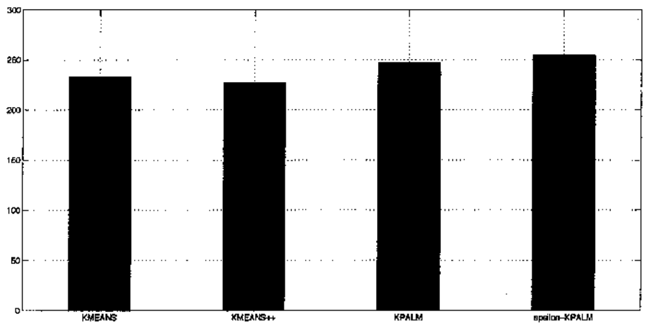


(b) Sparse Gaussians.

Figure 1: Two datasets, each 300 points.



(a) Dense Gaussians clustering.



(b) Sparse Gaussians clustering.

Figure 2: Results of clustering algorithms for dense and sparse datasets.

5 Returning to KMEANS

5.1 Similarity to KMEANS

The famous KMEANS algorithm has close relation 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

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

It is easy to see that if we take $\alpha_i(t) = 0$ for all $1 \leq i \leq m$ and $t \in \mathbb{N}$, then KPALM becomes KMEANS. We aim to employ the PALM theory once more and show that the sequence generated by KMEANS converges to a critical point of $\Psi(\cdot)$, as defined is (2.3). The sufficient decrease proof of Section 3 breaks down in this case, since it is based on Assumption 1(i), that is, $\alpha_i(t) > \underline{\alpha_i} > 0$, for all $t \in \mathbb{N}$ and $i = 1, 2, \ldots, m$. However, the proof of the subgradient lower bound for the iterates gap property follows through as is. In the following discussion we present the means to treat the case that $\alpha_i(t) = 0$, and prove the sufficient decrease property.

Lemma 5.0.1. *Let $\{z(t)\}_{t \in \mathbb{N}}$ be the sequence generated by KMEANS. Then, there exists $c > 0$ such that*

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

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, 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

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

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

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

Therefore, combining these arguments yields

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

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. □

Equipped with the last lemma we briefly prove the sufficient decrease property of KMEANS.

Proposition 5.1 (Sufficient decrease property for KMEANS sequence). *Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KMEANS. Then, there exists $\rho_1 > 0$ such that*

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

Proof. The function $x \mapsto H(w(t), x)$ remains strongly convex with parameter $\beta(w(t))$ (see (3.12)), hence we have a sufficient decrease in the x variable, namely,

$$\frac{\underline{\beta}}{2}\|x(t+1) - x(t)\|^2 \le H(w(t), x(t)) - H(w(t+1), x(t+1)). \tag{5.3}$$

Setting $\rho_1 = \underline{\beta}/2(1 + mc^2)$, we can write

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

where the first inequality follows from Lemma 5.0.1, the second follows from (5.3), and the last equality follows from the fact that $G(w(t)) = 0$, for all $t \in \mathbb{N}$. $\square$

5.2 KMEANS Local Minima Convergence Proof

In this section we present a simple and direct proof that KMEANS converges to local minima. We start with rewriting the KMEANS algorithm, in its most familiar form

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

$$C^l(t+1) = \left\{ a \in \mathcal{A} \mid \|a - x^l(t)\| \le \|a - x^j(t)\|, \quad \forall 1 \le l \le k \right\}. \tag{5.4}$$

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

$$x^l(t+1) = mean(C^l(t+1)) := \frac{1}{|C^l(t+1)|} \sum_{a \in C^l(t+1)} a. \tag{5.5}$$

(2.3) Stopping criteria: halt if

$$\forall 1 \le l \le k \quad C^l(t+1) = C^l(t) \tag{5.6}$$

As in KPALM, KMEANS needs Assumption 1(ii) for step (5.5) to be well defined. In order to prove the convergence of KMEANS to local minimum, we will need to following assumption.

Assumption 2. *Let $t \in \mathbb{N}$ be the final iteration of KMEANS run, then we assume that 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 $\overline{C^l}(x) = \{a \in \mathcal{A} \mid \|a - x^l\| \le \|a - x^j\|, \quad \forall j \ne l\}$, for all $1 \le l \le k$, and the sub-partition of $\mathcal{A}$ by $\underline{C^l}(x) = \{a \in \mathcal{A} \mid \|a - x^l\| < \|a - x^j\|, \quad \forall j \ne l\}$. Moreover, denote $R_{lj}(t) = \min_{a \in C^l(t)} \{\|a - x^j(t)\| - \|a - x^l(t)\|\}$ for all $1 \le l, j \le k$, and $r(t) = \min_{l \ne j} R_{lj}$.
Due to Assumption 2 we have that $\overline{C^l}(x(t)) = \underline{C^l}(x(t)) = C^l(t+1)$, for all $1 \le l \le k$, $t \in \mathbb{N}$, we also have that $r(t) > 0$ for all $t \in \mathbb{N}$.

Proposition 5.2. *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 $1 \le l \le k$.*

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 (5.6) 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

$$r(t) \le \|x^j(t) - a\| - \|x^l(t) - a\| \quad \forall j \ne l. \tag{5.7}$$

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 (5.7), 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 $1 \leq l \leq k$. $\square$

Proposition 5.3 (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 5.2.

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

6 Numeric Results

In this section we show the numeric results and compare the algorithms presented in this work with other algorithms that are commonly used to address the clustering problem.

The initialization points used within the implementation of the compared algorithms are as follows. KMEANS staring point is constructed by randomly choosing k different points from the dataset. The same technique is employed in the cases of KPALM and ε-KPALM, for the $x(0)$ variable. Whereas for the $w(0)$ variable, it is chosen at random from Δ^m. KMEANS++ takes also part in our comparison, and it is basically the same as KMEANS, with the exception of its starting point that is constructed in the following manner. The first center $x^1(0)$ is chosen randomly from the dataset $\mathcal{A}$. Suppose that $1 \leq l < k$ centers have already been chosen, set $x^{l+1}(0)$ to be the point in the dataset that is the furthest from its closest center.

6.1 Iris Dataset

We used the famous Iris dataset to test the performance of the KPALM algorithm. It is important to note that choosing the parameter α is left to the user, and as presented below, has a significant effect on the convergence rate and the quality of the achieved clustering, namely the value of the objective function over the generated series. All the plots in this section are made by averaging over 100 trials, each trial with random starting point.

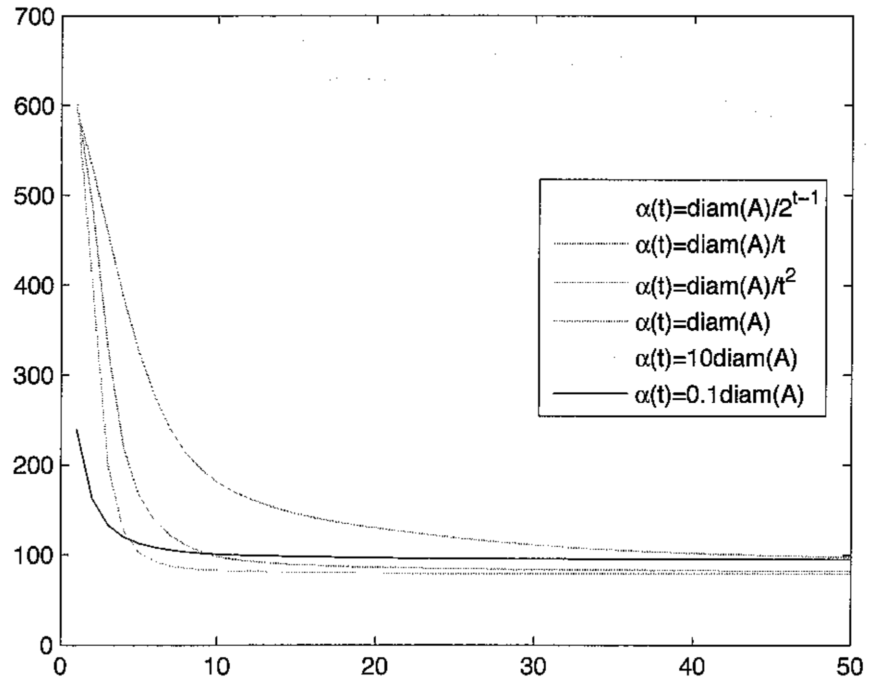
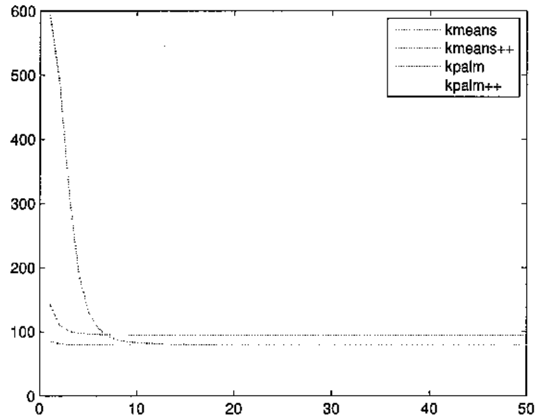


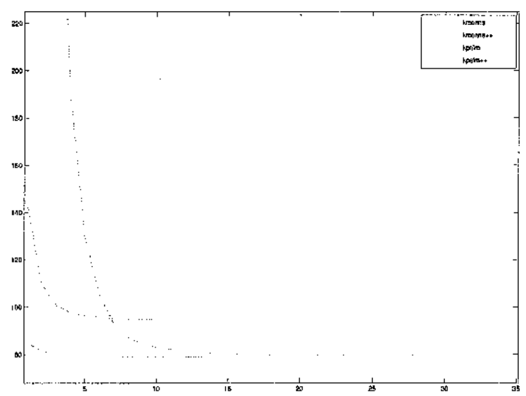
Figure 3: Comparison of the objective values for different values of α.

Figure 3 shows that dynamic values of the parameter α which decreases fast, such as $\alpha_i(t) = \frac{diam(\mathcal{A})}{2^{t-1}}$, achieve smaller function values.

In Figure 4 we made a comparison between KPALM with dynamic rule for choosing the parameter α, that is $\alpha_i(t) = \frac{diam(\mathcal{A})}{2^{t-1}}$, with KMEANS and KMEANS++. It demonstrates that KPALM can reach lower objective function values then KMEANS, and these are similar to the values achieved with KMEAN++. In addition, the KPALM++ are the objective function values achieved with KPALM when the x variable is initialized as in KMEANS++. Unlike KMEANS, the objective function values KPALM converge to are more stable and less sensitive to its starting point.



(a) Comparison of objective function values.

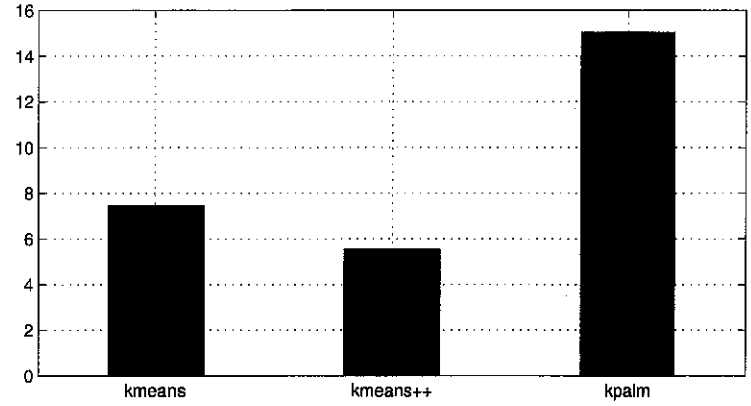


(b) Zoom of Figure 4a.

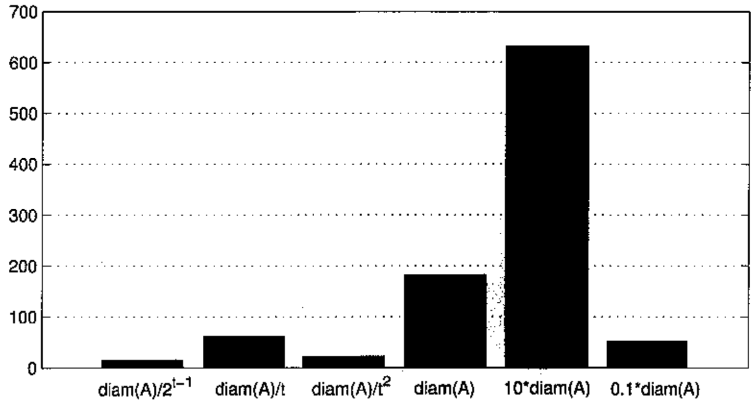
Figure 4: Comparison of objective function values for KMEANS, KMEANS++, KPALM and KMEANS++.

Figure 5 shows the number of iteration needed to reach precision of 1e-3 between consecutive objective function values.

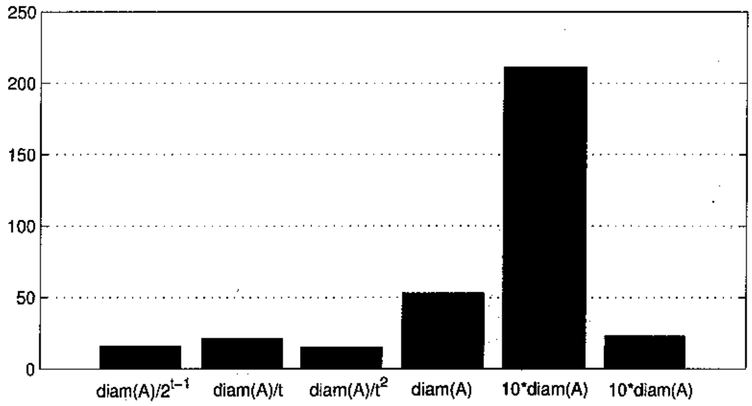
Similarly to Figure 3, in Figure 6 we can see a comparison of the objective values of Ψ_ε for different function values. The value of ε is set to be 1e-5.



(a) Number of iterations of KMEANS, KMEANS++ and KPALM with $\alpha(t) = diam(\mathcal{A})/2^{t-1}$.



(b) Number of iterations of KPALM with different updates of $\alpha(t)$.



(c) Number of iterations of ε-KPALM with different updates of $\alpha(t)$.

Figure 5: Comparison of number of iterations needed to reach 1e-3 precision of Ψ.