1 The Clustering Problem

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

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

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.

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

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

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

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

$\Delta = \left\{ u \in \mathbb{R}^k \mid \sum\limits_{l=1}^{k} u_l = 1, u_l \ge 0, l = 1, \ldots, k \right\}$

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

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

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

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

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

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

Finally, introducing few more useful definitions, for each $i = 1, \cdots, m$

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

Replacing the terms in (2.1) with the function above gives final equivalent clustering problem formulation

$$
\min \left\{ H(W, X) + G(W) \mid X \in (\mathbb{R}^n)^k, W \in (\mathbb{R}^k)^m \right\} \tag{2.3}
$$

3 Clustering via PALM approach

An equivalent smooth formulation to the clustering problem

PALM algorithms addresses nonconvex-nonsmooth problems of the form

$$
minimize_{x,y} \Psi(x, y) := f(x) + g(y) + H(x, y), \tag{3.1}
$$

and in the extended form for p blocks

$$
minimize \left\{ \Psi(x_1, \ldots, x_p) := \sum_{i=1}^{p} f_i(x_i) + H(x_1, \ldots, x_p) : x_i \in \mathbb{R}^{n_i} \right\}, \tag{3.2}
$$

where $H : \mathbb{R}^N \to \mathbb{R}$ with $N = \sum_{i=1}^{p} n_i$ is assumed to be C^1 and each $f_i, i = 1, \ldots, p$, is proper and lower-semicontinuous function.

Applying the PALM notations to the clustering problem formulation (1.2), with distance-like function $d(u, v) = \|u - v\|^2$, setting $f_l(x^l) = \delta_S(x^l)$, $l = 1, \ldots, k$, $g_i(w^i) = \delta_{\Delta^i}(w^i)$, $i = 1, \ldots, m$ and $H(x^1, \ldots, x^k, w^1, \ldots, w^m) = \sum_{i=1}^{m} v_i \sum_{l=1}^{k} w_l^i d(x^l, a^i)$.

Next, we confirm all requirements of f_l, g_i and H as listed in Assumptions 1 and 2 at (reference to PALM article). For simplicity, we introduce some notations $\mathbf{x} = (x^1, x^2, \ldots, x^k)$ and similarly $\mathbf{w} = (w^1, w^2, \ldots, w^m)$. Also $\mathbf{x}^{-l} = (x^1, \ldots, x^{l-1}, x^{l+1}, \ldots, x^k)$ and similarly $\mathbf{w}^{-i} = (w^1, \ldots, w^{i-1}, w^{i+1}, \ldots, w^m)$.

(i) Since $f_l, g_i, H \geq 0$ they all are proper. g_i and H are lower semicontinuous since Δ_i is closed and H in C^2. As for lower semicontinuity of f_l it requires S to be closed.

(ii) The partial gradient $\nabla_{x^l} H(\mathbf{x}, \mathbf{w})$ is globally Lipschitz with moduli $L_{x^l}(\mathbf{x}^{-l}, \mathbf{w}) = 2 \sum_{i=1}^{m} v_i w_l^i \leq$ $\leq 2 w_l^{max} \sum_{i=1}^{m} v_i = 2 w_l^{max}$, for $l = 1, \ldots, k$, where $w_l^{max} := \max_{i=1,\cdots,m} w_l^i$.

(iii) H is linear with respect to $\mathbf{w}$ thus $\nabla_{x^l} H(\mathbf{x}, \mathbf{w})$ is globally Lipschitz with moduli $L_{w^i}(\mathbf{x}, \mathbf{w}^{-i}) = 0$, for $i = 1, \ldots, m$. For PALM's proximal steps remain always well-defined, we set $L_{w^i}(\mathbf{x}, \mathbf{w}^{-i}) = \mu_i > 0$, for $i = 1, \ldots, m$ (see Remark 3 (iii)). Similarly, in case $L_{x^l}(\mathbf{x}^{-l}, \mathbf{w})$ is too close to 0, we set $L_{x^l}(\mathbf{x}^{-l}, \mathbf{w}) = \nu_l > 0$, for $l = 1, \cdots, k$.

(iv) $\inf\left\{L_{w^i}(\mathbf{x}, \mathbf{w}^{-i})\right\} = \sup\left\{L_{w^i}(\mathbf{x}, \mathbf{w}^{-i})\right\} = \mu_i, i = 1, \cdots, m$
and $\sup\left\{L_{x^l}(\mathbf{x}^{-l}, \mathbf{w})\right\} \le 2w_l^{max}$, $\inf\left\{L_{x^l}(\mathbf{x}^{-l}, \mathbf{w})\right\} \ge \nu_l, l = 1, \cdots, k$.

(v) ∇H is Lipschitz continuous on bounded subset, since H in C^2 (see Remark 3 (iv)).

(vi) PALM requires Ψ to be KL function. H is real polynomial function, thus satisfies the KL property. Δ_i is semi-algebraic set, and we require S to be semi-algebraic set.

Next, we formulate PALM's steps for the clustering problem, and explicitly compute the proximal formulas.

PALM-Clustering

(1) Initialization: Select random vectors $x^{l,0} \in S, l = 1, \cdots, k$ and $w^{i,0} \in \Delta^i, i = 1 \cdots, m$.

(2) For each $t = 0, 1, \cdots$ generate a sequence $\left\{(x^{1,t}, \cdots, x^{k,t}, w^{1,t}, \cdots, w^{m,t})\right\}_{t\in\mathbb{N}}$ as follows:

(2.1) For each $l = 1, \cdots, k$ compute:

(2.1.1) Take $\gamma_l > 1$, set $c_l^t = \gamma_l L_{x^l}(x^{1,t+1}, \cdots, x^{l-1,t+1}, x^{l+1,t}, \cdots, x^{k,t}, w^{1,t}, \cdots, w^{m,t})$ and compute

$$x^{l,t+1} \in prox_{c_l^t}^{f_l}(x^{l,t} - \tfrac{1}{c_l^t} \nabla_{x^l} H(x^{1,t+1}, \cdots, x^{l-1,t+1}, x^{l,t}, x^{l+1,t}, \cdots, x^{k,t}, w^{1,t}, \cdots, w^{m,t}))$$
$$= \Pi_S\left(x^{l,t} - \frac{\sum_{i=1}^m v_i w_l^{i,t} 2(x^{l,t} - a^i)}{\gamma_l \max\left\{\nu_l, 2\sum_{i=1}^m v_i w_l^{i,t}\right\}}\right) = \Pi_S\left(x^{l,t}\left(1 - \frac{\sum_{i=1}^m v_i w_l^{i,t}}{\gamma_l \max\left\{\frac{\nu_l}{2}, \sum_{i=1}^m v_i w_l^{i,t}\right\}}\right) + \frac{\sum_{i=1}^m v_i w_l^{i,t} a^i}{\gamma_l \max\left\{\frac{\nu_l}{2}, \sum_{i=1}^m v_i w_l^{i,t}\right\}}\right)$$

(2.2) For each $i = 1, \cdots, m$ compute:

(2.2.1) Take $\beta_i > 1$, set $d_i^t = \beta_i L_{w^i}(x^{1,t+1}, \cdots, x^{k,t+1}, w^{1,t+1}, \cdots, w^{i-1,t+1}, w^{i+1,t}, \cdots, w^{m,t})$ and compute

$$\begin{aligned} w^{i,t+1} &\in prox_{d_i^t}^{g_i}(w^{i,t} - \tfrac{1}{d_i^t} \nabla_{w^i} H(x^{1,t+1}, \cdots, x^{k,t+1}, w^{1,t+1}, \cdots, w^{i-1,t+1}, w^{i,t}, w^{i+1,t}, \cdots, w^{m,t}) \\ &= \Pi_{\Delta^i}(w^{i,t} - \tfrac{v_i}{\beta_i \mu_i}(w_1^{i,t}\|x^{1,t+1} - a^i\|^2, \cdots, w_k^{i,t}\|x^{k,t+1} - a^i\|^2)^T) \\ &= \Pi_{\Delta^i}((w_l^{i,t}(1 - \tfrac{v_i\|x^{l,t+1} - a^i\|^2}{\beta_i \mu_i}))_{1\le l\le k}) \end{aligned}$$

4 Clustering via ADMM approach

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

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

We present the augmented Lagrangian associated with the clustering problem

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

ADMM update:

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

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

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

$$w^{i,t+1} \in \left\{ w \in \mathbb{R}^k \mid w \in \Delta^i, \text{ such that if } l \notin Nearest(\mathbf{x}^{t+1}, a^i) \text{ then } w_l^i = 0 \right\}$$
$$\text{where } Nearest(\mathbf{x}, a^i) := \left\{ 1 \le l \le k \mid \|x^l - a^i\| = \min_{1 \le j \le k} \|x^j - a^i\| \right\}$$