

THE MULTI-FACILITY LOCATION PROBLEM: A PROBABILISTIC DECOMPOSITION METHOD

CEM IYIGUN AND ADI BEN-ISRAEL

ABSTRACT. A generalized Weiszfeld method is proposed for the multi-facility location problem. The problem is relaxed using probabilistic assignments, and is decomposed into single facility location problems, that are coupled by these assignments, and can be solved in parallel. The probabilistic assignments are updated at each iteration, using the distances to the current centers. The method thus iterates between assignments and centers updates, with the probabilistic assignments gradually approaching the pure assignments to the computed centers. The method also provides intrinsic criteria for the quality of the solution, and for the optimal number of facilities to serve the given customers. A duality theorem allows verifying the optimality of the cluster centers by solving a dual problem. Numerical experience with several problems from the literature is presented.

1. INTRODUCTION

The **Fermat–Weber** (also **single facility**) **location problem** is to locate a **facility** that will serve optimally a set of **customers**, given by their **locations** and **weights**, in the sense of minimizing the weighted sum of distances traveled by the customers, see [14] and [16]. A well known method for solving the problem is the **Weiszfeld method** [44], a gradient method that expresses and updates the sought center as a convex combination of the data points.

The **multi-facility location problem (MFLP)** is to locate a (given) number of facilities to serve the customers as above. Each customer is assigned to a single facility, and the problem (also called **location-allocation**) is to determine the optimal locations of the facilities, as well as the optimal assignments of customers (assignment is not an issue in the single facility case.)

In some MFLP cases the facility locations are constrained to lie in a given subset of the plane, in particular a given subset of the customers locations. This **constrained**, or **discrete, MFLP model** was first considered by Hakimi [20], and is often solved by mixed integer programming, see [39].

We study here the **continuous**, or **unconstrained, MFLP model**, where the facilities can be anywhere in the plane. We also assume that the facilities can handle all the customers assigned to them. The **capacitated** problem, where some facilities may have limited capacities, see, e.g., [45], will be treated elsewhere using the results of [23].

The **continuous MFLP** is a **clustering problem**, where customers and facility locations correspond to data points and centers, and the clustering criterion is the sum of distances from each data point to its

Date: July 29, 2012.

2000 Mathematics Subject Classification. Primary 90B85, 90C46, 62H30; Secondary 90C59, 92D40, 91B14, 94A17.

Key words and phrases. Facility location; clustering; Fermat–Weber problem; multi-facility location problem; Weiszfeld method; contour approximation; duality; Luce choice axiom; Kullback–Leibler divergence.

cluster center, to be minimized. We use interchangeably **center** = **facility location**, and **cluster** = **the set of customers assigned to a facility**.

This paper solves the continuous MFLP using the probabilistic clustering approach of [4], and is a continuation of [26]. We approximate the MFLP (which is NP hard, [37]) by replacing rigid assignments with probabilistic assignments, as in [5] and [12]. The resulting iteration is a natural generalization of the Weiszfeld method [44] to several facilities.

The plan of the paper is as follows. Section 2 describes the multi-facility location problem. The special case of a single facility is discussed in Section 3.

Section 4 introduces probabilistic assignments, with cluster membership probabilities that depend on the distances involved. Section 5 introduces the probabilistic location problem (P.K), an approximation of MFLP. The updates of the centers of (P.K) are explained in Section 6.

A generalized Weiszfeld algorithm for solving (P.K) is presented in Section 9. The cluster membership probabilities are used in Section 10 to chart the territories of the facilities, a Voronoi diagram of the centers. The gradient of the objective is modified in Section 11 to guarantee its existence everywhere. The modified gradient is used to characterize optimality and to prove convergence, Theorem 1.

Section 12 gives a duality theory for (P.K), based on the geometric duality of Kuhn ([29], [30]). A duality theorem, Theorem 2, allows verifying the optimality of any feasible solution (centers and their assignments) by solving a dual problem.

Section 13 recalls two useful concepts from probabilistic clustering, the joint distance function (JDF) and the classification uncertainty function (CUF), and applies them to the problem (P.K).

Section 14 presents numerical results from our method applied to several problems in the literature.

A proof of Theorem 2 is given in Appendix A. Appendix B shows the connection between the CUF (introduced in Section 13) and the Kullback-Leibler entropic distance. The probabilistic model used here is related to Luce's choice axiom [36] in Appendix C.

2. THE MULTI-FACILITY LOCATION PROBLEM

Notation: We denote an index set $\{1, 2, \dots, K\}$ by $1:K$. For a vector $\mathbf{x} = (x_j) \in \mathbb{R}^n$, $\|\mathbf{x}\|$ denotes the **Euclidean norm**, $\|\mathbf{x}\| := \sqrt{\sum_{j=1}^n |x_j|^2}$. The **distance** between the vectors \mathbf{x} and \mathbf{y} is denoted $d(\mathbf{x}, \mathbf{y})$. The **Euclidean distance**

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \tag{1}$$

is used throughout.

Consider

- a set $\mathbf{X} := \{\mathbf{x}_i : i \in 1:N\}$ of N **points** in \mathbb{R}^n ,
- a set $\mathbf{W} := \{w_i > 0 : i \in 1:N\}$ of their respective **weights**, and
- an integer $1 \leq K \leq N$.

Typically, the points $\{\mathbf{x}_i\}$ are the locations of **customers**¹, and the weights $\{w_i\}$ are their **demands**.

¹Most location problems are planar, i.e. $n = 2$, but the results below hold for general n , and are stated for \mathbb{R}^n .

The **multi-facility location problem (MFLP)**² defined next, uses the data

$$\text{data}(\mathbf{L}.K) := \{\mathbf{X}, \mathbf{W}, K\}, \quad (2)$$

and is to locate K facilities, and assign customers to these facilities, so as to minimize the sum of weighted distances

$$\min_{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K} \sum_{k=1}^K \sum_{\mathbf{x}_i \in \mathbf{X}_k} w_i d(\mathbf{x}_i, \mathbf{c}_k) \quad (\mathbf{L}.K)$$

where $\{\mathbf{c}_k : k \in 1:K\}$ are the centers (facility locations), and \mathbf{X}_k is the cluster of customers assigned to the k^{th} facility. For $K = 1$ this reduces to the Fermat–Weber location problem, see § 3, where assignment is absent. The other extreme case, $K = N$ (every point is a center), is of no interest.

The MFLP ($\mathbf{L}.K$) can be broken into two problems:

(a) the **assignment problem**: given K centers $\{\mathbf{c}_k\}$, assign each of the N points $\{\mathbf{x}_i\}$ to the nearest center, and

(b) the **location problem**: given the K clusters \mathbf{X}_k formed in (a), find the center of each cluster (using, say, the Weiszfeld method.)

A natural heuristic ([11], [33]) for solving ($\mathbf{L}.K$) is to iterate between these two problems: the clusters are updated (assignment), requiring updating the centers (location), then the points are re-assigned, etc.

A solution $\{(\mathbf{X}_k, \mathbf{c}_k) : k \in 1:K\}$ is **stable** if it cannot be improved by re-assigning any point \mathbf{x}_i to another cluster. The heuristic method described above may terminate in a non-optimal stable solution, declaring it to be optimal, as illustrated in Example 1.

Example 1. Given $N = 4$ points, $\mathbf{x}_1 = (0, 0)$, $\mathbf{x}_2 = (1, 0)$, $\mathbf{x}_3 = (1, 1)$, $\mathbf{x}_4 = (0, 1)$, and all weights $w_i = 1$, consider the problem ($\mathbf{L}.2$), and three of its solutions:

Solution 1: Each cluster is a pair of adjacent points, see e.g. Fig. 1(a). The centers may lie anywhere on the segments joining the points, and the value of the objective function is 2.

Solution 2: Unequal clusters with centers in opposite points, for example, $\mathbf{c}_1 = (0, 0)$, $\mathbf{c}_2 = (1, 1)$, see, e.g., Fig. 1(b). The value of the objective is 2.

Solution 3: Unequal clusters, see, e.g., Fig. 1(b) with centers $\mathbf{c}_1 = (\frac{1}{3+\sqrt{3}}, \frac{1}{3+\sqrt{3}})$, $\mathbf{c}_2 = (1, 1)$ and the value of the objective is 1.9318.

Solution 3 is optimal, but Solutions 1 and 2 are stable: for example, in Solution 1, the point \mathbf{x}_2 is not closer to the center of \mathbf{X}_1 than to the center of \mathbf{X}_2 , and should not be re-assigned (to get a better solution) on the basis of distance information alone.

A fourth solution is where each cluster is a pair of opposite points, the centers are anywhere on the diagonals, and the objective function value is $2\sqrt{2}$. This solution is unstable, except in the (uninteresting) case where the two centers coincide at $(\frac{1}{2}, \frac{1}{2})$.

If a simple example like this has stable solutions that are not optimal, one would expect many such solutions in bigger problems, say hundreds of customers and several facilities. \square

²Also called the **location-allocation problem**.

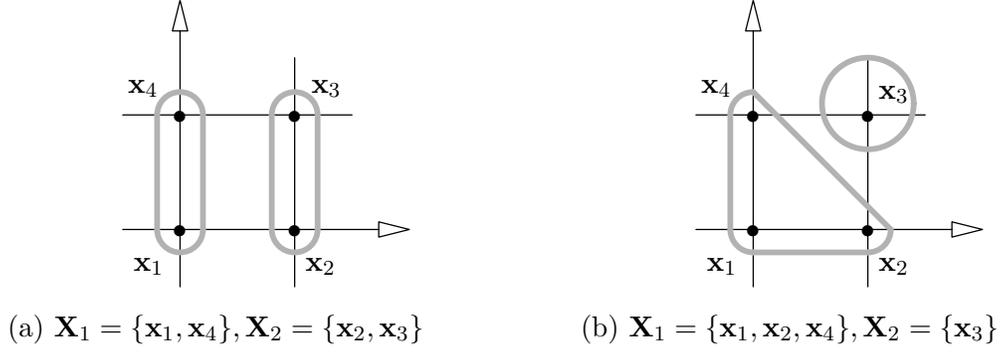


FIGURE 1. Illustration of Example 1

3. THE FERMAT–WEBER LOCATION PROBLEM

Given $\{\mathbf{X}, \mathbf{W}, 1\}$ as above, the **Fermat–Weber location problem** is to find a point $\mathbf{c} \in \mathbb{R}^n$ minimizing the sum of weighted distances,

$$\min_{\mathbf{c} \in \mathbb{R}^n} \sum_{i=1}^N w_i d(\mathbf{x}_i, \mathbf{c}), \quad (\text{L.1})$$

see [17], [35], [45] and their references.

If the points $\{\mathbf{x}_i\}$ are not collinear, as is assumed throughout, the objective function

$$f(\mathbf{c}) = \sum_{i=1}^N w_i d(\mathbf{x}_i, \mathbf{c}) \quad (3)$$

is strictly convex, and (L.1) has a unique optimal solution.

The gradient of (3) is undefined if $\mathbf{c} \in \mathbf{X}$, i.e. if \mathbf{c} coincides with one of the data points $\{\mathbf{x}_i\}$. For $\mathbf{c} \notin \mathbf{X}$,

$$\nabla f(\mathbf{c}) = - \sum_{i=1}^N w_i \frac{\mathbf{x}_i - \mathbf{c}}{\|\mathbf{x}_i - \mathbf{c}\|}, \quad (4)$$

and the optimal center \mathbf{c}^* , if not in \mathbf{X} , is characterized by $\nabla f(\mathbf{c}^*) = \mathbf{0}$, expressing it as a convex combination of the points \mathbf{x}_i ,

$$\mathbf{c}^* = \sum_{i=1}^N \lambda_i^* \mathbf{x}_i, \text{ with coefficients } \lambda_i^* = \frac{w_i / \|\mathbf{x}_i - \mathbf{c}^*\|}{\sum_{m=1}^N w_m / \|\mathbf{x}_m - \mathbf{c}^*\|} \text{ that depend on } \mathbf{c}^*.$$

This circular statement gives rise to the **Weiszfeld iteration**, [44],

$$\mathbf{c}_+ := T(\mathbf{c}) \quad (5)$$

where \mathbf{c}_+ is the updated center, \mathbf{c} is the current center, and

$$T(\mathbf{c}) := \begin{cases} \sum_{i=1}^N \left(\frac{w_i / \|\mathbf{x}_i - \mathbf{c}\|}{\sum_{m=1}^N w_m / \|\mathbf{x}_m - \mathbf{c}\|} \right) \mathbf{x}_i, & \text{if } \mathbf{c} \notin \mathbf{X}; \\ \mathbf{c}, & \text{if } \mathbf{c} \in \mathbf{X}. \end{cases} \quad (6)$$

In order to extend $\nabla f(\mathbf{c})$ to all \mathbf{c} , Kuhn [30] modified its definition as follows: $\nabla f(\mathbf{c}) := -R(\mathbf{c})$, where

$$R(\mathbf{c}) := \begin{cases} -\nabla f(\mathbf{c}), & \text{if } \mathbf{c} \notin \mathbf{X}; \\ \max\{0, \|R^j\| - w_j\} \frac{R^j}{\|R^j\|}, & \text{if } \mathbf{c} = \mathbf{x}_j \in \mathbf{X}, \end{cases} \quad (7)$$

and

$$R^j := \sum_{i \neq j} \frac{w_i}{\|\mathbf{x}_i - \mathbf{x}_j\|} (\mathbf{x}_i - \mathbf{x}_j) \quad (8)$$

is the **resultant force** of $N - 1$ forces of magnitude w_i and direction $\mathbf{x}_i - \mathbf{x}_j$, $i \neq j$.

3.1. The results of Kuhn, [30]. The following properties of the mappings $R(\cdot), T(\cdot)$, and the optimal center \mathbf{c}^* were proved by Kuhn [30]. By “ $= \mathbf{c}^*$ ” we mean “is optimal”.

- (a) $\mathbf{c} = \mathbf{c}^* \iff R(\mathbf{c}) = \mathbf{0}$.
- (b) $\mathbf{c}^* \in \text{conv } \mathbf{X}$ (the convex hull of \mathbf{X}).
- (c) If $\mathbf{c} = \mathbf{c}^*$ then $T(\mathbf{c}) = \mathbf{c}$. Conversely, if $\mathbf{c} \notin \mathbf{X}$, $T(\mathbf{c}) = \mathbf{c}$ then $\mathbf{c} = \mathbf{c}^*$.
- (d) If $T(\mathbf{c}) \neq \mathbf{c}$ then $f(T(\mathbf{c})) < f(\mathbf{c})$.
- (e) $\mathbf{x}_j = \mathbf{c}^* \iff w_j \geq \|R^j\|$.
- (f) If $\mathbf{x}_j \neq \mathbf{c}^*$, the direction of steepest descent of f at \mathbf{x}_j is $R^j / \|R^j\|$.
- (g) If $\mathbf{x}_j \neq \mathbf{c}^*$ there exists $\delta > 0$ such that

$$0 < \|\mathbf{c} - \mathbf{x}_j\| \implies \|T^s(\mathbf{c}) - \mathbf{x}_j\| > \delta \text{ for some } s.$$
- (h) $\lim_{\mathbf{c} \rightarrow \mathbf{x}_j} \frac{\|T(\mathbf{c}) - \mathbf{x}_j\|}{\|\mathbf{c} - \mathbf{x}_j\|} = \frac{\|R^j\|}{w_j}$.
- (i) For any \mathbf{c}_0 , if no $\mathbf{c}_r := T^r(\mathbf{c}_0) \in \mathbf{X}$, then $\lim_{r \rightarrow \infty} \mathbf{c}_r = \mathbf{c}^*$.

These results are generalized in Theorem 1 to the case of several facilities.

Remark 1. Another claim in [30], that

$$T^r(\mathbf{c}_0) \rightarrow \mathbf{c}^* \text{ for all but a denumerable number of initial centers } \mathbf{c}_0,$$

was refuted by Chandrasekaran and Tamir [9]. Convergence can be assured by modifying the algorithm (5)–(6) at a sticky, non-optimal, center that coincides with a data point \mathbf{x}_j . Balas and Yu [3] suggested moving from \mathbf{x}_j in the direction R^j (8) of steepest descent, assuring a decrease of the objective, and non-return to \mathbf{x}_j by (g) above. Vardi and Zhang [43] guaranteed leaving \mathbf{x}_j by adding to the objective function a quadratic term in the distances to the other data points. Convergence was also addressed by Brimberg [7], Drezner [15], Eckhardt [18], Katz [28] and others. \square

4. CLUSTER MEMBERSHIP PROBABILITIES

For $1 < K < N$, the problem (L.K) is NP hard, [37]. It can be solved polynomially in N for $K = 2$, see [13], and possibly for other given K .

We relax the assignment problem in (L.K) by using probabilistic assignments, see § 5. First we require the concept of **cluster membership probabilities**,

$$p_k(\mathbf{x}) := \text{Prob}\{\mathbf{x} \in \mathbf{X}_k\}, \quad k \in 1:K,$$

assumed to depend only on the distances $\{d(\mathbf{x}, \mathbf{c}_k) : k \in 1:K\}$ of the point \mathbf{x} from the K centers. A reasonable assumption is

$$\boxed{\text{membership in a cluster is more likely the closer is its center}} \quad (\text{A})$$

and a simple way to model this assumption is³,

$$w p_k(\mathbf{x}) d(\mathbf{x}, \mathbf{c}_k) = D(\mathbf{x}), \quad k \in 1:K, \quad (9)$$

where w is the weight of \mathbf{x} , and $D(\cdot)$ is a function of \mathbf{x} , that does not depend on k . $D(\mathbf{x})$ is called the **joint distance function** (JDF) at \mathbf{x} .

Equations (9) are optimality conditions for the extremum problem

$$\min \left\{ w \sum_{k=1}^K p_k^2(\mathbf{x}) d(\mathbf{x}, \mathbf{c}_k) : \sum_{k=1}^K p_k(\mathbf{x}) = 1, p_k(\mathbf{x}) > 0, k \in 1:K \right\} \quad (\text{E})$$

in the probabilities $\{p_1(\mathbf{x}), \dots, p_K(\mathbf{x})\}$. The squares of probabilities in the objective of (E) serve to smooth the underlying non-smooth problem (L.K), see the seminal paper by Teboulle [41].

Since probabilities add to one we get from (9),

$$p_k(\mathbf{x}) = \frac{\prod_{j \neq k} d(\mathbf{x}, \mathbf{c}_j)}{\sum_{\ell=1}^K \prod_{m \neq \ell} d(\mathbf{x}, \mathbf{c}_m)}, \quad k \in 1:K, \quad (10)$$

and the JDF at \mathbf{x} ,

$$D(\mathbf{x}) = w \frac{\prod_{j=1}^K d(\mathbf{x}, \mathbf{c}_j)}{\sum_{\ell=1}^K \prod_{m \neq \ell} d(\mathbf{x}, \mathbf{c}_m)}, \quad (11)$$

which is (up to a constant) the **harmonic mean** of the distances $\{d(\mathbf{x}, \mathbf{c}_k) : k \in 1:K\}$. The probabilities $\{p_k(\mathbf{x}) : k \in 1:K\}$ are determined by the centers $\{\mathbf{c}_k : k \in 1:K\}$, and the JDF also uses the weight w .

For example, in case $K = 2$,

$$p_1(\mathbf{x}) = \frac{d(\mathbf{x}, \mathbf{c}_2)}{d(\mathbf{x}, \mathbf{c}_1) + d(\mathbf{x}, \mathbf{c}_2)}, \quad p_2(\mathbf{x}) = \frac{d(\mathbf{x}, \mathbf{c}_1)}{d(\mathbf{x}, \mathbf{c}_1) + d(\mathbf{x}, \mathbf{c}_2)}, \quad (12a)$$

$$D(\mathbf{x}) = w \frac{d(\mathbf{x}, \mathbf{c}_1) d(\mathbf{x}, \mathbf{c}_2)}{d(\mathbf{x}, \mathbf{c}_1) + d(\mathbf{x}, \mathbf{c}_2)}, \quad (12b)$$

and for $K = 3$,

$$p_1(\mathbf{x}) = \frac{d(\mathbf{x}, \mathbf{c}_2) d(\mathbf{x}, \mathbf{c}_3)}{d(\mathbf{x}, \mathbf{c}_1) d(\mathbf{x}, \mathbf{c}_2) + d(\mathbf{x}, \mathbf{c}_1) d(\mathbf{x}, \mathbf{c}_3) + d(\mathbf{x}, \mathbf{c}_2) d(\mathbf{x}, \mathbf{c}_3)}, \quad \text{etc.},$$

$$D(\mathbf{x}) = w \frac{d(\mathbf{x}, \mathbf{c}_1) d(\mathbf{x}, \mathbf{c}_2) d(\mathbf{x}, \mathbf{c}_3)}{d(\mathbf{x}, \mathbf{c}_1) d(\mathbf{x}, \mathbf{c}_2) + d(\mathbf{x}, \mathbf{c}_1) d(\mathbf{x}, \mathbf{c}_3) + d(\mathbf{x}, \mathbf{c}_2) d(\mathbf{x}, \mathbf{c}_3)}.$$

³There are other ways to model Assumption (A), e.g. [4], but the simple model (9) works well enough for our purposes.

5. PROBABILISTIC APPROXIMATION OF (L.K)

There is a difference between the statements “ \mathbf{x} is a member of \mathbf{X}_k ” and “ \mathbf{x} is assigned to \mathbf{X}_k ”, making it necessary to introduce “probabilistic assignments”, of which the cluster membership probabilities of § 4 are a special case. For a given a point \mathbf{x} and clusters $\{\mathbf{X}_k : k \in 1:K\}$, a **probabilistic assignment** of \mathbf{x} is a vector $\mathbf{q}(\mathbf{x}) = (q_1(\mathbf{x}), \dots, q_K(\mathbf{x}))$, where $q_k(\mathbf{x})$ is the probability that \mathbf{x} is assigned to \mathbf{X}_k . A **pure assignment** (in short, an **assignment**) is a unit vector, say $\mathbf{q}(\mathbf{x}) = \mathbf{e}_k$, indicating that \mathbf{x} is assigned (deterministically) to \mathbf{X}_k .

A pure assignment of all points $\{\mathbf{x}_i\}$ corresponds to a clustering of \mathbf{X} ,

$$\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2 \cup \dots \cup \mathbf{X}_K, \text{ where } \mathbf{X}_k := \{\mathbf{x} \in \mathbf{X} : \mathbf{q}(\mathbf{x}) = \mathbf{e}_k\}. \quad (13)$$

Given the probabilistic assignments for each $\mathbf{x}_i \in \mathbf{X}$ we denote the set of KN numbers $\{q_k(\mathbf{x}_i)\}$ by

$$Q(\mathbf{X}) := \{q_k(\mathbf{x}_i) : k \in 1:K, i \in 1:N\}. \quad (14)$$

Probabilistic assignments are assumed to have the following property: If a data point \mathbf{x}_j coincides with a center \mathbf{c}_k , i.e. if $d(\mathbf{x}_j, \mathbf{c}_k) = 0$, then $\mathbf{q}(\mathbf{x}_j)$ is a pure assignment,

$$d(\mathbf{x}_j, \mathbf{c}_k) = 0 \implies q_k(\mathbf{x}_j) = 1, q_m(\mathbf{x}_j) = 0 \text{ for all } m \neq k. \quad (15)$$

For the cluster membership probabilities $\{p_k(\mathbf{x})\}$ this is assured by (10).

We use a probabilistic approximation of the problem (L.K) in the form

$$\min_{\{\mathbf{c}_1, \dots, \mathbf{c}_K\}} \sum_{k=1}^K \sum_{i=1}^N w_i q_k(\mathbf{x}_i) d(\mathbf{x}_i, \mathbf{c}_k), \quad (P.K)$$

$$\{\{q_1(\mathbf{x}_i), \dots, q_K(\mathbf{x}_i)\} : i \in 1:N\}$$

with two sets of variables, the **centers** $\{\mathbf{c}_k\}$ and **probabilistic assignments** $\{q_k(\mathbf{x}_i)\}$, that are updated iteratively. The problem (P.K) uses the same data, (2), as in problem (L.K).

For any centers and probabilistic assignments, the objective function of (P.K) is an upper bound on the optimal value of (L.K),

$$\sum_{k=1}^K \sum_{i=1}^N w_i q_k(\mathbf{x}_i) d(\mathbf{x}_i, \mathbf{c}_k) \geq \min(\text{L.K}), \quad (16)$$

and therefore so is the optimal value of (P.K),

$$\min(\text{P.K}) \geq \min(\text{L.K}). \quad (17)$$

The probabilistic assignments $\{q_k(\mathbf{x})\}$ used below are calculated by raising the cluster membership probabilities $\{p_k(\mathbf{x})\}$ to a power $\nu \geq 1$ and normalizing⁴,

$$q_k(\mathbf{x}_i) = \frac{p_k^\nu(\mathbf{x}_i)}{\sum_{j=1}^K p_j^\nu(\mathbf{x}_i)}, \quad i \in 1:N, k \in 1:K, \quad (18a)$$

⁴This is one of many possible constructions.

and can be calculated directly from the distances, since by (10),

$$q_k(\mathbf{x}_i) = \frac{\prod_{j \neq k} d(\mathbf{x}_i, \mathbf{c}_j)^\nu}{\sum_{\ell=1}^K \prod_{m \neq \ell} d(\mathbf{x}_i, \mathbf{c}_m)^\nu}. \quad (18b)$$

To record the dependence on $\{p_k(\mathbf{x})\}$ and the exponent ν , we denote the probabilistic assignments (18a) by $\{p_k^{(\nu)}(\mathbf{x}_i)\}$. In previous work ([4] and [23]–[26]) we used the exponent $\nu = 2$ without renormalization.

For a discussion of the exponent ν see § 8 below.

5.1. Bounds for (L.K) with given centers. Given a set of K centers $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$, not necessarily optimal, the minimal value of (L.K) is obtained by assigning each \mathbf{x}_i to its nearest center,

$$m(\mathbf{c}_1, \dots, \mathbf{c}_K) = \sum_{i=1}^N w_i \min\{d(\mathbf{x}_i, \mathbf{c}_k) : k \in 1:K\}. \quad (19)$$

This value (not necessarily optimal since the centers are arbitrary) can be bounded as follows.

Corollary 1. Given a set of centers $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$, the corresponding cluster membership probabilities $\{p_k(\mathbf{x}_i) : k \in 1:K, i \in 1:N\}$ of (10), and the values of the JDF (11), $\{D(\mathbf{x}_i) : i \in 1:N\}$, at the points $\{\mathbf{x}_i : i \in 1:N\}$,

$$\sum_{i=1}^N D(\mathbf{x}_i) \leq m(\mathbf{c}_1, \dots, \mathbf{c}_K) \leq \sum_{i=1}^N \frac{D(\mathbf{x}_i)}{\sum_{j=1}^K p_j^2(\mathbf{x}_i)}, \quad (20a)$$

$$\leq K \sum_{i=1}^N D(\mathbf{x}_i). \quad (20b)$$

Proof. The left inequality in (20a) follows from

$$D(\mathbf{x}_i) \leq w_i \min\{d(\mathbf{x}_i, \mathbf{c}_k) : k \in 1:K\}, \text{ by (11).}$$

To prove the right inequality, note that for any probabilistic assignments $\{q_k(\mathbf{x}_i) : i \in 1:N, k \in 1:K\}$,

$$\sum_{k=1}^K \sum_{i=1}^N w_i q_k(\mathbf{x}_i) d(\mathbf{x}_i, \mathbf{c}_k) \geq m(\mathbf{c}_1, \dots, \mathbf{c}_K) \quad (21)$$

In particular, for $q_k(\mathbf{x}) = p_k^{(2)}(\mathbf{x})$,

$$\begin{aligned}
\text{LHS (21)} &= \sum_{k=1}^K \sum_{i=1}^N w_i \frac{p_k^2(\mathbf{x}_i)}{\sum_{j=1}^K p_j^2(\mathbf{x}_i)} d(\mathbf{x}_i, \mathbf{c}_k), \\
&= \sum_{i=1}^N \frac{w_i}{\sum_{j=1}^K p_j^2(\mathbf{x}_i)} \sum_{k=1}^K p_k(\mathbf{x}_i)^2 d(\mathbf{x}_i, \mathbf{c}_k), \\
&= \sum_{i=1}^N \frac{1}{\sum_{j=1}^K p_j^2(\mathbf{x}_i)} \left(\sum_{k=1}^K p_k(\mathbf{x}_i) \right) D(\mathbf{x}_i), \text{ by (9),} \\
&= \sum_{i=1}^N \frac{D(\mathbf{x}_i)}{\sum_{j=1}^K p_j^2(\mathbf{x}_i)} = \text{RHS (20a)}.
\end{aligned}$$

Finally, (20b) follows from $\sum_{j=1}^K p_j^2(\mathbf{x}_i) \geq \frac{1}{K}$, for all i . \square

Note that the bounds (20a) are tighter as $\{p_j(\mathbf{x}_i)\}$ get closer to pure assignments.

6. UPDATES OF CENTERS

Given probabilistic assignments $\{q_k(\mathbf{x}_i)\}$, the objective function of (P.K) is a separable function of the centers,

$$f(\mathbf{c}_1, \dots, \mathbf{c}_K) := \sum_{k=1}^K f_k(\mathbf{c}_k), \quad (22a)$$

$$\text{where } f_k(\mathbf{c}) := \sum_{i=1}^N w_i q_k(\mathbf{x}_i) \|\mathbf{x}_i - \mathbf{c}\|, \quad k \in 1:K. \quad (22b)$$

The centers problem thus separates into K problems of type (L.1), coupled by the probabilistic assignments $\{q_k(\mathbf{x}_i)\}$. If the data points are not collinear, as we assume throughout, the centers problem has a unique optimal solution $\{\mathbf{c}_k^* : k \in 1:K\}$.

The gradient of (22b), wherever it exists, is

$$\nabla f_k(\mathbf{c}) = - \sum_{i=1}^N \frac{w_i q_k(\mathbf{x}_i)}{\|\mathbf{x}_i - \mathbf{c}\|} (\mathbf{x}_i - \mathbf{c}), \quad k \in 1:K. \quad (23)$$

Zeroing the gradients (23) we get, in analogy with the Weiszfeld method, the optimal centers $\{\mathbf{c}_1^*, \dots, \mathbf{c}_K^*\}$ as convex combinations of the data points,

$$\mathbf{c}_k^* = \sum_{i=1}^N \lambda_k^*(\mathbf{x}_i) \mathbf{x}_i, \quad (24a)$$

with coefficients $\lambda_k^*(\mathbf{x}_i)$ given by

$$\lambda_k^*(\mathbf{x}_i) = \frac{w_i q_k(\mathbf{x}_i) / \|\mathbf{x}_i - \mathbf{c}_k^*\|}{\sum_{j=1}^N w_j q_k(\mathbf{x}_j) / \|\mathbf{x}_j - \mathbf{c}_k^*\|}, \quad k \in 1:K, \quad i \in 1:N. \quad (24b)$$

Equations (24a)–(24b) induce K mappings $T_k : \mathbf{c} \rightarrow T_k(\mathbf{c})$, $k \in 1:K$,

$$T_k(\mathbf{c}) := \sum_{i=1}^N \lambda_{ki} \mathbf{x}_i, \quad \text{with } \lambda_{ki} := \frac{w_i q_k(\mathbf{x}_i) / \|\mathbf{x}_i - \mathbf{c}\|}{\sum_{j=1}^N w_j q_k(\mathbf{x}_j) / \|\mathbf{x}_j - \mathbf{c}\|}, \quad (25a)$$

for $\mathbf{c} \notin \{\mathbf{x}_j : j \in 1:N\}$, and by continuity,

$$T_k(\mathbf{x}_j) := \mathbf{x}_j, \quad \text{for } j \in 1:N. \quad (25b)$$

These mappings are used to update the centers $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ as follows,

$$\mathbf{c}_k^+ := T_k(\mathbf{c}_k), \quad \text{for } k \in 1:K. \quad (26)$$

7. THE MODIFIED GRADIENT

The gradient (23) is undefined (0/0) if \mathbf{c} coincides with any of the data points. In order to define the gradient everywhere, we modify its definition following Kuhn ([29]–[30]), and denote the modified gradient by $-\mathbf{R}_k$.

If a center $\mathbf{c}_k \notin \{\mathbf{x}_j : j \in 1:N\}$,

$$\mathbf{R}_k(\mathbf{c}_k) := \sum_{i=1}^N \frac{w_i q_k(\mathbf{x}_i)}{\|\mathbf{x}_i - \mathbf{c}_k\|} (\mathbf{x}_i - \mathbf{c}_k). \quad (27a)$$

Otherwise, if \mathbf{c}_k coincides with \mathbf{x}_j , define,

$$\mathbf{R}_k(\mathbf{x}_j) := \max \{ \|\mathbf{R}_k^j\| - w_j q_k(\mathbf{x}_j), 0 \} \frac{\mathbf{R}_k^j}{\|\mathbf{R}_k^j\|}, \quad (27b)$$

where

$$\mathbf{R}_k^j = \sum_{i \neq j} \frac{w_i q_k(\mathbf{x}_i)}{\|\mathbf{x}_i - \mathbf{x}_j\|} (\mathbf{x}_i - \mathbf{x}_j), \quad (27c)$$

and $q_k(\mathbf{x}_j) = 1$ in (27b), by (15). Therefore, $\mathbf{R}_k(\mathbf{x}_j) = \mathbf{0}$ if $\|\mathbf{R}_k^j\| < w_j$; otherwise, $\mathbf{R}_k(\mathbf{x}_j)$ is a vector with magnitude $\|\mathbf{R}_k^j\| - w_j$ and direction \mathbf{R}_k^j . As in (8), the vector \mathbf{R}_k^j is the **resultant** of $N - 1$ forces of magnitude $w_i q_k(\mathbf{x}_i)$ and direction $\mathbf{x}_i - \mathbf{x}_j$, $i \neq j$.

8. UPDATING THE EXPONENT ν

As is typically the case in gradient methods, the iterations (26) make big steps at first, approaching their fixed points, then the iterations slow down and movement at each iteration is small. The “fast” iterations are few in number, the number of the “slow” iterations is determined by the stopping rule.

Most progress towards solving the location problem occurs in the first few iterations. The slow iterations at the end deal mainly with the assignment problem.

With a low value of ν , such as $\nu = 2$, the distributions $\{p_k^{(\nu)}(\mathbf{x})\}$ may be far from pure assignments, and require rounding to the nearest integer, 0 or 1.

High values of the exponent ν in (18b), say $\nu > 10$, produce $\{p_k^{(\nu)}(\mathbf{x}_i)\}$ that are close to pure assignments. For this reason, high values of ν are useful in the slow iterations at the end. In contrast, using high values of ν at the beginning may cause premature convergence to a solution that is not necessarily optimal.

This suggests increasing the exponent ν at each iteration. We use here the simple update

$$\nu^+ := \nu + \Delta \quad (28)$$

where $\Delta > 0$ is the increment per iteration. If ν_0 is the initial exponent, the k^{th} -exponent is thus $\nu_0 + k\Delta$. This amounts to a simultaneous approximation of the location and assignment problems of § 2, concentrating on location first then switching gradually to assignment.

Increasing the exponent ν results in a decrease of the value of the objective function (22), as shown in following lemma.

Lemma 1. Let $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ be any centers, \mathbf{x} any point, $\{d_k(\mathbf{x}, \mathbf{c}_k) : k \in 1:K\}$ the distances, not all equal, and $\{p_k^{(\nu)}(\mathbf{x}) : k \in 1:K\}$ the probabilities in (18a). If $\nu_2 > \nu_1 \geq 1$ then

$$\sum_{k=1}^K p_k^{(\nu_2)}(\mathbf{x}) d(\mathbf{x}, \mathbf{c}_k) < \sum_{k=1}^K p_k^{(\nu_1)}(\mathbf{x}) d(\mathbf{x}, \mathbf{c}_k), \quad (29)$$

and

$$\lim_{\nu \rightarrow \infty} \sum_{k=1}^K p_k^{(\nu)}(\mathbf{x}) d(\mathbf{x}, \mathbf{c}_k) = \min \{d(\mathbf{x}, \mathbf{c}_k) : k \in 1:K\}. \quad (30)$$

Proof. Arrange the distances, say

$$d(\mathbf{x}, \mathbf{c}_1) \leq d(\mathbf{x}, \mathbf{c}_2) \leq \dots \leq d(\mathbf{x}, \mathbf{c}_K)$$

Then the corresponding probabilistic assignments (18a) satisfy, for any $\nu \geq 1$,

$$p_1^{(\nu)}(\mathbf{x}) \geq p_2^{(\nu)}(\mathbf{x}) \geq \dots \geq p_K^{(\nu)}(\mathbf{x})$$

and are not all equal. Raising the exponent from $\nu = \nu_1$ to $\nu = \nu_2$, increases the highest probabilities and decreases the lowest ones, in particular, there is an index j such that

$$p_k^{(\nu_2)}(\mathbf{x}) > p_k^{(\nu_1)}(\mathbf{x}) \text{ for } k \leq j, \text{ and } p_k^{(\nu_2)}(\mathbf{x}) \leq p_k^{(\nu_1)}(\mathbf{x}), \text{ for } k > j.$$

Let $\alpha_k := p_k^{(\nu_2)}(\mathbf{x}) - p_k^{(\nu_1)}(\mathbf{x})$. Then $\alpha_k > 0$ for $k \leq j$, and $\alpha_k \leq 0$ for $k > j$. Also, $\sum_{k=1}^K \alpha_k = 0$, and therefore

$$\sum_{k=1}^j \alpha_k = - \sum_{k=j+1}^K \alpha_k = \alpha, \text{ say.}$$

$$\therefore \frac{1}{\alpha} (\text{LHS}(29) - \text{RHS}(29)) = \sum_{k=1}^j \left(\frac{\alpha_k}{\alpha}\right) d(\mathbf{x}, \mathbf{c}_k) - \sum_{k=j+1}^K \left(-\frac{\alpha_k}{\alpha}\right) d(\mathbf{x}, \mathbf{c}_k) < 0$$

since $\sum_{k=1}^j \left(\frac{\alpha_k}{\alpha}\right) d(\mathbf{x}, \mathbf{c}_k)$ is a convex combination of low values, and $\sum_{k=j+1}^K \left(-\frac{\alpha_k}{\alpha}\right) d(\mathbf{x}, \mathbf{c}_k)$ is a convex combination of higher values.

To prove (30) let M be the index set of minimal distances,

$$M := \{i : d(\mathbf{x}, \mathbf{c}_i) = \min \{d(\mathbf{x}, \mathbf{c}_k) : k \in 1:K\}\}.$$

Then

$$\begin{aligned} \lim_{\nu \rightarrow \infty} p_i^{(\nu)} &= 0, \text{ for } i \notin M, \\ \lim_{\nu \rightarrow \infty} p_i^{(\nu)} &= \frac{1}{|M|}, \text{ for } i \in M, \end{aligned}$$

where $|M|$ is the number of elements of M . Therefore

$$\lim_{\nu \rightarrow \infty} \sum_{k=1}^K p_k^{(\nu)}(\mathbf{x}) d(\mathbf{x}, \mathbf{c}_k) = \sum_{k=1}^K \left(\lim_{\nu \rightarrow \infty} p_k^{(\nu)} \right) d(\mathbf{x}, \mathbf{c}_k) = \min \{d(\mathbf{x}, \mathbf{c}_k) : k \in 1:K\}.$$

□

9. A GENERALIZED WEISZFELD METHOD FOR THE MULTI-FACILITY LOCATION PROBLEM

The above results can be implemented in an algorithm for solving (P.K).

Algorithm 1. A generalized Weiszfeld method for several facilities

Data: $\mathbf{X} = \{\mathbf{x}_i : i \in 1:N\}$ data points (locations of customers),
 $\mathbf{W} = \{w_i : i \in 1:N\}$ weights,
 K the number of facilities
 $\epsilon > 0$ (stopping criterion)
 $\nu_0 > 0$ (initial value of the exponent ν)
 $\Delta > 0$ (the increment in (28))

Initialization: K arbitrary centers $\{\mathbf{c}_k : k \in 1:K\}$,

Iteration:

Step 1 **compute** distances $\{d(\mathbf{x}, \mathbf{c}_k) : k \in 1:K\}$ for all $\mathbf{x} \in \mathbf{X}$
Step 2 **update** the exponent ν (using (28))
and the probabilistic assignments $\{p_k^{(\nu)}(\mathbf{x}) : \mathbf{x} \in \mathbf{X}, k \in 1:K\}$ (using (18b))
Step 3 **update** the centers $\{\mathbf{c}_k^+ := T_k(\mathbf{c}_k) : k \in 1:K\}$ (using (25a))
Step 4 **if** $\sum_{k=1}^K d(\mathbf{c}_k^+, \mathbf{c}_k) < \epsilon$ **stop**
return to step 1

Remark 2.

(a) Convergence can be established as in the Weiszfeld algorithm, by forcing a move away from non-optimal data points (where the algorithm gets stuck by (25b)), using ideas of Balas and Yu [3], Vardi and Zhang

[43], and others, see Remark 1.

(b) As shown in [4], the above results also hold if the Euclidean distance (1) is replaced by an **elliptic distance**

$$d(\mathbf{x}, \mathbf{y}) := \langle \mathbf{x} - \mathbf{y}, T(\mathbf{x} - \mathbf{y}) \rangle, \quad (31)$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product, and T is a positive definite matrix. Algorithm 1 can thus be used with elliptic distances in Steps 1–4.

(c) The algorithm is robust since the coefficients (24b) are directly proportional to the weights, and inversely proportional to the distances. As a result, outliers or low weights are discounted, having little influence on the centers, see, e.g., Example 8 below. \square

Example 2. The dependence of Algorithm 1 on the exponent ν used in (18b) is illustrated here for the simple Example 1.

Fig. 2 shows three trajectories, starting at the same initial solution (a pair of close points at the center of the unit square $[0, 1]^2$), using a fixed exponent ν throughout the iterations, and converging to the three solutions of Example 1. For $\nu = 6$, $\nu = 12$, and $\nu = 18$, the iterates converged to Solutions 2, 3 and 1, respectively.

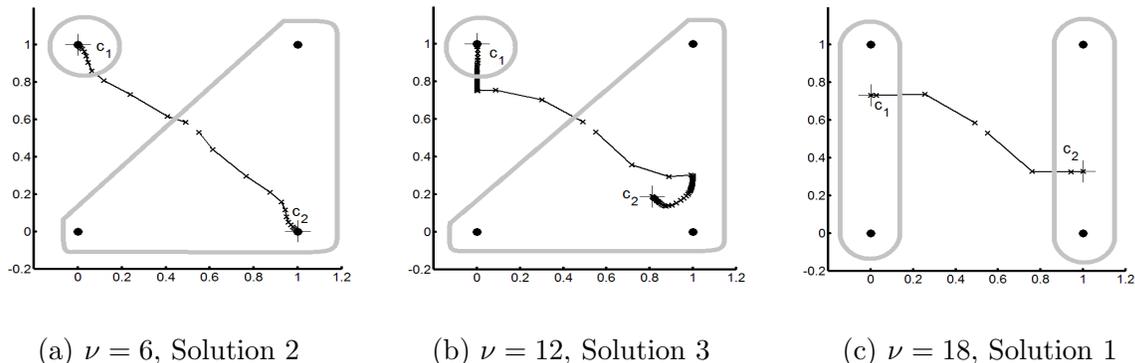


FIGURE 2. Trajectories of iterates in Example 1, for 3 different ν values

Table 1 shows the results for 1,000 initial solutions (each a pair of points chosen randomly in $[0, 1]^2$), using different values of ν (that are held constant throughout), and a stopping criterion of $\epsilon = 0.0001$. The table shows the percentage of cases ending with Solutions 1, 2 or 3 of Example 1, and the average number of iterations. The non-optimal Solution 2 is present only for small ν , and occurs in 100% of the time for $\nu = 6$. The optimal Solution 3 dominates for intermediate values of ν (it occurs 83% of the time for $\nu = 13$), and is present for all larger ν . The non-optimal Solution 1 dominates for large ν .

10. TERRITORIAL BOUNDARIES OF FACILITIES

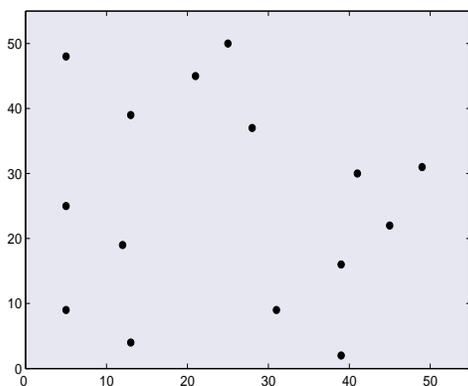
Each center \mathbf{c}_k serves the customers in a territory that is closer to it than to any other center, see, e.g., [21]–[22]. This territory is expressed naturally using the membership probabilities (10).

For any two (distinct) centers $\mathbf{c}_j, \mathbf{c}_k$ the locus of the points \mathbf{x} with $p_j(\mathbf{x}) = p_k(\mathbf{x})$ is by (10) represented by the equation

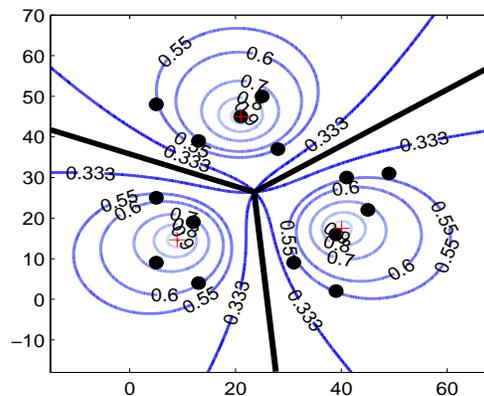
$$\|\mathbf{x} - \mathbf{c}_j\| = \|\mathbf{x} - \mathbf{c}_k\| \quad (32)$$

TABLE 1. Dependence of convergence on the exponent ν ($\epsilon = 0.0001$)

ν	Percentage of initial solutions resulting in			Average # of Iterations
	Solution 1	Solution 2	Solution 3	
6	0.0	100.0	0.0	35.88
13	9.9	7.0	83.1	124.62
14	15.7	4.2	80.1	97.51
16	28.8	1.6	69.6	79.34
18	41.3	0.9	57.8	61.56
20	51.0	0.6	48.4	47.34
22	58.8	0.2	41.0	34.39
30	72.1	0.0	27.9	13.82
40	75.9	0.0	24.1	9.00
50	76.8	0.0	23.2	6.97
100	76.9	0.0	23.1	5.91



(a) 15 data points



(b) Probability level sets and Voronoi diagram

FIGURE 3. Illustration of Example 3 ($N = 15$, $K = 3$)

and is thus a hyperplane⁵, the orthogonal bisector of the segment joining \mathbf{c}_j and \mathbf{c}_k . On this hyperplane one is indifferent between the two centers. The portion of the hyperplane (32) where

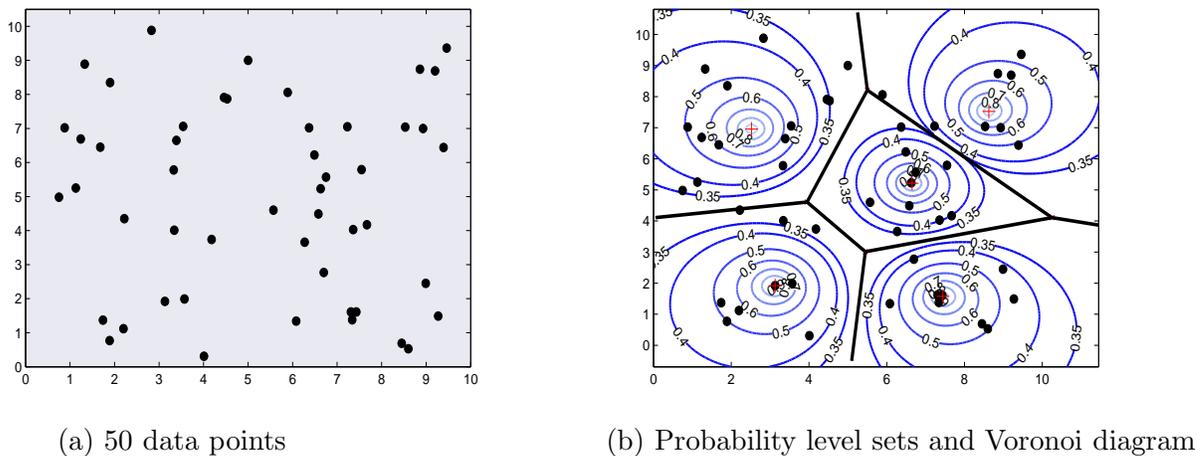
$$p_j(\mathbf{x}) = p_k(\mathbf{x}) \geq p_m(\mathbf{x}), \quad \forall m \neq j, k \quad (33)$$

is either empty, bounded or unbounded⁶. We call it the **common boundary** of the clusters \mathbf{X}_j and \mathbf{X}_k . The common boundaries constitute a **Voronoi diagram**, [2], of the centers. Each **Voronoi cell** is a polyhedron, and is the territory served best by its facility. \square

Example 3. Problem 9 in Table 2, due to Cooper [10], has $N = 15$ data points, see Figure 3(a), and $K = 3$ centers. Figure 3(b) shows the optimal centers, found by Algorithm 1, some level sets of the membership probability for each cluster, and the Voronoi diagram of the centers, partitioning the plane into territories served by each facility. \square

⁵A line in the typical case $n = 2$.

⁶If $n = 2$ it is either empty, a segment or a half-line.

FIGURE 4. Illustration of Example 4 ($N = 50$, $K = 5$)

Example 4. Figure 4(a) shows 50 data points of Problem 30 in Table 2 below. This problem, with $K = 5$, was solved by Eilon et al in [19, p. 83]. Figure 4(b) shows the optimal centers (found by Algorithm 1), some probability level sets for each cluster, and the common boundaries of the territories served by the facilities (2 common boundaries are empty). \square

11. CONVERGENCE

Algorithm 1 decomposes the problem (P.K) to K problems of type (L.1), coupled by the probabilistic assignments. The convergence proof of Kuhn [29]–[30] (with modifications, e.g., [3], [7] or [43]) can be adapted here.

While the algorithm converges from any initial set of centers $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$, it does not necessarily converge to an optimal solution of the original problem (L.K).

Writing (25a) as

$$T_k(\mathbf{c}) = \sum_{i=1}^N \lambda_{ki} \mathbf{x}_i - \mathbf{c} + \mathbf{c} = \sum_{i=1}^N \lambda_{ki} (\mathbf{x}_i - \mathbf{c}) + \mathbf{c}$$

we get, from (27a),

$$T_k(\mathbf{c}) = \mathbf{c} + \mathbf{h}_k(\mathbf{c}) \mathbf{R}_k(\mathbf{c}) \tag{34}$$

$$\text{with } \mathbf{h}_k(\mathbf{c}) = \frac{1}{\sum_{j=1}^N \frac{w_j q_k(\mathbf{x}_j)}{\|\mathbf{x}_j - \mathbf{c}\|}} \tag{35}$$

showing that Algorithm 1 is a gradient method, following the direction of the resultant $\mathbf{R}_k(\mathbf{c})$ with a step of length $\mathbf{h}_k(\mathbf{c})\|\mathbf{R}_k(\mathbf{c})\|$, except for data points $\{\mathbf{x}_j : j \in 1:N\}$ which are left fixed by T_k .

Results (a)–(i) of § 3.1 are next reproduced for several facilities.

Theorem 1. Let the data $\{\mathbf{X}, \mathbf{W}, K\}$, and a set $Q(\mathbf{X})$ of probabilistic assignments (14), be given.

(a) For any set of points $\{\mathbf{c}_k : k \in 1:K\} \subset \mathbb{R}^n$, the condition⁷

$$\mathbf{R}_k(\mathbf{c}_k) = \mathbf{0}, \text{ for all } k \in 1:K, \quad (36)$$

is necessary and sufficient for the points $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ to minimize $f(\mathbf{c}_1, \dots, \mathbf{c}_K)$ in (22a).

(b) The optimal centers $\{\mathbf{c}_k^* : k \in 1:K\}$ are in the convex hull of \mathbf{X} .

(c) If the point \mathbf{c} is an optimal location of the k^{th} -facility (for the given $Q(\mathbf{X})$), then it is a fixed point⁸ of T_k ,

$$T_k(\mathbf{c}) = \mathbf{c}, \text{ for some } k \in 1:K. \quad (37)$$

Conversely, if $\mathbf{c} \notin \mathbf{X}$ and satisfies (37), then \mathbf{c} is optimal.

(d) If $T_k(\mathbf{c}) \neq \mathbf{c}$, then the function $f_k(\cdot)$ of (22b) satisfies

$$f_k(T_k(\mathbf{c})) < f_k(\mathbf{c}), \quad k \in 1:K. \quad (38)$$

(e) If the data point \mathbf{x}_j is an optimal center of f_k then $w_j \geq \|\mathbf{R}_k^j\|$.

(f) If the data point \mathbf{x}_j is not an optimal center, the direction of steepest descent of f_k at \mathbf{x}_j is $\frac{\mathbf{R}_k^j}{\|\mathbf{R}_k^j\|}$.

(g) Let \mathbf{x}_i be a data point that is not an optimal center \mathbf{c}_k , for a given $k \in 1:K$. Then there exist $\delta > 0$ and an integer $s > 0$ such that

$$0 < \|\mathbf{x}_i - \mathbf{c}\| \leq \delta \quad \text{implies} \quad \|\mathbf{x}_i - T_k^{s-1}(\mathbf{c})\| \leq \delta \quad \text{and} \quad \|\mathbf{x}_i - T_k^s(\mathbf{c})\| > \delta$$

(h)
$$\lim_{\mathbf{c} \rightarrow \mathbf{x}_j} \frac{\|\mathbf{x}_j - T_k(\mathbf{c})\|}{\|\mathbf{x}_j - \mathbf{c}\|} = \frac{\|\mathbf{R}_k^j\|}{w_j q_k(\mathbf{x}_j)} \quad \text{for } k \in 1:K, j \in 1:N. \quad (39)$$

(i) Given any \mathbf{c}_k^0 , $k \in 1:K$, define the sequence $\{\mathbf{c}_k^r = T_k^r(\mathbf{c}_k^0) : r = 1, 2, \dots\}$. If no \mathbf{c}_k^r is a data point, then $\lim_{r \rightarrow \infty} \mathbf{c}_k^r = \mathbf{c}_k^*$, for some optimal centers $\{\mathbf{c}_1^*, \dots, \mathbf{c}_K^*\}$ (for the approximate problem (P.K)). \square

This theorem can be proved by adapting the proof of [26, Theorem 1] for probabilistic assignments. We give a proof in the supplementary material section at the end.

Remark 3.

(a) If the assignments $Q(\mathbf{X})$ are pure, then Theorem 1 reduces to the results (a)–(i) of § 3.1, for each of the clusters defined by (13).

(b) Theorem 1(d) guarantees a decrease of the function f_k by the iteration (26), but the function f_k on the left side of (38) may be different than the function f_k in RHS(38) if the probabilistic assignments $\{q_k(\mathbf{x}_i)\}$ are updated, as is the case in Algorithm 1. However, Lemma 1 guarantees a decrease if $\{p_k^{(\nu)}(\mathbf{x}_i)\}$ of (18b) are used, with the exponent ν raised at each iteration.

(c) Theorem 1(g) states that each non-optimal data point has a neighborhood that eventually repulses the sequence of approximations. Therefore the iterates may not get arbitrarily close to any non-optimal data point.

⁷Recall the definition of $\mathbf{R}_k(\cdot)$ in (27a)–(27b), and note that $\mathbf{R}_k(\cdot)$ and (22a)–(22b) use the same (given) probabilistic assignments $Q(\mathbf{X})$.

⁸Recall the definition in (25a)–(25b).

12. THE DUAL PROBLEM

We abbreviate the probabilistic assignments $q_k(\mathbf{x}_i)$ by q_{ki} , for $k \in 1:K$, $i \in 1:N$. A dual problem (D) for (P.K) is now given, following [29]. It uses the data

$$\text{data(D)} := \{\mathbf{X}, \mathbf{W}, K, Q(\mathbf{X})\} \quad (40)$$

consisting of data points $\mathbf{X} = \{\mathbf{x}_i\}$, weights $\mathbf{W} = \{w_i\}$, the number of facilities K , and the probabilistic assignments $Q(\mathbf{X}) = \{q_{ki}\}$. The dual variables are KN vectors $\{\mathbf{u}_{ki} : k \in 1:K, i \in 1:N\}$, one for each center and data point. We denote the set of dual variables by \mathbf{U} .

The **dual problem** is:

$$\max \quad g(\mathbf{U}) = \sum_{k=1}^K \sum_{i=1}^N \mathbf{u}_{ki} \cdot \mathbf{x}_i \quad (D)$$

$$\text{s.t.} \quad \sum_{i=1}^N \mathbf{u}_{ki} = \mathbf{0}, \quad k \in 1:K, \quad (41)$$

$$\|\mathbf{u}_{ki}\| \leq w_i q_{ki}, \quad i \in 1:N, k \in 1:K. \quad (42)$$

Problem (D) is a generalization of the dual problem for the single facility location problem, see also [29], [17, § 1.1.2], [27] and [42].

Remark 4.

(a) The dual problem (D) uses the same data as (P.K), namely the triple $\{\mathbf{X}, \mathbf{W}, K\}$, and in addition the probabilistic assignments $Q(\mathbf{X})$.

(b) Unlike problems (L.K) and (P.K) that have locally optimal solutions, the dual problem (D) has a unique optimal value (for the given data (40)), and its optimal solutions form a closed convex set.

(c) The dual problem (D) has a natural geometric interpretation. Let the weight w_i of the i^{th} -customer be split among the facilities, so that the k^{th} -facility gets the weight $q_{ki} w_i$. The vector \mathbf{u}_{ki} is interpreted as a force with direction $(\mathbf{x}_i - \mathbf{c}_k)$, and magnitude $\|\mathbf{u}_{ki}\| \leq q_{ki} w_i$. The location of the k^{th} -facility is then the equilibrium point of these forces, as expressed by the constraint (41). The objective of the problem (D) is

$$\sum_{k=1}^K \sum_{i=1}^N \mathbf{u}_{ki} \cdot \mathbf{x}_i = \sum_{k=1}^K \left(\sum_{i=1}^N \mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k) \right), \text{ by (41),}$$

and maximizing it is equivalent to minimizing the sum of terms $\mathbf{u}_{ki} \cdot (\mathbf{c}_k - \mathbf{x}_i)$ representing the work done by the force \mathbf{u}_{ki} along $(\mathbf{c}_k - \mathbf{x}_i)$.

Theorem 2.

(a) Let the data (40) be given. Then for any set of points $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$, and any set of feasible dual variables $\mathbf{U} = \{\mathbf{u}_{ki}\}$, the inequality

$$g(\mathbf{U}) \leq f(\mathbf{c}_1, \dots, \mathbf{c}_K), \quad (43)$$

holds, where $f(\mathbf{c}_1, \dots, \mathbf{c}_K)$ is the objective function (22a) of the primal problem (P.K).

(b) Given the data (40), and an optimal solution $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ of the primal problem (P.K) (with the same

$Q(\mathbf{X})$), there exist feasible dual variables \mathbf{U} such that

$$g(\mathbf{U}) = f(\mathbf{c}_1, \dots, \mathbf{c}_K). \quad (44)$$

(c) Let \mathbf{U} be an optimal solution of the dual problem (D). Then there exist $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ such that (44) holds. \square

A proof of this theorem is given in Appendix A.

Remark 5.

(a) Theorem 2 establishes duality between the centers–problem of (P.K) and (D), part (a) is a **weak duality** in the sense that any feasible solution \mathbf{U} of (D) gives a lower bound for the optimal value of (P.K), and conversely, any set of centers $\{\mathbf{c}_k\}$ for (P.K) gives an upper bound on the optimal value of (D). Parts (b)–(c) show that there is no duality gap.

(b) If $Q(\mathbf{X})$ are pure assignments, then Theorem 2(b) can be used to check the optimality of the centers for each of the clusters defined by (13). Indeed, if

$$g(\mathbf{U}) = m(\mathbf{c}_1, \dots, \mathbf{c}_K), \text{ see (19),} \quad (45)$$

then the centers $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ are optimal for their (not necessarily optimal) clusters, see, e.g., the following example.

Example 5. Consider the 4 points $\{\mathbf{x}_1, \dots, \mathbf{x}_4\}$ of Example 1, and the 3 solutions given there (for $K = 2$), starting with Solution 3.

Solution 3. Here the 2 clusters are $\mathbf{X}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4\}$, $\mathbf{X}_2 = \{\mathbf{x}_3\}$, see Fig. 1(b).

For $k \in 1:2$ and $i \in 1:4$ let \mathbf{u}_{ki} be the dual variables, and let q_{ki} be given by pure assignments, $q_{ki} = 1$ if \mathbf{x}_i is assigned to cluster \mathbf{X}_k , 0 otherwise. Then

$$q_{13} = q_{21} = q_{22} = q_{24} = 0$$

($q_{13} = 0$ since \mathbf{x}_3 does not belong to \mathbf{X}_1 , etc.) and therefore, by the constraints (42),

$$\mathbf{u}_{13} = \mathbf{u}_{21} = \mathbf{u}_{22} = \mathbf{u}_{24} = \mathbf{0}$$

The dual problem (D) can be written

$$\begin{aligned} \max \quad & \mathbf{u}_{11} \cdot \mathbf{x}_1 + \mathbf{u}_{12} \cdot \mathbf{x}_2 + \mathbf{u}_{14} \cdot \mathbf{x}_4 + \mathbf{u}_{23} \cdot \mathbf{x}_3 = \mathbf{u}_{12}[1] + \mathbf{u}_{14}[2] + \mathbf{u}_{23}[1] + \mathbf{u}_{23}[2] \\ \text{s.t.} \quad & \mathbf{u}_{11} + \mathbf{u}_{12} + \mathbf{u}_{14} = \mathbf{0}, \\ & \mathbf{u}_{23} = \mathbf{0}, \\ & \|\mathbf{u}_{ki}\| \leq 1, \text{ for } ki = 11, 12, 14 \text{ and } 23, \end{aligned} \quad (46)$$

where $\mathbf{u}_{ki}[j]$ is the j^{th} component of \mathbf{u}_{ki} , $j = 1, 2$. This problem has the unique solution

$$\mathbf{u}_{11} = -\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{u}_{12} = \frac{1}{4} \begin{pmatrix} \sqrt{2} + \sqrt{6} \\ \sqrt{2} - \sqrt{6} \end{pmatrix}, \quad \mathbf{u}_{14} = \frac{1}{4} \begin{pmatrix} \sqrt{2} - \sqrt{6} \\ \sqrt{2} + \sqrt{6} \end{pmatrix}, \quad \mathbf{u}_{23} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and the optimal value is

$$\mathbf{u}_{12}[1] + \mathbf{u}_{14}[2] = \frac{1}{2} (\sqrt{2} + \sqrt{6}) \approx 1.931852$$

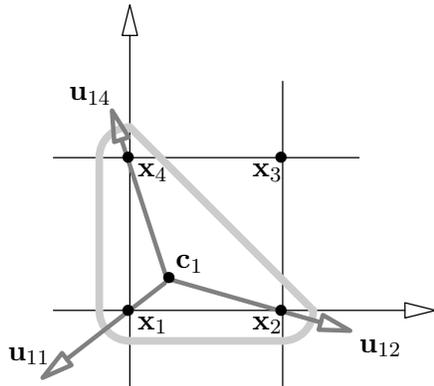


FIGURE 5. Illustration of Example 5: The resultant of the forces \mathbf{u}_{11} , \mathbf{u}_{12} and \mathbf{u}_{14} is zero

which equals the value of Solution 3 in Example 1. By Theorem 2, the centers in Solution 3 are optimal for their clusters. Figure 5 illustrates that the equilibrium condition (46) is satisfied.

Solution 2. We have the same clusters, pure assignments, and dual problem (D) as in Solution 3, and get the same optimal value of (D), which does not satisfy (45),

$$1.931852 < m(\mathbf{c}_1, \mathbf{c}_2) = 2$$

and therefore the centers in Solution 2 are not optimal for their clusters.

Solution 1. The dual problem for Solution 1 is different, and its solution $g(\mathbf{U})$ satisfies (45), showing the centers (arbitrarily chosen on the lines connecting the end points) are optimal for their clusters.

13. CONTOUR APPROXIMATION AND UNCERTAINTY

The **joint distance function (JDF)** at \mathbf{x} , see (11),

$$D(\mathbf{x}) = w \frac{\prod_{j=1}^K d(\mathbf{x}, \mathbf{c}_j)}{\sum_{\ell=1}^K \prod_{m \neq \ell} d(\mathbf{x}, \mathbf{c}_m)},$$

depends on the centers $\{\mathbf{c}_k : k \in 1:K\}$, and has the following useful property: for optimal centers, most of the data points are contained in the lower level sets of $D(\mathbf{x})$, see [1], [25]. We call this property **contour approximation** of the data.

The JDF of the set $\mathbf{X} = \{\mathbf{x}_i : i \in 1:N\}$ is defined as the sum of the values $D(\mathbf{x})$ over \mathbf{X} ,

$$D(\mathbf{X}) := \sum_{i=1}^N D(\mathbf{x}_i). \quad (47)$$

From the contour approximation property it follows that $D(\mathbf{X})$ is a measure of the proximity of customers to their respective facilities, and the lower the value of $D(\mathbf{X})$, the better is the set of centers $\{\mathbf{c}_k\}$.

The JDF has the dimension of weighted distance. Normalizing it, we get the dimensionless function

$$E(\mathbf{x}) = \frac{1}{w} K D(\mathbf{x}) / \left(\prod_{j=1}^K d(\mathbf{x}, \mathbf{c}_j) \right)^{1/K}, \quad (48)$$

with $0/0$ interpreted as zero. $E(\mathbf{x})$ is the harmonic mean of the distances $\{d(\mathbf{x}, \mathbf{c}_j) : j \in 1:K\}$ divided by their geometric mean. It follows that $0 \leq E(\mathbf{x}) \leq 1$, with $E(\mathbf{x}) = 0$ if any $d(\mathbf{x}, \mathbf{c}_j) = 0$, i.e. if \mathbf{x} is a cluster center, and $E(\mathbf{x}) = 1$ if and only if the distances $d(\mathbf{x}, \mathbf{c}_j)$ (and therefore the probabilities $p_j(\mathbf{x})$) are all equal.

$E(\mathbf{x})$ can be written, using (10)–(11), as the geometric mean of the cluster membership probabilities (up to a constant),

$$E(\mathbf{x}) = K \left(\prod_{j=1}^K p_j(\mathbf{x}) \right)^{1/K}. \quad (49)$$

In particular, for $K = 2$,

$$E(\mathbf{x}) = 2 \frac{\sqrt{d(\mathbf{x}, \mathbf{c}_1)d(\mathbf{x}, \mathbf{c}_2)}}{d(\mathbf{x}, \mathbf{c}_1) + d(\mathbf{x}, \mathbf{c}_2)} = 2 \sqrt{p_1(\mathbf{x})p_2(\mathbf{x})}. \quad (50)$$

The function $E(\mathbf{x})$ represents the **uncertainty** of classifying the point \mathbf{x} , with higher [lower] values of $E(\mathbf{x})$ corresponding to higher [lower] uncertainty, see Appendix B. We call $E(\mathbf{x})$ the **classification uncertainty function**, abbreviated **CUF**, at \mathbf{x} .

The CUF of the data set $\mathbf{X} = \{\mathbf{x}_i : i \in 1:N\}$ is defined as

$$E(\mathbf{X}) := \frac{1}{N} \sum_{i=1}^N E(\mathbf{x}_i). \quad (51)$$

$E(\mathbf{X})$ is a monotone decreasing function of K , the number of clusters, decreasing from $E(\mathbf{X}) = 1$ (for $K = 1$), to $E(\mathbf{X}) = 0$ (for $K = N$, the trivial case where every data point is a separate cluster, with no uncertainty.) Indeed, increasing the number of centers from K to $K + 1$ allows more customers access to closer centers, with increased probabilities for these centers.

The “right” number K of facilities to serve the given customers is in general determined by economic considerations (operating costs, etc.) An intrinsic criterion for determining the optimal K is provided by the rate of decrease of $E(\mathbf{X})$, see e.g. Figure 6(d).

Example 6. Figure 6(a) shows the dataset of Example 3. This problem was solved for $K = 3$, and Figure 6(b) shows the centers, and some level sets of the JDF (11), illustrating the contour approximation property, namely that lower contours of the JDF capture most of the data.

Figure 6(c) shows some level sets of the CUF (49), with darker shades indicating higher uncertainty. Uncertainty is minimal ($E(\mathbf{x}) = 0$) if \mathbf{x} is one of the centers. Note the patch of maximal uncertainty ($E(\mathbf{x}) \geq 0.99$) in the middle, where one is indifferent between the three centers⁹.

The “right” number of clusters (if not known a priori) can be determined by solving the problem (P.K) for values of $K = 1, 2, \dots$, calculating the CUF of the whole dataset (51), and stopping when the marginal decrease of uncertainty is negligible. The results are plotted in Figure 6(d), showing that most of the

⁹In analogy with **Lagrange points** in Astronomy, where the gravity pulls from several stars cancel.

uncertainty is reduced by $K = 3$, with a marginal reduction of uncertainty for higher values of K . If the data is amorphous with no clear number of clusters, then the graph of $E(\mathbf{X})$ does not give a clue.

The “right” number of clusters can be determined similarly from the JDF $D(\mathbf{X})$, (47), computed for $K = 1, 2, \dots$, etc., see [24]. □

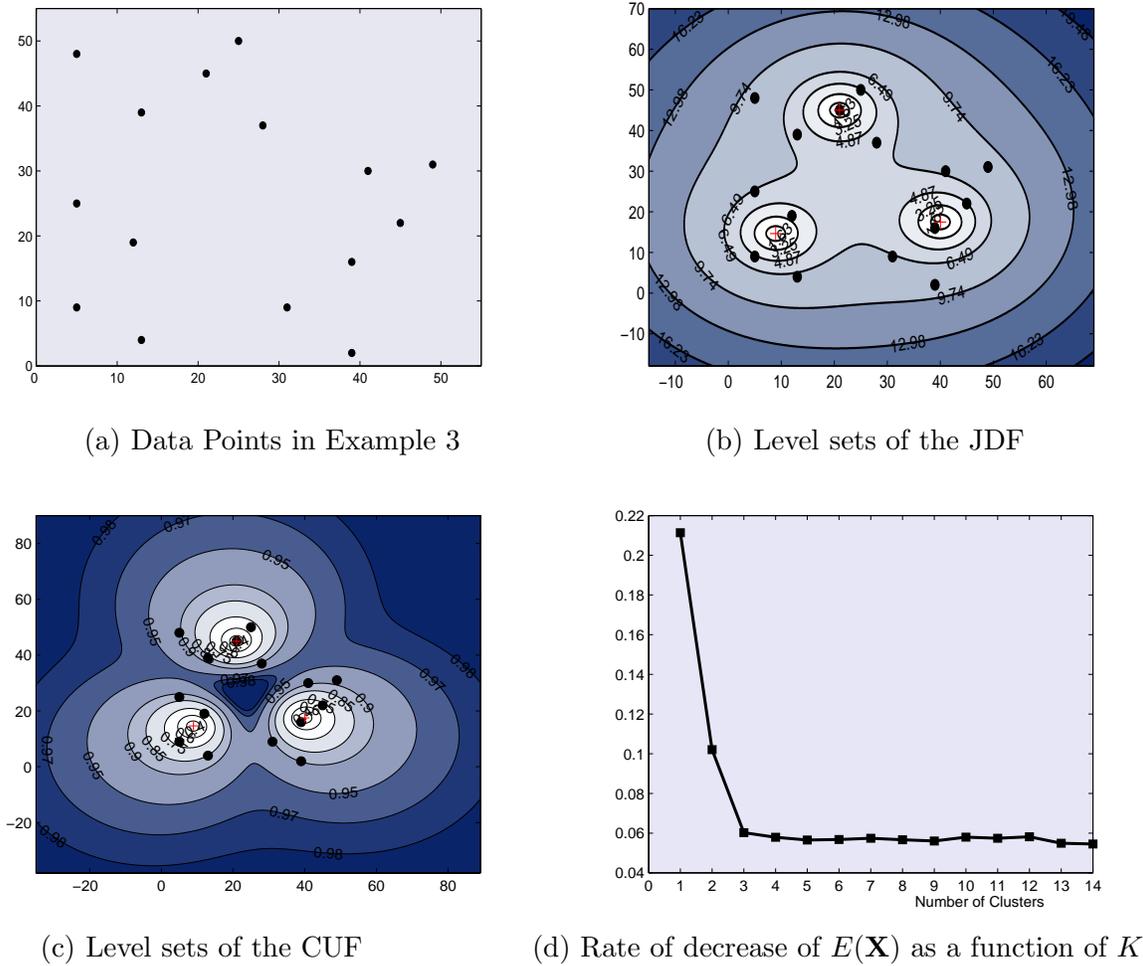


FIGURE 6. Illustration of Example 6

14. NUMERICAL EXAMPLES

Tables 2–5 below present results for the 31 test problems considered in [6]. For each of these problems, Table 2 gives the source (references in bibliography), the number N of customers, the number K of facilities, and the coordinates of the N customers.

Each problem was solved 500 times, using random initializations. The results, for $\nu_0 = 5$, $\Delta = 0.5$ and $\epsilon = 0.001$, are reported in Table 3.

Column 2 (“Best known f ”) gives the best values of the objective in the literature ([6], [8]), that are assumed optimal.

Column 3 (“Best f ”) gives the best values obtained with Algorithm 1. For all 31 problems, Algorithm 1

TABLE 2. Data for Problems 1–31

#	Source	N	K	Coordinates (x, y) of data points				
1	[10, Case 1]	7	2	(15, 15) (31, 23)	(5, 10) (22, 29)	(10, 27)	(16, 8)	(25, 14)
2	[10, Case 2]	7	2	(6, 8) (36, 8)	(6, 32) (36, 32)	(20, 8)	(20, 20)	(20, 32)
3	[10, Case 3]	7	2	(8, 12) (35, 26)	(5, 19) (35, 31)	(5, 26)	(5, 32)	(35, 20)
4	[10, Case 4]	7	2	(5, 23) (31, 32)	(9, 32) (16, 12)	(15, 23)	(21, 32)	(26, 23)
5	[10, Case 5]	7	2	(8, 10) (24, 17)	(8, 26) (31, 19)	(11, 20)	(17, 15)	(17, 22)
6	[10, Case 6]	7	2	(2, 19) (18, 16)	(35, 11) (11, 33)	(31, 33)	(25, 26)	(18, 23)
7	[10, Case 7]	7	2	(28, 6) (39, 33)	(28, 33) (6, 10)	(33, 17)	(33, 23)	(39, 6)
8	[40]	15	2					
9	[11], [40]	15	3	(5, 9)	(5, 25)	(5, 48)	(13, 4)	(12, 19)
10	[40]	15	4	(13, 39)	(28, 37)	(21, 45)	(25, 50)	(31, 9)
11	[40]	15	5	(39, 2)	(39, 16)	(45, 22)	(41, 30)	(49, 31)
12	[40]	15	6					
13	[40]	20	2					
14	[40]	20	3					
15	[40]	20	4	(53, 8)	(1, 34)	(33, 8)	(3, 26)	(17, 9)
16	[40]	20	5					
17	[40]	20	6					
18	[40]	25	2					
19	[40]	25	3					
20	[40]	25	4	(53, 20)	(24, 17)	(40, 22)	(22, 41)	(7, 13)
21	[40]	25	5					
22	[40]	25	6					
23	[40]	30	2					
24	[40]	30	3					
25	[40]	30	4	(5, 17)	(39, 3)	(50, 50)	(16, 40)	(22, 45)
26	[40]	30	5					
27	[19]	50	2	(1.33, 8.89)	(1.89, 0.77)	(9.27, 1.49)	(9.46, 9.36)	(9.20, 8.69)
28	[19]	50	3	(7.43, 1.61)	(6.08, 1.34)	(5.57, 4.60)	(6.70, 2.77)	(8.99, 2.45)
29	[19]	50	4	(8.93, 7.00)	(8.60, 0.53)	(4.01, 0.31)	(3.34, 4.01)	(6.75, 5.57)
30	[19]	50	5	(7.36, 4.03)	(1.24, 6.69)	(3.13, 1.92)	(8.86, 8.74)	(4.18, 3.74)
31	[38]	50	10	(2.22, 4.35)	(0.88, 7.02)	(8.53, 7.04)	(6.49, 6.22)	(4.53, 7.87)
				(4.46, 7.91)	(2.83, 9.88)	(3.39, 5.65)	(0.75, 4.98)	(7.55, 5.79)
				(8.45, 0.69)	(3.33, 5.78)	(6.27, 3.66)	(7.31, 1.61)	(6.37, 7.02)
				(7.23, 7.05)	(1.68, 6.45)	(3.54, 7.06)	(7.67, 4.17)	(2.20, 1.12)
				(3.57, 1.99)	(7.34, 1.38)	(6.58, 4.49)	(5.00, 9.00)	(6.63, 5.23)
				(5.89, 8.06)	(1.13, 5.25)	(1.90, 8.35)	(1.74, 1.37)	(9.37, 6.44)

obtained the optimal values. In problems 27–31, marked by “†” in Column 3, Algorithm 1 gave somewhat better results than the best known values, perhaps because of more precise cluster centers. In problems 19,

20 and 25, marked by a “*” in Column 3, getting the correct 3rd-decimal required a tolerance $\epsilon = 0.0001$. Column 4 (“Frequency (%)”) gives the percentage of the runs where Algorithm 1 gave the optimal solution. In problem 3 the best value was found in all runs.

Columns 5 and 6 give, respectively, the worst values (“Max. f Value”) found, and their frequency (“Frequency (%)”).

Columns 7 and 8 give, respectively, the average values (“Average f Value”) of all 500 solutions obtained, and the average number of iterations.

(Algorithm 1 is here tested using a program written in MATLAB R2010a and all experiments are conducted on a machine with 2.53 GHz Intel(R) Core(TM)2 Duo processor.)

Table 4 reports the same information as Table 3, for $\Delta = 0.1$ and $\epsilon = 0.0001$. The 2nd column of Table 3 is not repeated.

Remark 6. A comparison of Table 3 (with $\Delta = 0.5$, $\epsilon = 0.001$) and Table 4 ($\Delta = 0.1$, $\epsilon = 0.0001$) shows that the smaller increment Δ requires more iterations, but also increases the frequency of optimal values. To prevent early termination, a small tolerance ϵ is suggested in conjunction with the small Δ . This calls for further research, including other selections of the exponent ν , in particular an adaptive control where the exponent ν is updated according to the change $\sum_{k=1}^K d(\mathbf{c}_k^+, \mathbf{c}_k)$ in Step 4 of Algorithm 1.

Example 7. This example uses data from Bongartz et al [6, p. 308], consisting of 287 points, with different weights ranging from $w = 1$ to $w = 698$. The data points, shown in Fig. 7(a), are distinguished according to their weights. Customers with weight < 20 (called “minor customers”) are represented by dots. The squares show the “major customers”: the white squares corresponding to customers with weights $20 \leq w < 100$ and the gray squares to those with weights ≥ 100 .

The problem was solved for $K = 2, \dots, 5, 10$ using 200 random initializations for each K . The results are shown in Table 5. The optimal values in the 2nd column are in agreement with [8]. The 3rd and 5th columns give, respectively, the best and average values found in 200 random initializations. The errors, in columns 4 and 6, are relative to the optimal value. The last column gives the average number of iterations.

□

Example 8. To illustrate the robustness of Algorithm 1, see Remark 2(c), we solved Example 7 again just for the 61 major customers. The same center locations were found in both cases. The solution for $K = 3$ is shown in Fig. 7(b), together with the major customers. □

REFERENCES

- [1] M. Arav, Contour approximation of data and the harmonic mean, *J. Math. Inequalities* **2**(2008), 161–167
- [2] F. Aurenhammer, Voronoi diagrams - a survey of a fundamental geometric data structure, *ACM Computing Surveys* **23**(1991), 345–405
- [3] E. Balas and C-S Yu, A note on the Weiszfeld–Kuhn algorithm for the general Fermat problem, Tech. Report, Carnegie–Melon University, September 1982
- [4] A. Ben–Israel and C. Iyigun, Probabilistic distance clustering, *J. Classification* **25**(2008), 5–26
- [5] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981
- [6] I. Bongartz, E.H. Calamai and A.R. Conn, A projection method for ℓ_p norm location–allocation problems, *Math. Programming* **66**(1994), 283–312.

TABLE 3. Results for $\nu_0 = 5$, $\Delta = 0.5$ and $\epsilon = 0.001$

Problem #	Best known f	Min. f Value	Frequency (%)	Max. f Value	Frequency (%)	Average f Value	Average # of Iterations
1	50.449	50.449	49.80	58.500	3.40	51.451	53.086
2	72.000	72.000	71.40	87.122	0.80	74.680	30.978
3	38.318	38.318	100.00	38.318	n.a.	38.318	341.998
4	48.850	48.850	93.00	57.650	1.00	49.358	41.158
5	38.033	38.033	91.20	45.456	5.30	38.611	32.606
6	59.716	59.716	13.20	65.878	0.20	62.276	42.384
7	62.203	62.203	72.80	68.968	3.60	62.983	68.300
8	214.281	214.281	85.80	232.616	0.40	216.833	62.664
9	143.196	143.196	91.40	185.387	0.40	145.485	105.056
10	113.568	113.568	58.00	132.522	3.00	119.497	60.202
11	97.289	97.289	6.80	122.556	0.20	103.448	46.246
12	81.264	81.264	5.60	110.514	0.10	89.342	39.548
13	287.606	287.606	61.60	291.390	0.60	288.202	73.440
14	210.195	210.195	50.20	271.481	0.40	217.033	57.996
15	169.434	169.434	21.40	218.306	0.60	174.801	62.116
16	134.263	134.263	23.40	173.498	0.20	146.502	64.622
17	116.312	116.312	8.20	167.773	0.20	126.854	63.480
18	355.336	355.336	97.40	408.860	2.60	356.728	72.600
19	252.245	252.246*	77.40	334.403	0.20	257.194	71.084
20	207.402	207.403*	49.60	244.876	0.40	210.585	94.564
21	169.832	169.832	57.00	217.795	0.80	180.231	109.598
22	152.664	152.664	3.60	201.719	0.20	160.838	102.720
23	447.727	447.727	73.20	449.485	6.80	447.898	92.370
24	307.372	307.372	99.00	307.849	1.00	307.382	85.920
25	254.146	254.147*	62.60	285.587	0.60	260.664	114.000
26	220.056	220.056	27.80	270.470	0.20	229.815	87.400
27	135.522	135.513†	0.60	144.559	0.20	136.762	52.900
28	105.214	105.194†	1.20	113.798	0.20	106.572	51.950
29	84.154	84.134†	1.80	95.594	0.20	86.705	83.262
30	72.237	72.225†	1.40	89.901	0.40	73.983	45.156
31	41.685	41.673†	1.00	55.552	0.20	44.915	59.574

[7] J. Brimberg, Further notes on convergence of the Weiszfeld algorithm, *Yugoslav J. Operations Research* **13**(2003), 199–206[8] J. Brimberg, P. Hansen, N. Mladenović and E.D. Taillard, Improvements and comparison of heuristics for solving the uncapacitated multisource Weber problem, *Operations Research* **48**(2000), 444–460[9] R. Chandrasekaran and A. Tamir, Open questions concerning Weiszfeld's algorithm for the Fermat–Weber location problem, *Math. Programming* **44**(1989), 293–295[10] L. Cooper, Location–allocation problems, *Operations Research* **11**(1963), 331–343.

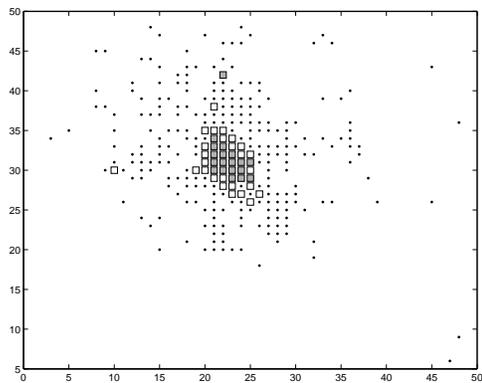
TABLE 4. Results for $\nu_0 = 5$, $\Delta = 0.1$ and $\epsilon = 0.0001$

Problem #	Min. f Value	Frequency (%)	Max. f Value	Frequency (%)	Average f Value	Average # of Iterations
1	50.449	55.40	58.501	3.40	51.410	411.688
2	72.000	74.40	81.764	4.00	74.308	61.832
3	38.318	100.00	38.318	100.00	38.318	973.155
4	48.850	95.40	57.650	4.60	49.255	121.252
5	38.033	86.80	48.068	0.20	38.748	31.042
6	59.716	17.40	65.871	0.20	61.973	188.906
7	62.203	79.60	68.968	3.40	62.798	62.342
8	214.281	89.40	232.226	10.60	216.183	362.698
9	143.196	97.80	172.188	2.20	143.834	192.270
10	113.568	59.00	132.522	6.20	119.426	135.108
11	97.289	9.60	123.306	0.40	103.133	172.550
12	81.264	10.80	107.289	0.20	88.511	179.946
13	287.606	95.80	289.123	4.20	287.670	379.588
14	210.195	65.00	234.675	13.40	215.914	217.610
15	169.434	27.40	218.306	0.80	173.303	205.930
16	134.263	55.40	178.164	0.20	144.668	206.210
17	116.312	14.20	153.638	0.20	126.082	197.458
18	355.336	99.00	408.860	1.00	355.871	371.790
19	252.246	93.20	305.447	0.80	254.077	392.712
20	207.403	62.60	241.876	1.20	210.344	483.272
21	169.832	63.40	217.193	0.20	179.344	366.894
22	152.664	4.60	193.662	0.20	160.075	310.980
23	447.727	100.00	447.727	100.00	447.727	567.000
24	307.372	100.00	307.372	100.00	307.372	408.000
25	254.146	65.80	280.952	8.60	260.910	673.274
26	220.056	29.80	271.407	0.20	229.094	471.118
27	135.520	100.00	135.520	100.00	135.520	283.514
28	105.199	81.60	110.533	18.40	106.177	283.328
29	84.134	74.60	94.274	0.60	86.225	537.374
30	72.226	4.60	89.898	0.40	73.556	357.688
31	41.674	0.40	55.218	0.20	44.725	231.574

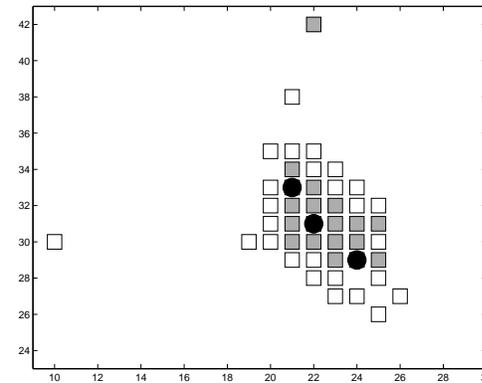
- [11] L. Cooper, Heuristic methods for location-allocation problems, *SIAM Review* **6**(1964), 37–53
- [12] T. Drezner, Z. Drezner, The gravity p -median model, *European J. Oper. Res.* **179**(2007), 1239–1251.
- [13] Z. Drezner, The planar two-center and two-median problems, *Transportation Science* **18**(1984), 351–361.
- [14] Z. Drezner (editor), *Location Decisions, Annals Oper. Res.* **40**(1992)
- [15] Z. Drezner, A note on the Weber location problem, pp. 153–161 in [14]
- [16] Z. Drezner and H.W. Hamacher (editors), *Facility Location: Applications and Theory*, Springer, 2002

TABLE 5. Results for Example 7, $\nu_0 = 5$ and $\Delta = 0.1$

K	Optimum Value [8]	Best f Value	Error (%)	Average f Value	Error (%)	Average # of Iterations
2	14427.593	14447.970	0.14	14497.313	0.33	297.22
3	12095.442	12115.823	0.17	12354.682	2.14	203.02
4	10661.477	10681.842	0.19	11066.700	3.80	192.83
5	9715.627	9735.993	0.21	10041.572	3.35	173.73
10	6705.036	6.714.277	0.14	7275.055	8.50	202.30



(a) 287 data points



(b) Major customers and centers

FIGURE 7. Illustration of Example 8 ($N = 287$, $K = 3$)

- [17] Z. Drezner, K. Klamroth, A. Schöbel and G.O. Wesolowsky, The Weber problem, Chapter 1 in [16]
- [18] I.N. Eckhardt, Weber's problem and Weiszfeld's algorithm in general spaces, *Math. Program.* **18**(1980), 186–196
- [19] S. Eilon, C.D.T. Watson–Gandy and N. Christofides, *Distribution Management: Mathematical Modelling and Practical Analysis*, Hafner, New York, 1971
- [20] S.L. Hakimi, Optimum distribution of switching centers in a communication network and some related graph theoretic problems, *Operations Resesarch* **13**(1965), 462–475
- [21] D.L. Huff, Defining and estimating a trading area, *J. Marketing* **28**(1964), 34–38
- [22] D.L. Huff, A programmed solution for approximating an optimum retail location, *Land Economics* **42**(1966), 293–303.
- [23] C. Iyigun and A. Ben–Israel, Probabilistic distance clustering adjusted for cluster size, *Probability Engrg. Info. Sci.* **22**(2008), 1–19
- [24] C. Iyigun and A. Ben–Israel, Probabilistic Distance Clustering, Algorithm and Applications, Chapter 2, *Clustering Challenges in Biological Networks*, S. Butenko, W.A. Chaovaitwongse and P.M. Pardalos (Editors), World Scientific, Singapore, 2009, ISBN-13 978-981-277-165-0
- [25] C. Iyigun and A. Ben–Israel, Contour approximation of data: a duality theory, *Lin. Algeb. and Appl.* **430**(2009), 2771–2780
- [26] C. Iyigun and A. Ben–Israel, A generalized Weiszfeld method for the multi–facility location problem, *Operations Research Letters* **38**(2010), 207–214
- [27] W. Kaplan and W.H. Yang, Duality theorem for a generalized Fermat–Weber problem, *Math. Programming* **76**(1997), 285–297
- [28] I.N. Katz, Local convergence in Fermat's problem, *Math. Program.* **6**(1974), 89–104
- [29] H.W. Kuhn, On a pair of dual nonlinear programs, in J. Abadie (ed.), *Methods of Nonlinear Programming*, Amsterdam, North-Holland, (1967), 38–54
- [30] H.W. Kuhn, A note on Fermat's problem, *Math. Programming* **4**(1973), 98–107

- [31] S. Kullback. *Information Theory and Statistics*. J. Wiley, New York, 1959.
- [32] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals Math. Statist.*, **22**(1951), 79–86
- [33] Y. Levin and A. Ben-Israhel, A heuristic method for large-scale multi-facility location problems, *Computers and Oper. Res.* **31**(2004), 257–272
- [34] R.F. Love and H. Juel, Properties and solution methods for large location–allocation problems, *J. Operational Research Society* **33**(1982), 443–452
- [35] R.F. Love, J.G. Morris and G.O. Wesolowsky, *Facilities Location: Models and Methods*, North-Holland, 1988
- [36] R.D. Luce, *Individual Choice Behavior*, Wiley, New York, 1959
- [37] N. Megiddo and K.J. Supowit, On the complexity of some common geometric location problems, *SIAM Journal on Computing* **13**(1984), 182–196
- [38] B.A. Murtagh and S.R. Niwattisyawong, An efficient method for the multi-depot location–allocation problem, *Journal of Operational Research Society* **33** (1982) 629–634.
- [39] C. ReVelle and R. Swain, Central facilities location, *Geographical Analysis* **2**(1970), 30–42
- [40] K.E. Rosing, An optimal method for solving the multi-Weber problems, *European Journal of Operational Research* **58** (1992) 414–426.
- [41] M. Teboulle, A unified continuous optimization framework for center-based clustering methods, *J. Machine Learning* **8**(2007), 65–102
- [42] H. Üster and R.F. Love, Duality in constrained multi-facility location models, *Naval Research Logistics Quart.* **49**(2002), 410–421
- [43] Y. Vardi and C-H Zhang, A modified Weiszfeld algorithm for the Fermat–Weber location problem, *Math. Programming* **A90**(2001), 559–566
- [44] E. Weiszfeld, Sur le point par lequel la somme des distances de n points donnés est minimum, *Tohoku Math. J.* **43** (1937), 355–386
- [45] G.O. Wesolowsky, The Weber problem: its history and perspectives, *Location Science* **1**(1993) 5–23
- [46] J.I. Yellott, Jr., Luce’s Choice Axiom, pp. 9094–9097 in *International Encyclopedia of the Social & Behavioral Sciences* (N.J. Smelser and P.B. Baltes, editors), ISBN 0-08-043076-7, 2001.

Appendix A: Proof of Theorem 2

Proof of part (a). The objective function $g(\mathbf{U})$ can be written, using (41), as

$$g(\mathbf{U}) = \sum_{k=1}^K \left(\sum_{i=1}^N \mathbf{u}_{ki} \cdot \mathbf{x}_i - \left(\sum_{i=1}^N \mathbf{u}_{ki} \right) \cdot \mathbf{c}_k \right) = \sum_{k=1}^K \sum_{i=1}^N \mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k). \quad (\text{A1})$$

$$\leq \sum_{k=1}^K \sum_{i=1}^N \|\mathbf{u}_{ki}\| \|\mathbf{x}_i - \mathbf{c}_k\| \quad (\text{A2})$$

$$\leq \sum_{k=1}^K \sum_{i=1}^N w_i q_{ki} \|\mathbf{x}_i - \mathbf{c}_k\|, \text{ by (42),} \quad (\text{A3})$$

$$= f(\mathbf{c}_1, \dots, \mathbf{c}_K).$$

Proof of part (b). We distinguish two cases.

Case 1. None of the centers $\{\mathbf{c}_k\}$ coincides with any of the data points $\{\mathbf{x}_i\}$.

For $k \in 1:K$ and $i \in 1:N$ define

$$\mathbf{u}_{ki} := \frac{w_i q_{ki}}{\|\mathbf{x}_i - \mathbf{c}_k\|} (\mathbf{x}_i - \mathbf{c}_k). \quad (\text{A4})$$

Then from (27a) and Theorem 1 it follows that,

$$\sum_{i=1}^N \mathbf{u}_{ki} = \mathbf{R}_k(\mathbf{c}_k) = \mathbf{0}, \text{ verifying (41).}$$

Also from (A4), we get, for all k, i ,

$$\|\mathbf{u}_{ki}\| = w_i q_{ki}, \quad (\text{A5})$$

proving that the inequalities (42) hold as equalities, and that $\{\mathbf{u}_{ki}\}$, defined by (A4), are feasible.

Finally, from (A4) and (A5) it follows that (A2) and (A3) are satisfied as equalities, proving (44).

Case 2. A center coincides with one of the data points, say

$$\mathbf{c}_k = \mathbf{x}_j, \text{ for some } k \in 1:K, j \in 1:N, \quad (\text{A6})$$

in which case (15) holds. Define

$$\mathbf{u}_{ki} := \frac{w_i q_{ki}}{\|\mathbf{x}_i - \mathbf{x}_j\|} (\mathbf{x}_i - \mathbf{x}_j), \quad \text{for } i \neq j, \quad (\text{A7a})$$

$$\mathbf{u}_{kj} := - \sum_{i \neq j} \mathbf{u}_{ki}. \quad (\text{A7b})$$

Then $\sum_i \mathbf{u}_{ki} = \mathbf{0}$ by definition, and $\|\mathbf{u}_{ki}\| = w_i q_{ki}$ for all $i \neq j$, as in (A5). Next,

$$\mathbf{u}_{kj} = -\mathbf{R}_k^j \text{ by (27c),}$$

and therefore by (27b),

$$\mathbf{R}_k(\mathbf{x}_j) = \mathbf{0} \text{ implies } w_j q_{kj} \geq \|\mathbf{R}_k^j\| = \|\mathbf{u}_{kj}\|,$$

proving that the variables \mathbf{U} defined by (A7a)–(A7b) are feasible.

From Case 1, the equality

$$\mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k) = w_i q_{ki} \|\mathbf{x}_i - \mathbf{c}_k\|,$$

holds for all \mathbf{x}_i that are not centers.

Also, for any center $\mathbf{c}_m \neq \mathbf{c}_k$, $q_{mj} = 0$ (by (15)), and therefore $\mathbf{u}_{mj} = \mathbf{0}$. In (A2)–(A3) the inequalities

$$\mathbf{u}_{kj} \cdot (\mathbf{x}_j - \mathbf{c}_k) \leq \|\mathbf{u}_{kj}\| \|\mathbf{x}_j - \mathbf{c}_k\| \leq w_j q_{kj} \|\mathbf{x}_j - \mathbf{c}_k\|,$$

reduce to trivial equalities, since by (A6) all three terms are zero.

The inequalities (A2)–(A3) therefore hold as equalities, proving (44).

Proof of part (c). Writing the objective function $g(\mathbf{U})$ as in (A1), the Lagrangian of (D) is

$$\sum_{k=1}^K \sum_{i=1}^N \mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k) - \sum_{k=1}^K \sum_{i=1}^N t_{ki} (\|\mathbf{u}_{ki}\| - w_i q_{ki})$$

with Lagrange multipliers $\{t_{ki}\}$. The Karush–Kuhn–Tucker conditions are

$$(\mathbf{x}_i - \mathbf{c}_k) - t_{ki} \frac{\mathbf{u}_{ki}}{\|\mathbf{u}_{ki}\|} = \mathbf{0}, \quad (\text{A8a})$$

$$\sum_{i=1}^N \mathbf{u}_{ki} = \mathbf{0}, \quad (\text{A8b})$$

$$\|\mathbf{u}_{ki}\| \leq w_i q_{ki}, \quad (\text{A8c})$$

$$t_{ki} \geq 0, \quad (\text{A8d})$$

$$t_{ki} (\|\mathbf{u}_{ki}\| - w_i q_{ki}) = 0, \quad (\text{A8e})$$

for all $k \in 1:K$, $i \in 1:N$. Again we distinguish two cases.

Case 1. All $t_{ki} > 0$. Then (A5) follows from (A8e) for all k, i , and from (A8a),

$$\|\mathbf{x}_i - \mathbf{c}_k\| = t_{ki}, \quad (\text{A9})$$

and by substituting (A5) and (A9) in (A8a), we get

$$\mathbf{u}_{ki} := \frac{w_i q_{ki}}{\|\mathbf{x}_i - \mathbf{c}_k\|} (\mathbf{x}_i - \mathbf{c}_k),$$

which is (A4), and the equality (44) follows as in the proof of Theorem 2(b), Case 1.

Case 2. Some Lagrange multipliers are zero, say $t_{kj} = 0$. Then by (A8a), $\mathbf{c}_k = \mathbf{x}_j$, and $t_{ki} > 0$ for $i \neq j$, by (A8a) and (A8d). Therefore, by (A8e),

$$\|\mathbf{u}_{ki}\| = w_i q_{ki}, \quad \text{for all } i \neq j.$$

From $\mathbf{c}_k = \mathbf{x}_j$ and (A8a) it follows that

$$(\mathbf{x}_i - \mathbf{x}_j) = t_{ki} \frac{\mathbf{u}_{ki}}{\|\mathbf{u}_{ki}\|}$$

and by taking norms,

$$\|\mathbf{x}_i - \mathbf{x}_j\| = t_{ki}, \quad \text{for all } i \neq j.$$

Substituting t_{ki} and $\|\mathbf{u}_{ki}\|$ in (A8a) gives,

$$\mathbf{u}_{ki} := \frac{w_i q_{ki}}{\|\mathbf{x}_i - \mathbf{x}_j\|} (\mathbf{x}_i - \mathbf{x}_j), \quad \text{for } i \neq j,$$

and from (A8b),

$$\mathbf{u}_{kj} := - \sum_{i \neq j} \mathbf{u}_{ki},$$

reproducing (A7a)–(A7b), and equality in (44) follows as in the proof of Theorem 2(b), Case 2. \square

Appendix B: The classification uncertainty function

Let \mathbb{P}^K be the set of K -dimensional probability vectors, denoted $\mathbf{p} = (p_i)$, $\mathbf{q} = (q_i)$. The **Kullback–Leibler divergence**,

$$I(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^K p_i \log \left(\frac{p_i}{q_i} \right), \quad (\text{B1})$$

is a well-known and useful distance function on \mathbb{P}^K , [31], [32].

Rewriting the CUF (49) as

$$E(\mathbf{x}) = \left(\prod_{j=1}^K \left(\frac{p_j(\mathbf{x})}{1/K} \right) \right)^{1/K}$$

and taking logarithms, we get

$$-\log E(\mathbf{x}) = \sum_{i \in \mathbf{K}} (1/K) \log \left(\frac{1/K}{p_i(\mathbf{x})} \right) = I(\mathbf{p}(\mathbf{x}), \frac{1}{K} \mathbf{1}), \quad (\text{B2})$$

the Kullback–Leibler distance between the distributions

$$\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), \dots, p_K(\mathbf{x})) \quad \text{and} \quad \frac{1}{K} \mathbf{1} = \left(\frac{1}{K}, \frac{1}{K}, \dots, \frac{1}{K} \right).$$

The latter distribution, $\frac{1}{K} \mathbf{1}$, is of maximal uncertainty in \mathbb{P}^K , and consequently the divergence $I(\mathbf{p}(\mathbf{x}), \frac{1}{K} \mathbf{1})$ in the right side of (B2) is a measure of the uncertainty of the distribution $\mathbf{p}(\mathbf{x})$, with smaller values corresponding to greater uncertainty.

Writing (B2) as

$$E(\mathbf{x}) = \exp \{ -I(\mathbf{p}(\mathbf{x}), \frac{1}{K} \mathbf{1}) \} \quad (\text{B3})$$

shows that $E(\mathbf{x})$ is an entropic measure of the uncertainty of classification, with higher values of $E(\mathbf{x})$ corresponding to a greater uncertainty in assigning \mathbf{x} to the given centers.

Appendix C: Cluster membership probabilities

In this appendix, $d_k(\mathbf{x})$ stands for $d(\mathbf{x}, \mathbf{c}_k)$, the distance of \mathbf{x} to the k^{th} -center \mathbf{c}_k .

The **membership probabilities** $\{p_k(\mathbf{x}) : k \in 1:K\}$ of a point \mathbf{x} are assumed to depend only on the **distances** $\{d_k(\mathbf{x}) : k \in 1:K\}$,

$$\mathbf{p}(\mathbf{x}) = \mathbf{f}(\mathbf{d}(\mathbf{x})) \quad (\text{C1})$$

where $\mathbf{p}(\mathbf{x}) \in \mathbb{R}^K$ is the vector of probabilities $(p_k(\mathbf{x}))$, and $\mathbf{d}(\mathbf{x})$ is the vector of distances $(d_k(\mathbf{x}))$. Desirable properties of the relation (C1) include

$$d_i(\mathbf{x}) < d_j(\mathbf{x}) \implies p_i(\mathbf{x}) > p_j(\mathbf{x}), \quad \text{for all } i, j \in 1:K, \quad (\text{C2a})$$

$$\mathbf{f}(\lambda \mathbf{d}(\mathbf{x})) = \mathbf{f}(\mathbf{d}(\mathbf{x})), \quad \text{for any } \lambda > 0, \quad (\text{C2b})$$

$$Q \mathbf{p}(\mathbf{x}) = \mathbf{f}(Q \mathbf{d}(\mathbf{x})), \quad \text{for any permutation matrices } Q. \quad (\text{C2c})$$

Condition (C2a) states that membership in a cluster is more probable the closer it is, which is Assumption (A) of § 4. Assuming continuity of \mathbf{f} it follows from (C2a) that

$$d_i(\mathbf{x}) = d_j(\mathbf{x}) \implies p_i(\mathbf{x}) = p_j(\mathbf{x}),$$

for any $i, j \in 1:K$. In particular, the probabilities $p_k(\mathbf{x})$ are all equal only if so are the distances $d_k(\mathbf{x})$.

The meaning of (C2b) is that the probabilities $p_k(\mathbf{x})$ do not depend on the scale of measurement, i.e., the function \mathbf{f} is homogeneous of degree 0. It follows that the probabilities $p_k(\mathbf{x})$ depend only on the ratios of the distances $\{d_k(\mathbf{x}) : k \in 1:K\}$, as is assumed in what follows.

The symmetry of \mathbf{f} , expressed by (C2c), guarantees that the probabilities $\{p_k(\mathbf{x})\}$ do not depend on the numbering of the clusters.

For any nonempty subset $\mathcal{S} \subset 1:K$, let

$$p_{\mathcal{S}}(\mathbf{x}) = \sum_{s \in \mathcal{S}} p_s(\mathbf{x}),$$

the probability that \mathbf{x} belongs to one of the clusters $\{\mathbf{X}_s : s \in \mathcal{S}\}$, and let $p_k(\mathbf{x}|\mathcal{S})$ denote the **conditional probability** that \mathbf{x} belongs to the cluster \mathbf{X}_k , given that it belongs to one of the clusters $\{\mathbf{X}_s : s \in \mathcal{S}\}$.

Since the probabilities $p_k(\mathbf{x})$ depend only on the ratios of the distances $\{d_k(\mathbf{x}) : k \in 1:K\}$, and these ratios are unchanged in subsets \mathcal{S} of the index set $1:K$, it follows that for all k and $\mathcal{S} \subset 1:K$ with $k \in \mathcal{S}$,

$$p_k(\mathbf{x}) = p_k(\mathbf{x}|\mathcal{S}) p_{\mathcal{S}}(\mathbf{x}) \tag{C3}$$

which is the **choice axiom** of Luce, [36, Axiom 1], and therefore, [46],

$$p_k(\mathbf{x}|\mathcal{S}) = \frac{v_k(\mathbf{x})}{\sum_{s \in \mathcal{S}} v_s(\mathbf{x})} \tag{C4}$$

where $v_k(\mathbf{x})$ is a scale function, in particular,

$$p_k(\mathbf{x}) = \frac{v_k(\mathbf{x})}{\sum_{s \in 1:K} v_s(\mathbf{x})}. \tag{C5}$$

Assuming $v_k(\mathbf{x}) \neq 0$ for all k , it follows that

$$p_k(\mathbf{x}) v_k(\mathbf{x})^{-1} = \frac{1}{\sum_{s \in 1:K} v_s(\mathbf{x})}, \tag{C6}$$

where the right hand side is a function of \mathbf{x} , and does not depend on k .

Property (C2a) implies that the function $v_k(\mathbf{x})$ decreases as $d_k(\mathbf{x})$ increases. A simple choice is

$$v_k(\mathbf{x}) = \frac{1}{w d_k(\mathbf{x})}, \tag{C7}$$

for which (C6) gives

$$w p_k(\mathbf{x}) d_k(\mathbf{x}) = \frac{1}{\sum_{s \in 1:K} \left(\frac{1}{d_s(\mathbf{x})} \right)}, \tag{C8}$$

in agreement with (9).

Supplementary Material: Proof of Theorem 1

The proof in [26, Appendix A] is adapted below for the case of probabilistic assignments (18a).

Proof of part (a). If \mathbf{c}_k is not one of the data points, then $-\mathbf{R}_k(\mathbf{c}_k)$ is the gradient (23) at \mathbf{c}_k , and (36) is both necessary and sufficient for a minimum, by the convexity of f_k .

If \mathbf{c}_k coincides with a data point \mathbf{x}_j , consider the change from \mathbf{x}_j to $\mathbf{x}_j + t \mathbf{z}$ where $\|\mathbf{z}\| = 1$. Then,

$$\frac{d}{dt} f_k(\mathbf{x}_j + t \mathbf{z}) \Big|_{t=0} = w_j q_k(\mathbf{x}_j) - \mathbf{R}_k^j \cdot \mathbf{z}. \tag{D1}$$

The greatest decrease of f_k is along \mathbf{R}_k^j , i.e., when

$$\mathbf{z} = \frac{\mathbf{R}_k^j}{\|\mathbf{R}_k^j\|}, \text{ (proving part (f)).}$$

Therefore \mathbf{c}_k (that coincides with \mathbf{x}_j) is a local minimum if and only if,

$$w_j q_k(\mathbf{x}_j) - \frac{\mathbf{R}_k^j \cdot \mathbf{R}_k^j}{\|\mathbf{R}_k^j\|} \geq 0,$$

which is equivalent to

$$\|\mathbf{R}_k^j\| \leq w_j q_k(\mathbf{x}_j),$$

or $\mathbf{R}_k(\mathbf{c}_k) = \mathbf{0}$, by (27b).

Proof of part (b). If \mathbf{c}_k^* is one of the data points, then it is trivially in the convex hull. Otherwise, the condition $\mathbf{R}_k(\mathbf{c}_k^*) = \mathbf{0}$, see (36), results in (24a)–(24b) as above.

Proof of part (c). Follows from part (a), since for $\mathbf{c} \notin \mathbf{X}$,

$$\mathbf{c} = T_k(\mathbf{c}) \iff R_k(\mathbf{c}) = \mathbf{0},$$

while for $\mathbf{c} \in \mathbf{X}$, $\mathbf{c} = T_k(\mathbf{c})$ for all k .

Proof of part (d). If $T_k(\mathbf{c}) \neq \mathbf{c}$ then \mathbf{c} is not one of the data points. It follows then from (25a) that $T_k(\mathbf{c})$ is the center of gravity of weights

$$\theta_{ki} := \frac{w_i q_k(\mathbf{x}_i)}{\|\mathbf{x}_i - \mathbf{c}\|}, \text{ placed at the data points } \mathbf{x}_i, \quad i \in 1:N. \quad (\text{D2})$$

By elementary calculus, $T_k(\mathbf{c})$ is the unique minimum of the strictly convex function

$$g(\mathbf{y}) = \sum_{i=1}^N \theta_{ki} \|\mathbf{x}_i - \mathbf{y}\|^2.$$

Since $\mathbf{c} \neq T_k(\mathbf{c})$,

$$g(T_k(\mathbf{c})) = \sum_{i=1}^N \theta_{ki} \|\mathbf{x}_i - T_k(\mathbf{c})\|^2 < g(\mathbf{c}) = \sum_{i=1}^N \theta_{ki} \|\mathbf{x}_i - \mathbf{c}\|^2 = f_k(\mathbf{c}).$$

On the other hand,

$$\begin{aligned} g(T_k(\mathbf{c})) &= \sum_{i=1}^N \theta_{ki} \left[\|\mathbf{x}_i - \mathbf{c}\| + \left(\|\mathbf{x}_i - T_k(\mathbf{c})\| - \|\mathbf{x}_i - \mathbf{c}\| \right) \right]^2 \\ &= f_k(\mathbf{c}) + 2 \left(f_k(T_k(\mathbf{c})) - f_k(\mathbf{c}) \right) + \sum_{i=1}^N \theta_{ki} \left[\|\mathbf{x}_i - T_k(\mathbf{c})\| - \|\mathbf{x}_i - \mathbf{c}\| \right]^2 \end{aligned}$$

Combining these results,

$$2f_k(T_k(\mathbf{c})) + \sum_{i=1}^N \theta_{ki} \left[\|\mathbf{x}_i - T_k(\mathbf{c})\| - \|\mathbf{x}_i - \mathbf{c}\| \right]^2 < 2f_k(\mathbf{c})$$

proving that $f_k(T_k(\mathbf{c})) < f_k(\mathbf{c})$.

Proof of part (e). Follows from part (a), the definition (27b)

$$\mathbf{R}_k(\mathbf{x}_j) := \max \left\{ \|\mathbf{R}_k^j\| - w_j q_k(\mathbf{x}_j), 0 \right\} \frac{\mathbf{R}_k^j}{\|\mathbf{R}_k^j\|},$$

and $q_k(\mathbf{x}_j) = 1$ since $\mathbf{x}_j = \mathbf{c}_k$.

Proof of part (f). This was shown in part (a).

Proof of part (g).

$$\begin{aligned} T_k(\mathbf{c}) - \mathbf{x}_i &= \mathbf{c} + h_k(\mathbf{c}) R_k(\mathbf{c}) - \mathbf{x}_i \\ &= h_k(\mathbf{c}) \sum_{j \neq i}^N \theta_{kj} (\mathbf{x}_j - \mathbf{c}) + \left(h_k(\mathbf{c}) \theta_{ki} - 1 \right) (\mathbf{x}_i - \mathbf{c}), \end{aligned}$$

with $h_k(\mathbf{c})$ as in (35) and θ_{ki} in (D2). Since \mathbf{x}_i is not optimal, we have, from (27c),

$$\left\| \sum_{j \neq i}^N \theta_{kj} (\mathbf{x}_j - \mathbf{x}_i) \right\| > w_i q_k(\mathbf{x}_i),$$

and therefore there exists $\delta' > 0$ and $\epsilon > 0$ such that

$$\left\| \sum_{j \neq i}^N \theta_{kj}(\mathbf{x}_j - \mathbf{c}) \right\| \geq (1 + 2\epsilon) w_i q_k(\mathbf{x}_i) \quad \text{for } \|\mathbf{x}_i - \mathbf{c}\| \leq \delta'.$$

By the definition of h_k , (35), we have

$$\lim_{\mathbf{c} \rightarrow \mathbf{x}_i} \frac{h_k(\mathbf{c}) w_i q_k(\mathbf{x}_i)}{\|\mathbf{x}_i - \mathbf{c}\|} = 1.$$

and therefore there exists $\delta'' > 0$ such that

$$\left| \frac{h_k(\mathbf{c}) w_i q_k(\mathbf{x}_i)}{\|\mathbf{x}_i - \mathbf{c}\|} - 1 \right| < \frac{\epsilon}{2(1 + \epsilon)} \quad \text{for } 0 < \|\mathbf{x}_i - \mathbf{c}\| \leq \delta''.$$

For $\delta := \min(\delta', \delta'')$, and $0 < \|\mathbf{x}_i - \mathbf{c}\| \leq \delta$, we have

$$\begin{aligned} \|\mathbf{x}_i - T_k(\mathbf{c})\| &> h_k(\mathbf{c})(1 + 2\epsilon) w_i q_k(\mathbf{x}_i) - \frac{\epsilon}{2(1 + \epsilon)} \|\mathbf{x}_i - \mathbf{c}\| \\ &> \left(1 - \frac{\epsilon}{2(1 + \epsilon)}\right) (1 + 2\epsilon) \|\mathbf{x}_i - \mathbf{c}\| - \frac{\epsilon}{2(1 + \epsilon)} \|\mathbf{x}_i - \mathbf{c}\| \\ &= (1 + \epsilon) \|\mathbf{x}_i - \mathbf{c}\|. \end{aligned}$$

If $\|\mathbf{x}_i - T_k(\mathbf{c})\| \leq \delta$, then

$$\|\mathbf{x}_i - T_k^2(\mathbf{c})\| > (1 + \epsilon)^2 \|\mathbf{x}_i - \mathbf{c}\|, \text{ etc.}$$

Since $\|\mathbf{x}_i - \mathbf{c}\| > 0$, $(1 + \epsilon)^t \|\mathbf{x}_i - \mathbf{c}\| > \delta$ for some positive integer t and hence

$$\|\mathbf{x}_i - T_k^s(\mathbf{c})\| > \delta \text{ for some positive integer } s, \text{ and } \|\mathbf{x}_i - T_k^{s-1}(\mathbf{c})\| \leq \delta.$$

Proof of part (h). For $\mathbf{c} \notin \{\mathbf{x}_i : i \in 1:N\}$,

$$\begin{aligned} T_k(\mathbf{c}) &= \sum_{i=1}^N \lambda_{ki} \mathbf{x}_i, \quad \text{with } \lambda_{ki} \text{ as in (25a),} \\ &= \sum_{i \neq j}^N \lambda_{ki} (\mathbf{x}_i - \mathbf{x}_j) + \mathbf{x}_j \sum_{i=1}^N \lambda_{ki}. \\ \therefore T_k(\mathbf{c}) - \mathbf{x}_j &= \sum_{i \neq j}^N \lambda_{ki} (\mathbf{x}_i - \mathbf{x}_j), \quad \text{since } \sum_{i=1}^N \lambda_{ki} = 1. \\ \therefore \frac{T_k(\mathbf{c}) - \mathbf{x}_j}{\|\mathbf{x}_j - \mathbf{c}\|} &= \sum_{i \neq j}^N \frac{\lambda_{ki}}{\|\mathbf{x}_j - \mathbf{c}\|} (\mathbf{x}_i - \mathbf{x}_j). \end{aligned} \tag{D3}$$

Using (25a),

$$\begin{aligned} \frac{\lambda_{ki}}{\|\mathbf{x}_j - \mathbf{c}\|} &= \frac{\frac{w_i q_k(\mathbf{x}_i)}{\|\mathbf{x}_i - \mathbf{c}\|}}{\sum_{m=1}^N \frac{w_m q_k(\mathbf{x}_m) \|\mathbf{x}_j - \mathbf{c}\|}{\|\mathbf{x}_m - \mathbf{c}\|}} = \frac{\frac{w_i q_k(\mathbf{x}_i)}{\|\mathbf{x}_i - \mathbf{c}\|}}{w_j q_k(\mathbf{x}_j) + \sum_{m \neq j} w_m q_k(\mathbf{x}_m) \frac{\|\mathbf{x}_j - \mathbf{c}\|}{\|\mathbf{x}_m - \mathbf{c}\|}} \\ \therefore \lim_{\mathbf{c} \rightarrow \mathbf{x}_j} \frac{\lambda_{ki}}{\|\mathbf{x}_j - \mathbf{c}\|} &= \frac{w_i q_k(\mathbf{x}_i)}{w_j q_k(\mathbf{x}_j) \|\mathbf{x}_i - \mathbf{c}\|}. \end{aligned}$$

Finally, by (D3),

$$\begin{aligned} \lim_{\mathbf{c} \rightarrow \mathbf{x}_j} \frac{\|T_k(\mathbf{c}) - \mathbf{x}_j\|}{\|\mathbf{x}_j - \mathbf{c}\|} &= \frac{1}{w_j q_k(\mathbf{x}_j)} \left\| \sum_{i \neq j}^N w_i q_k(\mathbf{x}_i) \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|} \right\| \\ &= \frac{\|\mathbf{R}_k^j\|}{w_j q_k(\mathbf{x}_j)}, \text{ by (27c)}. \end{aligned}$$

Proof of part (i). With the possible exception of \mathbf{c}_k^0 , the sequence $\{\mathbf{c}_k^r\}$ lies in the convex hull of vertices, a compact set. By the Bolzano–Weierstrass theorem, there exists a convergent subsequence $\{\mathbf{c}_k^\ell : \ell = 1, 2, \dots\}$, such that $\lim_{\ell \rightarrow \infty} \mathbf{c}_k^\ell = \mathbf{c}_k^\infty$. We prove that $\mathbf{c}_k^\infty = \mathbf{c}_k^*$.

If $\mathbf{c}_k^{\ell+1} = T_k(\mathbf{c}_k^\ell) = \mathbf{c}_k^\ell$ for some ℓ , then the sequence repeats from that point and $\mathbf{c}_k^\infty = \mathbf{c}_k^\ell$. Since \mathbf{c}_k^ℓ is not a data point, $\mathbf{c}_k^\infty = \mathbf{c}_k^*$ by Theorem 1(c).

Otherwise, by Theorem 1(d),

$$f_k(\mathbf{c}_k^0) > f_k(\mathbf{c}_k^1) > \dots > f_k(\mathbf{c}_k^r) > \dots > f_k(\mathbf{c}_k^*) \quad (\text{D4})$$

Hence

$$\lim_{\ell \rightarrow \infty} (f_k(\mathbf{c}_k^\ell) - f_k(T_k(\mathbf{c}_k^\ell))) = 0. \quad (\text{D5})$$

The continuity of T_k implies

$$\lim_{\ell \rightarrow \infty} T_k(\mathbf{c}_k^\ell) = T_k(\mathbf{c}_k^\infty) \quad (\text{D6})$$

we have

$$f_k(\mathbf{c}_k^\infty) - f_k(T_k(\mathbf{c}_k^\infty)) = 0. \quad (\text{D7})$$

Therefore, by Theorem 1(d), $\mathbf{c}_k^\infty = T_k(\mathbf{c}_k^\infty)$. If \mathbf{c}_k^∞ is not a data point, then $\mathbf{c}_k^\infty = \mathbf{c}_k^*$ by Theorem 1(c).

In any event, \mathbf{c}_k^∞ lies in the finite set of points $\{\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{c}_k^*\}$, where \mathbf{c}_k^* may be a data point.

The only case that remains is $\mathbf{c}_k^\infty = \mathbf{x}_j$ for some j . If $\mathbf{x}_j \neq \mathbf{c}_k^*$, we first isolate \mathbf{x}_j from the other data points (and \mathbf{c}_k^* if it is not a data point) by a δ -neighborhood that satisfies Theorem 1(g). Then it is clear that we can choose our subsequence $\mathbf{c}_k^\ell \rightarrow \mathbf{x}_j$ such that $\|\mathbf{x}_j - T_k(\mathbf{c}_k^\ell)\| > \delta$ for all ℓ . This means that the ratio $\frac{\|\mathbf{x}_j - T_k(\mathbf{c}_k^\ell)\|}{\|\mathbf{x}_j - \mathbf{c}_k^\ell\|}$ is unbounded. However, this contradicts Theorem 1(h). Hence $\mathbf{x}_j = \mathbf{c}_k^*$ and the theorem is proved. \square

CEM IYIGUN, DEPARTMENT OF INDUSTRIAL ENGINEERING, MIDDLE EAST TECHNICAL UNIVERSITY, ANKARA, TURKEY

E-mail address: iyigun@ie.metu.edu.tr

ADI BEN-ISRAEL, RUTCOR–RUTGERS CENTER FOR OPERATIONS RESEARCH, RUTGERS UNIVERSITY, 640 BARTHOLOMEW RD., PISCATAWAY, NJ 08854-8003, USA

E-mail address: adi.benisrael@gmail.com