A Contribution to Convergence Theory of Fuzzy c-Means and Derivatives*

Frank Höppner, *Member, IEEE* Frank Klawonn, *Member, IEEE*

Department of Computer Science
University of Applied Sciences BS/WF
Salzdahlumer Str. 46/48
D-38302 Wolfenbüttel, Germany
contact: frank.hoeppner@ieee.org

Abstract

In this paper we revisit the convergence and optimization properties of fuzzy clustering algorithms in general and the fuzzy c-means (FCM) algorithm in particular. Our investigation includes probabilistic and (a slightly modified implementation of) possibilistic memberships, which will be discussed under a unified view. We give a convergence proof for the axis-parallel variant of the algorithm by Gustafson and Kessel, that can be generalized to other algorithms more easily than in the usual approach. Using reformulated fuzzy clustering algorithms we apply Banach's classical contraction principle and establish a relationship between saddle points and attractive fixed points. For the special case of FCM we derive a sufficient condition for fixed points to be attractive, allowing identification of them as (local) minima of the objective function (excluding the possibility of a saddle point).

Keywords: fuzzy clustering, fuzzy c-means, possibilistic c-means, convergence, fixed point iteration, attractive fixed point, saddle-point

1 Introduction

The fuzzy c-means algorithm of Dunn [8] and Bezdek [2] is very popular and has been applied successfully in many areas. It partitions a data set $X = \{x_1, ..., x_n\} \subset \mathcal{X}$, $\mathcal{X} = \mathbb{R}^{\mathrm{DIM}}$, into c clusters that are characterized by representatives or prototypes $\mathbf{p} = (p_1, ..., p_c) \in \mathcal{X}^c$. The process of subdividing a data set X into distinct subsets with homogeneous elements is called clustering. With fuzzy clustering each datum x_j belongs to all clusters p_i simultaneously, but to different degrees $u_{i,j}$ with $U = [u_{i,j}] \in [0,1]^{c \times n}$. The fuzzy

*This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant Kl 648/1.

c-means algorithm performs clustering by minimizing the objective function of weighted distances

$$J_{FCM}(X;\mathbf{p},U)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{i,j}^{m}\|x_j-p_i\|^2 \tag{1}$$

taking the constraints

$$\forall i\in\mathbb{N}_{\leq c}\quad:\quad\sum_{j=1}^{n}u_{i,j}>0 \tag{2}$$

$$\forall j\in\mathbb{N}_{\leq n}\quad:\quad\sum_{i=1}^{c}u_{i,j}=1 \tag{3}$$

into account. The parameter m influences the fuzziness of the partition, and it is referred to as "fuzzifier". Constraint (2) makes sure that none of the clusters is empty and thus we really have a partition into no less than c clusters. Constraint (3) assures that every datum has the same overall weight in the data set. Fuzzy clustering under constraint (3) is often called "probabilistic clustering". Other fuzzy clustering techniques, using a relaxed constraint (3), are possibilistic clustering [7, 13] and noise clustering [7, 6], where modified objective functions J_{FCM} are used. Minimization of J_{FCM} is done by alternating optimization (AO); that is, J_{FCM} is optimized with respect to $u_{i,j}$ (assuming prototypes p_i to be constant) and with respect to prototype p_i (assuming memberships $u_{i,j}$ to be constant) alternatingly. Both minimization steps are repeated until the change in memberships and/or prototypes drops below a certain threshold. Bezdek has shown [2] that in each (half-) step of AO J_{FCM} is strictly minimized. By a theorem of Zangwill [14] convergence – or at least convergence of a subsequence – can be concluded. Since $\mathbf{p}$ and U are optimized independently, the algorithm may yield results representing jointly a saddle point.

There are only a few artificial examples where FCM has been proven to converge to a saddle point [5], but it is also difficult to decide in practice if such a case has occurred or not. Therefore, we consider convergence of probabilistic and possibilistic fuzzy clustering algorithms in general and contractive properties of FCM, as the most popular algorithm, in particular. Problems that are encountered when investigating these topics, as already mentioned by Bezdek, are the "two-part compositional nature of one iteration" and "the further difficulty of establishing local zones of convergence". The first problem can be solved easily by reformulating FCM [10]. Instead of addressing the second problem directly, we derive a sufficient condition for attractive fixed points. We show that it depends on the ratio of the number of data assigned almost unambiguously to those assigned ambiguously to the clusters, whether a fixed point is attractive or not. We establish a relationship between saddle points of J_{FCM} and attractive fixed points: If we find a particular solution of FCM to be attractive we have found a minimum (not a saddle-point) of J_{FCM}.

The paper is organized as follows. We complete the short review of fuzzy clustering algorithms in section 2 before we provide a motivation for reformu-

lated algorithms independently of the objective-function approach in section 3. Properties of fixed points of reformulated fuzzy clustering iterations are discussed in general in section 4. Then, we examine the monotonicity of J for probabilistic and possibilistic membership functions in section 5, where we also provide convergence proofs for another clustering algorithm besides FCM. Finally, in section 6 we discuss properties of FCM attractive fixed points, before we come to the conclusions in section 7. The appendices contain the proofs of the preceding sections.

2 Fuzzy Clustering Algorithms

Besides the cluster shape used by FCM, many other shapes have been proposed by different authors, see e.g. [11]. These algorithms usually need further prototype parameters and use modified distance measures. We denote the space of prototypes by $\mathcal{P}$ and introduce $d_s(x, \mathbf{p})$ to measure the distance between object x and prototype p_s. With FCM we have $\mathcal{P} = \mathcal{X}$ and $d_s(x, \mathbf{p}) = \|x - p_s\|^2$. Or with the AGK algorithm (an axis-parallel variant [12, 11] of the algorithm by Gustafson and Kessel [9]) the prototypes consist of a centre c_s and a diagonal matrix A_s with $\det(A_s) = 1$, that is $p_s = (c_s, A_s) \in \mathcal{P} := \mathcal{X} \times \{A \in \mathbb{R}^{\mathrm{DIM}\times\mathrm{DIM}} \,|\, A \text{ diagonal matrix}, \det(A) = 1\}$, and it uses a distance measure $d_s(x, \mathbf{p}) = (x - c_s)^\top A_s (x - c_s)$.

Objective-function based fuzzy clustering algorithms minimize the (constrained) objective function

$$J_{prob}(X; \mathbf{p}, U) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{i,j}^m d_s(x_j, \mathbf{p}) \tag{4}$$

by means of alternating optimization. For each half-step we obtain prototype and membership update equations from the necessary condition of a minimum

$$\frac{\partial J_{prob}^{(L)}}{\partial \mathbf{p}} = 0 \qquad \text{(assuming } U \text{ to be constant) and} \tag{5}$$

$$\frac{\partial J_{prob}^{(L)}}{\partial U} = 0 \qquad \text{(assuming } \mathbf{p} \text{ to be constant),} \tag{6}$$

respectively, where $J_{prob}^{(L)}$ additionally considers (3) by means of Lagrange multipliers. For the probabilistic membership model we obtain the update equation

$$u_{i,j} = \begin{cases} \left(\sum_{k=1}^{c} \left(\frac{\|x_j - p_i\|^2}{\|x_j - p_k\|^2} \right)^{\frac{1}{m-1}} \right)^{-1} & : \text{ in case } I_j = \emptyset \\ \frac{1}{|I_j|} & : \text{ in case } I_j \neq \emptyset, i \in I_j \\ 0 & : \text{ in case } I_j \neq \emptyset, i \notin I_j \end{cases} \tag{7}$$

where $I_j = \{i \in \mathbb{N}_{\leq c} \,|\, x_j = p_i\}$, and for the point-like FCM prototype model the update equation

$$p_i = \frac{\sum_{j=1}^{n} u_{i,j}^m x_j}{\sum_{j=1}^{n} u_{i,j}^m} \tag{8}$$

When switching to different membership functions than (7) the prototype update equations remain the same. We obtain alternative membership functions by relaxing (3), but then have to change the objective function J_{prob} in order to avoid the trivial solution $U \equiv 0$. We have to incorporate a term that increases J_{prob} with decreasing membership values, such as [13]

$$J_{poss}(X; \mathbf{p}, U) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^m \|x_j - p_i\|^2 + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - u_{i,j})^m, \tag{9}$$

where $\eta_i \in \mathbb{R}$ are constant parameters, leading to

$$u_{i,j} = \frac{1}{1 + (d_s(x, \mathbf{p})/\eta_i)^{1/(m-1)}} \tag{10}$$

When using memberships (10) instead of (7), FCM becomes possibilistic c-means (PCM). Although by PCM we always refer to (10), the possibilistic approach (relaxation of (3)) can be used to derive other membership functions as well [7].

3 Fuzzy Partitioning

Before we step into the details of the update equations, let us first analyse fuzzy clustering algorithms from an intuitive perspective. It is sometimes claimed that FCM memberships lack a meaningful model, e.g. when compared to probability functions in Gaussian mixture decomposition, since they have been obtained from an "arbitrary" objective function J_{prob} or J_{poss} and are not meaningful by themselves. In this section we develop the membership functions independently from J_{prob} or J_{poss} thereby refuting this claim.

Let us investigate how to define a *reasonable similarity measure* based on the given dissimilarity (or distance) values. The c-means algorithms can be summarized as iterative procedures that repeatedly replace the prototypes by the mean of *similar* data objects. With hard c-means (HCM) similarity is a binary measure that becomes 1 if a data object is closer to a prototype than to any other prototype, and 0 otherwise. What are other reasonable similarity measures? Suppose we have a tuple $\mathbf{p} \in \mathcal{P}^c$ of c reference objects (or prototypes) and dissimilarity measures

$$D_s : \mathcal{X} \times \mathcal{P}^c \to \mathbb{R}_{>0} \tag{11}$$

such that $D_s(x, \mathbf{p})$ increases as the dissimilarity of an object x to the s^{th} reference object increases, $1 \leq s \leq c$. We assume that the D_s functions do not only depend on p_s and x, but may also depend on the locations of other prototypes, which is why we introduce one D_s for each prototype and take all prototypes

$\mathbf{p} \in \mathcal{P}^c$ as their argument. We are interested in a similarity measure between x and the c prototypes based on the dissimilarities $D_s(x, \mathbf{p})$. Let us denote such a similarity measure by

$$u_s : \mathcal{X} \times \mathcal{P}^c \to \mathbb{R}_{>0} \tag{12}$$

The similarity $u_s(x, \mathbf{p})$ should increase as the dissimilarity $D_s(x, \mathbf{p})$ decreases; that is, for all $(x, \mathbf{p}) \in \mathcal{X} \times \mathcal{P}^c$:

$$u_s(x, \mathbf{p}) > u_i(x, \mathbf{p}) \Leftrightarrow D_s(x, \mathbf{p}) < D_i(x, \mathbf{p}) \tag{13}$$

From (13) a similarity measure u_s can be defined by means of $u_s(x, \mathbf{p}) = f(D_s(x, \mathbf{p}))$ with an arbitrary strictly decreasing function $f : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, e.g. $f(t) = \exp(-t)$. Although such an f may turn out to be useful in specific applications, it does not respect the duality between dissimilarity and similarity in the following sense: If we compare object x to p_1 and p_2 via their dissimilarity (e.g. $D_1(x, \mathbf{p}) = 2$ and $D_2(x, \mathbf{p}) = 4$) we see that the dissimilarity of x and p_2 is two times as high as the dissimilarity of x and p_1. However, if we compare the similarity values we find that x and p_1 are about seven times as similar as x and p_2 are. Intuitively we would expect p_1 and p_2 to be dissimilar to x to the same degree as p_2 and p_1 are similar to x. Making no further assumptions on the interpretation of the dissimilarity values there is no good reason for such a deviation between similarity and dissimilarity, and therefore their duality should be preserved: We say that u_s is dual to D_s if the following condition holds

$$\forall (x, \mathbf{p}) \in \mathcal{X} \times \mathcal{P}^c : \forall s, i \in \{1, .., c\} : \quad D_s(x, \mathbf{p}) \cdot u_s(x, \mathbf{p}) = D_i(x, \mathbf{p}) \cdot u_i(x, \mathbf{p}) \tag{14}$$

If the distance $D_1(x, \mathbf{p})$ is half of the distance $D_2(x, \mathbf{p})$ it follows from (14) that the similarity $u_1(x, \mathbf{p})$ is twice the similarity $u_2(x, \mathbf{p})$. To satisfy this condition we could use $f(t) = \frac{1}{t}$ or the similarity measure[1]

$$u_s(x, \mathbf{p}) = \frac{1}{D_s(x, \mathbf{p})}. \tag{15}$$

So far, we have compared the similarity of a data object to prototypes in a vector of reference objects $\mathbf{p}$. Now let us consider the case of having multiple prototype vectors $\mathbf{p}$ and $\mathbf{p}'$. Let us assume objects x with $D_1(x, \mathbf{p}) = 2$, $D_2(x, \mathbf{p}) = 4$ and x' with $D_1(x', \mathbf{p}') = 4$, $D_2(x', \mathbf{p}') = 8$ – how similar are both cases? Obviously, the data objects and prototypes have just been scaled by a factor of 2, their similarity values should ideally be the same. When we compare $u_1(x, \mathbf{p}) = \frac{1}{2}$ with $u_1(x', \mathbf{p}') = \frac{1}{4}$, this comparison lacks a common baseline since the overall membership of x to the reference objects $\mathbf{p}$ is $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ but the overall membership of x' to reference objects $\mathbf{p}'$ is only $\frac{3}{8}$. For a meaningful comparison, the membership values should be normalized: We say u_s is normalized if the following condition holds

$$\exists C \in \mathbb{R}_{>0} : \forall (x, \mathbf{p}) \in \mathcal{X} \times \mathcal{P}^c : \quad \sum_{s=1}^{c} u_s(x, \mathbf{p}) = C \tag{16}$$

[1] Here u_s is not a fuzzy similarity measure since u_s stays not necessarily within $[0, 1]$.

Requiring u_s to be dual to D_s and normalized, it is easy to see that only a unique similarity measure remains (up to a constant C):

$$u_s : \mathcal{X} \times \mathcal{P}^c \to [0,1], \qquad (x, \mathbf{p}) \mapsto \frac{C}{\sum_{i=1}^{c} \frac{D_s(x,\mathbf{p})}{D_i(x,\mathbf{p})}} \tag{17}$$

Now, if $D_i(x, \mathbf{p}) = \alpha D_i(x', \mathbf{p}')$ for all $1 \leq i \leq c$ we also have $u_i(x, \mathbf{p}) = u_i(x', \mathbf{p}')$ by using (17) and similarity is thus invariant to scaling.

Two reasonable candidates for similarity measures have evolved in this short discussion, (15) and (17). Let us now identify the relationship to the membership functions of fuzzy clustering algorithms. Let us define $D_s = d_s + \varepsilon$ ($\varepsilon \in \mathbb{R}_{>0}$)[2] and focus on the case $m = 2$ at the beginning. Then (17) becomes identical to probabilistic membership functions and replacing HCM's binary similarity measures by u_s leads to FCM. With (15) we also obtain a generalization of HCM, but u_s are not fuzzy memberships, since $u_s \in [0, \infty] \neq [0, 1]$. There are multiple ways how to get $u_s \in [0,1]$, we prefer to redefine d_s by $d_s/\eta_s + 1$. It is easy to show that this modification does not change the prototype update equations, but leads to PCM memberships and thus to PCM itself. Since the non-fuzzy case is interesting in its own right, we do not want to substitute d_s by $d_s/\eta_s + 1$ in general. Only if it is clear from the context that we address fuzzy memberships or PCM, the substitution applies.

In the previous paragraph we have considered the case $m = 2$ only, because a fuzzifier of 2 causes the distances to appear unchanged (no exponent) in the membership functions (7) and (10). We do not see an intuitive motivation for the introduction of a fuzzifier, it can only be motivated by considering the objective function itself. Nevertheless, if we set

$$D_s(x, \mathbf{p}) = d_s(x, \mathbf{p})^{\frac{1}{m-1}} \qquad + \varepsilon \tag{18}$$

we get probabilistic and possibilistic memberships for arbitrary fuzzifiers ($m > 1$) from (17) or (15). Note that our particular implementation of possibilistic memberships differs from the one used for PCM if $m \neq 2$. Figures 1 shows both membership functions for various fuzzifiers in the univariate case with a prototype located at the origin:

$$u_{PCM} = \frac{1}{1 + x^{\frac{1}{m-1}}} \qquad \text{and} \qquad u_{proposed} = \frac{1}{(1+x)^{\frac{1}{m-1}}}$$

It is often said that an increase of m increases the fuzziness of the partition. If we look at the PCM memberships (left subfigure) we see that an increase in m results in more fuzzy memberships far away from the prototype, but close to the prototype an increase in m makes the membership more and more look

[2]Why the ε in the definition of D_s? Formally, (11) requires $D_s > 0$ and without ε the d_s may become zero. More important, the definition (7) distinguishes the case of zero distance values explicitly. If we add a small constant, this does not significantly affect the memberships but we do not need a differentiable continuation of (17) in case one or more distance functions vanish. And by requiring $D_s > 0$ the similarity (15) is well-defined and remains continuous.

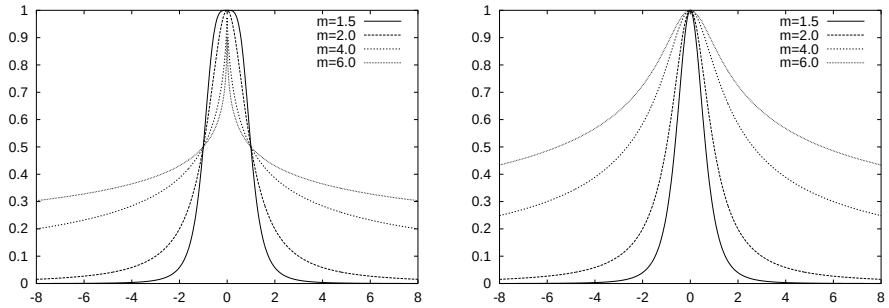


Figure 1: Comparison of different implementations of possibilistic memberships. Left: membership functions used by PCM, right: memberships functions developed in section 3.

like a sharp peak which is somewhat counterintuitive. If we look at the newly proposed membership definition (right subfigure), a more uniform increase in fuzziness can be observed.

In the following two sections we will examine reformulated fuzzy clustering algorithms that use one of these two membership functions, the probabilistic or the newly proposed implementation of possibilistic memberships. We will benefit from our particular implementation of possibilistic memberships in the next section, since it enables us to formulate theorems that hold for probabilistic as well as possibilistic membership functions. For the following, D_s is defined by (18) and u_s by either (15) or (17). When using (15) and replacing d_s by $d_s/\eta_s + 1$ we arrive (for $m = 2$) at PCM memberships.

4 Fixed Point Iteration

In this section we are interested in the AO iterations of fuzzy clustering algorithms. We reformulate the consecutive application of both half-steps by substituting the membership degrees $u_{i,j}$ in the prototype update equation by their definition. Let us fix a data set $X = \{x_1, ..., x_n\} \subset \mathcal{X}$, then the alternating optimization (AO) iteration of any fuzzy clustering algorithm can be reformulated by

$$\mathbf{p}^{(t+1)} = \Phi(\mathbf{p}^{(t)}), \qquad t \in \mathbb{N},$$

where Φ is an algorithm specific mapping and $\mathbf{p}^{(0)}$ are initial prototypes. In case of the fuzzy c-means algorithm we have

$$\Phi_{FCM} : \mathcal{P}^c \to \mathcal{P}^c, \quad \begin{pmatrix} p_1 \\ \vdots \\ p_c \end{pmatrix} \mapsto \begin{pmatrix} \frac{\sum_{j=1}^n u_1^m(x_j,p_1,..,p_c)x_j}{\sum_{j=1}^n u_1^m(x_j,p_1,..,p_c)} \\ \vdots \\ \frac{\sum_{j=1}^n u_c^m(x_j,p_1,..,p_c)x_j}{\sum_{j=1}^n u_c^m(x_j,p_1,..,p_c)} \end{pmatrix} \tag{19}$$

This reformulation is equivalent to the original two-stage FCM algorithm [10]. FCM terminates if consecutive $\mathbf{p}^{(t+1)}$ and $\mathbf{p}^{(t)}$ are almost identical, or in other words we have reached a fixed point $\mathbf{p}^{(t)} = \Phi(\mathbf{p}^{(t)})$. The reformulation (19) can be done with any fuzzy clustering algorithm and is not limited to FCM. However, with other algorithms we obtain definitions for Φ that are different from (19). With the AGK algorithm, for instance, besides the cluster centres p_i the diagonal elements $a_i = (a_{i,1}, ..., a_{i,\mathrm{DIM}}) \in \mathbb{R}^{\mathrm{DIM}}$ of A_i become part of the prototypes and have to be respected by Φ_{AGK}.

With all fuzzy clustering algorithms their prototype update equations are developed from $\frac{\partial J_{prob}}{\partial \mathbf{p}} = 0$ assuming U to be constant. Using this property as the only condition on Φ (see (21) in next theorem), we can show that – if such an iteration sequence has converged – we have found an extremum or saddle point of the objective function (4). This theorem is not specific to Euclidean distances or probabilistic memberships, it is therefore somewhat of a generalization of Bezdek's results.

Theorem 1 *Choose $m \in \mathbb{R}_{>1}$ and a data set $X = \{x_1, ..., x_n\} \subset \mathcal{X}$. Given differentiable functions $d_s : \mathcal{X} \times \mathcal{P}^c \to \mathbb{R}_{>0}$ for $s \in \mathbb{N}_{\leq c}$ we define D_s by (18) and u_s either by (15) or (17). Let*

$$J(\mathbf{p}) = \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m(x_j, \mathbf{p}) \cdot d_s(x_j, \mathbf{p}) \tag{20}$$

and $\Phi : \mathcal{P}^c \to \mathcal{P}^c$ be defined by

$$\forall \xi \in \mathcal{P}^c, \|\xi\| = 1 : \forall s \in \mathbb{N}_{\leq c} : \quad \sum_{j=1}^{n} u_s^m(x_j, \mathbf{p}) \frac{\partial d_s}{\partial \xi}(x_j, \Phi(\mathbf{p})) = 0 \tag{21}$$

If $\mathbf{p}$ is a fixed point of Φ, then $\mathbf{p}$ is an extremum or saddle point of J.

Note that condition (21) is the same as (5), only adapted to the reformulation. We will further investigate into the term "extremum" in the next section. Let us now briefly discuss the relationship between saddle points and attractive fixed points. We call a fixed point "attractive" if there is a neighbourhood of the fixed point on which Φ is a contraction. We intend to apply Banach's classical fixed point theorem that states that a contractive mapping has a unique fixed point. In addition any sequence of elements that we obtain by iterated application of the contractive mapping converges to the fixed point:

Theorem 2 (Banach) *Let V be a Banach space, $W \subset V$ closed. Let $\Phi : W \to W$ be a contraction, that is,*

$$(\exists 0 \leq \alpha < 1)\,(\forall x, y \in W) \qquad \|\Phi(x) - \Phi(y)\| \leq \alpha \|x - y\| \tag{22}$$

Then Φ has exactly one fixed point and the iteration $x^{(t+1)} = \Phi(x^{(t)})$ converges to this fixed point for any $x^{(0)} \in W$.

By applying Banach's theorem to Φ we can show

Theorem 3 *Given a continuous function $J : Y \to \mathbb{R}$ and let $\Phi : Y \to Y$ be a continuous mapping such that J is monotone under Φ, that is,*

$$\forall \mathbf{p} \in Y : J(\Phi(\mathbf{p})) \leq J(\mathbf{p}) \qquad \text{or} \qquad \forall \mathbf{p} \in Y : J(\Phi(\mathbf{p})) \geq J(\mathbf{p}). \tag{23}$$

If $\mathbf{p}$ is a saddle point of J, $\mathbf{p}$ cannot be an attractive fixed point of Φ.

This theorem can be applied to FCM since (23) has been shown by Bezdek for FCM [2, 5] and it is also true for the reformulated FCM. (We will investigate into (23) for other cases in the next section.) This theorem gives theoretical considerations about attractive properties of fixed points a more practical meaning: if we find a particular fixed point $\mathbf{p}$ to be attractive, $\mathbf{p}$ is a (local) extremum of J. As a user of FCM it would be interesting to have a criterion telling us whether the final partition of FCM represents an attractive fixed point (i.e. a local minimum) or whether we got stuck in a saddle point so that another application with a different initialization is definitely recommendable. The theorem also complements the general theorem on alternating optimization provided by Bezdek et al. [4]. Applying their more general main theorem for our purposes of FCM, we have that when the Hessian of J_{FCM} is positive at $(\mathbf{p}^{(0)}, U^{(0)})$, then there is a neighbourhood of $(\mathbf{p}^{(0)}, U^{(0)})$ for which the FCM iteration scheme converges to $(\mathbf{p}^{(0)}, U^{(0)})$.

Finally, to see that we do not have to distinguish between J_{prob} and J in this concern, we need the following two theorems:

Theorem 4 *Let us assume that the parameters $u_{c,j}$ $(j = 1, \ldots, n)$ are eliminated in the objective function (4) by replacing $u_{c,j} = 1 - \sum_{i=1}^{c-1} u_{i,j}$ according to the constraint (3). Let $\mathcal{U}$ be the function induced by (7). Then for $\mathbf{p} \in \mathcal{P}^c$*

$$\nabla J(\mathbf{p}) = \mathbf{0} \;\Leftrightarrow\; \nabla J_{prob}(\mathbf{p}, \mathcal{U}(\mathbf{p})) = \mathbf{0}.$$

Theorem 5 *A $\mathbf{p} \in \mathcal{P}^c$ is a strict local minimum of J if and only if $(\mathbf{p}, U(\mathbf{p}))$ is a strict local minimum of J_{prob}.*

Theorems 4 and 5 establish the equivalence of minima of J_{prob} and J when using (7). Therefore, the results obtained for J in Theorem 3 also apply to J_{prob}. It is known for FCM that J has no maxima, since memberships and prototypes are always chosen such that J is minimized (we will also discuss this point in the next section). From the equivalences in Theorems 4 and 5 we can

thus conclude that saddle points of J_{prob} are mapped to either (local) maxima or saddle points of J. At this point, we cannot state similar theorems for the possibilistic case, since we need some results from section 5.2. Afterwards, the proofs can be easily adapted and the theorems hold for the possibilistic case, too.

Before we develop a condition for attractive fixed points of FCM in section 6, let us first investigate further into condition (23).

5 Monotonicity of J under Φ-Iterations

5.1 Probabilistic Memberships

In this subsection we assume that u_s is defined by (17) and (18). We begin with a relationship between probabilistic fuzzy clustering (AO) algorithms and steepest descent of J:

Theorem 6 *Using the notation of Theorem 1, for the case of u_s defined by (17) a steepest descent algorithm minimizing the objective function (20) is given by*

$$\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)} - \gamma(\nabla_{\mathbf{p}} J)(\mathbf{p}^{(t)}) \tag{24}$$

where γ is the step size and

$$(\nabla_{\mathbf{p}} J)(\mathbf{p}^{(t)}) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_i^m(x_j, \mathbf{p}^{(t)})(\nabla_{\mathbf{p}} d_i)(x_j, \mathbf{p}^{(t)}) \tag{25}$$

Let us compare theorems 1 and 6 when applied to FCM (see also [1]). Theorem 1 requires (21), which becomes

$$-2\sum_{j=1}^{n} u_{s,j}^m (x_j - p_s^{(t+1)}) = 0 \quad \Leftrightarrow \quad p_s^{(t+1)} = \frac{\sum_{j=1}^{n} u_{s,j}^m x_j}{\sum_{j=1}^{n} u_{s,j}^m}$$

and thus defines Φ as given in (19). Theorem 6 requires (24), which becomes

$$p_s^{(t+1)} = p_s^{(t)} + 2\gamma \sum_{j=1}^{n} u_{s,j}^m (x_j - p_s^{(t)})$$

Choosing $\gamma = \left(2\sum_{j=1}^{n} u_{s,j}^m\right)^{-1}$ we obtain the same Φ iteration as before:

$$p_s^{(t+1)} = p_s^{(t)} + 2\gamma \left(\sum_{j=1}^{n} u_{s,j}^m x_j \right) - p_s^{(t)} = \frac{\sum_{j=1}^{n} u_{s,j}^m x_j}{\sum_{j=1}^{n} u_{s,j}^m}$$

The steepest descent approach does not provide a value for γ, however, the fact that the Φ iteration can be interpreted as a steepest descent algorithm

(with automatic step size-adjustment[3]) confirms that FCM converges to a (local) minimum. For probabilistic memberships, we can speak of minima instead of extrema of J in Theorem 1. Other fuzzy clustering algorithms can also be interpreted as a steepest descent procedure (with automatic step size-adjustment). Appendix B contains a proof that AGK can be interpreted in this sense. This is not necessarily the case for all fuzzy clustering algorithms and for some of them it may be difficult to show due to the complexity of their definition of d_s.

Nevertheless, we assume that all (probabilistic) fuzzy clustering algorithms strictly decrease J in each iteration step. This property has been proven for FCM by Bezdek [2], but is missing for many other algorithms. Here, we show this property for the AGK algorithm. The idea of the proof can also be applied to other algorithms.

Theorem 7 *Let $\mathcal{P} = \mathbb{R}^{\mathrm{DIM}} \times \mathbb{R}^{\mathrm{DIM}-1}$ be the space of AGK prototypes, let X have a non-zero variance in each component. Let J be given by (20) with $\Phi = \Phi_{AGK}$ (see appendix B for update of matrix elements). If $\mathbf{p} \in \mathcal{P}^c$ is not a fixed point of Φ_{AGK} we have*

$$J(\Phi_{AGK}(\mathbf{p})) < J(\mathbf{p}). \tag{26}$$

The sketch of the proof is as follows: We argue on J_{prob} and show that

$$\begin{aligned} J(p) = J_{prob}(\mathbf{p}, \mathcal{U}(\mathbf{p})) &\overset{(\star)}{>} J_{prob}(\Phi(\mathbf{p}), \mathcal{U}(\mathbf{p})) \\ &\overset{(\star\star)}{>} J_{prob}(\Phi(\mathbf{p}), \mathcal{U}(\Phi(\mathbf{p}))) = J(\Phi(\mathbf{p})) \end{aligned}$$

where $\mathcal{U}(\mathbf{p})$ denotes the membership matrix obtained by (17) from $\mathbf{p}$. Let us now fix $U = \mathcal{U}(\mathbf{p})$ to show $(\star)$. We argue that if we shift one of the prototype parameters to infinity, J_{prob} approaches infinity if we are beyond the convex hull H of the data set. Thus we know that there *must* be a global minimum of J_{prob} somewhere. Since there is a *unique* solution for (21) in case of AGK (and many other fuzzy clustering algorithms as well) we know that there is only a single extremum or saddle-point of J_{prob} (with U fixed), which therefore must be the global minimum. Thus the prototype update step leads to a strict local minimum of J_{prob} and establishes $(\star)$. Bezdek has shown that a membership update step of J_{prob} also leads us to a strict local minimum, and his proof was independent of the used distance measure, which guarantees $(\star\star)$. The only difficulty that may arise when transferring the proof to other algorithms is to show that J strictly increases if we move a prototype parameter to infinity, which might be difficult if there are constraints on the prototype parameters. See appendix B for the details in case of AGK.

The strict descent of the Φ iteration in J together with the boundedness of $J \geq 0$ guarantees termination of AGK.

[3] Strictly speaking, a steepest descent algorithm has a constant γ, but many extensions have been proposed with varying γ to overcome the problem of getting trapped in local minima. FCM selects γ automatically, so it is more than simple steepest descent.

5.2 Possibilistic Memberships

Here we consider the case of u_s defined by (15) and (18). To get fuzzy similarity values we also have to replace d_s by $d_s/\eta_s + 1$ in J. Recall that this yields an implementation of possibilistic memberships that is different from PCM memberships if $m \neq 2$. By $\mathcal{U}(\mathbf{p})$ we denote the resulting possibilistic membership matrix for $\mathbf{p}$.

Despite the fact that our implementation of possibilistic memberships has been derived without establishing a (modified) objective function in advance, it is nevertheless possible to provide such a modified $J_{poss'}$. We only have to select an appropriate function that decreases monotonically in $u_{i,j}$ to avoid the trivial solution $U \equiv 0$. The most simple choice for such a function is $-u_{i,j}$.

$$J_{poss'}(X; \mathbf{p}, U) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^m d_i(x_k, \mathbf{p}) - m \sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j} \tag{27}$$

It is easily seen that minimization of $J_{poss'}$ with respect to $u_{i,j}$ assuming $\mathbf{p}$ to be constant yields $u_{i,j} = \frac{1}{d_i(x_j,\mathbf{p})^{1/(m-1)}}$. If we replace $u_{i,j}$ in (27) we obtain

$$J_{poss'}(X; \mathbf{p}, \mathcal{U}(\mathbf{p})) = (1-m) \sum_{i=1}^{c} \sum_{j=1}^{n} \frac{1}{d_i(x_j, \mathbf{p})^{1/(m-1)}} = (1-m) J(\mathbf{p}) \tag{28}$$

Up to a constant factor we have $J_{poss'}(\mathbf{p}, \mathcal{U}(\mathbf{p})) = J(\mathbf{p})$ again, and since this factor is negative ($m > 1$) minimization of $J_{poss'}$ corresponds to maximization of J. Thus, the extrema in Theorem 1 denote maxima of J. Consequently possibilistic clustering algorithms (like PCM or P-AGK) can be interpreted as a steepest-ascent algorithm (with automatic step size-adjustment):

Theorem 8 *Given D_s by (18), u_s by (15), and J by (20), a steepest ascent algorithm maximizing J is given by*

$$\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)} + \gamma (\nabla J)(\mathbf{p}^{(t)})$$

where γ is the step size and $(1-m)\nabla J$ is given by (25).

To show that $J(\Phi(\mathbf{p})) > J(\mathbf{p})$, which is the necessary condition (23) in Theorem 3, it is sufficient to provide the following

Theorem 9 *For arbitrary d_s, $m > 1$, $(x, \mathbf{p}) \in \mathcal{X} \times \mathcal{P}^c$, and membership matrix U we have*

$$J_{poss'}(X; \mathbf{p}, U) \geq J_{poss'}(X; \mathbf{p}, \mathcal{U}(\mathbf{p})) \tag{29}$$

where $\mathcal{U}(\mathbf{p})$ denotes the possibilistic membership matrix for prototypes $\mathbf{p}$. The equality holds iff $U = U^*$.

In order to show that the monotonicity

$$\begin{aligned} (1-m)J(p) &= J_{poss'}(\mathbf{p}, \mathcal{U}(\mathbf{p})) \\ &\overset{(\star)}{>} J_{poss'}(\Phi(\mathbf{p}), \mathcal{U}(\mathbf{p})) \\ &\overset{(\star\star)}{>} J_{poss'}(\Phi(\mathbf{p}), \mathcal{U}(\Phi(\mathbf{p}))) = (1-m)J(\Phi(\mathbf{p})) \end{aligned}$$

holds, Theorem 9 replaces Bezdek's theorem for probabilistic memberships in showing ($\star\star$). As far as ($\star$) is concerned, the same arguments that have been used before for the probabilistic case remain valid, since the memberships are considered as constants in this step. Thus, we immediately have for FCM and AGK that $J(\Phi(\mathbf{p})) > J(\mathbf{p})$ with $\Phi = \Phi_{FCM}$ and $\Phi = \Phi_{AGK}$, respectively.

We would like to note in passing that unlike the probabilistic case, here it is possible to define a $J_{poss'}$ for $m = 1$ by

$$J_{poss'}(X; \mathbf{p}, U) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j} d_i(x_k, \mathbf{p}) - \sum_{i=1}^{c} \sum_{j=1}^{n} \log(u_{i,j})$$

leading again to $u_s = 1/d_s$ as in (15).

6 Attractive FCM Fixed Points

In this section we deal solely with the (probabilistic) fuzzy c-means algorithm, that is $d_s(x, \mathbf{p}) = ||x - p_s||^2 + \varepsilon$, $\Phi = \Phi_{FCM}$, u_s defined by (17) and (18), $m \in \mathbb{R}_{>1}$. We seek for conditions for fixed points to be attractive, since such fixed points exclude solutions that represent saddle points. The following theorem states a sufficient condition for a fixed point $\mathbf{p}$ to be attractive. For brevity, if the prototypes $\mathbf{p}$ are clear from the context, we will use $u_{s,j}$ instead of $u_s(x_j, \mathbf{p})$.

Theorem 10 *Let $\mathbf{p}$ be a fixed point of Φ. If $\|a\|_2 < 1$ holds, where $a \in \mathbb{R}^c$ with*

$$\forall s \in \mathbb{N}_{\leq c}: \quad a_s = \frac{2m}{m-1} \left(\frac{\sum_{j=1}^{n} u_{s,j}^m \left(1 - u_{s,j} + \sum_{i=1, i \neq s}^{c} u_i \frac{\|x_j - p_s\|}{\|x_j - p_i\|}\right)}{\sum_{j=1}^{n} u_{s,j}^m} \right) \tag{30}$$

then $\mathbf{p}$ is an attractive fixed point. (By $\|\cdot\|_2$ we denote the Euclidean norm.)

Corollary 1 *Let $\mathbf{p}$ be a fixed point of Φ and $1 < m < 3$. If $\|a\|_2 < 1$ holds, where $a \in \mathbb{R}^c$ with*

$$\forall s \in \mathbb{N}_{\leq c}: \quad a_s = \frac{2m}{m-1} \left(\frac{\sum_{j=1}^{n} u_{s,j}^{m-1} - u_{s,j}^{m+1}}{\sum_{j=1}^{n} u_{s,j}^m} \right) \tag{31}$$

then $\mathbf{p}$ is an attractive fixed point.

Condition (30) is more general than (31) and should be preferred when deciding about the attractive property of a fixed point. However, we stated (31) because it contains only occurrences of membership degrees (data and prototypes are eliminated). This enables us to discuss the nature of attractive fixed points below independent from data set X and prototypes $\mathbf{p}$.

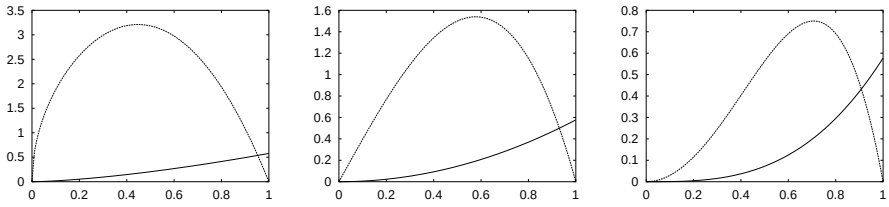
Whether a specific fixed point is attractive or not depends on the membership values and thus on the data set. What kind of data sets satisfy (31)? Instead

of $\|a\|_2 < 1$ let us consider the condition $|a_s| < \frac{1}{\sqrt{c}}$ from which $\|a\|_2 < 1$ can be concluded. Then $|a_s| < \frac{1}{\sqrt{c}}$ is equivalent to

$$\frac{2m}{m-1}\left(\sum_{j=1}^{m} u_{s,j}^{m-1} - u_{s,j}^{m+1}\right) < \frac{1}{\sqrt{c}}\sum_{j=1}^{n} u_{s,j}^{m}$$

$$\text{or} \qquad \sum_{j=1}^{n} g(u_{s,j}) > 0 \tag{32}$$

$$\text{where} \qquad g(u) = \left(\frac{u^m}{\sqrt{c}} - \frac{2m}{m-1}(u_{s,j}^{m-1} - u_{s,j}^{m+1})\right)$$



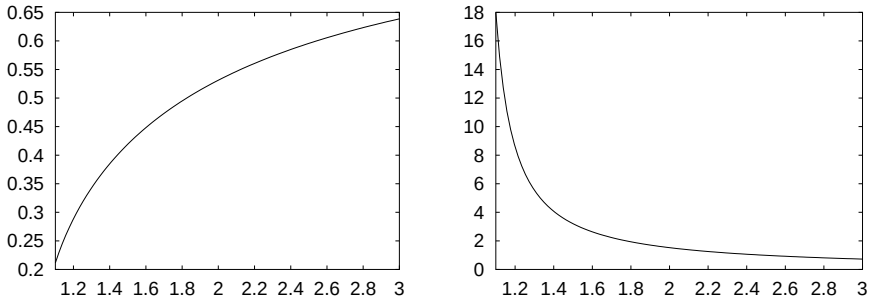
(a) $m = 1.5$ (b) $m = 2$ (c) $m = 3$

Figure 2: Positive and negative terms in $g(u)$ for different values of m.

As long as $g(u) > 0$ for each datum x_j we have for sure a contraction. Figure 2 shows the positive ($\frac{u_s^m}{\sqrt{c}}$) and the negative terms ($\frac{2m}{m-1}(u_s^{m-1} - u_s^{m+1})$ – note that $\forall u \in (0,1) : u^{m-1} > u^{m+1}$) of $g(u)$ for different values of m ($c = 3$). We can see from the figures, that we have a positive difference for memberships near 1 and a negative difference for "ambiguous" memberships. Loosely speaking, if we have clusters with many data vectors near the prototypes and only some ambiguous data in between, then the lemma tells us that Φ is a contraction near $\mathbf{p}$. It may look hard to satisfy (31) in case $m = 1.5$ and easier in case $m = 3$, however, we must be aware of the fact that with an increasing (resp. decreasing) fuzzifier fuzzy (resp. crisp) memberships are more likely to occur. The following observation gives a hint about the required ratio of the number of unambiguous data vectors to the number of ambiguous data vectors.

Observation 1 *Using the notation of Theorem 10,* $\mathbf{p}$ *is an attractive fixed point, if there are* $-\sqrt{c} \cdot g(\hat{u})$ *times more unambiguous data vectors* ($u \approx 1$) *than ambiguous data vectors* ($u \approx \frac{1}{2}$)*, where*

$$\hat{u} = \frac{\sqrt{(m-1)^2 + 16c(m^2-1)} - (m-1)}{4\sqrt{c}(m+1)} \tag{33}$$



(a) Membership degree $\hat{u}$ where loss becomes maximal versus m.

(b) Maximum loss $g(\hat{u})$ versus m.

Figure 3: Loss versus fuzzifier m.

Figure 3 depicts the results of the observation. From figure 3(a) we can see that the membership degree $\hat{u}$ with maximum loss $g(\hat{u})$ when using a fuzzifier $m = 2$ is approximately 0.53. And from figure 3(b) we see that this maximum loss is about 1.53 for $m = 2$. Having $c = 3$ clusters we can conclude, if a final FCM partition has at least $\sqrt{c} \cdot 1.53 \approx 2.64$ unambiguously assigned data objects for each ambiguous data object, FCM has found an attractive fixed point. And from Theorem 3 we thus know **for sure** that we have found a minimum – and not a saddle point. (Be aware that this is a very coarse estimation and by means of (30) you will find many more fixed points being attractive.)

As an example, let us revisit Sabin's data set $X = \{-3, -1, 1, 3\}$ with initial prototypes $\mathbf{p} = (-1, 0, 1)$. This data set has been used in [5] as an example for FCM converging into a saddle point of J_{prob}, as shown in figure 4. The final prototypes are computed to be $\mathbf{p} \approx (-2.9, 0, 2.9)$. Application of theorem 10 indicates that this fixed point may not be attractive. Following the discussion in the previous section, we could expect this since the prototype in the middle has no unambiguously assigned data vectors at all. If we change the initialization slightly to $\mathbf{p} = (-1, \varepsilon, 1)$, $\varepsilon \neq 0$, FCM consequently terminates into a different fixed point (which is no saddle point but a minimum).

Figure 5 shows an example of FCM converging into an attractive fixed point with $\|a\| < 1$, the partition therefore represents a minimal solution of J_{prob}. However, one can also find examples where FCM has found a minimal solution but $\|a\| \geq 1$. (We used (30) for the examples.) It might be interesting to observe some similarity of (31) with validity measures for cluster partitions [3]. Not surprisingly, in case of figure 5 the value of $\|a\|$ becomes minimal for $c = 4$, and we have $\|a\| > 1$ for all other cases.

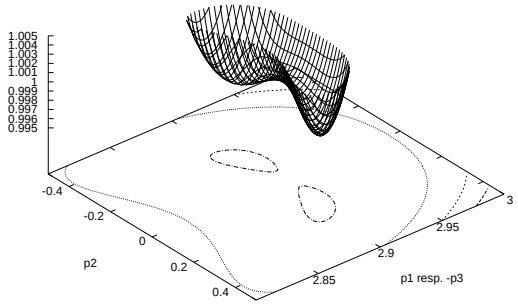


Figure 4: Sabin's example demonstrates FCM converging to a saddle point of J_{FCM} with $p_1 = -p_3 \approx 2.9$ and $p_2 = 0$. (J_{FCM} has been scaled appropriately to make the saddle point clearly visible.)

7 Conclusion

In this paper, we have presented an intuitive explanation for using probabilistic and possibilistic memberships that is not motivated by objective-function minimization only. Probabilistic memberships are obtained uniquely if we want to preserve duality of dissimilarity measures D_s and similarity measures u_s and require scale-invariance. We have also proposed non-fuzzy membership functions which are in some sense dual to probabilistic membership functions, since in both cases fixed points are extrema of the same[4] objective function J: extrema are minima in the case of probabilistic and maxima in case of possibilistic memberships. If we replace d_s by $d_s/\eta_s + 1$ our non-fuzzy memberships become fuzzy and identical to possibilistic memberships of PCM in case of $m = 2$.

Then, we have discussed convergence of fuzzy clustering algorithms in general, and attractive fixed points of FCM in particular. It was known before, that FCM can be interpreted as a steepest descent algorithm [1]. We have shown that this is also the case with other fuzzy clustering algorithms (AGK), and that possibilistic clustering can be interpreted as gradient ascent. We have further shown that other results which have been formulated only for FCM in the past (equivalence of reformulated FCM functional [10], minimization of J_m in probabilistic case [2]) in fact are true for other fuzzy clustering algorithms resp. similar in the possibilistic case. Besides that, the main contributions to the known theory may be summarized as follows:

- We have provided new proofs for the monotonicity of J under AO iter-

[4]Both objective functions are identical in the sense that they can be written as $\sum_j \sum_i u_{i,j}^m d_{i,j}$, but the membership functions u are different for possibilistic and probabilistic clustering, of course.

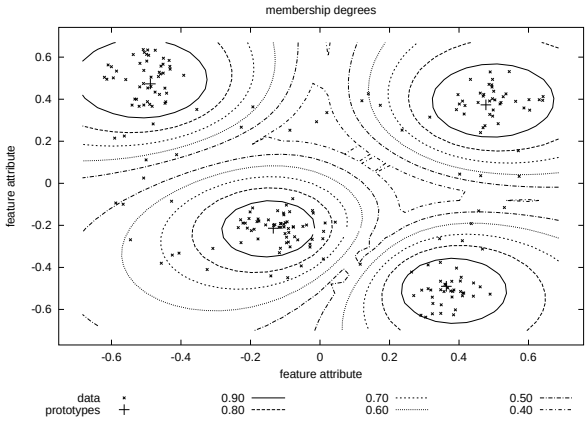


Figure 5: An example for an attractive fixed point. The contour lines indicate the membership degrees.

ations for the axis-parallel Gustafson-Kessel algorithm. The idea of the proof is quite simple and thus can also be transferred to other fuzzy clustering algorithms.

- We have established a relationship between saddle-points and extrema of the objective function J that holds for both types of memberships and all fuzzy clustering algorithms: An attractive fixed point cannot be a saddle-point.

- We have developed sufficient conditions for attractive fixed points of FCM. In the best case, this condition can be used to decide that FCM has terminated in a (local) minimum of J and was not trapped in a saddle point.

A Proofs

The following lemma is needed for the following proofs.

Lemma 1 *Given D_s by (18) and u_s by (17), we obtain for any directional derivative of u_s with respect to ξ*

$$\frac{\partial u_s}{\partial \xi} = u_s \left(\sum_{i=1}^{c} \frac{u_i}{D_i} \frac{\partial D_i}{\partial \xi} \right) - \frac{u_s}{D_s} \frac{\partial D_s}{\partial \xi} \tag{34}$$

Proof of Lemma 1: We rewrite u_s as follows

$$u_s = \frac{1}{D_s \sum_{i=1}^{c} \frac{1}{D_i}}$$

$$
\begin{aligned}
&= \frac{1}{D_s \frac{\sum_{i=1}^{c} \prod_{t=1,t\neq i}^{c} D_t}{\prod_{i=1}^{c} D_i}} \\
&= \frac{\prod_{i=1}^{c} D_i}{D_s \sum_{i=1}^{c} \prod_{t=1,t\neq i}^{c} D_t} \\
&= \frac{\prod_{i=1,i\neq s}^{c} D_i}{\sum_{i=1}^{c} \prod_{t=1,t\neq i}^{c} D_t} \qquad (35)
\end{aligned}
$$

Then we have $u_s = \frac{A}{B}$ with

$$
\begin{aligned}
A &= \prod_{i=1,i\neq s}^{c} D_i \\
B &= \sum_{i=1}^{c} \prod_{t=1,t\neq i}^{c} D_t
\end{aligned}
$$

From the directional derivatives

$$
\begin{aligned}
\frac{\partial A}{\partial y} &= \sum_{i=1,i\neq s}^{c} \frac{\partial D_i}{\partial y} \prod_{t=1,t\neq i,s}^{c} D_t \\
\frac{\partial B}{\partial y} &= \sum_{i=1}^{c} \sum_{t=1,t\neq i}^{c} \frac{\partial D_t}{\partial y} \prod_{j=1,j\neq i,t}^{c} D_j
\end{aligned}
$$

we obtain

$$
\begin{aligned}
&\frac{\partial \frac{A}{B}}{\partial y} \\
= \; & \frac{1}{B^2} \left(\sum_{i=1,i\neq s}^{c} \frac{\partial D_i}{\partial y} \prod_{t=1,t\neq i,s}^{c} D_t \right) \left(\sum_{i=1}^{c} \prod_{t=1,t\neq i}^{c} D_t \right) \\
- \; & \frac{1}{B^2} \left(\prod_{i=1,i\neq s}^{c} D_i \right) \left(\sum_{i=1}^{c} \sum_{t=1,t\neq i}^{c} \frac{\partial D_t}{\partial y} \prod_{j=1,j\neq i,t}^{c} D_j \right) \\
= \; & \frac{1}{B^2} \left(\sum_{i=1,i\neq s}^{c} \sum_{j=1}^{c} \frac{\partial D_i}{\partial y} \left(\prod_{t=1,t\neq i,s}^{c} D_t \right) \left(\prod_{l=1,l\neq j}^{c} D_l \right) \right) \\
- \; & \frac{1}{B^2} \left(\sum_{i=1}^{c} \sum_{t=1,t\neq i}^{c} \frac{\partial D_t}{\partial y} \left(\prod_{j=1,j\neq i,t}^{c} D_j \right) \left(\prod_{l=1,l\neq s}^{c} D_l \right) \right) \\
= \; & \frac{1}{B^2} \left(\sum_{i=1,i\neq s}^{c} \sum_{j=1}^{c} \frac{\partial D_i}{\partial y} \frac{1}{D_i D_s D_j} \right) \left(\prod_{l=1}^{c} D_l \right)^2
\end{aligned}
$$

$$
\begin{aligned}
- & \frac{1}{B^2}\left(\sum_{i=1}^{c}\sum_{t=1,t\neq i}^{c}\frac{\partial D_t}{\partial y}\frac{1}{D_iD_sD_t}\right)\left(\prod_{l=1}^{c}D_l\right)^2 \\
= & \frac{1}{D_sB^2}\left(\prod_{l=1}^{c}D_l\right)^2\left(\sum_{i=1,i\neq s}^{c}\sum_{j=1}^{c}\frac{\partial D_i}{\partial y}\frac{1}{D_iD_j}-\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\frac{\partial D_j}{\partial y}\frac{1}{D_iD_j}\right) \\
= & \frac{1}{D_sB^2}\left(\prod_{l=1}^{c}D_l\right)^2\left(-\sum_{j=1}^{c}\frac{\partial D_s}{\partial y}\frac{1}{D_sD_j}+\sum_{i=1}^{c}\frac{\partial D_i}{\partial y}\frac{1}{D_i^2}\right) \\
= & \frac{1}{D_sB^2}\left(\prod_{l=1}^{c}D_l\right)^2\left(\sum_{i=1}^{c}\frac{D_s\frac{\partial D_i}{\partial y}-\frac{\partial D_s}{\partial y}D_i}{D_sD_i^2}\right) \\
= & \frac{1}{B^2}\left(\prod_{l=1}^{c}D_l\right)^2\left(\sum_{i=1}^{c}\frac{D_s\frac{\partial D_i}{\partial y}-\frac{\partial D_s}{\partial y}D_i}{D_s^2D_i^2}\right)
\end{aligned}
$$

which can be rewritten as

$$
\frac{\partial u_s}{\partial \xi}=\left(\sum_{i=1}^{c}u_pu_qD_pD_q\frac{D_s\frac{\partial D_i}{\partial \xi}-\frac{\partial D_s}{\partial \xi}D_i}{D_s^2D_i^2}\right) \tag{36}
$$

with $p,q\in\mathbb{N}_{\leq c}$ using (35). For equation (34) we start from (36) with $p=s$ and $q=i$ and obtain

$$
\begin{aligned}
\frac{\partial u_s}{\partial y} &= u_s\sum_{i=1}^{c}u_i\left(\frac{\frac{\partial D_i}{\partial y}}{D_i}-\frac{\frac{\partial D_s}{\partial y}}{D_s}\right) \\
&= u_s\left(\left(\sum_{i=1}^{c}\frac{u_i}{D_i}\frac{\partial D_i}{\partial y}\right)-\left(\frac{1}{D_s}\frac{\partial D_s}{\partial y}\sum_{i=1}^{c}u_i\right)\right) \\
&\overset{(3)}{=} u_s\left(\left(\sum_{i=1}^{c}\frac{u_i}{D_i}\frac{\partial D_i}{\partial y}\right)-\frac{1}{D_s}\frac{\partial D_s}{\partial y}\right)
\end{aligned}
$$

which is identical to (34). ■

Proof of Theorem 1: Consider $\mathbf{p}$ such that $\Phi(\mathbf{p})=\mathbf{p}$ holds ($\star$). Let $\xi\in\mathcal{P}^c$ with $\|\xi\|=1$. With $\frac{\partial D_i}{\partial \xi}=\frac{1}{m-1}d_i^{\frac{2-m}{m-1}}\frac{\partial d_i}{\partial \xi}$ we have

$$
\frac{1}{D_i}\frac{\partial D_i}{\partial \xi}=d_i^{\frac{-1}{m-1}}\frac{\partial D_i}{\partial \xi}=\frac{1}{m-1}d_i^{\frac{2-m-1}{m-1}}\frac{\partial d_i}{\partial \xi}=\frac{1}{m-1}\frac{1}{d_i}\frac{\partial d_i}{\partial \xi} \tag{37}
$$

We have to show that $\frac{\partial J}{\partial \xi}=0$:

$$
\begin{aligned}
\frac{\partial J}{\partial \xi} &= \sum_{j=1}^{n}\sum_{s=1}^{c}\frac{\partial u_s^m}{\partial \xi}d_s+\sum_{j=1}^{n}\sum_{s=1}^{c}u_s^m\frac{\partial d_s}{\partial \xi} \\
&= \sum_{j=1}^{n}\sum_{s=1}^{c}mu_s^{m-1}\frac{\partial u_s}{\partial \xi}d_s+\sum_{s=1}^{c}\sum_{j=1}^{n}u_s^m\frac{\partial d_s}{\partial \xi} \qquad (38)
\end{aligned}
$$

First, we consider the simpler case where u_s is defined by (15). Then $d_s \frac{\partial u_s}{\partial \xi} = d_s \frac{\partial}{\partial \xi} d_s^{-1/(m-1)} = \frac{-1}{m-1} d_s^{-1/(m-1)} \frac{\partial d_s}{\partial \xi} = \frac{-1}{m-1} u_s \frac{\partial d_s}{\partial \xi}$ and we continue

$$\begin{aligned}
&= -\frac{m}{m-1} \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m \frac{\partial d_s}{\partial \xi} + \sum_{s=1}^{c} \sum_{j=1}^{n} u_s^m \frac{\partial d_s}{\partial \xi} \\
&= -\frac{1}{m-1} \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m \frac{\partial d_s}{\partial \xi} \\
\overset{(\star)(21)}{=} &\sum_{s=1}^{c} 0
\end{aligned}$$

Let us now consider the case where u_s is defined by (17). We continue from (38):

$$\begin{aligned}
\overset{(\star)(21)}{=} &\sum_{j=1}^{n} \sum_{s=1}^{c} m u_s^{m-1} \frac{\partial u_s}{\partial \xi} d_s + \sum_{s=1}^{c} 0 \\
\overset{(34)}{=} &\sum_{j=1}^{n} \sum_{s=1}^{c} m u_s^m d_s \left(\left(\sum_{i=1}^{c} \frac{u_i}{D_i} \frac{\partial D_i}{\partial \xi} \right) - \frac{1}{D_s} \frac{\partial D_s}{\partial \xi} \right) \\
= &\; m \left(\sum_{j=1}^{n} \sum_{s=1}^{c} \sum_{i=1}^{c} u_s^m u_i \frac{d_s}{D_i} \frac{\partial D_i}{\partial \xi} - \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m \frac{d_s}{D_s} \frac{\partial D_s}{\partial \xi} \right) \\
\overset{(37)}{=} &\; \frac{m}{m-1} \left(\sum_{j=1}^{n} \sum_{s=1}^{c} \sum_{i=1}^{c} u_s^m u_i \frac{d_s}{d_i} \frac{\partial d_i}{\partial \xi} - \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m \frac{\partial d_s}{\partial \xi} \right) \\
\overset{(14)}{=} &\; \frac{m}{m-1} \left(\sum_{j=1}^{n} \sum_{s=1}^{c} \sum_{i=1}^{c} u_s u_i^m \frac{D_i^{m-1}}{D_s^{m-1}} \frac{d_s}{d_i} \frac{\partial d_i}{\partial \xi} - \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m \frac{\partial d_s}{\partial \xi} \right) \\
= &\; \frac{m}{m-1} \left(\sum_{j=1}^{n} \sum_{s=1}^{c} \sum_{i=1}^{c} u_s u_i^m \frac{d_i}{d_s} \frac{d_s}{d_i} \frac{\partial d_i}{\partial \xi} - \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m \frac{\partial d_s}{\partial \xi} \right) \\
= &\; \frac{m}{m-1} \left(\sum_{j=1}^{n} \sum_{i=1}^{c} u_i^m \frac{\partial d_i}{\partial \xi} \sum_{s=1}^{c} u_s - \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m \frac{\partial d_s}{\partial \xi} \right) \\
\overset{(3)}{=} &\; \frac{m}{m-1} \left(\sum_{j=1}^{n} \sum_{i=1}^{c} u_i^m \frac{\partial d_i}{\partial \xi} - \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^m \frac{\partial d_s}{\partial \xi} \right) \\
= &\; 0
\end{aligned}$$

Thus, in both cases we have a zero in the first derivative and J must therefore have an extremum or saddle-point at $\mathbf{p}$. ■

Proof of Theorem 3: Let $\mathbf{p}$ be a fixed point of Φ that represents a saddle point of J. Assume that $\mathbf{p}$ is an attractive fixed point. Since Φ is continuous

we can find an $\varepsilon > 0$ and a closed ball $B(\mathbf{p},\varepsilon) := \{x \in \mathcal{P}^c \,|\, \|x - \mathbf{p}\| \leq \varepsilon\}$ such that Φ satisfies (22) with $W = B(\mathbf{p},\varepsilon)$. So $\Phi|_{B(\mathbf{p},\varepsilon)}$ is a contraction and the iterative application of Φ to any $x \in B(\mathbf{p},\varepsilon)$ converges towards $\mathbf{p}$ within $B(\mathbf{p},\varepsilon)$. Since $\mathbf{p}$ is a saddle point of J we find at least one point $\mathbf{q} \in B(\mathbf{p},\varepsilon)$ with $J(\mathbf{q}) < J(\mathbf{p})$. Because J is continuous we can find a $\delta > 0$ such that for all $\mathbf{r} \in B(\mathbf{p},\delta) : J(\mathbf{r}) > J(\mathbf{q})$ (resp. $(J(\mathbf{r}) < J(\mathbf{q}))$. Since Φ never increases (resp. decreases) J, we can never enter $B(\mathbf{p},\delta)$ and therefore cannot converge to $\mathbf{p}$ when starting from $\mathbf{q}$. On the other hand, according to Banach's theorem we must converge to $\mathbf{p}$. This is a contradiction and therefore the assumption was false. Thus $\mathbf{p}$ is a non-attractive fixed point. ■

Proof of Theorem 4: Let x be a variable occurring in $\mathcal{P}$. Then we have according to the chain rule

$$\frac{\partial J}{\partial x} = \frac{\partial J_{prob}}{\partial x} + \sum_{i,j} \frac{\partial J_{prob}}{\partial u_{i,j}} \cdot \frac{\partial u_{i,j}}{\partial x} \tag{39}$$

Note that in the first term on the right-hand side of (39) the partial derivative refers to the variable x that also occurs in J_{prob}.

For a $\mathbf{p} \in \mathcal{P}$, let us now assume $\nabla J(\mathbf{p}) = 0$, i.e. the left-hand side of (39) is zero. Since the $u_{i,j}$ are chosen by $\mathcal{U}(\mathbf{p})$ in such a way that $\frac{J_{prob}}{\partial u_{i,j}} = 0$, the sum on the right-hand side of (39) is zero as well. Therefore, the first term on the right-hand side of (39) must be zero, too. Together this implies $\nabla J_{prob} = \mathbf{0}$.

For the other implication let us now assume $\nabla J_{prob}(\mathbf{p},\mathcal{U}(\mathbf{p})) = 0$. Then the first term on the right-hand side of (39) is zero and for the same reason as before the values $\frac{J_{prob}}{\partial u_{i,j}} = 0$ are zero so that the sum also becomes zero. ■

Proof of Theorem 5: In [2] it was shown that $\mathcal{U}(\mathbf{p})$ is a strict local minimum for the function J_{prob} (for fixed $\mathbf{p}$). Since $J(\mathbf{p}) = J_{prob}(\mathbf{p},\mathcal{U}(\mathbf{p}))$ is continuous, this implies that if $\mathbf{p}$ is a strict local minimum of J, then $(\mathbf{p},\mathcal{U}(\mathbf{p}))$ is a strict local minimum of J_{prob}.

For the other direction let $(\mathbf{p},\mathcal{U}(\mathbf{p}))$ be a local minimum of J_{prob}, i.e.

$$(\exists \varepsilon > 0)(\forall (\tilde{\mathbf{p}},\tilde{U})) : \Big(\|(\tilde{\mathbf{p}},\tilde{U}) - (\mathbf{p},\mathcal{U}(\mathbf{p}))\|_1 < \varepsilon \;\Rightarrow\; J_{prob}((\tilde{\mathbf{p}},\tilde{U}) > J_{prob}(\mathbf{p},\mathcal{U}(\mathbf{p}))\Big)$$

Since the update function $\mathcal{U}$ is continuous, we have

$$(\exists \delta > 0)(\forall \tilde{\mathbf{p}}) \;\left(\|\tilde{\mathbf{p}} - \mathbf{p}\|_1 < \delta \;\Rightarrow\; \|\mathcal{U}(\tilde{\mathbf{p}}) - \mathcal{U}(\mathbf{p})\|_1 < \frac{\varepsilon}{2} \right).$$

Let $\tilde{\mathbf{p}}$ be s.t. $\|\tilde{\mathbf{p}} - \mathbf{p}\|_1 < \min\left\{\delta, \frac{\varepsilon}{2}\right\}$. This implies $\|(\tilde{\mathbf{p}},\tilde{U}) - (\mathbf{p},\mathcal{U}(\mathbf{p}))\|_1 < \varepsilon$ and we have

$$J(\tilde{\mathbf{p}}) = J_{prob}(\tilde{\mathbf{p}},\tilde{U}) > J_{prob}(\mathbf{p}^{(0)},\mathcal{U}(\mathbf{p}^{(0)})) = J(\mathbf{p}^{(0)})$$

■

Proof of Theorem 6: Let us revisit the proof of Theorem 1 for the case that u_s is defined by (17). Equations ($\star$) and (21) are applied only once. Not making use of them we obtain

$$\frac{\partial J}{\partial \xi} = \sum_{j=1}^{n} \sum_{s=1}^{c} u_s^c \frac{\partial d_s}{\partial \xi} \quad \text{and thus} \quad \nabla_{\mathbf{p}} J = \sum_{s=1}^{c} \sum_{j=1}^{n} u_s^m \nabla_{\mathbf{p}} d_s$$

■

For *Proof of Theorem 7* see appendix B.

Proof of Theorem 8: Same line of proof as in Theorem 6, using the fact that $J_{poss'} = (1-m)J$ holds. ■

Proof of Theorem 9: To show that $J_{poss'}$ decreases with every membership update step we may consider all $u_{i,j}$ in $J_{poss'}$ independently, since there is no condition like (3) as with probabilistic memberships. Choose an arbitrary $u_{i,j}$ (denoted by u), then we collect all occurrences of $u_{i,j}$ in $J_{poss'}$ in f:

$$f(u) := u^m d - mu = u\left(u^{m-1} d - m\right).$$

Thus, we have up to three zero crossings at $u = 0$ and $u = \pm w$ with $w := \left(\frac{m}{d}\right)^{1/(m-1)}$. For $u \in (w, \infty]$ we have $f(u) > 0$ and for $u \in (0, w)$ we have $f(u) < 0$. Within the relevant range $[0, \infty]$ we have $f(0) = 0 = f(w)$ and $f(u) < 0$ for $u \in (0, w)$, thus we can conclude that there is at least one local minimum within $[0, w]$. Since $\frac{\partial f}{\partial u} = 0$ yields a single positive solution at $u = \left(\frac{1}{d}\right)^{1/(m-1)} < w$ we know that the possibilistic memberships strictly minimize $f(u)$. Thus, with fixed distance values, J is minimized by setting $U = \mathcal{U}(\mathbf{p})$. ■

Proof of Theorem 10: Consider a fixed point $\mathbf{p}$ of Φ with $\|\frac{\partial \Phi}{\partial \mathbf{p}}(\mathbf{p})\| < 1$. If Φ has a continuous derivative, then we conclude that there is even a small ball B around $\mathbf{p}$ such that (22) holds for all elements of the ball. Furthermore, from (22) we know that $\Phi(B) \subset B$ because $\alpha < 1$. Then, according to Banach's contraction principle, $\Phi|_B$ is a contraction with a unique fixed point and the iteration scheme converges to this fixed point for any starting value $\mathbf{p}^{(0)} \in B$. Therefore $\|\frac{\partial \Phi}{\partial \mathbf{p}}(\mathbf{p})\| < 1$ is a sufficient condition for $\mathbf{p}$ being an attractive fixed point.

Let $\mathbf{p}$ be a fixed point of Φ, that is

$$\forall 1 \le s \le c: \qquad \frac{\sum_{j=1}^{n} u_s^m x_j}{\sum_{j=1}^{n} u_s^m} = p_s \tag{40}$$

Since Φ has continuous derivatives we can show (22) by means of $\|D\Phi(\mathbf{p})\| < 1$. Equivalently we may consider $\|D\Phi_\xi(\mathbf{p})\| < 1$ for any direction $\xi \in \mathcal{P}^c$, $\|\xi\| = 1$, where $\Phi_\xi : \mathbb{R} \to \mathcal{P}^c, t \mapsto \Phi(\mathbf{p} + t\xi)$. (By $\|\cdot\|$ we denote the Euclidean norm.) Denoting the c prototype components of $\Phi_\xi(t)$ by $\Phi_\xi(t)_i$ we have

$$\|D\Phi_\xi(t)\| = \|(\|D\Phi_\xi(t)_1\|, ..., \|D\Phi_\xi(t)_c\|)\|$$

and therefore examine $\|D\Phi_\xi(t)_s\|$ for $s \in \mathbb{N}_{\leq c}$. (Carefully distinguish D_i as the dissimilarity measure and $D\Phi$ as the total derivative of Φ.) For better readability we abbreviate $u_s(x_j, \mathbf{p} + t\xi)$ by u_s. We have

$$\begin{aligned}
D\Phi_\xi(t)_s &= \left(\frac{\partial}{\partial t}\frac{\sum_{j=1}^n u_s^m x_{j,l}}{\sum_{j=1}^n u_s^m}\right)_{1\leq l\leq \mathrm{DIM}} \\
&= \frac{\left(\sum_{j=1}^n \frac{\partial u_s^m}{\partial t} x_j\right)\left(\sum_{j=1}^n u_s^m\right) - \left(\sum_{j=1}^n \frac{\partial u_s^m}{\partial t}\right)\left(\sum_{j=1}^n u_s^m x_j\right)}{\left(\sum_{j=1}^n u_s^m\right)^2} \\
&\overset{(40)}{=} \frac{\left(\sum_{j=1}^n \frac{\partial u_s^m}{\partial t} x_j\right) - \left(\sum_{j=1}^n \frac{\partial u_s^m}{\partial t}\right) p_s}{\sum_{j=1}^n u_s^m} \\
&= \frac{m\sum_{j=1}^n u_s^{m-1}\frac{\partial u_s}{\partial t}(x_j - p_s)}{\sum_{j=1}^n u_s^m}
\end{aligned}$$

With $\frac{\partial u_s(\mathbf{p}+t\xi)}{\partial t} = \frac{\partial u_s}{\partial \xi}(\mathbf{p})$ we continue

$$\begin{aligned}
&\overset{(34)}{=} \frac{m\sum_{j=1}^n u_s^m \left(\sum_{i=1}^c u_i \left(\frac{1}{D_i}\frac{\partial D_i}{\partial \xi} - \frac{1}{D_s}\frac{\partial D_s}{\partial \xi}\right)\right)(x_j - p_s)}{\sum_{j=1}^n u_s^m} \\
&\overset{(37)}{=} \frac{\frac{m}{m-1}\sum_{j=1}^n u_s^m \left(\sum_{i=1}^c u_i \left(\frac{1}{d_i}\frac{\partial d_i}{\partial \xi} - \frac{1}{d_s}\frac{\partial d_s}{\partial \xi}\right)\right)(x_j - p_s)}{\sum_{j=1}^n u_s^m} \\
&= \frac{\frac{m}{m-1}\sum_{j=1}^n u_s^{m-1}\alpha_{s,j}}{\sum_{j=1}^n u_s^m} \qquad (41)
\end{aligned}$$

with

$$\alpha_{s,j} = \left(\sum_{i=1}^c u_i u_s \left(\frac{-2(x_j - p_i)^\top \xi_i}{d_i} - \frac{-2(x_j - p_s)^\top \xi_s}{d_s}\right)\right)(x_j - p_s) \qquad (42)$$

Obviously, in case $i = s$ the summand evaluates to zero, we can therefore exclude $i = s$ from the summation. Let us now consider $\|\alpha_{s,j}\|$:

$$\begin{aligned}
\|\alpha_{s,j}\| &= 2\left|\sum_{i=1, i\neq s}^c u_i u_s \left(\frac{(x_j - p_i)^\top \xi_i}{\|x_j - p_i\|^2} - \frac{(x_j - p_s)^\top \xi_s}{\|x_j - p_s\|^2}\right)\right| \|x_j - p_s\| \\
&\leq 2\sum_{i=1, i\neq s}^c u_i u_s \left(\frac{\left|(x_j - p_i)^\top \xi_i\right|}{\|x_j - p_i\|}\frac{\|x_j - p_s\|}{\|x_j - p_i\|} + \frac{\left|(x_j - p_s)^\top \xi_s\right|}{\|x_j - p_s\|}\right)
\end{aligned}$$

With $x^\top y = \cos(\angle(x,y))\|x\| \cdot \|y\|$ for any $x, y \in \mathcal{X}$ and $\|\xi\| = 1$ we continue

$$= 2\sum_{i=1, i\neq s}^c u_i u_s \left(|\cos(\angle(x_j - p_i, \xi_i))|\,\|\xi_i\|\frac{\|x_j - p_s\|}{\|x_j - p_i\|} + |\cos(\angle(x_j - p_s, \xi))|\,\|\xi_s\|\right)$$

$$\leq \quad 2 \sum_{i=1, i\neq s}^{c} u_i u_s \left(\frac{\|x_j - p_s\|}{\|x_j - p_i\|} + 1 \right) \tag{43}$$

From (3) we obtain $\sum_{i=1, i\neq s}^{c} u_i = 1 - u_s$ and therefore

$$\|\alpha_{s,j}\| = 2u_s \left(1 - u_s + \sum_{i=1, i\neq s}^{c} u_i \frac{\|x_j - p_s\|}{\|x_j - p_i\|} \right)$$

We obtain (30) directly when substituting this result in (41). ■

Proof of Corollary 1: To get (31), we continue from (43) in the proof of theorem 10 by considering the two cases

- **Case 1:** $\|x_j - p_s\| \leq \|x_j - p_i\|$. Then we have $u_s \frac{\|x_j - p_s\|}{\|x_j - p_i\|} \leq u_s \leq 1$.
- **Case 2:** $\|x_j - p_s\| > \|x_j - p_i\|$. From (14) we obtain

  $$u_s = u_i \frac{D_i}{D_s} = u_i \left(\frac{d_i}{d_s} \right)^{\frac{1}{m-1}} = u_i \left(\frac{\|x_j - p_i\|}{\|x_j - p_s\|} \right)^{\frac{2}{m-1}}$$

  leading us to

  $$u_s \frac{\|x_j - p_s\|}{\|x_j - p_i\|} = u_i \left(\frac{\|x_j - p_i\|}{\|x_j - p_s\|} \right)^{\frac{2}{m-1}} \frac{\|x_j - p_s\|}{\|x_j - p_i\|} = u_i \left(\frac{\|x_j - p_i\|}{\|x_j - p_s\|} \right)^{q}$$

  with $q = \frac{2}{m-1} - 1 = \frac{3-m}{m-1}$. From $1 < m < 3$ we can conclude $q > 0$ and therefore $\left(\frac{\|x_j - p_i\|}{\|x_j - p_s\|} \right)^{q} \leq 1$.

Thus in both cases we have $u_s \frac{\|x_j - p_s\|}{\|x_j - p_i\|} \leq 1$ and we can continue

$$\begin{aligned} \|\alpha_{s,j}\| \quad &\leq \quad 2 \sum_{i=1, i\neq s}^{c} u_i (1 + u_s) = 2(1 + u_s) \sum_{i=1, i\neq s}^{c} u_i \\ &\overset{(3)}{=} \quad 2(1 + u_s)(1 - u_s) = 2(1 - u_s^2) \end{aligned} \tag{44}$$

So we have found finally

$$\left\| \frac{\partial \Phi(\mathbf{p})_s}{\partial \xi} \right\| \leq \frac{2m}{m-1} \frac{\sum_{j=1}^{n} u_s^{m-1} (1 - u_s^2)}{\sum_{j=1}^{n} u_s^m} = \frac{2m}{m-1} \left(\frac{\sum_{j=1}^{n} u_s^{m-1} - u_s^{m+1}}{\sum_{j=1}^{n} u_s^m} \right)$$

Since $\mathbf{p}$ is an attractive fixed point, if $\|\frac{\partial \Phi(\mathbf{p})}{\partial \xi}\| < 1$, the statement (31) has been proven. ■

Proof of Observation 1: We require $|a_s| < \frac{1}{\sqrt{c}}$ or (32) in order to satisfy (31). A maximal *saving* is obtained for data vectors with membership degree $u = 1$ with $g(u) = \frac{1}{\sqrt{c}}$. Next we calculate the greatest *loss* a membership degree may

cause. The necessary condition for a maximal loss of a certain membership u is given by $\frac{\partial g(u)}{\partial u} = 0$ $(u \neq 0)$:

$$\begin{aligned}
\frac{\partial g(u)}{\partial u} &= \frac{m}{\sqrt{c}} u^{m-1} - \frac{2m}{m-1}(m-1)u^{m-2} + \frac{2m}{m-1}(m+1)u^m = 0 \\
\Leftrightarrow \quad 0 &= u^2 + \frac{m-1}{2\sqrt{c}(m+1)}u - \frac{m-1}{m+1} \\
\Rightarrow \quad u &= \sqrt{\frac{(m-1)^2 + 16c(m^2-1)}{16c(m+1)^2}} - \frac{m-1}{4\sqrt{c}(m+1)} \\
&= \frac{\sqrt{(m-1)^2 + 16c(m^2-1)} - (m-1)}{4\sqrt{c}(m+1)}
\end{aligned}$$

Therefore the maximal loss is $g(\hat{u})$. To compensate a maximal loss $g(\hat{u})$ (caused by a fuzzy assignment) we need $n = -\sqrt{c} \cdot g(\hat{u})$ unambiguous data objects with maximal gain $\frac{1}{\sqrt{c}}$. ■

B Proofs concerning AGK

In this section we consider the axis-parallel variant [12, 11] of the algorithm by Gustafson and Kessel (GK) [9], which will be called AGK algorithm in the following. A cluster prototype p_i in AGK consists of the cluster centre $c_i \in \mathcal{X}$ and a diagonal norm matrix $A_i \in \mathbb{R}^{\mathrm{DIM} \times \mathrm{DIM}}$ with $\det(A_i) = 1$. The algorithm uses the distance $d_i(x, \mathbf{p}) = (x_j - c_i)^\top A_i (x_j - c_i)$. In the prototype update step, the centres are adjusted just like the FCM centres (8), and the diagonal elements $a_{i,l}$ (denoting $A_i[l,l]$) according to

$$a_{i,l} = \frac{\left(\prod_{k=1}^{\mathrm{DIM}} \sum_{j=1}^{n} u_{i,j}^m (x_{j,k} - c_{i,k})^2\right)^{1/p}}{\sum_{j=1}^{n} u_{i,j}^m (x_{j,l} - c_{i,l})^2} \tag{45}$$

A necessary condition for AGK (and GK) to yield valid results is that the data set X has non-zero variance in each component.

Observation 2 *The AGK algorithm can be interpreted as a grouped-variable steepest descent algorithm (where the centres and each diagonal element are updated successively).*

Proof of Observation 2: The constraint $\det(A_i) = 1$ on prototype $p_i = (c_i, A_i)$ can be eliminated by considering only $\mathrm{DIM} - 1$ independent diagonal elements $a_{i,l}$ and defining $a_{i,\mathrm{DIM}} = 1/\prod_{l=1}^{\mathrm{DIM}-1} a_{i,l}$. The fact that the centre update can be interpreted as a steepest descent has already been discussed in the text for FCM and remains true for AGK.

Let us now consider the diagonal elements of A_i. To calculate the gradient ∇J we use (25) and therefore need ∇d_i with respect to the diagonal elements

$$\frac{\partial d_i}{\partial a_{i,l}}(x_j, \mathbf{p}) = (x_{j,l} - c_{i,l})^2 - \frac{1}{\prod_{s=1}^{\mathrm{DIM}-1} a_s a_l}(x_{j,\mathrm{DIM}} - c_{i,\mathrm{DIM}})^2$$

Then, according to (24), we have

$$a_{i,l}^{(t+1)} = a_{i,l}^{(t)} - \gamma \sum_{j=1}^{n} u_{i,j}^m \left((x_{j,l} - c_{i,l})^2 - \frac{(x_{j,\mathrm{DIM}} - c_{i,\mathrm{DIM}})^2}{\prod_{s=1}^{\mathrm{DIM}-1} a_{i,s}^{(t)} a_{i,l}^{(t)}} \right)$$

Choosing the step size $\gamma = \left(a_{i,l}^{(t)} \sum_{j=1}^{n} u_{i,j}^m (x_{j,l} - c_{i,l})^2 \right)^{-1}$ we get

$$\begin{aligned} a_{i,l}^{(t+1)} &= a_{i,l}^{(t)} - a_{i,l}^{(t)} + \frac{1}{\prod_{s=1}^{\mathrm{DIM}-1} a_{i,s}^{(t)}} \frac{\sum_{j=1}^{n} u_{i,j}^m (x_{j,\mathrm{DIM}} - c_{i,\mathrm{DIM}})^2}{\sum_{j=1}^{n} u_{i,j}^m (x_{j,l} - c_{i,l})^2} \\ &= \frac{a_{i,\mathrm{DIM}}^{(t)} \sum_{j=1}^{n} u_{i,j}^m (x_{j,\mathrm{DIM}} - c_{i,\mathrm{DIM}})^2}{\sum_{j=1}^{n} u_{i,j}^m (x_{j,l} - c_{i,l})^2} \end{aligned}$$

If we use (45) to replace $a_{i,\mathrm{DIM}}^{(t)}$ in the last equation, we obtain (45) for the considered diagonal element $a_{i,l}^{(t)}$. Thus, the $a_{i,l}$ update can be interpreted as a steepest descent. ■

Proof of Theorem 7: We use the notation from the proof of observation 2, in particular the diagonal matrices of the prototypes are characterized by $\mathrm{DIM} - 1$ elements due to the constraint $\det(A_i) = 1$. Let $\mathbf{p} \in Y$ not be a fixed point of Φ_{AGK}. Let $\mathcal{U}(\mathbf{p})$ denote the probabilistic membership matrix of $\mathbf{p}$. By definition, $J(\mathbf{p}) = J_m(\mathbf{p}, \mathcal{U}(\mathbf{p}))$ holds.

Now, the application of $\Phi_{AGK}(\mathbf{p})$ yields new prototypes $\mathbf{p}'$ and we show that $J_m(\mathbf{p}, \mathcal{U}(\mathbf{p})) > J_m(\mathbf{p}', \mathcal{U}(\mathbf{p}))$ holds. Note that the membership matrix is fixed to $\mathcal{U}(\mathbf{p})$. Choose $\xi \in Y$ with $\|\xi\| = 1$. By $\mathbf{p}_t$ we denote $\mathbf{p} + t\xi$. Let H be the convex hull of the data set. As t increases at least one prototype element or diagonal matrix element approaches infinity. Thus, there is always a T such that for all $t > T$ we leave H ($\mathbf{p}_t \in \mathcal{X} \backslash H$).

- *Case 1: With $\mathbf{p}_t$ a prototype centre is shifted to infinity.* From the definition of $d_i(x, \mathbf{p}) = (x - c_i)^\top A_i (x - c_i)$ then the distance of this prototype to each of the data objects approaches infinity and thus we have $\lim_{t \to \infty} J_m(\mathbf{p}_t, \mathcal{U}(\mathbf{p})) = \infty$.

- *Case 2: With $\mathbf{p}_t$ a diagonal matrix element $a_{i,l}$ is shifted to infinity.* Due to the non-zero variance in each component of the data set we then find a data object $x_j \in X$ such that the distance $x_{j,l} - c_{i,l}$ is not zero. While the increase of $a_{i,l}$ causes $a_{i,\mathrm{DIM}}$ to decrease the sum of distances cannot become smaller than 0. But with $t \to \infty$, we have $d_i(x_j, \mathbf{p}_i) \to \infty$, and therefore still $\lim_{t \to \infty} J_m(\mathbf{p}_t, \mathcal{U}(\mathbf{p})) = \infty$.

From the continuity of J_m we thus conclude that J_m must have at least one (global) minimum (and may have other zeros of the first derivatives indicating other local minima/maxima or saddle points). But from the fact that there is a *unique* solution for (21) we know that there is only a single extremum or saddle point of J, which therefore *must* be a strict local minimum. Since $\mathbf{p}$ is not a fixed point ($\mathbf{p} \neq \mathbf{p}'$) we have $J(\mathbf{p}') < J(\mathbf{p})$.

Secondly, we use Bezdek's results to get $J_m(\mathbf{p}', \mathcal{U}(\mathbf{p})) > J_m(\mathbf{p}', \mathcal{U}(\mathbf{p}')) = J(\mathbf{p}')$. For the case of probabilistic memberships Bezdek has shown [2] that $J_m(\mathbf{p}', U) \leq J_m(\mathbf{p}', \mathcal{U}(\mathbf{p}'))$ and $J_m(\mathbf{p}', U) = J_m(\mathbf{p}', \mathcal{U}(\mathbf{p}')) \Leftrightarrow U = \mathcal{U}(\mathbf{p}')$. His proof is independent of the chosen distances, which appear as constants in his calculations. Here, we have $U = \mathcal{U}(\mathbf{p}) \neq \mathcal{U}(\mathbf{p}')$, otherwise $\mathbf{p}$ would be a fixed point of Φ_{AGK}. ■

References

[1] A. J. Abrantes and J. S. Marques. A class of constrained clustering algorithms for object boundary extraction. *IEEE Trans. on Image Processing*, 5(11):1507–1521, Nov. 1996.

[2] J. C. Bezdek. A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(1):1–8, Jan. 1980.

[3] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[4] J. C. Bezdek, R. J. Hathaway, R. E. Howard, C. A. Wilson, and M. P. Windham. Local convergence analysis of a grouped variable version of coordinate descent. *Journal of Optimization Theory and Applications*, 54(3):471–477, Sept. 1987.

[5] J. C. Bezdek, R. J. Hathaway, M. J. Sabin, and W. T. Tucker. Convergence theory for fuzzy c-means: Counterexamples and repairs. *IEEE Trans. on Systems, Man, and Cybernetics*, 17(5):873–877, Sept. 1987.

[6] R. N. Davé. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12:657–664, Nov. 1991.

[7] R. N. Davé and R. Krishnapuram. Robust clustering methods: A unified view. *IEEE Trans. on Fuzzy Systems*, 5(2):270–293, May 1997.

[8] J. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[9] D. E. Gustafson and W. C. Kessel. Fuzzy clustering with a fuzzy covariance matrix. In *Proc. of the IEEE Conference on Decision and Control*, pages 761–766, Jan. 1979.

[10] R. J. Hathaway and J. C. Bezdek. Optimization of clustering criteria by reformulation. *IEEE Trans. on Fuzzy Systems*, 3(2):241–245, May 1995.

[11] F. Höppner, F. Klawonn, R. Kruse, and T. A. Runkler. *Fuzzy Cluster Analysis*. John Wiley & Sons, Chichester, England, 1999.

[12] F. Klawonn and R. Kruse. Constructing a fuzzy controller from data. *Fuzzy Sets and Systems*, 85:177–193, 1997.

[13] R. Krishnapuram and J. M. Keller. A possibilistic approach to clustering. *IEEE Trans. on Fuzzy Systems*, 1(2):98–110, May 1993.

[14] W. Zangwill. *Nonlinear Programming: A Unified Approach*. Prentice-Hall, Englewood Cliffs, NJ, 1969.