A Stochastic Model for Analyzing the Interpretability-Accuracy Trade-off in Interpretable Fuzzy Systems Using Nested Hyperball Structures

Krisztián Balázs[1] **László T. Kóczy**[1,2]

[1]Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Magyar tudósok körútja 2., Budapest, H-1117, Hungary
balazs@tmit.bme.hu, *koczy@tmit.bme.hu*

[2]Department of Automation
Széchenyi István University
Egyetem tér 1., Győr, H-9026, Hungary
koczy@sze.hu

Abstract

Our recent work proposed a new meaning preservation approach together with a parameterizable nested hyperball structured search space for interpretable fuzzy systems in order to solve a problem of inconsistency observed in conventional interpretable fuzzy knowledge bases and simultaneously to address the adjustment of the trade-off between interpretability and accuracy.

Based on intuitive reasonings and simulation results a conjecture was formulated about favorable trade-off adjustment properties of the proposed method.

The aim of the present paper is to construct a mathematical model, in which the conjectured properties can be analyzed and formally verified. Some computational considerations about the interpretation of the resulting knowledge bases are also made.

Keywords: Interpretable fuzzy systems, Knowledge extraction, Interpretability-accuracy trade-off, Formal analysis

1. Introduction

Fuzzy systems use fuzzy sets to describe domains of values of certain variables. Similarly to human thinking, linguistic terms can be used for this purpose. This property makes fuzzy systems rather unique among modeling systems, because while maintaining some intuitive conditions about the collection of fuzzy sets, they possess the ability to be easily interpreted, i.e. easily understandable even for laymen users.

If the knowledge base of a fuzzy system is constructed by experts using their own knowledge, i.e. the fuzzy sets in the rules of the rule base are defined manually, the mentioned conditions can be met easier than if the knowledge base is built automatically via machine learning processes. However, there are a number of methods to deal with this problem (e.g. [1], [2], [3]).

In case of interpretable fuzzy systems within *conventional approaches* (throughout this paper conventional approaches denote the ones being widely accepted within the fuzzy research community and discussed e.g. in [1], [2], [3]) a rule base is constructed considering particular restrictions. After the learning process the resulting fuzzy sets are labeled with linguistic terms. The restrictions of using only a bounded set of labeled fuzzy sets are necessary in order the linguistic terms to have more or less intuitive meanings. There are some generally accepted guidelines within the community for these restrictions as follows (see e.g. [1] – [4]):

(1) *Distinguishability*: the sets must be distinguishable from each other, i.e. the allowed overlap of the sets is limited.
(2) *Justifiable number of sets*: the number of sets must be at most as many that a human can deal with (e.g. considering the well-known Miller's number, 7 ± 2 [5]).
(3) *Normality*: each set must be normal (the height of the sets must be 1).
(4) *Coverage* (applies only for dense rule bases): the sets must form a cover of the whole input space, i.e. all elements of the input space must belong to at least one set with at least a predefined $\alpha > 0$ membership value.

It must be mentioned that other restrictions can also be considered (e.g. *convexity*, *unimodality*, *complementarity*, *uniform granulation*, *leftmost/rightmost fuzzy sets*, *natural zero positioning* [4]), or some conditions might be omitted from the list, for example, the *coverage* property can be sub-

stituted by a weaker condition in case of interpolative fuzzy systems as they use sparse rule bases (cf. [6]). Since these restrictions may vary, henceforth in this paper they are only referred to as *interpretability conditions*, and the proposed approaches will be independent of these specific restrictions.

Our past works [7] – [10] dealt with the construction of various types of fuzzy rule based knowledge extraction architectures by applying several evolutionary optimization approaches. These researches mainly focused on the efficiency of the established systems in terms of the achieved accuracy of the extracted knowledge base. However, as the outstanding inherent interpretability possibility of fuzzy systems can be a strong reason for their application, this paper deals with interpretability issues when fuzzy rule based systems are used for knowledge extraction.

Due to the trade-off between interpretability and accuracy, conventional approaches have a huge disadvantage. Depending on the resulting rule base of the learning process, totally different sets can be labeled with the very same linguistic terms, i.e. the vocabulary is not persistent throughout the wide range of problems. It might even occur, when two partitions of the input space in two contexts differ essentially, that the same sets are labeled with different linguistic terms (see Figure 1).

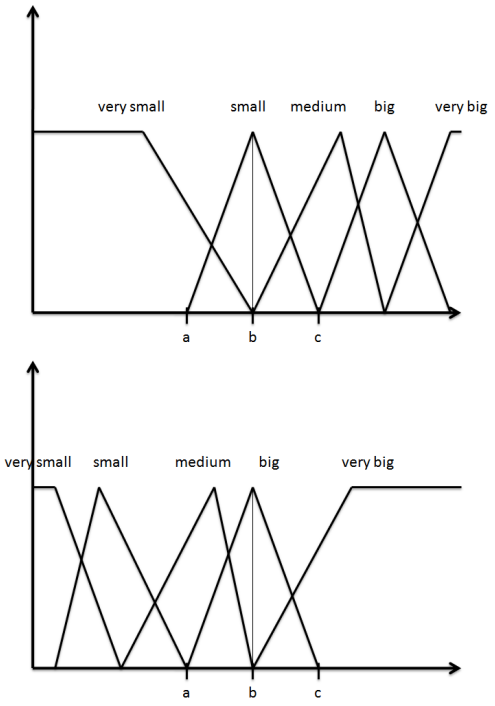

Figure 1: The same fuzzy set (characterised by points a, b and c) is labeled with different linguistic terms in two different covers

However, even if the sets belonging to the linguistic terms were defined exactly, and thus if the terms denoted the same sets in each resulting rule base, i.e. if the vocabulary was persistent, the interpretation of the result could be significantly different for two different persons, because due to the ambiguity of natural languages the meaning of a natural language term may differ for different people.

This is the reason why our recent work [11] proposed a new, personalized spectrum of approaches for constructing interpretable fuzzy systems. The main idea of these approaches is to use the linguistic terms in the same sense as the user uses them, i.e. to have a *common vocabulary* with the user.

Together with this new meaning preservation approach, a parameterizable nested hyperball structured search space was proposed for interpretable fuzzy systems in order to solve the above mentioned problem of inconsistency observed in conventional interpretable fuzzy knowledge bases and simultaneously to address the adjustment of the trade-off between interpretability and accuracy.

Based on intuitive reasonings and simulation results a conjecture was formulated about favorable trade-off adjustment properties of the proposed method. Namely, if the search space of the rule base parameters is restricted to narrower hyperballs, then although the resulting knowledge base becomes less accurate, after interpretation it will be more accurate than in case of the application of broader hyperballs.

This paper aims at constructing a mathematical model, in which this conjecture can be analyzed and formally verified.

The next section briefly describes the recently proposed approaches. The third section establishes a stochastic model as the base of formal discussions and then verifies the previously conjectured properties. Some computational considerations about the interpretation of the resulting knowledge bases will be made in the fourth section. Finally, the last section draws some conclusions and highlights some open questions concerning the new approaches.

2. The recently proposed approaches

This section gives an overview of the recently proposed approaches [11], on which the results of the present work are based.

2.1. Meaning preservation technique

As interpretability means that the knowledge is formulated in a manner that makes the information directly understandable for the user, the easiest way to meet the requirements of interpretability is to hold the information in a representation being familiar to the user. Trivially, such representations can be natural languages. However, there are some difficulties with them due to their imprecision. If people hear or read something being formulated in a natural language, they associate a meaning to the heard or read text. However, a person may associate

a certain meaning, whereas another person may associate something else, because there are no exact definitions of phrases in natural languages.

In order to deploy the terms in the same sense as the user applies them, they must be *interview*ed. A simple interview could be to ask the user to define the fuzzy sets. However, supposing someone not being familiar with fuzzy sets at all (which is a rather realistic assumption), the interview can be worked out by using fuzzy membership elicitation techniques (see e.g. [12]). After that, the adequate fuzzy sets can be constructed easily.

Linguistic terms may not involve only adjectives (e.g. 'hot'), but modifiers, so-called linguistic hedges, too (e.g. 'a bit', or 'very'). These modifiers can be considered to be transformations of the sets of the adjectives being under modification. Therefore, if the user is interviewed about 'cold', 'hot' and about how the user modifies the meaning of an adjective if it is combined with the linguistic hedge 'very', the meanings of 'very cold' and 'very hot' need not be interviewed, because they can be computed by applying the transformation of 'very' on the fuzzy sets of 'cold' and 'hot'. This may lead to a complexity reduction. (Obviously, the transformations should be carefully defined based on a well-designed interview, because e.g. in case of 'very' shifting the certain sets with a positive value may be suitable for 'hot', but it is surely not a proper action for 'cold'.)

The whole procedure may work in reverse, too. Instead of interviewing, the user could be *train*ed, i.e. the user could be told about the meaning of certain terms (adjectives and modifiers) in a similar way to the interview.

Based on the user defined linguistic terms, fuzzy rules and rule bases can be constructed easily. However, not every rule base constructed from these terms will fulfill the interpretability conditions, and thus not all of them will be interpretable (e.g. due to lack of consistency). These ones will be called invalid, whereas the ones fulfilling the interpretability conditions will be referred to as valid *interpretable solutions*.

2.2. Nested hyperball structured search space

The interpreted information can be characterized by a finite, but in practice, a limited amount of features, because a human can deal with only a limited number of information units. Furthermore, a human cannot distinguish between units of information being too close to each other in meaning, i.e. the granularity of distinguishable information is not infinitely small, and hence the space of possible solutions is bounded. Thus, the set of interpretable solutions will be considered finite and will be denoted by X_0, hereafter.

If a fuzzy system is constructed from samples by applying supervised machine learning techniques and interpretability is the main objective of this process, the task of learning is to determine an $x_0^* \in X_0$, such that $\forall x \in X_0 : A(x_0^*) \geq A(x)$, where $A(.)$ is the measure of (relative) accuracy, which is a strictly monotonic decreasing function of the error, which can be calculated e.g. based on the differences between the outputs of the system and the desired outputs. This x_0^* can be achieved by global searching numerical optimization algorithms after a sufficient time. The result of the learning process is the most accurate knowledge base among interpretable solutions. Clearly, the stress is on interpretability in this case.

Let X_∞ denote the largest considerable set of parameter vectors of the particular fuzzy rule base. If a fuzzy system is constructed from samples by applying supervised machine learning techniques and accuracy is the main objective of this process, the task of learning is to determine an $x_\infty^* \in X_\infty$, such that $\forall x \in X_\infty : A(x_\infty^*) \geq A(x)$, where $A(.)$ is the same (relative) accuracy function as it was above. This x_∞^* can be approximated with arbitrary accuracy by global searching numerical optimization algorithms (recall that global search methods stochastically converge to the global optimum). The result of the learning process is the most accurate knowledge base regardless of interpretability.

It is obvious that if a sequence of search spaces being nested into each other $X_0 \subset X_{r_1} \subset X_{r_2} \subset \cdots \subset X_{r_n} \subset X_\infty$ (where the sequence of r_i indices is a strictly monotonic increasing sequence) is defined, then it has a positive probability that an optimal solution in a broader space has higher accuracy than all the elements of a narrower space.

However, if $r_i > 0$ and $x_{r_i}^* \notin X_0$ (where $x_{r_i}^*$ is the optimal solution within X_{r_i}), an interpretation can also be given, if there is an interpreter function $\mathfrak{I} : X_\infty \mapsto X_0$, such that $\mathfrak{I}(x_{r_i}^*)$ is somehow the 'closest' element from X_0 to $x_{r_i}^*$, i.e. $\forall x \in X_0 : d(x_{r_i}^*, \mathfrak{I}(x_{r_i}^*)) \leq d(x_{r_i}^*, x)$, where $d : X_\infty \times X_\infty \mapsto \mathbb{R}^+ \cup \{0\}$ is a metric. That is, the interpretation of a solution $x_{r_i}^* \notin X_0$ is the closest interpretable solution $x_0 \in X_0$ to $x_{r_i}^*$ according to a distance function.

It is clear, that $\mathfrak{I}(x_\infty^*)$ can never be more accurate than x_0^* by definition as well as x_0^* can never be more accurate than x_∞^*.

This shows (matching intuitive expectations) that interpretability and accuracy are conflicting requirements: if an interpretable knowledge base is constructed, it is less accurate, and if a more accurate one is constructed, expectedly, after interpretation it becomes less accurate than if interpretation had been the main, and accuracy had only been a secondary objective.

These conflicting approaches can be combined with different weights to intermediate approaches if both the accuracy of the non-interpreted knowledge base and the accuracy of the interpreted one are important. Such combinations can be achieved

by narrowing the search space of possible knowledge bases and producing a sequence of nested search spaces $X_0 \subset X_{r_1} \subset X_{r_2} \subset \cdots \subset X_{r_n} \subset X_\infty$.

It would be greatly favorable, if, as a benefit, there was a tendency showing that the interpreted solution was expectedly more accurate in case of a narrower search space, because this way by choosing a narrower search space from the sequence, although, the accuracy of the non-interpreted knowledge base would be lower, the accuracy of the interpreted knowledge base would be higher. Therefore, roughly spoken, one could balance between interpretability and accuracy by selecting the proper search space.

As it was confirmed experimentally in [11], this tendency holds, if the search spaces are unions of hyper-balls constructed around all the elements of X_0, where the hyper-balls have the same radius values, furthermore a broader and a narrower search space differ from each other only in the radius value.

Henceforth, r_1, r_2, etc. will not only denote indices here, but they will also stand for the radius of the corresponding balls. Furthermore, X_0 can be considered as the union of balls having zero radius around the interpretable solutions, i.e. the interpretable solutions themselves, and X_∞ can be considered as the union of infinitely large balls that cover the whole problem space. This is the reason of the indices 0 and ∞.

Formally, the search spaces can be defined as follows:

$$X_r = \{x \in X_\infty | \exists x_0 \in X_0 : d(x, x_0) < r\} \qquad (1)$$

Clearly, this way the search spaces are nested into each other ($X_{r_i} \subset X_{r_j}$, if $r_i < r_j$).

3. Mathematical model and formal analysis

In order to verify the above property theoretically, it is necessary to establish a mathematical model as a ground for the formal analysis. Since there are basically two significant unknown components in a machine learning process, namely the learning problem and the (quasi-)optimal knowledge base (i.e. the parameter set of the learning architecture) obtained, the mathematical model definitely has to be stochastic. Therefore, in the following the basic notions of the model will be defined in the light of this requirement.

It is also obvious that beyond the stochastic behavior the model has to be established by assuming the fulfillment of some intuitive conditions. These assumptions will follow after the basic definitions.

After the construction of the model, the expected properties will be derived formally.

3.1. Basic definitions

The definitions of interpretable and valid solutions emphasizing their relation, furthermore the definitions of the given translation invariant metric, the accuracy function, the interpreter function and the narrowed search spaces are repeated first.

Definition 1. *X_0 is the set of* interpretable solutions*, X_∞ is the set of* valid solutions*, for which sets the relations $X_0 \subseteq X_\infty \subseteq \mathbb{R}^n$ hold true, where n is the number of numerical parameters of the modeling architecture.*

Definition 2. *$d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_0^+$ is an arbitrary translation invariant metric (where $\mathbb{R}_0^+$ is the set of non-negative real numbers).*

Definition 3. *$A : X_\infty \to \mathbb{R}$ is the* accuracy function*, which is continuous over X_∞.*

Definition 4. *$\mathfrak{I} : X_\infty \to X_0$ is the* interpreter function*, for which $\forall x_\infty \in X_\infty \forall x_0 \in X_0 : d(\mathfrak{I}(x_\infty), x_\infty) \leq d(x_0, x_\infty)$ (w.r.t. the given metric d).*

Definition 5. *$X_r := \{x \in X_\infty | \exists x_0 \in X_0 : d(x, x_0) < r\} \subseteq \mathbb{R}^n$ is the* narrowed search space *w.r.t. a given $r \in \mathbb{R}_0^+ \cup \{\infty\}$.*

Henceforth, r_1, r_2, etc. will not only denote indices, but they will also stand for the radius of the corresponding balls. Furthermore, X_0 can be considered as the union of balls having zero radius around the interpretable solutions, i.e. the interpretable solutions themselves, and X_∞ can be considered as those parts of the union of infinitely large balls covering the whole problem space. This is the reason of the indices '0' and '∞'.

Definition 6. *$X_r^* := \{x \in X_r | \forall y \in X_r : A(x) \geq A(y)\}$ is the* set of optimal solutions *w.r.t. a given $r \in \mathbb{R}_0^+ \cup \{\infty\}$.*

In order to obtain a stochastic model the event space, the event algebra and the probability measure have to be defined.

Definition 7. *$\mathcal{L}$ is the set of* possible learning problems *(defined by e.g. input-output samples) and $\mathcal{C}$ is the set of* choice functions *on $\{X_r^* | 0 \leq r \leq \infty\}$. (That is, every $f \in \mathcal{C}$ assigns an element $f(X_r^*)$ of X_r^* to each set X_r^* for every $0 \leq r \leq \infty$).*

Then $\Omega := \mathcal{L} \times \mathcal{C}$ is the event space*, which contains "(learning problem, function determining the optima chosen by the optimization process for all radius values)" pairs as elementary events.*

The event algebra $\mathcal{F}$ is the σ-algebra containing all the elements of $\mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ denotes the powerset of the event space, $\mathcal{F} = \mathcal{P}(\Omega)$.

There is also given a probability measure $\mathbb{P}$ over $\mathcal{F}$ defining how frequently the events arise.

The (quasi-)optimal solutions found by the optimization algorithms are defined as random variables.

Definition 8. *$x_r^* : \Omega \to X_r^*$ is a random variable for each r determined by the choice functions of the elementary events (the optimum chosen by the optimization algorithm).*

Since for every r the set X_r is a subset of X_∞, X_r is not necessarily the union of balls. In the following definition R is the radius of the largest open balls contained in X_∞ around all the interpretable solutions.

Definition 9. *$R := \sup\{r \in \mathbb{R}_0^+ \cup \{\infty\} | \forall x_0 \in X_0 : B_r(x_0) \subseteq X_r\}$ is the* critical radius *w.r.t. the metric d, where $B_r(x_0) = \{x \in \mathbb{R}^n | d(x, x_0) < r\}$ is the open ball around x_0 in $\mathbb{R}^n$ with radius r.*

In later formal argumentations different events will be considered. Some of them are defined here.

Definition 10. *Given $0 \leq r_1 < r_2 \leq \infty$, then the following two events form a partition of Ω. $E_1 := \{\omega \in \Omega | x^*_{r_2} \in X_{r_1}\}$, i.e. $x^*_{r_2}$ lies inside the narrower search space X_{r_1}, $E_2 := \{\omega \in \Omega | x^*_{r_2} \notin X_{r_1}\}$, i.e. $x^*_{r_2}$ lies outside the narrower search space X_{r_1}.*

3.2. Intuitive assumptions

The intuitively defined conditions about the model are listed below.

It is a reasonable assumption that the following event has a positive probability: "the optimum of a broader search space has a higher accuracy than the optimum of a narrower one".

Assumption 1. *If $0 \leq r_1 < r_2 \leq \infty$, then*

$$\mathbb{P}(A(x^*_{r_1}) < A(x^*_{r_2})) > 0 \tag{2}$$

Assume that if an optimization process running in a broader search space finds a solution within a narrower set, then the distribution of this solution is equal to the distribution of the one found by the optimization process when running in the narrower search space.

Assumption 2. *If $0 \leq r_1 < r_2 \leq \infty$, then in case of the event E_1, the conditional distribution of the random variables $x^*_{r_1}$ and $x^*_{r_2}$ are equal, i.e. for each Borel set $B \in \mathcal{B}(X_{r_1})$: $\mathbb{P}(x^*_{r_1} \in B | E_1) = \mathbb{P}(x^*_{r_2} \in B | E_1)$.*

It is also assumed that if two points within the set of valid solutions having the same accuracy value are translated by the same vector, the expected accuracy will remain the same.

Assumption 3. *$\forall x_1, x_2 \in X_\infty, \forall v \in \mathbb{R}^n : (A(x_1) = A(x_2)) \wedge (x_1 + v \in X_\infty) \wedge (x_2 + v \in X_\infty) \rightarrow \mathbb{E}(A(x_1 + v)|E) = \mathbb{E}(A(x_2 + v)|E)$, for both events $E = E_{222}$ and $E = E_6$ (defined later in Lemma 2 and in Theorem 2).*

The following assumption declares the positive probability of an event defined later.

Assumption 4. *The event E_{221}, which will be defined in Lemma 2, has positive probability, i.e. $\mathbb{P}(E_{221}) > 0$.*

3.3. Theoretical results

The expected property will be proved in parts through lemmas. The first lemma applies the following proposition.

Proposition 1. *If $0 \leq r_1 < r_2 \leq \infty$, then:*

$$\forall \omega \in \Omega : A(x^*_{r_1}) \leq A(x^*_{r_2}) \tag{3}$$

Proof. It follows from Definition 5 that $X_{r_1} \subset X_{r_2}$, hence due to Definition 8 and Definition 6 the proposition holds. □

Lemma 1. *If $0 \leq r_1 < r_2 \leq \infty$, then $\mathbb{E}A(x^*_{r_1}) < \mathbb{E}A(x^*_{r_2})$.*

Proof. Let us define two events:

$$E_3 := \{\omega \in \Omega | A(x^*_{r_1}) = A(x^*_{r_2})\} \tag{4}$$

$$E_4 := \{\omega \in \Omega | A(x^*_{r_1}) < A(x^*_{r_2})\} \tag{5}$$

It is clear that due to Proposition 1 these events form a partition of Ω.

Applying the "Tower Law" [13]:

$$i \in \{1, 2\} : \mathbb{E}A(x^*_{r_i}) = \mathbb{E}(\mathbb{E}(A(x^*_{r_i} | \sigma(E_3, E_4)))) =$$

$$\mathbb{E}(A(x^*_{r_i} | E_3))\mathbb{P}(E_3) + \mathbb{E}(A(x^*_{r_i} | E_4))\mathbb{P}(E_4), \tag{6}$$

where $\sigma(E_3, E_4)$ is the σ-algebra generated by events E_3 and E_4.

It follows from Eq. 4 and Eq. 5 that $\mathbb{E}(A(x^*_{r_1}|E_3)) = \mathbb{E}(A(x^*_{r_2}|E_3))$ and $\mathbb{E}(A(x^*_{r_1}|E_4)) < \mathbb{E}(A(x^*_{r_2}|E_4))$ hold true, respectively.

Therefore, considering Assumption 1, i.e. $\mathbb{P}(E_4) > 0$, it follows that $\mathbb{E}A(x^*_{r_1}) < \mathbb{E}A(x^*_{r_2})$. □

Lemma 2. *If $0 \leq r_1 < r_2 \leq R$, then $\mathbb{E}A\mathfrak{I}(x^*_{r_2}) < \mathbb{E}A\mathfrak{I}(x^*_{r_1})$.*

Proof. Let us distinguish three cases:

(1) In case of the event E_1, due to Assumption 2, the interpreted solutions $\mathfrak{I}(x^*_{r_1})$ and $\mathfrak{I}(x^*_{r_2})$ have the same distributions, i.e. for each Borel set $B \in \mathcal{B}(X_0)$: $\mathbb{P}(\mathfrak{I}(x^*_{r_1}) \in B|E_1) = \mathbb{P}(\mathfrak{I}(x^*_{r_2}) \in B|E_1)$, thus $\mathbb{E}(A\mathfrak{I}(x^*_{r_1})|E_1) = \mathbb{E}(A\mathfrak{I}(x^*_{r_2})|E_1)$.

(2) In case of the event E_2 if $\mathfrak{I}(x^*_{r_1}) = \mathfrak{I}(x^*_{r_2})$ (event E_{21}), then $\mathbb{E}(A\mathfrak{I}(x^*_{r_1})|E_{21}) = \mathbb{E}(A\mathfrak{I}(x^*_{r_2})|E_{21})$, obviously.

(3) In case of the event E_2 if $\mathfrak{I}(x^*_{r_1}) \neq \mathfrak{I}(x^*_{r_2})$ (event E_{22}), then let us define x' such that $x' := \mathfrak{I}(x^*_{r_2}) + (x^*_{r_1} - \mathfrak{I}(x^*_{r_1}))$ (see Figure 2). Then $A(x^*_{r_1}) \geq A(x')$, otherwise $x^*_{r_1}$ would not be in $X^*_{r_1}$, which would be a contradiction. Now, consider the functions $f, g : [0, 1] \rightarrow \mathbb{R}$

$$f(\lambda) = A(\lambda\mathfrak{I}(x^*_{r_1}) + (1 - \lambda)x^*_{r_1}) \tag{7}$$

$$g(\lambda) = A(\lambda\mathfrak{I}(x^*_{r_2}) + (1 - \lambda)x') \tag{8}$$

If $\nexists\lambda \in [0,1] : f(\lambda) = g(\lambda)$ (event $E_{221} \subseteq E_{22}$), then since A is continuous over X_∞, due to the Intermediate Value Theorem [14] and the fact that $f(0) = A(x^*_{r_1}) \geq A(x') = g(0)$, $\forall\lambda \in [0,1] : f(\lambda) > g(\lambda)$ holds. Hence $\mathfrak{I}(x^*_{r_1}) = f(1) > g(1) = \mathfrak{I}(x^*_{r_2})$. Thus, $\mathbb{E}(A\mathfrak{I}(x^*_{r_1})|E_{221}) > \mathbb{E}(A\mathfrak{I}(x^*_{r_2})|E_{221})$ holds.
If $\exists\lambda \in [0,1] : f(\lambda) = g(\lambda)$ (event $E_{222} \subseteq E_{22}$), then applying Assumption 3 with the substitutions $x_1 := \lambda\mathfrak{I}(x^*_{r_1}) + (1-\lambda)x^*_{r_1}, x_2 := \lambda\mathfrak{I}(x^*_{r_2}) + (1-\lambda)x'$ and $v := (1-\lambda)(\mathfrak{I}(x^*_{r_1}) - x^*_{r_1})$ (see Figure 2), equation $\mathbb{E}(A\mathfrak{I}(x^*_{r_1})|E_{222}) = \mathbb{E}(A\mathfrak{I}(x^*_{r_2})|E_{222})$ holds.

Since $\{E_1, E_{21}, E_{221}, E_{222}\}$ is a partition of Ω and according to Assumption 4 event E_{221} has positive probability, the statement follows. □

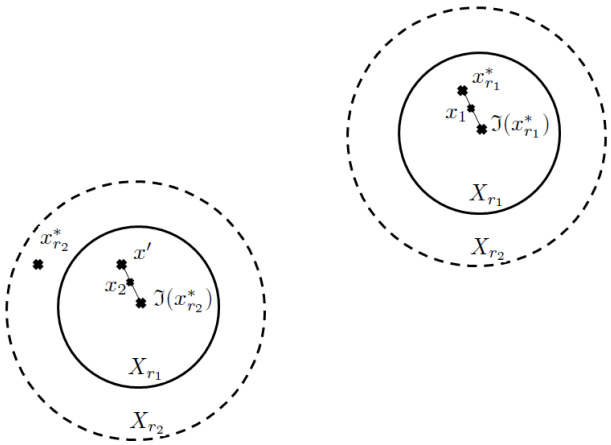

Figure 2: Illustration of the variables defined in Lemma 2.

Based on the above lemmas it can be seen that the formal analysis verifies the expected property formulated in the following theorem.

Theorem 1. *If $0 \leq r_1 < r_2 \leq R$, then*

$$\mathbb{E}A\mathfrak{I}(x^*_{r_2}) < \mathbb{E}A\mathfrak{I}(x^*_{r_1}) \leq \mathbb{E}A(x^*_{r_1}) < \mathbb{E}A(x^*_{r_2}). \quad (9)$$

The equality in the middle holds exactly when $r_1 = 0$.

Proof. From Lemma 1 and Lemma 2 it is straightforward to see that the theorem holds true. □

Remark 1. *Without Assumption 1 and Assumption 4 the inequalities of Eq. 9 within Theorem 1 would not be strict:*

$$\mathbb{E}A\mathfrak{I}(x^*_{r_2}) \leq \mathbb{E}A\mathfrak{I}(x^*_{r_1}) \leq \mathbb{E}A(x^*_{r_1}) \leq \mathbb{E}A(x^*_{r_2}). \quad (10)$$

Proof. This follows from the proofs of Lemma 1 and Lemma 2. □

The next theorem shows why the interpreter function should choose the closest interpretable solution.

Theorem 2. *If $0 \leq r \leq R$, then the closest interpretable solution $\mathfrak{I}(x^*_r)$ gives the highest expected accuracy among the interpretable solutions.*

Proof. The proof will be similar to the third case in Lemma 2.

Take an arbitrary interpretable solution $x_0 \in X_0$. Let us define x' such that $x' := x_0 + (x^*_r - \mathfrak{I}(x^*_r))$. Then $A(x^*_r) \geq A(x')$, otherwise x^*_r would not be in X^*_r, which would be a contradiction. Now, consider the functions $f, g : [0,1] \to \mathbb{R}$

$$f(\lambda) = A(\lambda\mathfrak{I}(x^*_r) + (1-\lambda)x^*_r) \quad (11)$$

$$g(\lambda) = A(\lambda x_0 + (1-\lambda)x') \quad (12)$$

If $\nexists\lambda \in [0,1] : f(\lambda) = g(\lambda)$ (event E_5), then since A is continuous over X_∞, due to the Intermediate Value Theorem and the fact that $f(0) = A(x^*_r) \geq A(x') = g(0)$, $\forall\lambda \in [0,1] : f(\lambda) > g(\lambda)$ holds. Hence $\mathfrak{I}(x^*_r) = f(1) > g(1) = x_0$. Thus, $\mathbb{E}(A\mathfrak{I}(x^*_r)|E_5) > \mathbb{E}(A(x_0))|E_5)$ holds.

If $\exists\lambda \in [0,1] : f(\lambda) = g(\lambda)$ (event E_6), then applying Assumption 3 with the substitutions $x_1 := \lambda\mathfrak{I}(x^*_r) + (1-\lambda)x^*_r, x_2 := \lambda x_0 + (1-\lambda)x'$ and $v := (1-\lambda)\mathfrak{I}(x^*_r) - (1-\lambda)x^*_r$, the equation $\mathbb{E}(A\mathfrak{I}(x^*_r)|E_6) = \mathbb{E}(A(x_0))|E_6)$ holds.

Since $\{E_5, E_6\}$ is a partition of Ω, the statement of the theorem follows. □

Naturally, depending on the applied metric the balls may not only be balls (like in case of Euclidean-metric), but they can be e.g. cubes (Maximum-metric), octahedrons (Manhattan-metric), ellipsoids (Mahalanobis-metric), or many others. In other words, every object being a ball according to any translation invariant metric can be applied instead of Euclidean balls.

The two original and the uncountable intermediate approaches (since the radius can take arbitrary values from $(0, \infty)$) form a whole spectrum between the two opposite ends, i.e. between the interpretable-oriented, and the accuracy-oriented approaches. Selecting an approach closer to the interpretable-oriented end results in a knowledge base being less accurate before and more accurate after the interpretation, whereas selecting an approach closer to the accuracy-oriented end results in a knowledge base being more accurate before and less accurate after the interpretation, expectedly.

If the newly proposed approaches are compared to the conventional ones from the point of view of expected accuracy, the following observations can be made. Conventional techniques search in a narrowed solution space $X_{conv} \subset X_\infty$ due to the interpretability conditions, but after the learning process the resulting fuzzy sets are labeled, which is the same as applying an interpreter function $\mathfrak{I}_{conv}$. Obviously, in general the thus obtained interpretable solution will not be equal to the one given by $\mathfrak{I}$, because in the conventional case not necessary the closest interpretable solutions are selected, which leads to a sub-optimal solution (cf. Theorem 2).

4. Interpretation of the extracted knowledge base — computational considerations

In case of the interpretable-oriented approach, there is no need for interpretation, since the result is already interpretable ($\mathfrak{I}(x_0^*) = x_0^*$). Otherwise, three cases can be distinguished:

1. The value of the radius of the balls r is less than or equal to the minimum distance between interpretable solutions. In this case there is exactly one interpretable solution around x_r^* within the distance of r. Therefore, a ball with radius r should be constructed around x_r^*, and the one and only interpretable solution within the ball will be $\mathfrak{I}(x_r^*)$.
2. The value of the radius of the balls r is greater than the minimum distance between interpretable solutions and at most r_0, which is a predefined limit. In this case there are intersecting balls, hence there can be more than one (but at least one) interpretable solution around x_r^* within the distance of r. Thus after a ball with radius r is constructed around x_r^*, within the ball all the distances between the interpretable solutions and x_r^* should be computed and compared to each other. Then the closest interpretable solution should be chosen.
3. The value of the radius of the balls r is greater than r_0. In this case there can be so many interpretable solutions around x_r^* within the distance of r, that it would result a significant computational demand to find all of the interpretable solutions within the ball and compute their distances from x_r^*. A better choice to construct a ball with radius r_0, and if there are no interpretable solutions in the ball, to construct another one with radius $r_1 > r_0$, and so forth, iteratively, until there is at least one interpretable solution within the ball. Definitely, in case of X_∞ this method should be applied instead of constructing such a large ball that certainly contains at least one interpretable solution and risking the possibility to construct a ball containing all interpretable solutions. Because in this unfavorable case each interpretable solution should be found and all distances should be computed, which has the same (huge) time complexity (apart from a constant factor) as to evaluate all of the interpretable solutions, i.e. to find x_0^* with exhaustive search, whereas x_0^* is the best interpretable solution by definition, but $\mathfrak{I}(x_\infty^*)$ is not certainly.

5. Conclusions

In the first part of the present paper a brief overview was given about our recently proposed approaches for interpretable fuzzy systems, where a meaning preservation technique together with a new parameterizable search space narrowing method was applied in order to simultaneously deal with an inconsistency problem of conventional interpretable fuzzy systems and the adjustability of the interpretability-accuracy trade-off. The conjecture formulated in [11] announcing favorable properties of the trade-off adjustment approach was also recalled.

The second part of the paper mainly aimed at formally analyzing and proving the mentioned conjecture based on an intuitively established stochastic model. After the successful formal verification of the expected favorable properties, some computational considerations about the interpretation of the extracted knowledges are made.

Despite the successful formal reasonings, there are a number of open questions left about the recently proposed approaches. What type of metrics would be practical for certain problems? How could the optimum be searched in the narrowed spaces efficiently? What about the time complexity of the proposed approaches? And so forth...

Future work may aim at finding answers to such questions.

Acknowledgments

The research was supported by the National Scientific Research Fund Grants OTKA K75711, OTKA K105529 and a Széchenyi István University Main Research Direction Grant.

References

[1] J. Espinosa and J. Vandewalle, Constructing Fuzzy Models with Linguistic Integrity from Numerical Data — AFRELI Algorithm, *IEEE Transactions on Fuzzy Systems*, 8(5):591–600, 2000.

[2] J. M. Alonso, L. Magdalena and S. Guillaume, HILK: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism, *International Journal of Intelligent Systems*, 23(7):761–794, 2008.

[3] J. M. Alonso, L. Magdalena, and G. Gonzalez-Rodriguez, Looking for a good fuzzy system interpretability index: An experimental approach, *Int. Journal of Approximate Reasoning*, 51:115–134, 2009.

[4] C. Mencar and A. M. Fanelli, Interpretability constraints for fuzzy information granulation, *Information Sciences*, 178(24):4585–4618, 2008.

[5] G. A. Miller, The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, *The Psychological Review*, 63:81–97, 1956.

[6] L. T. Kóczy and K. Hirota, Approximate reasoning by linear rule interpolation and general approximation, *Internat. J. Approx. Reason.*, 9:197–225, 1993.

[7] K. Balázs, J. Botzheim and L. T. Kóczy, Comparative Analysis of Interpolative and Non-interpolative Fuzzy Rule Based Machine Learning Systems Applying Various Numerical Optimization Methods, *World Congress on Computational Intelligence* (WCCI 2010), pages 875–982, Barcelona (Spain), 2010.

[8] K. Balázs, J. Botzheim and L. T. Kóczy, "Hierarchical Fuzzy System Modeling by Genetic and Bacterial Programming Approaches", *World Congress on Computational Intelligence* (WCCI 2010), pages 1866–1871, Barcelona (Spain), 2010.

[9] K. Balázs and L. T. Kóczy, Constructing dense, sparse and hierarchical fuzzy systems by applying evolutionary optimization techniques, *Applied and Computational Mathematics*, 11(1):81–101, 2012.

[10] K. Balázs and L. T. Kóczy, Hierarchical-Interpolative Fuzzy System Construction By Genetic And Bacterial Memetic Programming Approaches, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(supp02):105–131, 2012.

[11] K. Balázs and L. T. Kóczy, New Parameterizable Search Space Narrowing Technique for Adjusting between Accuracy and Interpretability in Fuzzy Systems, 13^{th} *IEEE International Symposium on Computational Intelligence and Informatics* (CINTI 2012), pages 323–328, Budapest (Hungary), 2012.

[12] D. Dubois and H. Prade. *Fuzzy Sets and Systems, Theory and Applications*, Academic Press Inc., Chestnut Hill, MA, USA, 1980.

[13] R. Durrett. *Probability: Theory and Examples, Fourth edition*, Cambridge University Press, 2010.

[14] W. Rudin. *Real and complex analysis*, McGraw-Hill, 1987.