

1 The Clustering Problem

Let $\mathcal{A} = \{a^1, \dots, a^m\}$ be a given set of points in \mathbb{R}^n , and let $1 < k < m$ be a fixed given number of clusters. The clustering problem consists of partitioning the data \mathcal{A} into k subsets $\{A^1, \dots, A^k\}$, called clusters. For each $l = 1, \dots, k$, the cluster A_l is represented by its center x^l , and we want to determine k cluster centers $\{x_1, \dots, x_k\}$ such that the sum of proximity measures from each point a^i to a nearest cluster center x^l is minimized.

The clustering problem formulation is given by

$$\min_{x^1, \dots, x^k \in \mathbb{R}^n} \sum_{i=1}^m \min_{1 \leq l \leq k} d(x^l, a^i), \quad (1.1)$$

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

2 Problem Reformulation and Notations

We introduce some notations that will be used throughout this document.

$A = (a^1, \dots, a^m) \in (\mathbb{R}^n)^m$, where $a^i \in \mathbb{R}^n, i = 1, \dots, m$

$W = (w^1, \dots, w^m) \in (\mathbb{R}^k)^m$, where $w^i \in \mathbb{R}^k, i = 1, \dots, m$

$X = (x^1, \dots, x^k) \in (\mathbb{R}^n)^k$, where $x^l \in \mathbb{R}^n, l = 1, \dots, k$

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

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

$$\text{For some } S \subseteq \mathbb{R}^n, \delta_S(p) = \begin{cases} 0 & \text{if } p \in S \\ \infty & \text{if } p \notin S \end{cases}$$

$$\langle u, v \rangle = \sum_{l=1}^k u_l \cdot v_l, \text{ for } u, v \in \mathbb{R}^k$$

Using the functional optimization representation of minimum of k values, i.e. $\min_{1 \leq l \leq k} u_l = \min \{\langle u, v \rangle \mid v \in \Delta\}$, and applying it over (1.1), gives a smooth reformulation of the clustering problem

$$\min_{X \in (\mathbb{R}^n)^k} \sum_{i=1}^m \min_{w^i \in \Delta} \langle w^i, d^i(X) \rangle \quad (2.1)$$

Further replacing the constrain over w^i with $\delta_\Delta(\cdot)$ function results in a equivalent formulation

$$\min_{X \in (\mathbb{R}^n)^k, W \in (\mathbb{R}^k)^m} \left\{ \sum_{i=1}^m \langle w^i, d^i(X) \rangle + \delta_\Delta(w^i) \right\} \quad (2.2)$$

Finally, introducing several more useful definitions, for each $i = 1, \dots, m$

$$\begin{aligned} H_i(W, X) &= \langle w^i, d^i(X) \rangle \\ G(w^i) &= \delta_\Delta(w^i) \\ H(W, X) &= \sum_{i=1}^m H_i(W, X) \\ G(W) &= \sum_{i=1}^m G(w^i) \end{aligned}$$

Replacing the terms in (2.1) with the functions above gives a compact form that is equivalent to the original clustering problem

$$\min \left\{ \Psi(Z) := H(W, X) + G(W) \mid Z := (W, X) \in (\mathbb{R}^k)^m \times (\mathbb{R}^n)^k \right\} \quad (2.3)$$

3 Clustering via PALM Approach

3.1 Introduction to PALM Theory

Presentation of PALM's requirements and of the algorithm steps \dots

3.2 Clustering with PALM for $d(u, v) = \|u - v\|^2$

In this section we tackle the clustering problem with distance-like function $d(u, v) = \|u - v\|^2$. Using the discussion about PALM, we will construct a semi-PALM algorithm. Since the clustering problem has a specific structure, we are ought to exploit it in the following manner. First we notice that the map $W \mapsto H(W, X) = \sum_{i=1}^m \langle w^i, d^i(X) \rangle$ is linear in W , so there is no need to linearize it. In

addition, the map $X \mapsto H(W, X) = \sum_{i=1}^m \langle w^i, d^i(X) \rangle = \sum_{i=1}^m \sum_{l=1}^k w_l^i \|x^l - a^i\|^2 = \sum_{l=1}^k \sum_{i=1}^m w_l^i \|x^l - a^i\|^2$ is convex in X , hence we can drop the proximal term in PALM algorithm.

Now we propose the semi-PALM algorithm for clustering.

(1) Initialization: Set $t = 0$, and pick random vectors $(W(0), X(0)) \in \Delta^m \times (\mathbb{R}^n)^k$

(2) For each $t = 0, 1, \dots$ generate a sequence $\{(W(t), X(t))\}_{t \in \mathbb{N}}$ as follows:

(2.1) Cluster Assignment: Take $\nu \in (0, 1]$, compute $\beta(t) = \min_{1 \leq l \leq k} \left\{ \sum_{i=1}^m w_l^i(t) \right\}$, set $\alpha(t) = \nu \beta(t)$ and for $i = 1, \dots, m$ compute

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

(2.2) Centers Update: For $l = 1, \dots, k$ compute $x^l \in \mathbb{R}^n$ via

$$X(t+1) = \arg \min \left\{ H(W(t+1), X) \mid X \in (\mathbb{R}^n)^k \right\} \quad (3.2)$$

At each step t , the PALM-Clustering algorithm alternates between cluster assignment and centers update. The explicit formulas for step t are given below

$$w^i(t+1) = \Pi_{\Delta} \left(w^i(t) - \frac{d^i(X(t))}{\alpha(t)} \right) \quad i = 1, \dots, m \quad (3.3)$$

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

Remark 1. (i) $\alpha(t)$ is the step-size, and it must be positive. If for some step $t \in \mathbb{N}$ $\alpha(t) = 0$ then $\exists l' \in [1, k]$ such that $\beta(t) = \sum_{i=1}^m w_{l'}^i = 0$, since $\forall l \in [1, k], \forall i \in [1, m] : w_l^i \geq 0$ then $\forall i \in [1, m] w_{l'}^i = 0$. Thus, none of the points in \mathcal{A} belong to cluster l' , in that case the algorithm can halt. Hence from now on we assume that $\forall t \in \mathbb{N}, \beta(t) = \min_{1 \leq l \leq k} \left\{ \sum_{i=1}^m w_l^i(t) \right\} > 0$, and it follows that $\alpha(t) > 0$.

(ii) $\forall t \in \mathbb{N} \quad W(t) \in \Delta^m \Rightarrow \Psi(Z(t)) = H(W(t), X(t)) + G(W(t)) = H(W(t), X(t))$.

Lemma 3.0.1 (Strong convexity of $H(W, X)$ in X). At step t , the mapping $X \mapsto H(W(t), X)$ is strongly convex $\Leftrightarrow \beta(t) > 0$.

Proof. Since the mapping $X \mapsto H(W(t), X) = \sum_{l=1}^k \sum_{i=1}^m w_l^i \|x^l - a^i\|^2$ is C^2 , it is strongly convex \Leftrightarrow its Hessian matrix smallest eigenvalue is positive.

$$\nabla_{x^j} \nabla_{x^l} H(W(t), X) = \begin{cases} 0 & \text{if } j \neq l, \quad j, l \in [1, k] \\ 2 \sum_{i=1}^m w_l^i(t) & \text{if } j = l, \quad j, l \in [1, k] \end{cases}$$

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

Now we are ready to prove the decrease property of the PALM-Clustering algorithm.

Proposition 3.1 (Sufficient decrease property).

$\exists \rho_1 > 0$ such that $\rho_1 \|Z(t+1) - Z(t)\|^2 \leq \Psi(Z(t)) - \Psi(Z(t+1))$, $\forall t \in \mathbb{N}$.

Proof. From (3.1) we derive the following inequality

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

Hence, we obtain

$$\frac{\alpha(t)}{2} \|w^i(t+1) - w^i(t)\|^2 \leq H_i(W(t), X(t)) - H_i(W(t+1), X(t))$$

Summing this inequality over $i = 1, \dots, m$ gives

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

Recall that the mapping $X \mapsto H(W(t), X)$ is strongly convex with parameter $2\beta(t) > 0$, hence we have

$$\begin{aligned} &H(W(t+1), X(t)) - H(W(t+1), X(t+1)) \\ &\geq \nabla_X H(W(t+1), X(t+1))^T (X(t) - X(t+1)) + \frac{2\beta(t)}{2} \|X(t) - X(t+1)\|^2 \\ &= \beta(t) \|X(t) - X(t+1)\|^2 \end{aligned}$$

where the last equality follows from (3.2).

Set $\rho_1 = \min \{\alpha(t), \beta(t)\} = \alpha(t)$, combined with the previous inequalities, we have

$$\begin{aligned} \rho_1 \|Z(t+1) - Z(t)\|^2 &= \rho_1 (\|W(t+1) - W(t)\|^2 + \|X(t+1) - X(t)\|^2) \\ &\leq [H(W(t), X(t)) - H(W(t+1), X(t))] + [H(W(t+1), X(t)) - H(W(t+1), X(t+1))] \\ &= H(Z(t)) - H(Z(t+1)) = \Psi(Z(t)) - \Psi(Z(t+1)). \end{aligned}$$

□

Next, we aim to prove the subgradient lower bound for iterates property. We start with few preliminary results.

Proposition 3.2 (Sufficient Lipschitz continuity condition). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be C^1 function, if $\|\nabla f(z)\|$ is bounded on $z \in \Omega$, then $\exists M > 0$ such that $\forall x, y \in \Omega$ $\|f(x) - f(y)\| \leq M\|x - y\|$, where $M = \sup_{z \in \Omega} \|\nabla f(z)\|$.*

Lemma 3.2.1. $\{\|\nabla_X d^i(X(t+1))\|\}_{t \in \mathbb{N}}$ is bounded set.

Proof.

$$\begin{aligned}
\|\nabla_X d^i(X(t+1))\| &= 2\| (x^1(t+1) - a^i, \dots, x^k(t+1) - a^i) \| \\
&\leq 2 \sum_{l=1}^k \|x^l(t+1) - a^i\| \leq 2 \sum_{l=1}^k \left\| \frac{\sum_{j=1}^m w_l^j(t+1) a^j}{\sum_{j=1}^m w_l^j(t+1)} - a^i \right\| \\
&\leq 2 \left(\left(\sum_{l=1}^k \sum_{j=1}^m \frac{w_l^j(t+1)}{\sum_{j=1}^m w_l^j(t+1)} \|a^j\| \right) + k \|a^i\| \right) \leq 2k \left(\|a^i\| + \sum_{j=1}^m \|a^j\| \right)
\end{aligned}$$

□

From the last two results it follows that

$$\exists M > 0 \quad \|d^i(X(t+1) - d^i(X(t)))\| \leq M \|X(t+1) - X(t)\|, \quad \forall t \in \mathbb{N}. \quad (3.5)$$

Proposition 3.3 (Subgradient lower bound for iterates property).

$\exists \rho_2 > 0$ and $\gamma(t+1) \in \partial \Psi(Z(t+1))$ such that $\|\gamma(t+1)\| \leq \rho_2 \|Z(t+1) - Z(t)\|^2$, $\forall t \in \mathbb{N}$.

Proof. $\Psi = H + G$, then

$$\begin{aligned}
\partial \Psi &= \nabla H + \partial G = (\nabla_W H, \nabla_X H) + \left((\partial_{w^i} \delta_\Delta)_{i=1, \dots, m}, (\vec{0})_{l=1, \dots, k} \right) \\
&= \left((\nabla_{w^i} H_i + \partial_{w^i} \delta_\Delta)_{i=1, \dots, m}, \nabla_X H \right)
\end{aligned}$$

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

$$\begin{aligned}
\partial \Psi(Z(t+1)) &= \left((\nabla_{w^i} H_i(W(t+1), X(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)))_{i=1, \dots, m}, \nabla_X H(W(t+1), X(t+1)) \right) \\
&= \left((d^i(X(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)))_{i=1, \dots, m}, \nabla_X H(W(t+1), X(t+1)) \right) \\
&= \left((d^i(X(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)))_{i=1, \dots, m}, \vec{0} \right)
\end{aligned}$$

where the last equality follows from (3.2), that is the optimality condition of $X(t+1)$.

Taking the norm of the last equation yields

$$\|\partial \Psi(Z(t+1))\| \leq \sum_{i=1}^m \|d^i(X(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1))\|. \quad (3.6)$$

The optimality condition of $w^i(t+1)$ that is derived from (3.1), yields $\forall i = 1, \dots, m$ that there $\exists u^i(t+1) \in \partial \delta_\Delta(w^i(t+1))$ such that

$$d^i(X(t+1)) + \alpha(t) (w^i(t+1) - w^i(t)) + u^i(t+1) = 0 \quad (3.7)$$

Setting $\gamma(t+1) = \left((d^i(X(t+1)) + u^i(t+1))_{i=1, \dots, m}, \vec{0} \right) \in \partial\Psi(Z(t+1))$, and plugging (3.7) into (3.6) we have

$$\begin{aligned}
\|\gamma(t+1)\| &\leq \sum_{i=1}^m \|d^i(X(t+1)) - d^i(X(t)) - \alpha(t)(w^i(t+1) - w^i(t))\| \\
&\leq \sum_{i=1}^m \|d^i(X(t+1)) - d^i(X(t))\| + m\alpha(t)\|Z(t+1) - Z(t)\| \\
&\leq \sum_{i=1}^m M\|X(t+1) - X(t)\| + m\alpha(t)\|Z(t+1) - Z(t)\| \\
&\leq m(M + \alpha(t))\|Z(t+1) - Z(t)\|
\end{aligned}$$

where the third inequality follows from (3.5).

Define $\rho_2 = m(M + \alpha(t))$ and the result follows. \square

4 Clustering via ADMM Approach

First we add new variables $z^l, l = 1, \dots, k$, and formulate an equivalent problem to the clustering problem (see (1.2)):

$$\min_{x^1, \dots, x^k \in \mathbb{R}^n} \min_{w^1, \dots, w^m \in \mathbb{R}^k} \min_{z^1, \dots, z^k \in S} \left\{ \sum_{i=1}^m v_i \sum_{l=1}^k w_l^i d(x^l, a^i) \mid w^i \in \Delta^i, i = 1, \dots, m, x^l = z^l, l = 1, \dots, k \right\} \quad (4.1)$$

We present the augmented Lagrangian associated with the clustering problem

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{w}) = H(\mathbf{x}, \mathbf{w}) + \sum_{l=1}^k (y^l)^T (x^l - z^l) + \frac{\rho}{2} \sum_{l=1}^k \|x^l - z^l\|^2 \quad (4.2)$$

ADMM update:

$$x^{l,t+1} := \frac{\rho z^{l,t} - y^{l,t} + 2 \sum_{i=1}^m v_i w_l^{i,t} a^i}{\rho + 2 \sum_{i=1}^m v_i w_l^{i,t}}$$

$$z^{l,t+1} := \Pi_S(x^{l,t+1} + \frac{y^{l,t}}{\rho})$$

$$y^{l,t+1} := y^{l,t} + \rho(x^{l,t+1} - z^{l,t+1})$$

$$w^{i,t+1} \in \{w \in \mathbb{R}^k \mid w \in \Delta^i, \text{ such that if } l \notin \text{Nearest}(\mathbf{x}^{t+1}, a^i) \text{ then } w_l^i = 0\}$$

$$\text{where } \text{Nearest}(\mathbf{x}, a^i) := \left\{ 1 \leq l \leq k \mid \|x^l - a^i\| = \min_{1 \leq j \leq k} \|x^j - a^i\| \right\}$$