Normalization Invariance: A New Approach to Foundations of Fuzzy Control

Olga Kosheleva and Vladik Kreinovich
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
olgak@utep.edu, vladik@utep.edu

Abstract

In many practical situations – e.g., in many driving situations – skilled humans are still performing better than automated control systems. It is therefore desirable to incorporate the knowledge of these skilled human controllers into an automatic systems. This is not easy, since for a significant part of this knowledge, experts cannot describe it in precise computer-understandable terms, they can only describe it by using imprecise ("fuzzy") words from natural language like "small". Techniques for translating such "fuzzy" knowledge into precise computer-understandable form are known as *fuzzy* techniques. There are many versions of fuzzy techniques, some work better, some perform worse – and the difference between the best and worst performances is often drastic. At present, the best techniques are often selected by time-consuming trial-and-error. In this paper, we show that in many cases, such a procedure can be sped up if we use a natural requirement that these techniques should not change under appropriate normalizations.

1 Formulation of the Problem

The ubiquity and importance of imprecise ("fuzzy") knowledge. In the 1960s, Lotfi Zadeh, who was, at that time, one of the world's recognized leaders in control and a co-author of the most popular control textbook, noticed that control by human experts was often better than the supposedly optimal control based on the mathematical models.

How can this be? A natural explanation – which the experts themselves confirmed – was that the experts possess some knowledge that is not yet reflected in the mathematical models. So why not incorporate this knowledge into an automatic control systems to improve their performance?

The problem was that this additional expert knowledge was not formulated in precise computer-understandable terms. The experts could only describe it by using imprecise ("fuzzy") words from natural language.

This is normal: many people around the world can drive, and many of them are willing to explain how they drive – and even teach others to drive – but a perfect self-driving car, a car that is, in all situations, at least as reliable as an average driver, remains a dream. Why? Because if you ask a driver what to do when he/she is on the freeway driving at a speed of 65 miles per hours, and a car 10 feet in front slows down to 60 miles per hour, a natural answer is: "Slow down a little bit". This is perfectly understandable to humans, this is how many of use learned to drive, but to implement such advice in an automatic system, we need to know for how many milliseconds and with what exactly force to press on the brake – and a normal human driver cannot describe his/her recommendation in this form.

Enter fuzzy techniques. Lotfi Zadeh called techniques that utilize such missing "fuzzy" knowledge *fuzzy techniques*, and he came up with several version of such techniques; see, e.g., [2, 3, 4, 5, 6, 7].

Zadeh's fuzzy techniques have been very successful. Fuzzy methodology has led to many successful applications. It helps to design control for trains, for elevators, for rice cookers, for cars' automatic transmissions, and for many other objects.

Some options of fuzzy techniques work better than others. As we will explain in the following sections, fuzzy methodology has many options: we can select different membership functions, different "and"- and "or"-operations, different hedge functions, etc.

Empirically, some options lead to more effective control, some are less effective.

Resulting problem. As of now, in most cases, the option that leads to a better control is selected empirically, by trial and error. Trial and error often succeeds, but it usually takes a lot of tries to find a good combination of options.

It is therefore desirable to come up with some ideas that will guide us to the good option faster.

What we do in this paper. In this paper, we show that it is indeed possible to limit possible choices – and thus, to speed up the search – if we take into account one aspect of the original Zadeh's scheme, an aspect that is often overlooked in textbooks – the need for normalization.

The structure of this chapter. We start, in Section 2, with explaining the main notions behind Zadeh's methodology – the notion of degree of confidence, and the resulting notions of a membership function and a fuzzy set. In describing this notion, we will pay special attention to the idea of normalization.

After that, in Section 3, we will introduce other components of fuzzy methodology: "and"- and "or"-operations and hedges. Then, in Section 4, we introduced the notion of normalization-invariance and show:

- that many empirically successful choices of fuzzy methodology are normalization-invariant, and

- that, in many cases, the requirement that these choices be normalization-invariant drastically narrows down the set of possible choices – and thus, potentially speeds up the search for the best option.

2 Degree of confidence, membership function, fuzzy set, and normalization: a brief reminder

What is a degree of confidence. How can we describe an imprecise notion like "small" or "tall"?

To describe a precise notion, such as "positive" or "smaller than 10", it is sufficient to be able to describe, for each possible value x of the corresponding quantity, whether this value satisfies the given property or not. In mathematical terms, to describe a precise property, we need to describe the *set* of all the values that satisfy this property. A set can be alternatively described by its *membership function* $m(x)$ which is:

- equal to 1 when the number x belongs to the set (i.e., the value x satisfies the given property), and
- equal to 0 when the number x does not belong to the set (i.e., the value x does not satisfy the given property).

Unfortunately, such a description is not possible for properties like "small". Yes:

- for some very small values x, we are indeed absolutely sure that x is small, and
- for some very large values x, we are indeed absolutely sure that x is not small.

However, for many intermediate values x, we can only say that x is small to some extent.

How can we describe this "extent"? To most of us, transforming your degree into a number is a normal practice: this is what we usually do when, after visiting a bank or a restaurant, we are asked to fill a survey, with questions like: "On the scale from 0 to 10, estimate how satisfied you are with this visit."

It is convenient to describe all degrees of confidence in the same scale. In principle, we can use different scales. In surveys, it is usually 0 to 10; in student evaluations of faculty, it is 0 to 4, etc.

To make degrees compatible, Zadeh suggested to re-scale everything to the scale from 0 to 1. The reason behind selecting this 0-to-1-scale is that it is in perfect accordance of the usual computer representations. Indeed, the maximal degree of confidence corresponds to absolute truth – and in a computer, "true" is usually represented as 1. Similarly, the smallest degree of confidence means that the expert is absolutely sure that the corresponding statement is false –

and in a computer, "false" is usually represented as 0. It is therefore reasonable to describe intermediate degrees of confidence by numbers between 0 and 1.

How shall we transform from one scale to another? How can we transform from the original scale to the scale from 0 to 1?

In general, a transformation that reduces different scales to a single scale – the scale that serves as a kind of a norm – is called *normalization*. In these terms, the question is: what normalization should be use?

A natural idea is to make sure that larger values correspond to larger values, and that equally spaced values remain equally spaced. Let us describe this requirement in precise terms.

Definition 1. *Let $N > 0$ be a number. We say that a mapping $f : [0, N] \mapsto [0, 1]$ for which $f(0) = 0$ and $f(N) = 1$ is a* natural re-scaling *if it has the following two properties:*

- *if $a < b$, then $f(a) < f(b)$, and*
- *if $b - a = c - b$, then $f(b) - f(a) = f(c) - f(b)$.*

Proposition 1. *For each $N > 0$, there is only one natural re-scaling $f(a) = a/N$.*

Proof. It is easy to see that the function $f(a) = a/N$ is a natural re-scaling. Let us show that it is the only one. Indeed, let $f(a)$ be a natural re-scaling. Then for every integer n, we have

$$0 < N/n < 2N/n < \ldots < N$$

and

$$N/n - 0 = 2N/n - N/n = \ldots = N - (n - 1) \cdot N/n.$$

Thus, because of Definition 1, we have

$$f(N/n) - f(0) = f(2N/n) - f(N/n) = \ldots = f(N) - f((n - 1) \cdot N/n).$$

Since $f(0) = 0$ and $f(N) = 1$, we get

$$f(N/n) - 0 = f(2N/n) - f(N/n) = \ldots = 1 - f((n - 1) \cdot N/n)$$

So, the interval $[0, 1]$ is divided into n equal subintervals. Therefore, we have $f(N/n) = 1/n$, $f(2N/n) = 2/n$, and, in general, $f(a) = a/N$ for all values a of the type $(m/n) \cdot N$, i.e., for all values a for which a/N is a rational number.

When the ratio a/N is not rational, we can, for every n, take the integer part m of the value $a \cdot n/N$ and get $(m/n) \cdot N \leq a < ((m + 1)/n) \cdot N$. Since a natural re-scaling preserves monotonicity, we have

$$f((m/n) \cdot N) \leq f(a) \leq f(((m + 1)/n) \cdot N).$$

We already know the values $f((m/n) \cdot N)$, so we conclude that $m/n \leq f(a) \leq (m+1)/n$. When n tends to infinity, we have $m/n \to a/N$ and $1/n \to 0$, so both left- and right-side of the above triple inequality tend to the same limit a/N. Thus, the intermediate value tends to the same limit, i.e., indeed $f(a) = a/N$. The proposition is proven.

Discussion. So, we get a linear transformation. For example, if we start with a scale from 0 to 10, then, to get to the values from 0 to 1, we just divide all the degrees by 10, so that 1 becomes 0.1, 2 becomes 0.2, etc.

The notions of a membership function and a fuzzy set. According to the above, to describe a natural-language property, we need to have a function $m(x)$ that assigns, to each value x of the corresponding quantity, a value $m(x)$ from the interval $[0, 1]$.

For "crisp" (= well-defined) sets this is exactly what we described earlier as a membership function. For such sets, the function $m(x)$ always takes the values 0 or 1. It is natural to use the same term to describe functions whose values may be sometimes intermediate between 0 and 1. To emphasize the relation with crisp sets – and to emphasize the difference between this concept and the concept of a crisp set – membership functions are also called *fuzzy sets.*

Empirical fact. In most applications, the most effective are piece-wise linear membership functions. This include *triangular* membership functions, i.e., functions that start with 0, then linearly increase from 0 to 1, the linearly decrease from 1 to 0, and then stay at 0. Another example is *trapezoid* functions, i.e., functions for which, after increasing to 1, the function stays at 1 for some time, and only then starts decreasing to 0.

The notion of normalization. At first glance, what we described may sound like it is all we need. However, there are some subtle issues here – issues that are often skipped in popular descriptions of fuzzy techniques, but that are important in many applications.

To emphasize this importance, let us start with an example. Suppose that a person – who has never stayed in a fancy hotel before – stays there for the first time. Everything there is better than in his previous stays, so it is natural to expect that he/she will assign the highest grade 10 to this stay – which results in the membership degree 1. So, in this case, the largest value of his/her membership function $m(x)$ describing the quality of different hotels x is 1. This person understands that there could potentially be even better hotels – where he eats caviar for breakfast and famous signers sign lullabies in his/her hotel room to help him/her fall asleep, but he makes a realistic estimate.

On the other hand, we may have a picky customer who know that theoretically, there can be better hotels – although he/she never stayed in such better hotels and is not even sure that better hotels exist. Such customer will mark his/her degree of confidence that this hotel is good by, e.g., 7 – which corresponds to the degree of confidence 0.7 – and he/she will mark other hotels even lower. For this person, the largest value of his/her membership function $m(x)$ describing the quality of different hotels x is 0.7.

These two customers may have exactly the same opinion about all the hotels in which they stayed, but they mark them differently. This may sound like a hypothetical situation, but we had a colleague, originally from a very good India university, with whom we had a similar different in grading students. We usually give the top grade of A (100%) to a student who performs very well, but that colleague, at first, only gave the best students 60% – as was a tradition at his university. We both understood the strengths and weaknesses of all the students, the difference was in grading.

To decrease the effect of such a difference, Zadeh proposed to use the same normalization idea as before – namely, to divide all the degrees $m(x)$ by the largest of such degrees. For the resulting membership function, the largest value of $m(x)$ is always equal to 1. Zadeh called such fuzzy sets *normalized*.

3 Negation operations, "and"- and "or"-operations and hedges: a brief reminder

Need for negation operations. Suppose that we have a degree of confidence $m(x)$ that x is small. This will help us formalize the expert rules describing what to do when x is small. But what if we also have an expert rule describing what to do when x is not small? To describe such a rule in precise terms, we need to know the degree to which x is not small. How can we find such a degree?

When we are absolutely sure that x is small, i.e., if $m(x) = 1$, then clearly we do not have any degree of confidence that x is not small, so the corresponding degree should be 0. We can describe this by saying that $f_\neg(1) = 0$. Similarly, if we are absolutely sure that x is not small, i.e., if $m(x) = 0$, then clearly our degree of confidence that x is small should be 1. We can describe this by saying that $f_\neg(0) = 1$ But what can we say about the intermediate values $m(x) \in (0,1)$?

In this case, we can use arguments similar to the ones we had in the previous section, the only difference is that monotonicity now goes in a different direction: the more we are confident that x satisfied the given property, the less confident we are that it does not satisfy this property.

Definition 2. *We say that a mapping* $f_\neg : [0,1] \mapsto [0,1]$ *for which* $f(0) = 1$ *and* $f(1) = 0$ *is a* negation operation *if it has the following two properties:*

- *if* $a < b$*, then* $f_\neg(a) > f_\neg(b)$*, and*
- *if* $b - a = c - b$*, then* $f_\neg(b) - f_\neg(a) = f_\neg(c) - f_\neg(b)$.

Proposition 2. *There is only one negation operation* $f_\neg(a) = 1 - a$.

Proof is similar to the proof of Proposition 1.

Need for "and"-operations. Similarly to the case of negation, many expert rules use two or more conditions. For example, the driving example with which we started this paper is a particular case of a more general rule: "if a car in front of you is close and start decelerating a little bit, brake a little bit."

We can elicit, from the experts, the values of the membership function corresponding to "close" and the values of the membership function corresponding to degrees of deceleration. If we use N values for each of the quantities, then, to solicit these two membership functions, we need to ask $2N$ questions. But what we need for this rule us to get the degrees of confidence, for all possible values x and y, of the statement "x is close and y means decelerating a little bit". With N possible values x and N possible values y, we need to ask N^2 questions to the expert. If we have three conditions, we will need to ask N^3 questions, etc. Already for $N = 10$ this becomes not very realistic.

Since we cannot directly elicit the degree of confidence in all such complex "and"-statements of the type $A \,\&\, B$, we need to estimate these degrees based on information that we have, i.e., based on the degrees of confidence a and b in statements A and B. In other words, we need a function $f_\&(a, b)$ that transforms our degrees of confidence a and b in statements A and B into an estimate $f_\&(a, b)$ or the expert's degree of confidence in the composite statement $A \,\&\, B$. Such a function is known as an *"and"-operation*, or, for historical reasons, a *t-norm*.

A usual definition of an "and"-operation include some reasonable properties. For example, small changes in a and b should lead to small changes in $f_\&(a, b)$ – i.e., in precise terms, the function $f_\&(a, b)$ must be *continuous*.

Also, since $A \,\&\, B$ means the same as $B \,\&\, A$, our estimates for the degrees of confidence should be the same for both formulas, i.e., we should have $f_\&(a, b) = f_\&(b, a)$. In mathematical terms, this means that the operation $f_\&(a, b)$ must be *commutative*.

Similarly, since $A \,\&\, (B \,\&\, C)$ means the same as $(A \,\&\, B) \,\&\, C$, our estimates for the degrees of confidence should be the same for both formulas, i.e., we should have $f_\&(a, f_\&(b, c)) = f_\&(f_\&(a, b), c)$. In mathematical terms, this means that the operation $f_\&(a, b)$ must be *associative*.

Need for "or"-operations. Similarly, we need a function $f_\vee(a, b)$ that transforms our degrees of confidence a and b in statements A and B into an estimate $f_\vee(a, b)$ or the expert's degree of confidence in the composite statement $A \vee B$. Such a function is known as an *"or"-operation*, or, for historical reasons, a *t-conorm*.

Usually, "or"-operations are also assumed to be continuous, commutative, and associative.

Empirical fact. In most cases, the most effective are either piece-wise linear "and"- and "or"-operations, such as

$$f_\&(a, b) = \min(a, b), \quad f_\&(a, b) = \max(a + b - 1, 0),$$

$$f_\vee(a, b) = \max(a, b), \quad f_\vee(a, b) = \min(a + b, 1),$$

or operations $f_\&(a, b) = a \cdot b$ and $f_\vee(a, b) = a + b - a \cdot b$.

Need for hedges. Often, in describing their opinions, experts use *hedges* like "very", "somewhat", etc. An expert may say that the value is small, or that it very small, or that is somewhat small, etc. To describe these opinions in precise terms, we need to find, for each value x of the corresponding quantity, not only the degree that this value is small, but also the degree to which it is very small, somewhat small, etc. We cannot do this by asking the expert about all such statements: there are many possible hedges, and, similarly to the case of "and"-operations, the resulting number of questions becomes unrealistically large.

Since we cannot elicit all these degree directly from the expert, a natural idea is to estimate these degrees based on the information that we have – i.e., in this case, based on the degree of confidence that x is small. In other words, to describe each hedge h like "very", we need to have a function $h(m)$ that transform the degree m – to which some property is satisfied – into an estimate $h(m)$ of the degree to which this property is h satisfied. For example, the function $h(m)$ corresponding to "very" transforms the degree m to which a given value is small to an estimate $h(m)$ for the degree to which this value is very small.

Usually, it is assumed that small changes in m should lead to small changes in $h(m)$ – i.e., in precise terms, that the function $h(m)$ must be *continuous*.

Empirical fact. The most effective hedges are of the form $h(m) = m^a$ for some a. For example, "very" corresponds to $h(m) = m^2$ and "somewhat" corresponds to $h(m) = \sqrt{m} = m^{0.5}$.

4 Normalization-invariance provides a theoretical explanation for empirical successes

What we do in this section. In this section, we show that the above empirical successes can be explained by the following natural requirement: that since membership functions are defined modulo a normalization, the results of applying different operations should not change when we apply this normalization. We will explain this first on the example of hedges, then on the example of membership functions and finally, on the example of "and"-operations.

4.1 Case of hedges

Analysis of the problem. Suppose that we know that a quantity x lies between 0 and 1, and for each value x between 0 and 1, we have a degree of confidence $m(x)$ that this value is significant. Clearly, we should have $m(0) = 0$, and the function $m(x)$ should be increasing with x. Since this membership function is increasing, its largest value $m(1)$ is attained when $x = 1$. Since we are considering normalized membership functions, we should have $m(1) = 1$.

Suppose now that we learned an additional information: that the value x is bounded by some number $x_0 < 1$. In this situation, we need a new membership function $m'(x)$ that is limited to the values $x \in [0, x_0]$. For these values, the expert's degrees of confidence are the same as before, but since we only consider normalized membership functions, we now need to normalize the resulting function. The new largest value is now $m_0 \stackrel{\text{def}}{=} m(x_0)$, so we get a new membership function $m'(x) = m(x)/m_0$.

Supposed now that we are also interested in the degree to which x is very significant. For the original membership function defined on the interval $[0, 1]$, these degrees are equal to $h(m(x))$, where $h(m)$ is the hedge function corresponding to "very". One way to get a similar hedged membership function corresponding to values $x \in [0, x_0]$ is to normalize the function $h(m(x))$, i.e., to divide all its values by its largest value $h(m(x_0)) = h(m)$ on the interval $[0, x_0]$, and get $h(m(x))/h(m_0)$.

There is also another way to describe the degree to which x is very significant when x is limited to the interval $[0, x_0]$. Namely, we start with the membership function $m'(x) = m(x)/m_0$ and apply hedging to its values, resulting in $h(m'(x)) = h(m(x)/m_0)$.

These are two different way to estimate the same degrees, so these two estimates should be the same: $h(m(x))/h(m_0) = h(m(x)/m_0)$. This must be true for all x, so we must have $h(m)/h(m_0) = h(m/m_0)$ for all m and m_0.

Similar arguments can be repeated for all other hedges. Thus, we arrive at the following definition.

Definition 3. *We say that a continuous function $h(m)$ is* normalization-invariant *if $h(m)/h(m_0) = h(m/m_0)$ for all $m \le m_0$.*

Proposition 3. *For every continuous function $h(m)$ for which $h(0) = 0$, the following two conditions are equivalent to each other:*

- *the function $h(m)$ is normalization-invariant, and*
- *the function $h(m)$ has the form $h(m) = m^a$ for some $a > 0$.*

Proof. It is easy to check that all the functions $h(m) = m^a$ are normalization-invariant. Vice versa, suppose the function $h(m)$ is normalization-invariant, i.e., satisfies the above equality. If we denote m/m_0 by z, so that $m = m_0 \cdot z$, then this equality takes the form $h(m_0 \cdot z)/h(m_0) = h(z)$, i.e., equivalently, the form $h(m_0 \cdot z) = h(m_0) \cdot h(z)$. It is known (see, e.g., [1]) that every continuous function that satisfies this equality has the desired form $h(m) = m^a$. The proposition is proven.

Conclusion. So, for hedge functions, normalization-invariance indeed explains why the functions m^a are so effective: because they are consistent – two estimate of the same degree are the same – and thus, they better describe the intuitive meaning of these hedges.

4.2 Case of membership functions

Analysis of the problem. Let us now consider the same problem as before, but with a different twist. We are again interested in describing the degree $m_0(x)$ to which a positive value x is significant. In the hedges case, we assumed that possible values x form the interval $[0,1]$. However, the numerical value of a quantity depends on the choice of a measuring unit. For example, what was 1 meter becomes 100 centimeters. In general, if we change the measuring unit to a one that is λ times smaller, then all numerical values are multiplied by λ: $x \mapsto x' = \lambda \cdot x$. In particular, the interval of possible values of x' is now the interval $[0,\lambda]$. So, each new value x' corresponds to the old value $x = x'/\lambda$. Thus, the degree $m'(x')$ that the value x' in the new units is significant is simply the degree $m'(x') = m_0(x'/\lambda)$ to which the value $x = x'/\lambda$ in the old units is significant.

So, it makes sense to conclude that when we consider the property "being significant" on any interval $[0,\lambda]$, then we should use the membership function $m'(x') = m_0(x'/\lambda)$.

On the other hand, as we have mentioned in the hedges case, if we restrict ourselves to the interval $[0,x_0]$, then we get the membership function $m_0(x')/m_0(x_0)$. In particular, for $\lambda = x_0$, we get $m_0(x')/m_0(\lambda)$.

We have two expressions for the same degree of confidence, so these two expressions must be equal, i.e., we must have $m_0(x'/\lambda) = m_0(x')/m_0(\lambda)$. Thus, we arrive at the following definition.

Definition 4. *We say that a continuous function $m_0(x)$ is* normalization-invariant *if $m(x'/\lambda) = m(x')/m(\lambda)$ for all x' and λ.*

Proposition 4. *For every continuous function $m_0(x)$ for which $m_0(0) = 0$, the following two conditions are equivalent to each other:*

- *the function $m_0(x)$ is normalization-invariant, and*
- *the function $m_0(x)$ has the form $m_0(x) = x^a$ for some $a > 0$.*

Proof: this is, in effect, the same result as Proposition 3 that we have already proven. Notations are different, meaning is different, but from the mathematical viewpoint, this is exactly the same result.

Discussion. So, the basic membership function $m_0(x)$ – defined on the interval $[0,1]$ – should have the form $m_0(x) = x^a$, and a general membership function located on an interval $[0,\lambda]$ should have the form $m_0(x/\lambda)$, i.e., the form $m(x) = C \cdot x^a$,w here we denoted $C = \lambda^{-a}$.

What about shifts. The numerical value of a physical quantity also depends on the starting point – e.g., depending on the starting point, we get different years in the usual Western calendar and in calendars of different religions: Muslim, Buddhist, Jewish, etc. When we change a starting point to a new one which is x_0 units lower, then we get a shift: the value x_0 is added to all numerical values. Any interval $[A,B]$ can be reduced to an interval of the type $[0,\lambda]$ if we replace the original values x with new values $x' = x - A$. We have shown that on such intervals, the membership function has to have a form $m'(x') = C \cdot (x')^a$. So, the degree of confidence $m(x)$ when x is expressed in the old units can be obtained as the membership degree of the same quantity when expressed in the new units $x' = x - A$: $m(x) = m'(x') = C \cdot (x')^a = C \cdot (x - A)^a$.

What about negations. The above result deals with membership functions defined on an interval. On other intervals, we may have different expressions. So, from the viewpoint of normalization-invariance, a reasonable membership function $m(x)$ on an interval $[A, B]$ – e.g., the membership function describing significance – should – at least piece-wise – consist of fragments of the type $C \cdot (x - A)^a$. And, correspondingly, this should also be true for its negation $m'(x) = 1 - m(x)$ for which $m'(B) = 0$. In other words, the membership function corresponding to negation should also have to have the similar form $m'(x) = C' \cdot (B - x)^a$. Interestingly, these two conditions lead to linearity:

Definition 5. *We say that a function $m(x)$ defined on an interval $[A, B]$ for which $m(A) = 0$ and $m(B) = 1$ is* strongly normalization-invariant *if $m(x)$ has the form $m(x) = C \cdot (x - A)^a$ and the function $1 - m(x)$ has the form $1 - m(x) = C' \cdot (B - x)^a$.*

Proposition 5. *For every continuous function $m(x)$ on an interval $[A, B]$, the following two conditions are equivalent to each other:*

- *the function $m(x)$ is strongly normalization-invariant, and*
- *the function $m(x)$ is linear.*

Proof. It is easy to check that a linear function $m(x)$ for which $m(A) = 0$ and $m(B) = 1$ is strongly normalization-invariant. Vice versa, assume that the function $m(x)$ is normalization-invariant. This means that for all x, we must have $C \cdot (x - A)^a + C' \cdot (B - x)^a = 1$. Since this sum is a constant, its derivative should be equal to 0 for all x. This derivative is equal to

$$C \cdot a \cdot (x - A)^{a-1} - C' \cdot a \cdot (B - x)^{a-1}.$$

When $a < 1$, then for $x = A$, the first term in this expression is infinite and the second is finite – so the sum cannot be equal to 0. When $a > 1$, then for $x = A$, the first term is 0 and the second term is non-zero, so the sum also cannot be equal to 0. So, the only cases when this equality can be satisfied is when a is neither smaller than 1 nor larger than 1 – i.e., when $a = 1$. In this case, the function $m(x)$ is linear. The proposition is proven.

Discussion. Remember that we are talking about membership functions on an interval, and there may be several such interval. So, the actual membership function should be piece-wise linear.

Conclusion. Thus, for membership functions, strong normalization-invariance indeed explains why piece-wise linear functions are so effective: because they are consistent – two estimate of the same degree are the same – and thus, they better describe the intuitive meaning.

4.3 Case of "and"- and "or"-operations

Analysis of the problem. Since it is reasonable to restrict ourselves to piece-wise linear membership functions, it makes sense to select "and"- (and "or"-)

operations so that for every two linear functions $m(x)$ and $m'(x)$, the function $f_\&(m(x), m'(x))$ (or, correspondingly, $f \vee (m(x), m'(x))$) should also be piece-wise linear.

This time, to describe the fact that small changes in a and b should lead to small changes in $f(a,b)$, we will use a stronger formalization of analyticity rather than continuity.

Definition 6. *We say that an analytical function $f(a,b)$ is* consistent with normalization invariance *if for every two piece-wise linear functions $m(x)$ and $m'(x)$, the function $f(m(x), m'(x))$ is also piece-wise linear.*

Proposition 6. *For every analytical function $f(a,b)$, the following two conditions are equivalent to each other:*

- *the function $f(a,b)$ is consistent with normalization invariance, and*
- *the function $f(a,b)$ is piece-wise linear.*

Proof. It is easy to check that every piece-wise linear function $f(a,b)$ has the desired property. Vice versa, suppose that a function $f(a,b)$ is consistent with normalization-invariance. Let us prove, by contradiction, that in this case, for each point (a_0, b_0). the corresponding Taylor expansion of the function $f(a,b)$ consists only of linear terms. We will prove this by contradiction. Let us assume that this expansion has at least one non-zero non-linear term

$$c \cdot (a - a_0)^m \cdot (b - b_0)^m.$$

Out of such terms, let us select the one which has the smallest overall degree m and, if there are several such terms, the one with the smallest n. Then, for $m(x) = a_0 + k_a \cdot x$, for sufficiently small k_a, terms with higher orders in a can be safely ignored and similarly, for $m'(x) = b_0 + k_b \cdot x$, terms corresponding to higher orders in b can also be safely ignored. So, for these membership functions, the value $f(m(x), m'(x))$ should have, in addition to linear terms and ignored terms, an expression $c \cdot k_a^m \cdot x^m \cdot k_b^n \cdot x^n = \text{const} \cdot x^{n+m}$. The resulting expression is clearly non-linear. The proposition is proven.

First conclusion. Thus, for "and"- and "or"-operations, normalization-invariance indeed explains why piece-wise linear operations are so effective: because the only ones for which the resulting membership functions are strongly normalization-invariant (i.e., consistent) – and thus, they better describe the intuitive meaning of these operations.

What about the product? We have explained the effectiveness of piece-wise linear operations, but, as we have mentioned earlier, there is another "and"-operation which is effective but not piece-wise linear: the product operation $f_\&(a,b) = a \cdot b$. Let us show that it effectiveness can also be explained by normalization-invariance.

Indeed, suppose that started with two increasing-from-0-to-1 membership functions $m(x)$ and $m'(x')$ corresponding to two properties. Then, by definition of the "and"-operation, the degree of confidence that both properties are satisfied is estimated as $f_\&(m(x), m'(x'))$. Suppose now that we get additional information about both the range of x and the range of x', reducing their upper bounds of, correspondingly, x_0 and x'_0. What will be the new degree of confidence that both properties are satisfied?

There are two ways to estimate this new degree of confidence. One way is to first normalize the membership functions, to $m(x)/m_0$ and $m'(x)/m'_0$, where we denoted $m_0 \stackrel{\text{def}}{=} m(x_0)$ and $m'_0 \stackrel{\text{def}}{=} m'(x'_0)$, and then apply the "and"-operation, resulting in $f_\&(m(x)/m_0, m'(x)/m'_0)$.

Another idea is to first apply the "and"-operation to the original membership functions, and then to normalize the resulting function $f(m(x), m'(x')$ by dividing it to the largest possible value when $x \le x_0$ and $x' \le x'_0$, i.e., by the value $f(m_0, m'_0)$.

These are two estimates for the same quantity, so it is reasonable to require that these two estimates should be equal: $f_\&(m(x)/m_0, m'(x)/m'_0) = f(m(x), m'(x))/f(m_0, m'_0)$. This should be true for all possible values $m = m(x)$ and $m' = m'(x')$. So, we arrive at the following definition.

Definition 7. *We say that a non-constant continuous commutative and associative operation $f_\&(a, b)$ is* normalization-invariant *if $f_\&(m/m_0, m'/m'_0) = f(m, m')/f(m_0, m'_0)$ for all $m \le m_0$ and $m' \le m'_0$.*

Proposition 7. *For every non-constant continuous commutative and associative operation $f_\&(a, b)$, the following two conditions are equivalent to each other:*

- *the operation $f_\&(a, b)$ is normalization-invariant, and*
- *the operation $f_\&(a, b) = a \cdot b$.*

Proof. It is easy to check that the operation $f_\&(a, b) = a \cdot b$ is normalization-invariant. Vice versa, suppose the the operation $f_\&(a, b) = a \cdot b$ is normalization-invariant, i.e., satisfies the above equality. If we denote m/m_0 by z and m'/m'_0 by z', so that $m = m_0 \cdot z$ and $m' = m'_0 \cdot z'$, then this equality takes the form $f_\&(m_0 \cdot z, m'_0 \cdot z')/f_\&(m_0, m'_0) = f_\&(z, z')$, i.e., equivalently, the form $f_\&(m_0 \cdot z, m'_0 \cdot z') = f_\&(m_0, m'_0) \cdot f_\&(z, z')$. It is known (see, e.g., [1]) that every continuous function that satisfies this equality has the form $f_\&(m, m') = m^a \cdot (m')^{a'}$ for some a and a'.

From commutativity, we conclude that $a = a'$, so $f_\&(m, m') = (m \cdot m')^a$. Associativity now means that for all m, m', and m'', we have $(m \cdot (m' \cdot m'')^a)^a = ((m \cdot m')^a \cdot m'')^a$, i.e., that $m^a \cdot (m')^{a^2} \cdot (m'')^{a^2} = m^{a^2} \cdot (m')^{a^2} \cdot (m'')^a$. These two expression must coincide for all m, m', and m''. So, the powers of m in both expressions should be the same: $a^2 = a$. This implies that either $a = 1$ or $a = 0$. For $a = 0$, we get a constant function – and we only consider non-constant functions $f_\&(a, b)$. So the only remaining option is $a = 1$, for which $f_\&(a, b) = a \cdot b$. The proposition is proven.

Second conclusion. So, for "and"-operations, normalization-invariance indeed explains why the operation $f_\&(a, b) = a \cdot b$ is so effective: because they are consistent – two estimate of the same degree are the same – and thus, they better describe the intuitive meaning of these operations.

Observation. Interestingly, to explain piece-wise linear "and"-operations, we used a combination of different properties of the same object, while to explain the product, we used a combination of properties of different objects. So maybe indeed piece-wise linear "and"-operations are more appropriate when we combine properties of the same object, while the product is more appropriate when we combine properties of different objects?

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.

It was also supported by a grant from the Hungarian National Research, Development and Innovation Office (NRDI), and by the Institute for Risk and Reliability, Leibniz Universitaet Hannover, Germany.

References

[1] J. Aczél and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, 2008.

[2] R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, Oxford University Press, New York, 2017.

[3] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[4] J. M. Mendel, *Explainable Uncertain Rule-Based Fuzzy Systems*, Springer, Cham, Switzerland, 2024.

[5] H. T. Nguyen, C. L. Walker, and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2019.

[6] V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical Principles of Fuzzy Logic*, Kluwer, Boston, Dordrecht, 1999.

[7] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.