A Novel Class of Globally Convergent Algorithms For Clustering Problems

Sergey Voldman

Raymond and Beverly Sackler Faculty of Exact Sciences

Tel-Aviv University

April 6, 2016

Research conducted under the supervision of
Prof. Marc Teboulle (Tel-Aviv University)
and Prof. Shoham Sabach (Technion)

Goal and Outline

Develop and analyze two globally convergent center-based clustering algorithms each with a different distance-like function.

Goal and Outline

Develop and analyze two globally convergent center-based clustering algorithms each with a different distance-like function.

Outline

- Introduction to the clustering problem.

Goal and Outline

Develop and analyze two globally convergent center-based clustering algorithms each with a different distance-like function.

Outline

- Introduction to the clustering problem.
- Exploiting problem structure: Basic scheme and convergence methodology.

Goal and Outline

Develop and analyze two globally convergent center-based clustering algorithms each with a different distance-like function.

Outline

- Introduction to the clustering problem.
- Exploiting problem structure: Basic scheme and convergence methodology.
- Clustering with the squared Euclidean norm: KPALM algorithm and its analysis.

Goal and Outline

Develop and analyze two globally convergent center-based clustering algorithms each with a different distance-like function.

Outline

- Introduction to the clustering problem.
- Exploiting problem structure: Basic scheme and convergence methodology.
- Clustering with the squared Euclidean norm: KPALM algorithm and its analysis.
- Clustering with the Euclidean norm: ε-KPALM algorithm and its analysis.

Goal and Outline

Develop and analyze two globally convergent center-based clustering algorithms each with a different distance-like function.

Outline

- Introduction to the clustering problem.
- Exploiting problem structure: Basic scheme and convergence methodology.
- Clustering with the squared Euclidean norm: KPALM algorithm and its analysis.
- Clustering with the Euclidean norm: ε-KPALM algorithm and its analysis.
- Numerical results of the proposed algorithms.

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem has attracted a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem has attracted a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- The clustering problem is NP-hard.

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem has attracted a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- The clustering problem is NP-hard.
- Let $\mathcal{A} = \{a^1, a^2, \ldots, a^m\} \subset \mathbb{R}^n$ set of points, and $1 < k < m$ a given number of clusters.

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem has attracted a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- The clustering problem is NP-hard.
- Let $\mathcal{A} = \{a^1, a^2, \ldots, a^m\} \subset \mathbb{R}^n$ set of points, and $1 < k < m$ a given number of clusters.
- The goal is to partition the data $\mathcal{A}$ into k subsets $\{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^k\}$ called clusters.

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem has attracted a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- The clustering problem is NP-hard.
- Let $\mathcal{A} = \{a^1, a^2, \ldots, a^m\} \subset \mathbb{R}^n$ set of points, and $1 < k < m$ a given number of clusters.
- The goal is to partition the data $\mathcal{A}$ into k subsets $\{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^k\}$ called clusters.
- Each cluster $\mathcal{C}^l$ is represented by its center $x^l \in \mathbb{R}^n$.

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem has attracted a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- The clustering problem is NP-hard.
- Let $\mathcal{A} = \{a^1, a^2, \ldots, a^m\} \subset \mathbb{R}^n$ set of points, and $1 < k < m$ a given number of clusters.
- The goal is to partition the data $\mathcal{A}$ into k subsets $\{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^k\}$ called clusters.
- Each cluster $\mathcal{C}^l$ is represented by its center $x^l \in \mathbb{R}^n$.
- The clustering problem is given by

$$(P_0) \qquad \min_{x \in \mathbb{R}^{nk}} \left\{ F(x) := \sum_{i=1}^{m} \min_{1 \le l \le k} d(x^l, a^i) \right\},$$

with $d(\cdot, \cdot)$ is a distance-like function. Here we focus on:

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem has attracted a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- The clustering problem is NP-hard.
- Let $\mathcal{A} = \{a^1, a^2, \ldots, a^m\} \subset \mathbb{R}^n$ set of points, and $1 < k < m$ a given number of clusters.
- The goal is to partition the data $\mathcal{A}$ into k subsets $\{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^k\}$ called clusters.
- Each cluster $\mathcal{C}^l$ is represented by its center $x^l \in \mathbb{R}^n$.
- The clustering problem is given by

$$(P_0) \qquad \min_{x \in \mathbb{R}^{nk}} \left\{ F(x) := \sum_{i=1}^{m} \min_{1 \le l \le k} d(x^l, a^i) \right\},$$

  with $d(\cdot, \cdot)$ is a distance-like function. Here we focus on:
  - $d(u, x) := \|u - v\|^2$

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem has attracted a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- The clustering problem is NP-hard.
- Let $\mathcal{A} = \{a^1, a^2, \ldots, a^m\} \subset \mathbb{R}^n$ set of points, and $1 < k < m$ a given number of clusters.
- The goal is to partition the data $\mathcal{A}$ into k subsets $\{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^k\}$ called clusters.
- Each cluster $\mathcal{C}^l$ is represented by its center $x^l \in \mathbb{R}^n$.
- The clustering problem is given by

$$(P_0) \qquad \min_{x \in \mathbb{R}^{nk}} \left\{ F(x) := \sum_{i=1}^{m} \min_{1 \le l \le k} d(x^l, a^i) \right\},$$

with $d(\cdot, \cdot)$ is a distance-like function. Here we focus on:

- $d(u, x) := \|u - v\|^2$
- $d(u, x) := \|u - v\|$

Problem Reformulation

- Using the fact that

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

where $\Delta := \left\{ u \in \mathbb{R}^d : \sum_{l=1}^{d} u_l = 1, \ u \ge 0 \right\}$ is the simplex in $\mathbb{R}^k$, problem (P_0) can be transformed into

$$(P_1) \qquad \min_{x \in \mathbb{R}^{nk}} \left\{ \sum_{i=1}^{m} \min_{w^i \in \Delta} \langle w^i, d^i(x) \rangle \right\},$$

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

Problem Reformulation

- Using the fact that
$$\min_{1 \le l \le k} u_l = \min \left\{ \langle u, v \rangle : v \in \Delta \right\},$$
where $\Delta := \left\{ u \in \mathbb{R}^d : \sum_{l=1}^{d} u_l = 1,\ u \ge 0 \right\}$ is the simplex in $\mathbb{R}^k$, problem (P_0) can be transformed into
$$(P_1) \qquad \min_{x \in \mathbb{R}^{nk}} \left\{ \sum_{i=1}^{m} \min_{w^i \in \Delta} \langle w^i, d^i(x) \rangle \right\},$$
with $d^i(x) = (d(x^1, a^i), d(x^2, a^i), \ldots, d(x^k, a^i)) \in \mathbb{R}^k, \quad i = 1, 2, \ldots, m.$
- Replacing the constrain $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$ results in
$$(P_2) \qquad \min_{x \in \mathbb{R}^{nk}, w \in \mathbb{R}^{km}} \left\{ \sum_{i=1}^{m} \left(\langle w^i, d^i(x) \rangle + \delta_\Delta(w^i) \right) \right\},$$
where $w = (w^1, w^2, \ldots, w^m) \in \mathbb{R}^{km}$.

Problem Reformulation

- Using the fact that

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

where $\Delta := \left\{ u \in \mathbb{R}^d : \sum_{l=1}^{d} u_l = 1,\ u \geq 0 \right\}$ is the simplex in $\mathbb{R}^k$, problem (P_0) can be transformed into

$$(P_1) \qquad \min_{x \in \mathbb{R}^{nk}} \left\{ \sum_{i=1}^{m} \min_{w^i \in \Delta} \langle w^i, d^i(x) \rangle \right\},$$

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

- Replacing the constrain $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$ results in

$$(P_2) \qquad \min_{x \in \mathbb{R}^{nk}, w \in \mathbb{R}^{km}} \left\{ \sum_{i=1}^{m} \left(\langle w^i, d^i(x) \rangle + \delta_\Delta(w^i) \right) \right\},$$

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

- The final version is given by

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

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

Clustering: Problem Structure

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

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

We are interested in two important cases for clustering:

(a) $d^i(x) := \left\|x - a^i\right\|^2$ (the common squared Euclidean distance)

(b) $d^i(x) := \left\|x - a^i\right\|$ (relevant for problems with outliers)

Clustering: Problem Structure

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

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

We are interested in two important cases for clustering:

(a) $d^i(x) := \left\|x - a^i\right\|^2$ (the common squared Euclidean distance)

(b) $d^i(x) := \left\|x - a^i\right\|$ (relevant for problems with outliers)

- In case (a) $H(w,x)$ is a <u>smooth</u> function.

Clustering: Problem Structure

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

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

We are interested in two important cases for clustering:

(a) $d^i(x) := \left\|x - a^i\right\|^2$ (the common squared Euclidean distance)

(b) $d^i(x) := \left\|x - a^i\right\|$ (relevant for problems with outliers)

- In case (a) $H(w,x)$ is a smooth function.
- Whereas, in case (b) $H(w,x)$ is a nonsmooth.

Clustering: Problem Structure

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

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

We are interested in two important cases for clustering:

(a) $d^i(x) := \left\|x - a^i\right\|^2$ (the common squared Euclidean distance)

(b) $d^i(x) := \left\|x - a^i\right\|$ (relevant for problems with outliers)

- In case (a) $H(w,x)$ is a <u>smooth</u> function.
- Whereas, in case (b) $H(w,x)$ is a <u>nonsmooth</u>.
- In both cases the problem is <u>nonconvex</u>.

Clustering: Problem Structure

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

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

We are interested in two important cases for clustering:

(a) $d^i(x) := \left\|x - a^i\right\|^2$ (the common squared Euclidean distance)

(b) $d^i(x) := \left\|x - a^i\right\|$ (relevant for problems with outliers)

- In case (a) $H(w,x)$ is a smooth function.
- Whereas, in case (b) $H(w,x)$ is a nonsmooth.
- In both cases the problem is nonconvex.
- We want to exploit the special structure to devise simple schemes. Attractive approach is via alternating minimization.

Alternating Minimization

Very simple idea (goes back to Gauss-Seidel),

$$w^{k+1} \in \operatorname{argmin}_{w \in \Delta^m} \sigma\left(w, x^k\right), \quad x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^{nk}} \sigma\left(w^{k+1}, x\right)$$

- In fact this is the essence of <u>k-means</u> (for smooth H) and of <u>k-median</u> (for nonsmooth H) algorithms.

Alternating Minimization

Very simple idea (goes back to Gauss-Seidel),

$$w^{k+1} \in \operatorname{argmin}_{w \in \Delta^m} \sigma\left(w, x^k\right), \quad x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^{nk}} \sigma\left(w^{k+1}, x\right)$$

- In fact this is the essence of <u>k-means</u> (for smooth H) and of <u>k-median</u> (for nonsmooth H) algorithms.
- These algorithms are not known to globally converge to stationary (critical) point.

Alternating Minimization

Very simple idea (goes back to Gauss-Seidel),

$$w^{k+1} \in \operatorname{argmin}_{w \in \Delta^m} \sigma\left(w, x^k\right), \quad x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^{nk}} \sigma\left(w^{k+1}, x\right)$$

- In fact this is the essence of k-means (for smooth H) and of k-median (for nonsmooth H) algorithms.
- These algorithms are not known to globally converge to stationary (critical) point.
- Here we follow the recent algorithm (PALM) of (Bolte-Sabach-Teboulle (2014)) to tackle these issues and devise simple schemes.

The Optimization Model

$$(M) \qquad \text{minimize}_{x,y}\sigma(x,y) := f(x) + g(y) + H(x,y)$$

The Optimization Model

$$(M) \qquad \text{minimize}_{x,y}\sigma(x,y) := f(x) + g(y) + H(x,y)$$

Assumption 1.

(i) *$f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ are proper and lsc functions.*

(ii) *$H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a C^1 function.*

(iii) *Partial gradients of H are Lipshitz continuous: $H(\cdot, y) \in C^{1,1}_{L(y)}$ and likewise $H(x, \cdot) \in C^{1,1}_{L}(x)$.*

The Optimization Model

$$(M) \qquad \text{minimize}_{x,y}\sigma(x,y) := f(x) + g(y) + H(x,y)$$

Assumption 1.

(i) *$f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ are proper and lsc functions.*

(ii) *$H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a C^1 function.*

(iii) *Partial gradients of H are Lipshitz continuous: $H(\cdot, y) \in C^{1,1}_{L(y)}$ and likewise $H(x, \cdot) \in C^{1,1}_{L}(x)$.*

- **NO convexity** will be assumed in the objective or/and the constraints (built-in through f and g extended valued).

The Optimization Model

$$(M) \qquad \text{minimize}_{x,y}\sigma(x,y) := f(x) + g(y) + H(x,y)$$

Assumption 1.

(i) *$f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ are proper and lsc functions.*

(ii) *$H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a C^1 function.*

(iii) *Partial gradients of H are Lipshitz continuous: $H(\cdot, y) \in C^{1,1}_{L(y)}$ and likewise $H(x, \cdot) \in C^{1,1}_{L}(x)$.*

- **NO convexity** will be assumed in the objective or/and the constraints (built-in through f and g extended valued).

The Optimization Model

$$(M) \qquad \text{minimize}_{x,y}\sigma(x,y) := f(x) + g(y) + H(x,y)$$

Assumption 1.

(i) *$f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ are proper and lsc functions.*

(ii) *$H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a C^1 function.*

(iii) *Partial gradients of H are Lipshitz continuous: $H(\cdot, y) \in C^{1,1}_{L(y)}$ and likewise $H(x, \cdot) \in C^{1,1}_{L}(x)$.*

- **NO convexity** will be assumed in the objective or/and the constraints (built-in through f and g extended valued).

Let $\sigma : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and lsc function. Given $x \in \mathbb{R}^n$ and $t > 0$, the proximal map defined by:

$$\mathrm{prox}_t^{\sigma}(x) := \text{argmin}\left\{\sigma(u) + \frac{t}{2}\|u - x\|^2 : \; u \in \mathbb{R}^n\right\}.$$

The Algorithm: Proximal Altertnating Linearization Minimization (PALM)

- PALM blends alternating minimization with the classical proximal gradient over the two blocks (x, y).

The Algorithm: Proximal Altertnating Linearization Minimization (PALM)

- PALM blends alternating minimization with the classical proximal gradient over the two blocks (x, y).

The Algorithm: Proximal Altertnating Linearization Minimization (PALM)

- PALM blends alternating minimization with the classical proximal gradient over the two blocks (x, y). This leads towards the following approximations:

$$\widehat{\sigma}(x, y^k) = \left\langle x - x^k, \nabla_x H(x^k, y^k) \right\rangle + \tfrac{c_k}{2} \left\| x - x^k \right\|^2 + f(x),$$

$$\widetilde{\sigma}(x^{k+1}, y) = \left\langle y - y^k, \nabla_y H(x^{k+1}, y^k) \right\rangle + \tfrac{d_k}{2} \left\| y - y^k \right\|^2 + g(y).$$

The Algorithm: Proximal Altertnating Linearization Minimization (PALM)

- PALM blends alternating minimization with the classical proximal gradient over the two blocks (x, y). This leads towards the following approximations:

$$\widehat{\sigma}(x, y^k) = \left\langle x - x^k, \nabla_x H(x^k, y^k) \right\rangle + \tfrac{c_k}{2} \left\| x - x^k \right\|^2 + f(x),$$

$$\widetilde{\sigma}(x^{k+1}, y) = \left\langle y - y^k, \nabla_y H(x^{k+1}, y^k) \right\rangle + \tfrac{d_k}{2} \left\| y - y^k \right\|^2 + g(y).$$

1. Initialization: start with any $\left(x^0, y^0\right) \in \mathbb{R}^n \times \mathbb{R}^m$.
2. For each $k = 0, 1, \ldots$ generate a sequence $\left\{ \left(x^k, y^k\right) \right\}_{k \in \mathbb{N}}$:
   - 2.1. Take $\gamma_1 > 1$, set $c_k = \gamma_1 L_1 \left(y^k\right)$ and compute

     $$x^{k+1} \in \operatorname{argmin} \left\{ \widehat{\sigma}(x, y^k) : \ x \in \mathbb{R}^n \right\} = \operatorname{prox}^f_{c_k} \left(x^k - {c_k}^{-1} \nabla_x H \left(x^k, y^k \right) \right).$$

   - 2.2. Take $\gamma_2 > 1$, set $d_k = \gamma_2 L_2 \left(x^{k+1}\right)$ and compute

     $$y^{k+1} \in \operatorname{argmin} \left\{ \widetilde{\sigma}(x^{k+1}, y) : \ y \in \mathbb{R}^m \right\} = \operatorname{prox}^g_{d_k} \left(y^k - {d_k}^{-1} \nabla_y H \left(x^{k+1}, y^k \right) \right).$$

KPALM Algorithm for Clustering with the Squared Norm Distance

Recalling the Clustering Problem for the squared norm distance:

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

$$H(w,x) = \sum_{i=1}^{m} H^i(w,x) = \sum_{i=1}^{m} \langle w^i, d^i(x)\rangle = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i \left\|x^l - a^i\right\|^2.$$

- Thus, here $H(w,x)$ is $C^{1,1}$ and fits the optimization model.

Inspired by PALM, we devise the KPALM algorithm, exploiting the specific structure of the model, namely

KPALM Algorithm for Clustering with the Squared Norm Distance

Recalling the Clustering Problem for the squared norm distance:

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

$$H(w,x) = \sum_{i=1}^{m} H^i(w,x) = \sum_{i=1}^{m} \langle w^i, d^i(x)\rangle = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i \left\|x^l - a^i\right\|^2.$$

- Thus, here $H(w,x)$ is $C^{1,1}$ and fits the optimization model.

Inspired by PALM, we devise the KPALM algorithm, exploiting the specific structure of the model, namely

KPALM Algorithm for Clustering with the Squared Norm Distance

Recalling the Clustering Problem for the squared norm distance:

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

$$H(w, x) = \sum_{i=1}^{m} H^i(w, x) = \sum_{i=1}^{m} \langle w^i, d^i(x)\rangle = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i \left\|x^l - a^i\right\|^2.$$

- Thus, here $H(w, x)$ is $C^{1,1}$ and fits the optimization model.

Inspired by PALM, we devise the KPALM algorithm, exploiting the specific structure of the model, namely

- 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.

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

KPALM Algorithm for Clustering with the Squared Norm Distance

Recalling the Clustering Problem for the squared norm distance:

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

$$H(w,x) = \sum_{i=1}^{m} H^i(w,x) = \sum_{i=1}^{m} \langle w^i, d^i(x)\rangle = \sum_{i=1}^{m}\sum_{l=1}^{k} w_l^i \left\|x^l - a^i\right\|^2.$$

- Thus, here $H(w,x)$ is $C^{1,1}$ and fits the optimization model.

Inspired by PALM, we devise the KPALM algorithm, exploiting the specific structure of the model, namely

- 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.

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

- The function $x \mapsto H(w,x)$, for fixed w, is quadratic and convex. Hence, <u>there is no need</u> to add a proximal term as suggested in PALM.

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

KPALM Algorithm for Clustering with the Squared Norm Distance

KPALM performs alternation between the cluster assignment and the center update steps. Solving the previously described minimization sub-problems yields the KPALM algorithm.

KPALM Algorithm for Clustering with the Squared Norm Distance

KPALM performs alternation between the cluster assignment and the center update steps. Solving the previously described minimization sub-problems yields the KPALM algorithm.

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

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

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

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

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

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

KPALM Algorithm for Clustering with the Squared Norm Distance

KPALM performs alternation between the cluster assignment and the center update steps. Solving the previously described minimization sub-problems yields the KPALM algorithm.

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

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

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

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

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

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

- Setting $\alpha_i(t) = 0$ yields the popular k-means algorithm.

Convergence Analysis of KPALM

We will use the methodology developed in (Bolte-Sabach-Teboulle (2014)).

Definition 1.

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. A sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ is called a gradient-like descent sequence for σ if for all $k \in \mathbb{N}$ the following two conditions hold:

(C1) Sufficient decrease property: There exists a positive scalar ρ_1 such that

$$\rho_1 \left\|z^{k+1} - z^k\right\|^2 \leq \sigma\left(z^k\right) - \sigma\left(z^{k+1}\right).$$

(C2) A subgradient lower bound for the iterates gap:

- $\left\{z^k\right\}_{k\in\mathbb{N}}$ is bounded.
- There exists a positive scalar ρ_2 such that

$$\left\|w^{k+1}\right\| \leq \rho_2 \left\|z^{k+1} - z^k\right\|, \; w^{k+1} \in \partial\sigma\left(z^{k+1}\right).$$

Convergence Analysis of KPALM

We will use the methodology developed in (Bolte-Sabach-Teboulle (2014)).

Definition 1.

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. A sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ is called a gradient-like descent sequence for σ if for all $k \in \mathbb{N}$ the following two conditions hold:

(C1) Sufficient decrease property: There exists a positive scalar ρ_1 such that

$$\rho_1 \left\|z^{k+1} - z^k\right\|^2 \leq \sigma\left(z^k\right) - \sigma\left(z^{k+1}\right).$$

(C2) A subgradient lower bound for the iterates gap:

- $\left\{z^k\right\}_{k\in\mathbb{N}}$ is bounded.
- There exists a positive scalar ρ_2 such that

$$\left\|w^{k+1}\right\| \leq \rho_2 \left\|z^{k+1} - z^k\right\|, \; w^{k+1} \in \partial\sigma\left(z^{k+1}\right).$$

Theorem 2 (Bolte-Sabach-Teboulle (2014)).

Let $\sigma : \mathbb{R}^d \to (-\infty, \infty]$ be a proper, lower semicontinuous and semi-algebraic function with $\inf \sigma > -\infty$, and assume that $\left\{z^k\right\}_{k\in\mathbb{N}}$ is a gradient-like descent sequence for σ. If $\omega\left(z^0\right) \subset \mathrm{crit}(\sigma)$ then the sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ convergences to a critical point z^* of σ.

Semi-Algebraic Functions

Definition 3.

(i) A subset S of $\mathbb{R}^n$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{ u \in \mathbb{R}^n : \; g_{ij}(u) = 0 \text{ and } \; h_{ij}(u) < 0 \right\}.$$

(ii) A function $\sigma : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called semi-algebraic if its graph

$$\left\{ (u, t) \in \mathbb{R}^{n+1} : \; \sigma(u) = t \right\}$$

is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

Semi-Algebraic Functions

Definition 3.

(i) A subset S of $\mathbb{R}^n$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{ u \in \mathbb{R}^n : \; g_{ij}(u) = 0 \text{ and } h_{ij}(u) < 0 \right\}.$$

(ii) A function $\sigma : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called semi-algebraic if its graph

$$\left\{ (u, t) \in \mathbb{R}^{n+1} : \; \sigma(u) = t \right\}$$

is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

A wealth of semi-algebraic functions:

- Real polynomial functions.

Semi-Algebraic Functions

Definition 3.

(i) A subset S of $\mathbb{R}^n$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{ u \in \mathbb{R}^n : \; g_{ij}(u) = 0 \text{ and } h_{ij}(u) < 0 \right\}.$$

(ii) A function $\sigma : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called semi-algebraic if its graph

$$\left\{ (u, t) \in \mathbb{R}^{n+1} : \; \sigma(u) = t \right\}$$

is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

A wealth of semi-algebraic functions:

- Real polynomial functions.
- Indicator functions of semi-algebraic sets.

Semi-Algebraic Functions

Definition 3.

(i) A subset S of $\mathbb{R}^n$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{ u \in \mathbb{R}^n : \; g_{ij}(u) = 0 \text{ and } h_{ij}(u) < 0 \right\}.$$

(ii) A function $\sigma : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called semi-algebraic if its graph

$$\left\{ (u, t) \in \mathbb{R}^{n+1} : \; \sigma(u) = t \right\}$$

is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

A wealth of semi-algebraic functions:

- Real polynomial functions.
- Indicator functions of semi-algebraic sets.
- Finite sums and product of semi-algebraic functions.

Semi-Algebraic Functions

Definition 3.

(i) A subset S of $\mathbb{R}^n$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{ u \in \mathbb{R}^n : \; g_{ij}(u) = 0 \text{ and } h_{ij}(u) < 0 \right\}.$$

(ii) A function $\sigma : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called semi-algebraic if its graph

$$\left\{ (u, t) \in \mathbb{R}^{n+1} : \; \sigma(u) = t \right\}$$

is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

A wealth of semi-algebraic functions:

- Real polynomial functions.
- Indicator functions of semi-algebraic sets.
- Finite sums and product of semi-algebraic functions.
- Composition of semi-algebraic functions.

KPALM Analysis - $H \in C^1$

- Semi-Algebraicity: H is a weighted sum of squared Euclidean norms hence semi-algebraic, and Δ is semi-algebraic set, thus $\delta_\Delta(\cdot)$ is semi-algebraic, and in turn σ since it is sum of these functions.

KPALM Analysis - $H \in C^1$

- Semi-Algebraicity: H is a weighted sum of squared Euclidean norms hence semi-algebraic, and Δ is semi-algebraic set, thus $\delta_\Delta(\cdot)$ is semi-algebraic, and in turn σ since it is sum of these functions.
- Boundedness: $w^i(t) \in \Delta$ and

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

KPALM Analysis - $H \in C^1$

- Semi-Algebraicity: H is a weighted sum of squared Euclidean norms hence semi-algebraic, and Δ is semi-algebraic set, thus $\delta_\Delta(\cdot)$ is semi-algebraic, and in turn σ since it is sum of these functions.
- Boundedness: $w^i(t) \in \Delta$ and

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

KPALM Analysis - $H \in C^1$

- Semi-Algebraicity: H is a weighted sum of squared Euclidean norms hence semi-algebraic, and Δ is semi-algebraic set, thus $\delta_\Delta(\cdot)$ is semi-algebraic, and in turn σ since it is sum of these functions.
- Boundedness: $w^i(t) \in \Delta$ and

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

Assumption 2.

(i) *The chosen sequences of parameters* $\{\alpha_i(t)\}_{t \in \mathbb{N}}$, $1 \leq i \leq m$, *are bounded:* $0 < \underline{\alpha_i} \leq \alpha_i(t) \leq \overline{\alpha_i} < \infty, \quad \forall\, t \in \mathbb{N}.$

(ii) *For all* $t \in \mathbb{N}$ *there exists* $\underline{\beta} > 0$ *such that* $2 \min_{1 \leq l \leq k} \sum_{i=1}^m w_l^i(t) := \beta(w(t)) \geq \underline{\beta}$.

KPALM Analysis - $H \in C^1$

- Semi-Algebraicity: H is a weighted sum of squared Euclidean norms hence semi-algebraic, and Δ is semi-algebraic set, thus $\delta_\Delta(\cdot)$ is semi-algebraic, and in turn σ since it is sum of these functions.
- Boundedness: $w^i(t) \in \Delta$ and

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

Assumption 2.

(i) *The chosen sequences of parameters* $\{\alpha_i(t)\}_{t\in\mathbb{N}}$, $1 \le i \le m$, *are bounded:* $0 < \underline{\alpha_i} \le \alpha_i(t) \le \overline{\alpha_i} < \infty, \quad \forall\, t \in \mathbb{N}$.

(ii) *For all* $t \in \mathbb{N}$ *there exists* $\underline{\beta} > 0$ *such that* $2 \min_{1\le l\le k} \sum_{i=1}^m w_l^i(t) := \beta(w(t)) \ge \underline{\beta}$.

Denote $\underline{\alpha} = \min_{1\le i\le m} \underline{\alpha_i}$ and $\overline{\alpha} = \max_{1\le i\le m} \overline{\alpha_i}$.

Global Convergence of KPALM

Suppose that Assumption 2 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM.

To apply the general methodology of [BST] we proved the following results:

Global Convergence of KPALM

Suppose that Assumption 2 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM.

To apply the general methodology of [BST] we proved the following results:

Proposition 1 (Sufficient decrease property).

There exists $\rho_1 > 0$ *such that*

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

Global Convergence of KPALM

Suppose that Assumption 2 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM.

To apply the general methodology of [BST] we proved the following results:

Proposition 1 (Sufficient decrease property).

There exists $\rho_1 > 0$ such that

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

Proposition 2 (Subgradient lower bound for the iterates gap).

For each $t \in \mathbb{N}$ define

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

Then $\gamma(t) \in \partial\sigma(z(t))$ and there exists $\rho_2 > 0$ such that

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

Global Convergence of KPALM

Suppose that Assumption 2 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by KPALM.

To apply the general methodology of [BST] we proved the following results:

Proposition 1 (Sufficient decrease property).

There exists $\rho_1 > 0$ *such that*

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

Proposition 2 (Subgradient lower bound for the iterates gap).

For each $t \in \mathbb{N}$ *define*

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

Then $\gamma(t) \in \partial\sigma(z(t))$ *and there exists* $\rho_2 > 0$ *such that*

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

Theorem 4 (KPALM convergence result).

The sequence $\{z(t)\}_{t\in\mathbb{N}}$ *converges to a critical point of* σ.

ε-KPALM Algorithm for Clustering with the Norm Distance

Recalling the Clustering Problem for the Euclidean norm distance:

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

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

- <u>H is not smooth, hence PALM cannot be applied as is</u>. Therefore approximating:

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

ε-KPALM Algorithm for Clustering with the Norm Distance

Recalling the Clustering Problem for the Euclidean norm distance:

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

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

- H is not smooth, hence PALM cannot be applied as is. Therefore approximating:

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

- This leads towards the following approximation of the Clustering Problem

$$(P_\varepsilon) \qquad min\left\{\sigma_\varepsilon(z) := H_\varepsilon(w,x) + G(w) \mid z := (w,x) \in \mathbb{R}^{km} \times \mathbb{R}^{nk}\right\}.$$

ε-KPALM Algorithm for Clustering with the Norm Distance

Recalling the Clustering Problem for the Euclidean norm distance:

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

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

- H is not smooth, hence PALM cannot be applied as is. Therefore approximating:

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

- This leads towards the following approximation of the Clustering Problem

$$(P_\varepsilon) \qquad min\left\{\sigma_\varepsilon(z) := H_\varepsilon(w,x) + G(w) \mid z := (w,x) \in \mathbb{R}^{km} \times \mathbb{R}^{nk}\right\}.$$

- Also $d_\varepsilon^i(x)$ is the smoothed version of $d^i(x)$, whose $1 \le l \le k$ coordinate is $\left(\|x^l - a^i\|^2 + \varepsilon^2\right)^{1/2}$.

ε-KPALM Algorithm for Clustering with the Norm Distance

- With respect to w we apply the same step as in KPALM:

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

ε-KPALM Algorithm for Clustering with the Norm Distance

- With respect to w we apply the same step as in KPALM:

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

- With respect to x we tackle the subproblem differently than in KPALM, linearizing the function $x \to H(w, \cdot)$, for fixed w, and adding a regularizing term

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

where

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

ε-KPALM Algorithm for Clustering with the Norm Distance

- With respect to w we apply the same step as in KPALM:

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

- With respect to x we tackle the subproblem differently than in KPALM, linearizing the function $x \to H(w, \cdot)$, for fixed w, and adding a regularizing term

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

where

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

ε-KPALM Algorithm for Clustering with the Norm Distance

- With respect to w we apply the same step as in KPALM:

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

- With respect to x we tackle the subproblem differently than in KPALM, linearizing the function $x \to H(w, \cdot)$, for fixed w, and adding a 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

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

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

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

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

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

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

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

ε-KPALM Analysis - Using an Auxiliary Function

- The situation is more difficult, since $H \in C^1$ but not in $C^{1,1}$ for fixed x or w.

ε-KPALM Analysis - Using an Auxiliary Function

- The situation is more difficult, since $H \in C^1$ but not in $C^{1,1}$ for fixed x or w.
- We cannot obtain directly the descent property.

ε-KPALM Analysis - Using an Auxiliary Function

- The situation is more difficult, since $H \in C^1$ but not in $C^{1,1}$ for fixed x or w.
- We cannot obtain directly the descent property.
- We need to use an auxiliary function and as shown in the thesis we proposed two approaches.

ε-KPALM Analysis - Using an Auxiliary Function

- The situation is more difficult, since $H \in C^1$ but not in $C^{1,1}$ for fixed x or w.
- We cannot obtain directly the descent property.
- We need to use an auxiliary function and as shown in the thesis we proposed two approaches.

ε-KPALM Analysis - Using an Auxiliary Function

- The situation is more difficult, since $H \in C^1$ but not in $C^{1,1}$ for fixed x or w.
- We cannot obtain directly the descent property.
- We need to use an auxiliary function and as shown in the thesis we proposed two approaches.

Motivated by the recent work on Weber problem (Beck-Sabach(2015)), we develop several auxiliary results. We define the function $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ by

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

$v^i \in \mathbb{R}_+$ fixed weights. We also need the auxiliary function $h_\varepsilon : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

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

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

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

Key Properties for Sufficient Decrease and Subgradient Bound Proofs

Lemma 5.

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

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

Key Properties for Sufficient Decrease and Subgradient Bound Proofs

Lemma 5.

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

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

Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM.

Proposition 3 (Bounds for L_ε^l).

(i) *Denote* $d_{\mathcal{A}} = \mathrm{diam}(Conv(\mathcal{A}))$. *For all* $t \in \mathbb{N}$ *and* $l = 1, 2, \ldots, k$ *we have*

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

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

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

Key Properties for Sufficient Decrease and Subgradient Bound Proofs

Lemma 5.

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

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

Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM.

Proposition 3 (Bounds for L_ε^l).

(i) *Denote $d_{\mathcal{A}} = \mathrm{diam}(Conv(\mathcal{A}))$. For all $t \in \mathbb{N}$ and $l = 1, 2, \ldots, k$ we have*

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

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

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

Proposition 4 (Sufficient Decrease w.r.t. x).

For all $t \in \mathbb{N}$, we have

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

Global Convergence of ε-KPALM

Suppose that Assumption 2 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM.

Proposition 5 (Sufficient decrease property).

There exists $\rho_1 > 0$ *such that*

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

Global Convergence of ε-KPALM

Suppose that Assumption 2 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM.

Proposition 5 (Sufficient decrease property).

There exists $\rho_1 > 0$ *such that*

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

Proposition 6 (Subgradient lower bound for the iterates gap).

For each $t \in \mathbb{N}$ *define*

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

Then $\gamma(t) \in \partial\sigma_\varepsilon(z(t))$ *and there exists* $\rho_2 > 0$ *such that*

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

Global Convergence of ε-KPALM

Suppose that Assumption 2 holds true and let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM.

Proposition 5 (Sufficient decrease property).

There exists $\rho_1 > 0$ such that

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

Proposition 6 (Subgradient lower bound for the iterates gap).

For each $t \in \mathbb{N}$ define

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

Then $\gamma(t) \in \partial\sigma_\varepsilon(z(t))$ and there exists $\rho_2 > 0$ such that

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

Theorem 6 (ε-KPALM convergence result).

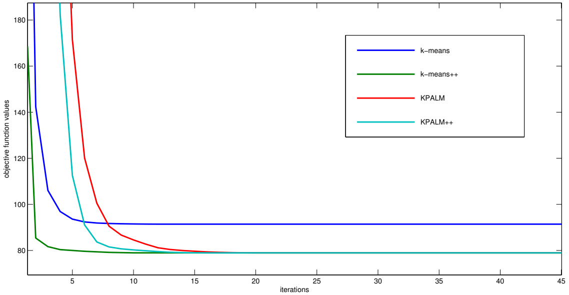
The sequence $\{z(t)\}_{t\in\mathbb{N}}$ converges to a critical point of σ_ε.

Numerical Results: The Squared Norm

- Initialization issues: randomly picking k data points as starting centers vs. k-means++.

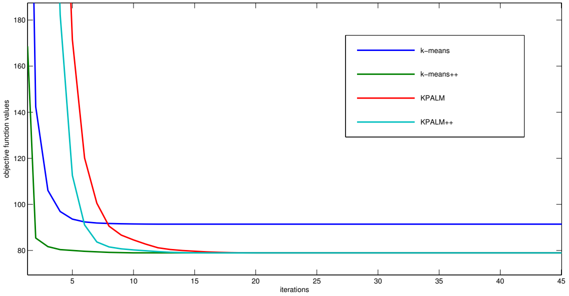
Numerical Results: The Squared Norm

- Initialization issues: randomly picking *k* data points as starting centers vs. k-means++.
- Comparison of the objective function values performed on the Iris dataset for squared norm algorithms



Numerical Results: The Squared Norm

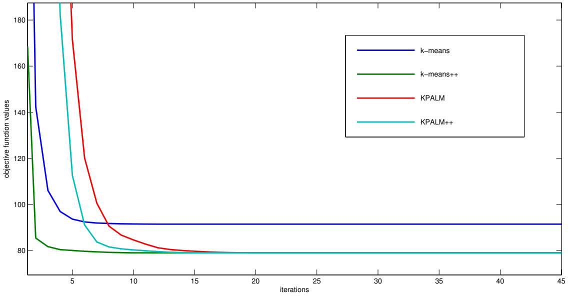
- Initialization issues: randomly picking *k* data points as starting centers vs. k-means++.
- Comparison of the objective function values performed on the Iris dataset for squared norm algorithms



- In the squared Euclidean setting, KPALM achieves lower objective function values than k-means. When using a more sophisticated initialization step, such as the one in k-means++, then k-means++ and KPALM++ achieve similar objective function values.

Numerical Results: The Squared Norm

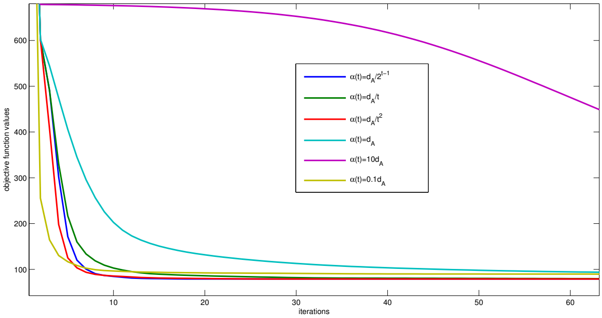
- Initialization issues: randomly picking *k* data points as starting centers vs. k-means++.
- Comparison of the objective function values performed on the Iris dataset for squared norm algorithms



- In the squared Euclidean setting, KPALM achieves lower objective function values than k-means. When using a more sophisticated initialization step, such as the one in k-means++, then k-means++ and KPALM++ achieve similar objective function values.
- k-means needs less number of iterations then KPALM to reach a certain precision.

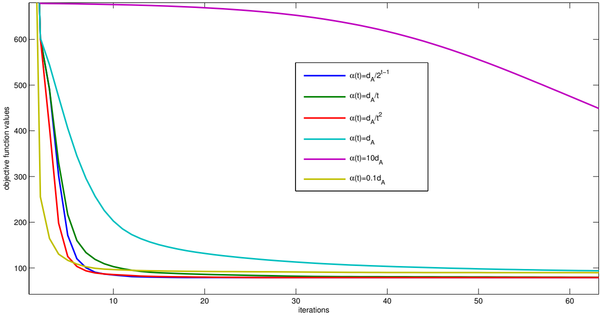
Numerical Results: The Squared Norm

- Comparison of the objective function values performed on the Iris dataset for KPALM algorithm with different α parameter updates.



Numerical Results: The Squared Norm

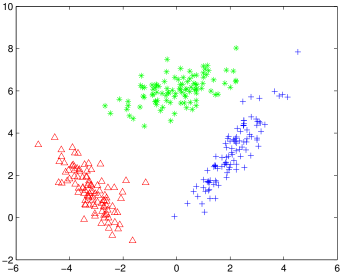
- Comparison of the objective function values performed on the Iris dataset for KPALM algorithm with different α parameter updates.



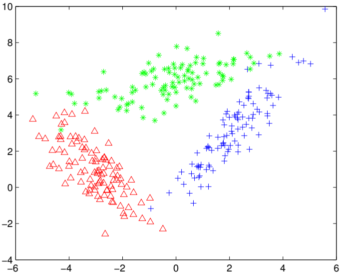
- It is preferable to use dynamic update of $\alpha(t)$ parameter to achieve a faster convergence, both in KPALM and ϵ-KPALM. Example for suitable choices can be $\alpha(t) = \mathrm{diam}(\mathcal{A})/t$ and $\alpha(t) = \mathrm{diam}(\mathcal{A})/2^t$.

Numerical Results: Squared Norm Algorithms vs. ε-KPALM

- Generated two synthetic datasets, each 300 points



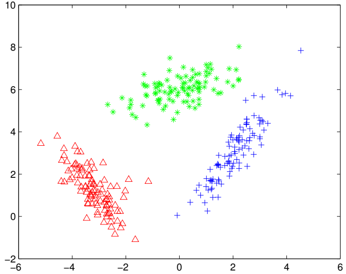
(a) Dense Gaussians



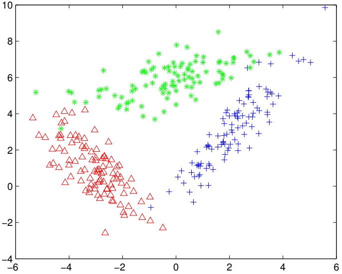
(b) Sparse Gaussians

Numerical Results: Squared Norm Algorithms vs. ε-KPALM

- Generated two synthetic datasets, each 300 points



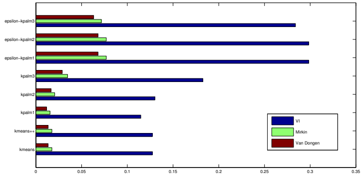
(a) Dense Gaussians



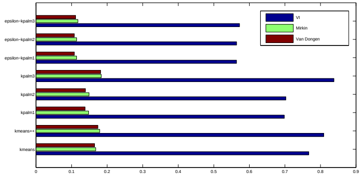
(b) Sparse Gaussians

- Compared the resulting clustering using metrics, such as Variation of Information defined by: $VI(\mathcal{C}^1, \mathcal{C}^2) := -\sum_{i,j} r_{i,j}\left[\log(r_{i,j}/p_i) + \log(r_{i,j}/q_j)\right]$, where $\mathcal{C}^1 = \left\{C_1^1, C_2^1, \ldots C_k^1\right\}$ and $\mathcal{C}^2 = \left\{C_1^2, C_2^2, \ldots C_l^2\right\}$ are two clusterings of $\mathcal{A}$, and $m = |\mathcal{A}|, \quad p_i = |C_i^1|/m, \quad q_j = |C_j^2|/m, \quad r_{i,j} = |C_i^1 \cap C_j^2|/m.$

Numerical Results: Squared Norm Algorithms vs. ε-KPALM

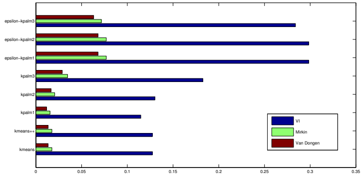


(a) Dense Gaussians metrics comparison

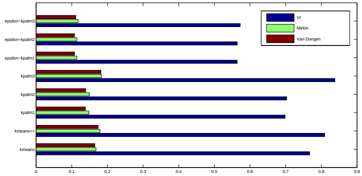


(b) Sparse Gaussians metrics comparison

Numerical Results: Squared Norm Algorithms vs. ε-KPALM



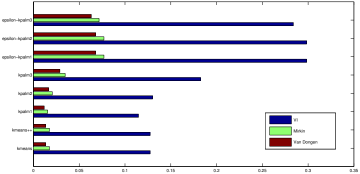
(a) Dense Gaussians metrics comparison



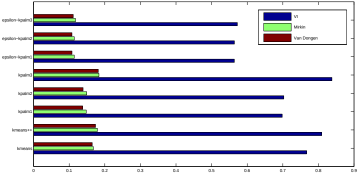
(b) Sparse Gaussians metrics comparison

- When the convex hulls of the desired clusters are mutually exclusive, algorithms which solve the clustering problem with the squared Euclidean distance are preferable to ε-KPALM.

Numerical Results: Squared Norm Algorithms vs. ε-KPALM



(a) Dense Gaussians metrics comparison



(b) Sparse Gaussians metrics comparison

- When the convex hulls of the desired clusters are mutually exclusive, algorithms which solve the clustering problem with the squared Euclidean distance are preferable to ε-KPALM.
- In datasets with outliers, the clustering obtained with ε-KPALM is more similar to the desired clustering, in terms of clustering metrics, than the clusterings obtained via the squared Euclidean algorithms. Therefore, as expected, for data with outliers, the choice of a norm instead of the squared norm is a more natural choice, and the ε-KPALM algorithm appears to be a promising algorithm to handle such data.

Thank you
for your attention!

KPALM Sufficient Decrease Proof

Since $x \mapsto H(w, x) = \sum_{l=1}^{k} \sum_{i=1}^{m} w_l^i \left\| x^l - a^i \right\|^2$ is C^2, and it Hessian is given by

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

then it is strongly convex with parameter $\beta(w)$, whenever $\beta(w) = 2 \min_{1 \leq l \leq k} \sum_{i=1}^{m} w_l^i > 0$.

KPALM Sufficient Decrease Proof

Since $x \mapsto H(w,x) = \sum_{l=1}^{k} \sum_{i=1}^{m} w_l^i \left\| x^l - a^i \right\|^2$ is C^2, and it Hessian is given by

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

then it is strongly convex with parameter $\beta(w)$, whenever $\beta(w) = 2 \min_{1 \leq l \leq k} \sum_{i=1}^{m} w_l^i > 0$.

Assumption 2(ii) ensures that $x \mapsto H(w(t),x)$ is strongly convex with parameter $\beta(w(t))$, hence

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

where the equality follows from $\nabla_x H(w(t+1),x(t+1)) = 0$.

KPALM Sufficient Decrease Proof-Contd.

From the w update step we derive

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

Summing the last inequality over all $1 \leq i \leq m$ yields

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

KPALM Sufficient Decrease Proof-Contd.

From the w update step we derive

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

Summing the last inequality over all $1 \leq i \leq m$ yields

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

Set $\rho_1 = \frac{1}{2}\min\left\{\underline{\alpha}, \underline{\beta}\right\}$, by combining the sufficient decrease in x and w variables we get

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

KPALM Subgradient Lower Bound The Iterates Gap Proof

Recalling the steps of KPLAM algorithm:

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

$$x(t+1) = \text{argmin} \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\} \tag{2}$$

KPALM Subgradient Lower Bound The Iterates Gap Proof

Recalling the steps of KPLAM algorithm:

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

$$x(t+1) = \operatorname{argmin} \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\} \tag{2}$$

The subgradient of σ is given by

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

KPALM Subgradient Lower Bound The Iterates Gap Proof

Recalling the steps of KPLAM algorithm:

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

$$x(t+1) = \operatorname{argmin} \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\} \tag{2}$$

The subgradient of σ is given by

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

Evaluating the last relation at $z(t+1)$ and using (2)

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

KPALM Subgradient Lower Bound The Iterates Gap Proof

Recalling the steps of KPLAM algorithm:

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

$$x(t+1) = \operatorname{argmin} \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\} \tag{2}$$

The subgradient of σ is given by

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

Evaluating the last relation at $z(t+1)$ and using (2)

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

The optimality condition of $w^i(t+1)$ (see (1)), implies that there exists $u^i(t+1) \in \partial \delta_\Delta(w^i(t+1))$ such that

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

KPALM Subgradient Lower Bound The Iterates Gap Proof-Contd.

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

KPALM Subgradient Lower Bound The Iterates Gap Proof-Contd.

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

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

where the third inequality follows from the inequality

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

with $M = \max_{1 \leq i \leq m} \|a^i\|$ and the result follows with $\rho_2 = m(4M + \overline{\alpha})$.

Properties of the Auxiliary Functions h_ε and f_ε

Lemma 7.

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

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

Properties of the Auxiliary Functions h_ε and f_ε

Lemma 7.

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

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

Lemma 8.

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

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

Preliminaries for Sufficient Decrease Proof

Proposition (Bounds for L_ε^l).

Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM.

(i) *Denote $d_{\mathcal{A}} = \mathrm{diam}(Conv(\mathcal{A}))$. For all $t \in \mathbb{N}$ and $l = 1, 2, \ldots, k$ we have*

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

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

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

Preliminaries for Sufficient Decrease Proof

Proposition (Bounds for L_ε^l).

Let $\{z(t)\}_{t\in\mathbb{N}}$ be the sequence generated by ε-KPALM.

(i) *Denote $d_{\mathcal{A}} = \mathrm{diam}(Conv(\mathcal{A}))$. For all $t \in \mathbb{N}$ and $l = 1, 2, \ldots, k$ we have*

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

(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) Using $\underline{\beta}$ as in Assumption 2(ii) and the fact that $x^l(t) \in Conv(\mathcal{A})$, we obtain

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

where the first inequality follows from the fact that $\left\|x^l(t) - a^i\right\| \leq d_{\mathcal{A}}$.

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

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

Preliminaries for Sufficient Decrease Proof

Proposition (Sufficient Decrease w.r.t. x).

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

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

Preliminaries for Sufficient Decrease Proof

Proposition (Sufficient Decrease w.r.t. x).

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.

By definition of f_ε, using the weights $v_i = w_l^i(t+1)$, we obtain

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

Therefore, by applying Lemma 7(iii) with $x = x^l(t+1)$ and $y = x^l(t)$, we get

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

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

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

The result follows by replacing the last term with the following compact form

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

ε-KPALM Sufficient Decrease Proof

With respect to w ε-KPALM sufficient decrease is similar to that of KPALM, which yields

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

Applying Proposition 4 with the center update step in ε-KPALM we get for all $t \in \mathbb{N}$ that

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

where the second inequality follows from Proposition 3(i).

ε-KPALM Sufficient Decrease Proof

With respect to w ε-KPALM sufficient decrease is similar to that of KPALM, which yields

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

Applying Proposition 4 with the center update step in ε-KPALM we get for all $t \in \mathbb{N}$ that

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

where the second inequality follows from Proposition 3(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 (1) and (2) yields

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

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof

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

$$\gamma(t+1) := \left(\left(d_\varepsilon^i(x(t+1)) + u^i(t+1) \right)_{i=1,\ldots,m}, \nabla_x H_\varepsilon(w(t+1), x(t+1)) \right) \in \partial \sigma_\varepsilon(z(t+1)), \qquad (3)$$

where $u^i(t+1) \in \partial \delta_\Delta(w^i(t+1)),\ i = 1, 2, \ldots, m$.

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof

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

$$\gamma(t+1) := \left(\left(d_\varepsilon^i(x(t+1)) + u^i(t+1) \right)_{i=1,\ldots,m}, \nabla_x H_\varepsilon(w(t+1), x(t+1)) \right) \in \partial \sigma_\varepsilon(z(t+1)), \tag{3}$$

where $u^i(t+1) \in \partial \delta_\Delta(w^i(t+1)),\ i = 1, 2, \ldots, m$.
Writing the optimality condition of the cluster assignment step yields

$$d_\varepsilon^i(x(t)) + \alpha_i(t) \left(w^i(t+1) - w^i(t) \right) + u^i(t+1) = \mathbf{0}. \tag{4}$$

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof

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

$$\gamma(t+1) := \left(\left(d_\varepsilon^i(x(t+1)) + u^i(t+1) \right)_{i=1,\ldots,m}, \nabla_x H_\varepsilon(w(t+1), x(t+1)) \right) \in \partial\sigma_\varepsilon(z(t+1)), \tag{3}$$

where $u^i(t+1) \in \partial\delta_\Delta(w^i(t+1)),\ i = 1, 2, \ldots, m$.
Writing the optimality condition of the cluster assignment step yields

$$d_\varepsilon^i(x(t)) + \alpha_i(t) \left(w^i(t+1) - w^i(t) \right) + u^i(t+1) = \mathbf{0}. \tag{4}$$

Plugging (4) into (3), and taking the norm yields

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

where the last inequality follows from the inequality

$$\|d_\varepsilon^i(x) - d_\varepsilon^i(y)\| \leq \frac{d_{\mathcal{A}}}{\varepsilon} \|x - y\|, \text{ whenever } x^l, y^l \in Conv(\mathcal{A}) \quad \forall\, 1 \leq l \leq k,$$

and the fact that $\overline{\alpha} = \max_{1 \leq i \leq m} \overline{\alpha_i}$.

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof-Contd.

It is sufficient to show that $\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| \leq c\|x(t+1) - x(t)\|$, for some constant $c > 0$.

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof-Contd.

It is sufficient to show that $\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| \leq c\|x(t+1) - x(t)\|$, for some constant $c > 0$. First, note that for fixed $w \in \Delta^m$ and any $x \in \mathbb{R}^{nk}$ the following relation holds

$$\nabla_{x^l} H_\varepsilon(w, x) = \nabla f_\varepsilon\left(x^l\right), \quad \forall\, l = 1, 2, \ldots, k. \tag{5}$$

Now, for all $l = 1, 2, \ldots, k$, we have

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

where the last equality follows from center update step.

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof-Contd.

It is sufficient to show that $\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| \leq c\|x(t+1) - x(t)\|$, for some constant $c > 0$. First, note that for fixed $w \in \Delta^m$ and any $x \in \mathbb{R}^{nk}$ the following relation holds

$$\nabla_{x^l} H_\varepsilon(w, x) = \nabla f_\varepsilon\left(x^l\right), \quad \forall\, l = 1, 2, \ldots, k. \tag{5}$$

Now, for all $l = 1, 2, \ldots, k$, we have

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

where the last equality follows from center update step.Therefore,

$$\begin{aligned}
\|\nabla_x H_\varepsilon(w(t+1), x(t+1))\| &\leq \sum_{l=1}^{k} \|\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1))\| \qquad (6) \\
&\leq \sum_{l=1}^{k} L_\varepsilon^l(w(t+1), x(t))\|x^l(t+1) - x^l(t)\| + \sum_{l=1}^{k} \|\nabla_{x^l} H_\varepsilon(w(t+1), x(t+1)) - \nabla_{x^l} H_\varepsilon(w(t+1), x(t))\| \\
&\leq \frac{m}{\varepsilon} \sum_{l=1}^{k} \|x^l(t+1) - x^l(t)\| + \sum_{l=1}^{k} \kappa^l(t)\|x^l(t+1) - x^l(t)\|,
\end{aligned}$$

where the last inequality follows from Proposition 3(ii) and Lemma 8 combined with the (5) observation using

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

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof-Contd.

From Proposition 3(ii) we obtain that

$$\kappa^l(t) = \frac{2}{\frac{1}{L^l_\varepsilon(w(t+1),x(t))} + \frac{1}{L^l_\varepsilon(w(t+1),x(t+1))}} \leq \frac{2}{\frac{\varepsilon}{m} + \frac{\varepsilon}{m}} = \frac{m}{\varepsilon}.$$

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof-Contd.

From Proposition 3(ii) we obtain that

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

Hence, from (6), we have

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

ε-KPALM Subgradient Lower Bound The Iterates Gap Proof-Contd.

From Proposition 3(ii) we obtain that

$$\kappa^l(t) = \frac{2}{\frac{1}{L^l_\varepsilon(w(t+1),x(t))} + \frac{1}{L^l_\varepsilon(w(t+1),x(t+1))}} \leq \frac{2}{\frac{\varepsilon}{m} + \frac{\varepsilon}{m}} = \frac{m}{\varepsilon}.$$

Hence, from (6), we have

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

Therefore, setting $\rho_2 = \frac{md_{\mathcal{A}}}{\varepsilon} + \overline{\alpha}\sqrt{m} + \frac{2m\sqrt{k}}{\varepsilon}$, yields the final result

$$\|\gamma(t+1)\| \leq \rho_2 \|z(t+1) - z(t)\| \text{ with } \gamma(t+1) \in \partial\sigma_\varepsilon(z(t+1)).$$