# A Novel Class of Globally Convergent Algorithms For Clustering Problems

Sergey Voldman

Raymond and Beverly Sackler Faculty of Exact Sciences

Tel-Aviv University

06.04.2016

Under the supervision of

Prof. Marc Teboulle (Tel Aviv University)
and Prof. Shoham Sabach (Technion)

# Goal and Outline

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

## Outline

- Introduction to the clustering problem.
- Introduction to the convergence methodology.
- Clustering with the squared Euclidean norm: KPALM algorithm and its analysis.
- Clustering with the Euclidean norm: $\varepsilon$-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 focused a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- It has been shown that 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

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

with $d(\cdot, \cdot)$ being a distance-like function, such as the squared Euclidean norm.

# 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 $\Delta$ 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
$$(P) \qquad min\left\{\Psi(z) := H(w,x) + G(w) \mid z := (w,x) \in \mathbb{R}^{km}\times\mathbb{R}^{nk}\right\},$$
with $H(w,x) = \sum_{i=1}^{m} H^i(w,x) = \sum_{i=1}^{m}\langle w^i, d^i(x)\rangle$ and $G(w) = \sum_{i=1}^{m} G^i(w^i) = \sum_{i=1}^{m}\delta_\Delta(w^i)$.

# Convergence Methodology: Definitions

**Definition (Limiting Subdifferential $\partial\sigma(x)$)**

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.
The (Limiting) Subdifferential $\partial\sigma(x)$ is defined via:

$$\begin{aligned} u^* \in \partial\sigma(x) \quad & \text{iff} \quad (x_k, u_k^*) \to (x, u^*) \text{ s.t. } \sigma(x_k) \to \sigma(x) \text{ and} \\ \sigma(u) \quad & \geq \quad \sigma(x_k) + \langle u_k^*, u - x_k \rangle + o(\|u - x_k\|) \end{aligned}$$

- $x \in \mathbb{R}^d$ is a critical point of $\sigma$ if $\partial\sigma(x) \ni 0$.
- The set of critical points of $\sigma \equiv \text{crit}\,\sigma$.
- $r \in \mathbb{R}$ is a critical value if $\exists x \in \text{crit}\,\sigma : \sigma(x) = r$.

# KL property

Denote the following class of concave functions

$$\Phi_\eta = \left\{ \varphi \in C\left([0,\eta), \mathbb{R}_+\right) : \ \varphi \in C^1\left((0,\eta)\right), \ \varphi' > 0, \ \varphi(0) = 0 \right\}.$$

## Definition (Kurdyka-Łojasiewicz property)

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and lower semicontinuous.

(i) $\sigma$ admits the KL property at $\overline{u} \in \mathit{dom}\, \partial\sigma := \{u \in \mathbb{R}^d : \ \partial\sigma \neq \emptyset\}$ if there exist $\eta \in (0, +\infty]$, a neighborhood $U$ of $\overline{u}$ and a function $\varphi \in \Phi_\eta$, such that for all

$$u \in U \cap \left\{ x \in \mathbb{R}^d : \ \sigma(\overline{u}) < \sigma(x) < \sigma(\overline{u}) + \eta \right\},$$

the following inequality holds

$$\varphi'(\sigma(u) - \sigma(\overline{u}))\, \mathrm{dist}(0, \partial\sigma(u)) \geq 1,$$

where $\mathrm{dist}(x, S) := \inf\{\|y - x\| : \ y \in S\}$ denotes the distance from $x \in \mathbb{R}^d$ to $S \subset \mathbb{R}^d$.

(ii) If $\sigma$ satisfy the KL property at each point of *dom* $\sigma$ then $\sigma$ is called a *KL function*.

# Semi-Algebraic Functions

## Theorem (Bolte-Daniilidis-Lewis (2006))

*Let $\sigma : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper and lsc function. If $\sigma$ is semi-algebraic then it satisfy the KL property at any point of $\operatorname{dom} \sigma$.*

## Definition

(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 $\mathbb{R}^{n+1}$.

# The 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.
- Sup/Inf type function, *e.g.*, $\sup \{g(u, v) : \ v \in C\}$ is semi-algebraic when $g$ is a semi-algebraic function and $C$ a semi-algebraic set.
- The function $x \to \mathrm{dist}(x, S)^2$ is semi-algebraic whenever $S$ is a nonempty semi-algebraic subset of $\mathbb{R}^n$.
- $\|\cdot\|_0$ (counts the non-zero values) is semi-algebraic.
- $\|\cdot\|_p$ is semi-algebraic whenever $p > 0$ is rational.

In particular, for distance-like functions $d(x, y) = \|x - y\|^2$ and $d(x, y) = \|x - y\|$ the resulting clustering function defined in (P) is semi-algebraic.

# Gradient-Like Descent Sequence

## Definition

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

(C1) Sufficient decrease property: There exists a positive scalar $\rho_1$ 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:

- $\{z^k\}_{k\in\mathbb{N}}$ is bounded.
- There exists a positive scalar $\rho_2$ 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 (Bolte-Shabach-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 $\{z^k\}_{k\in\mathbb{N}}$ is a gradient-like descent sequence for $\sigma$. If $\omega\left(z^0\right) \subset crit(\sigma)$ then the sequence $\{z^k\}_{k\in\mathbb{N}}$ convergences to a critical point $z^*$ of $\sigma$.*

# The Optimization Model

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

## Assumption

(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 choice of two blocks of variables is ONLY for the sake of simplicity.
- The optimization model (M) covers many applications: signal/image processing, machine learning, etc....

## Building the Algorithm

- Simplest Approach: Alternating Minimization (AM)

$$x^{k+1} \in \operatorname{argmin}_x \Psi\left(x, y^k\right); \qquad y^{k+1} \in \operatorname{argmin}_y \Psi\left(x^{k+1}, y\right).$$

However, this scheme does not converge, unless very restrictive assumptions are made, such as the uniqueness of the minimizer in each step

- To overcome the above difficulty: Regularization with “prox"

$$x^{k+1} \in \operatorname{argmin}_x \left\{\Psi\left(x, y^k\right) + \frac{c_k}{2}\left\|x - x^k\right\|^2\right\},$$

$$y^{k+1} \in \operatorname{argmin}_y \left\{\Psi\left(x^{k+1}, y\right) + \frac{d_k}{2}\left\|y - y^k\right\|^2\right\}.$$

- However, the above scheme is only “conceptual" in the sense that
  1. It requires solving excatly two nonconvex difficult problems (does not exploit special structure/data info of $f, g$ and $H$.)
  2. Involves *Nested* optimization...Needs an optimal $x$ to proceed computation of $y$!

To overcome all the above difficulties, we take one further very simple step which exploits the data information on $H$.

# The Proximal-Forward Backward Scheme/Proximal Gradient

Suitable for the composite smooth + nonsmooth model:

$$\min\left\{h(u)+\sigma(u):\ u\in\mathbb{R}^d\right\},\quad h\in C^{1,1}$$

$$u^{k+1}\in\operatorname{argmin}_{u\in\mathbb{R}^d}\left\{\left\langle u-u^k,\nabla h\left(u^k\right)\right\rangle+\frac{t}{2}\left\|u-u^k\right\|^2+\sigma(u)\right\}.$$

- Useful for smooth and "simple" (i.e., easy prox) nonsmooth $\sigma(\cdot)$. Convex case well studied, convergence and complexity [Lions-Mercier (79), Nesterov (07), Beck-Teboulle (09)]
- Nonconvex case: Convergence of the whole sequence to a critical point! Very recent in [Attouch-Bolte-Svaiter (11)].

Can we preserve the qualities of both building blocks: Simplicity of AM and Global Convergence of PFB to tackle our general model (M)?

# The Algorithm: Proximal Altertnating Linearization Minimization (PALM)

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:

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

Replacing $\Psi$ in AM scheme with its first-order approximation in each block, given by:

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

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

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

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

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

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

**Main computational step:** Computing prox of a nonconvex function.

# Proximal Map for Nonconvex Functions

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:

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

## Proposition (Rockafellar and Wets)

*If* $\inf_{\mathbb{R}^n} \sigma > -\infty$*, then, for every* $t \in (0, \infty)$*, the set* $\operatorname{prox}_{1/t}^{\sigma}(x)$ *is nonempty and compact.*

- Here $\operatorname{prox}_t^{\sigma}$ is a set-valued map.
- When $\sigma := \delta_X$, the indicator function of a nonempty and closed set $X$, the proximal map reduces to the set-valued projection operator onto $X$.

# PALM is Well Defined

Thanks to the prox properties, since PALM is defined by two proximal computations, all we need is:

$$\text{Assume:} \quad \inf_{\mathbb{R}^n \times \mathbb{R}^m} \Psi > -\infty, \quad \inf_{\mathbb{R}^n} f > -\infty \quad \text{and} \quad \inf_{\mathbb{R}^m} g > -\infty.$$

$\boxed{\text{Thus Problem } (M) \text{ is inf-bounded and PALM is well defined!!!}}$

**Quick Recall on Nonsmooth Analysis – [Rockafellar-Wets (98)]**

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.

- (Limiting) Subdifferential $\partial\sigma(x)$:

$$\begin{aligned} u^* \in \partial\sigma(x) \quad &\text{iff} \quad (x_k, x_k^*) \to (x, x^*) \text{ s.t. } \sigma(x_k) \to \sigma(x) \text{ and} \\ \sigma(u) \quad &\geq \quad \sigma(x_k) + \langle x_k^*, u - x_k \rangle + o(\|u - x_k\|) \end{aligned}$$

- $x \in \mathbb{R}^d$ is a critical point of $\sigma$ if $\partial\sigma(x) \ni 0$.
- The set of critical points of $\sigma \equiv \text{crit}\, \sigma$.
- $r \in \mathbb{R}$ is a critical value if $\exists x \in \text{crit}\, \sigma : \sigma(x) = r$.

# An Informal General Convergence Proof Procedure

**Given:** Let $\Psi : \mathbb{R}^N \to \overline{\mathbb{R}}$ be a proper, lsc and bounded from below function.

$$(P) \qquad \inf \left\{ \Psi(z) : \; z \in \mathbb{R}^N \right\}.$$

Suppose $\mathcal{A}$ is a generic algorithm which generates a sequence $\{z^k\}_{k \in \mathbb{N}}$ via:

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

**Goal:** To prove that whole $\{z^k\}_{k \in \mathbb{N}}$ converges to a critical point of $\Psi$.

Basically, the methodology consists of three main steps.

(i) **Sufficient decrease property:** Find a positive constant $\rho_1$ such that

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

(ii) **A subgradient lower bound for the iterates gap:** Assume that $\{z^k\}_{k \in \mathbb{N}}$ is bounded. Find another positive constant $\rho_2$, such that

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

These two steps are typical for *any descent* type algorithms but lead ONLY to *convergence of limit points*.

# PALM: First Convergence Properties

From now on we assume that the generated sequence is bounded. Let $\{z^k\}_{k\in\mathbb{N}}$ $(z^k = (x^k, y^k))$ be a sequence generated by PALM.

The set of all limit points is denoted by $\omega(z^0)$, where $z^0$ is the starting point.

**Lemma (Properties of the limit point set $\omega(z^0)$)**

*Let $\{z^k\}_{k\in\mathbb{N}}$ be a sequence generated by PALM. Then*

(i) $\emptyset \neq \omega(z^0) \subset \operatorname{crit}\Psi$.

(ii) *We have*

$$\lim_{k\to\infty} \operatorname{dist}\left(z^k, \omega\left(z^0\right)\right) = 0.$$

(iii) $\omega(z^0)$ *is a nonempty, compact and connected set.*

(iv) *The objective function $\Psi$ is finite and constant on $\omega(z^0)$.*

$$\boxed{\text{Does the whole } \{z^k\}_{k\in\mathbb{N}} \text{ converge to a critical point of Problem } (M)?}$$

# YES! Using the Kurdyka-Łojasiewicz Property (Third Step)

(iii) **Using the KL property:** Assume that $\Psi$ is a KL function and show that the generated sequence $\{z^k\}_{k\in\mathbb{N}}$ is a *Cauchy sequence*.

If, $\Psi$ is a KL function, we get:

## Theorem (A finite length property)

*Let $\{z^k\}_{k\in\mathbb{N}}$ be a sequence generated by PALM. The following assertions hold.*

(i) *The sequence $\{z^k\}_{k\in\mathbb{N}}$ has finite length, that is,*

$$\sum_{k=1}^{\infty} \left\| z^{k+1} - z^k \right\| < \infty.$$

(ii) *The sequence $\{z^k\}_{k\in\mathbb{N}}$ converges to a critical point $z^* = (x^*, y^*)$.*

# Proof of Theorem

For simplicity, assume $\Psi^* = 0$ and that $\Psi$ is smooth. From sufficient decease and subgradient bound for iterates gaps properties we have

$$\exists \rho_1 > 0 : \ \rho_1 \left\| z^{k+1} - z^k \right\|^2 \leq \Psi\left(z^k\right) - \Psi\left(z^{k+1}\right) \tag{1}$$

$$\exists \rho_2 > 0 : \ \rho_2 \left\| \nabla\Psi\left(z^k\right) \right\| \leq \left\| z^{k+1} - z^k \right\| \tag{2}$$

Combining (1) with (2) yields

$$\rho_1\rho_2 \left\| z^{k+1} - z^k \right\| \cdot \left\| \nabla\Psi\left(z^k\right) \right\| \leq \Psi\left(z^k\right) - \Psi\left(z^{k+1}\right). \tag{3}$$

Since $\Psi$ is a KL-function there exists a concave desingularizing function $\varphi$ such that

$$\varphi'\left(\Psi\left(z^k\right)\right) \left\| \nabla\Psi\left(z^k\right) \right\| \geq 1 \tag{4}$$

From the concavity of $\varphi$ it follows that

$$\varphi(u) - \varphi(v) \leq (u - v)\varphi'(v)$$

Substituting $u = \Psi\left(z^{k+1}\right)$ and $v = \Psi\left(z^k\right)$ in the last equation yields

$$\left(\Psi\left(z^k\right) - \Psi\left(z^{k+1}\right)\right)\varphi'\left(\Psi\left(z^k\right)\right) \leq \varphi\left(\Psi\left(z^k\right)\right) - \varphi\left(\Psi\left(z^{k+1}\right)\right). \tag{5}$$

Combining (3) with (5) and applying (4) yields

$$\rho_1\rho_2 \left\| z^{k+1} - z^k \right\| \leq \rho_1\rho_2 \left\| z^{k+1} - z^k \right\| \cdot \left\| \nabla\Psi\left(z^k\right) \right\| \varphi'\left(\Psi\left(z^k\right)\right) \leq \varphi\left(\Psi\left(z^k\right)\right) - \varphi\left(\Psi\left(z^{k+1}\right)\right).$$

# Proof of Theorem-Contd.

Summing the last inequality over all $k \in \mathbb{N}$ yields part (i).
Now, take $q > p > l$ we have

$$z^q - z^p = \sum_{k=p}^{q-1} \left( z^{k+1} - z^k \right)$$

hence

$$\left\| z^q - z^p \right\| = \left\| \sum_{k=p}^{q-1} \left( z^{k+1} - z^k \right) \right\| \leq \sum_{k=p}^{q-1} \left\| z^{k+1} - z^k \right\| \leq \sum_{k=l}^{\infty} \left\| z^{k+1} - z^k \right\| \xrightarrow[l \to \infty]{} 0,$$

implying that $\left\{ z^k \right\}$ is a Cauchy sequence and hence a convergent sequence. The fact that $\emptyset \neq \omega \left( z^0 \right) \subset crit\Psi$ yields part (ii).

KL property ensures us that $\left\{ z^k \right\}_{k \in \mathbb{N}}$ is a Cauchy sequence! Thus converges!

The main question now is: Are there many KL functions?

# KPALM Algorithm For The Squared Euclidean Norm

We devise a PALM-like algorithm, exploiting the specific structure of $H$, 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, there is no need to add a proximal term as suggested in PALM.

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

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

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

# KPALM Analysis

In order to prove the convergence of the sequence that is generated by KPALM, $\{z(t) := (w(t), x(t))\}_{n \in \mathbb{N}}$, to a critical point, we need to show the properties required by PALM theory.

- Boundedness: $w^i(t) \in \Delta$ and

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

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

## Assumption

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

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

## 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

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

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

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

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

# Subgradient Lower Bound The Iterates Gap Proof

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

$$w^i(t+1) = \arg\min_{w^i \in \Delta} \left\{ \langle w^i, d^i(x(t)) \rangle + \frac{\alpha_i(t)}{2} \|w^i - w^i(t)\|^2 \right\} \tag{6}$$

$$x(t+1) = \operatorname{argmin} \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\} \tag{7}$$

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

Evaluating the last relation at $z(t+1)$ and using (7)

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

The optimality condition of $w^i(t+1)$ (see (6)), implies that there exists $u^i(t+1) \in \partial \delta_\Delta(w^i(t+1))$ such that

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

# Subgradient Lower Bound The Iterates Gap Proof-Contd.

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

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

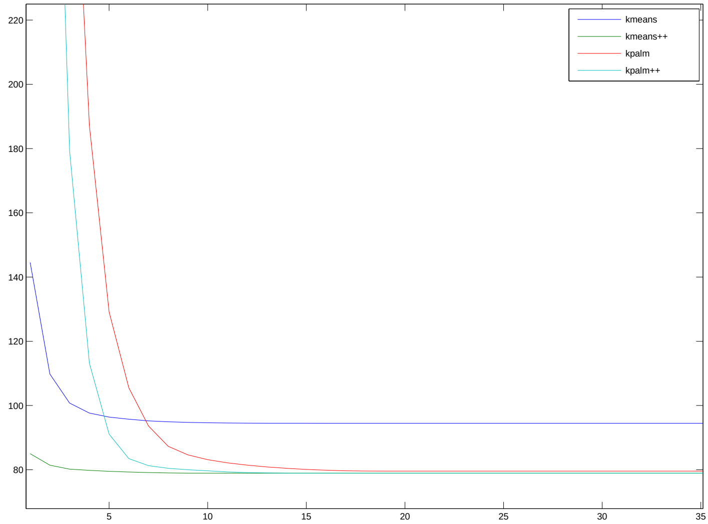
where the third inequality follows from the inequality

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

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

# Some Graphics

Below is a comparison of the objective function values performed on the Iris data-set for KMEANS, KMEANS++, KPALM and KMEANS++ algorithms.



# Some Graphics

Below is a comparison of the objective function values performed on the Iris data-set for KPALM algorithm with different $\alpha$ parameter updates.