Lecture 1: Why Explainable AI? Why Fuzzy Explainable AI? What Is Fuzzy?

1 Why Explainable AI?

What is one of the main purposes of science. One the main objectives of science is to predict future events based on the information that we have. For example, we know the temperature, wind speed, wind direction, and humidity at different locations, and, based on this information, we want to predict tomorrow's weather – e.g., tomorrow's temperature at different locations in El Paso and at different moments of time.

In general, we know the values of some quantities $x_1, \ldots, x_n$, and we want to predict the value of some quantity y. To predict this value, we need to apply some algorithm to the available data $x_1, \ldots, x_n$. Let us denote this algorithm by f. Then, the predicted value is computed as $y = f(x_1, \ldots, x_n)$.

Where does this algorithm come from?

Where does the desired prediction algorithm come from? Often, the desired algorithm comes from the theoretical analysis, but this theoretical analysis has to be on some empirical dependencies. So, the big question is: how do we find an empirical dependence $y = f(x)$?

For this purpose, we need to perform several measurements, in which we measure both the values x and the corresponding value y.

- We perform the first measurement, and get the values $x^{(1)}$ and $y^{(1)}$.
- We perform the second measurement, and get the values $x^{(2)}$ and $y^{(2)}$, etc.

Let us denote the overall number of measurements by K. Under this notation, we know the values $x^{(k)}$ and $y^{(k)}$ for $k = 1, \ldots, K$.

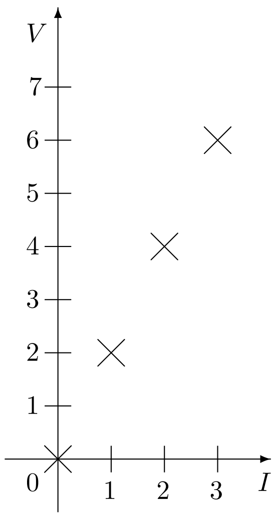
Based on this data, we need to find an algorithm $y = f(x)$ for which $y^{(k)} = f(x^{(k)})$ for all k. This algorithm is what a usual description of the scientific method calls a *hypothesis*. Then, we *test* this hypothesis – by checking that the relation $y = f(x)$ holds for future measurements, and, if the hypothesis is confirmed, we can use this algorithm.

Let us give examples.

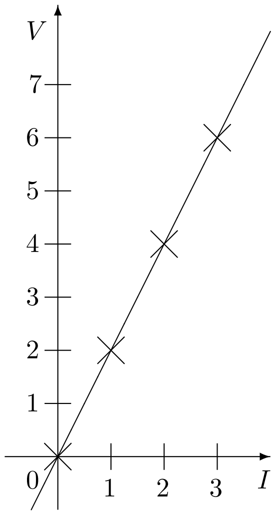
First example: Ohm's Law. One example is Ohm's Law, according to which the voltage V is proportional to the current I: $V = I \cdot R$, for some coefficient R (known as *resistance*). How did Ohm come up with this formula?

- He measured the voltage $V^{(1)}$ corresponding to no current $I^{(1)} = 0$, and came up with $V^{(1)} = 0$.
- Then, he measured the voltage $V^{(2)}$ corresponding to no current $I^{(2)} = 1$, and came up with $V^{(2)} = 2$.
- After that, he measured the voltage $V^{(3)}$ corresponding to no current $I^{(3)} = 2$, and came up with $V^{(3)} = 4$.
- Finally, he measured the voltage $V^{(4)}$ corresponding to no current $I^{(4)} = 3$, and came up with $V^{(4)} = 6$.

He then plotted the results:



Ohm guessed that this data can be described by a the simplest possible formula – a linear formula, in this case by the formula $V = 2I$:

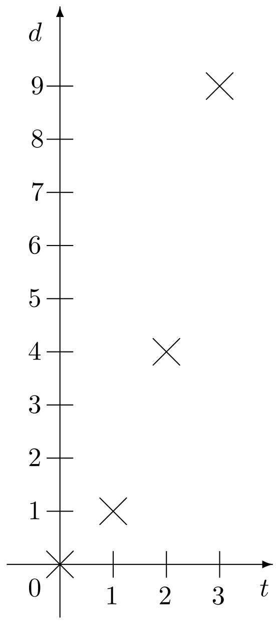


He then performed additional measurements that confirmed his hypothesis, and his formula became what we now call Ohm's Law.

Second example: Galileo's formula. How did Galileo come up with a formula according to which the distance d traveled by a free falling body depends on time as $d = c\cdot t^2$ for some constant c? He measured the distance d for different times.

- He measured the distance $d^{(1)}$ corresponding to no time $t^{(1)} = 0$, and came up with $d^{(1)} = 0$.
- Then, he measured the distance $d^{(2)}$ after one second $t^{(2)} = 1$, and came up with $d^{(2)} = 1$ unit.
- After that, he measured the distance $d^{(3)}$ after two seconds $t^{(3)} = 2$, and came up with $d^{(3)} = 4$.
- Finally, he measured the distance $d^{(4)}$ after three seconds $t^{(4)} = 3$, and came up with $d^{(4)} = 9$.

He then plotted the results:



He then guessed that these measurement results fit the quadratic formula $y = x^2$. He then performed other measurement, checked that this formula holds for other measurement results as well, and this became the law of physics.

How do we go from examples to the formula? In all these cases:

- we know the values $x^{(k)}$ and $y^{(k)}$ corresponding to $k = 1, \ldots, K$, and
- we want to find a function $y = f(x)$ for which $y^{(k)} = f\left(x^{(k)}\right)$ for all k.

In reality, measurements are approximate, so we only have $y^{(k)} \approx f\left(x^{(k)}\right)$, but for now, we will ignore this difference.

Originally, this task was performed by guesses, but eventually, algorithms appeared that solve this problem.

In mathematics, this problem is known as *interpolation* or *extrapolation*:

- it is called interpolation if the value x is in between some of the values $x^{(k)}$, and
- it is called extrapolation if the value x is outside the region of all the values $x^{(k)}$.

The names are different, but the algorithm is usually the same in both cases.

In computer science, this problem is known as *machine learning:*

- we have some examples $x^{(k)}$ and $y^{(k)}$, $k = 1, \ldots, K$, and
- we want to come up with an algorithm $y = f(x)$ that, given x, would predict the value y; this algorithm must be consistent with all the observations, so we must have $y = f(x)$ for which $y^{(k)} = f\left(x^{(k)}\right)$ for all k.

Successes of machine learning, especially of deep learning. Lately, machine learning – especially a complex of machine learning techniques known as *deep learning* – has been spectacularly successful. A few year ago, a deep learning algorithm learned how to play Go – a complex game for which AI could not achieve any reasonable level – so well that it easily defeated a human world champion.

Deep learning algorithms make the current autonomous vehicles very successful. They help with medical diagnostics: e.g., they diagnose a lung disease based on an X-ray much more accurately than a human radiologist. Machine learning algorithms are used by banks to decide who to give loans to – and its use drastically has decreased the banks' losses.

But there is a problem: we need to make AI explainable. It would all be great if the machine learning results were perfect, but they are not.

- Autonomous vehicles, while they are, on average, much safer than cars driven by humans – did have accidents, and two people were killed.
- A medical system for analyzing X-rays is more accurate than human radiologists, but it still sometime misdiagnoses.

And herein lies a problem:

- If a human radiologist is not 100% sure of the diagnosis, he/she can consult with colleagues: they exchange their motivations, and hopefully, come up with a more accurate diagnosis.
- If a bank refuses someone a loan based on the human analysis, the applicant can ask why his/her application was declined, and, based on the bank's explanation, argue to change the bank's decision – or at least understand what is needed to be more successful next year. For example, the bank may say that the applicant does not have enough credit history, in which case time will help.
- If a student is not happy with the grade on a test, the student can ask the instructor (or the TA) and, if possible, argue that more partial credit is due – and often, such arguments succeed.

On the other hand, many computer-based systems offer no explanation at all. So even when the answer is wrong, there is no way we can know that it is wrong o to try to get a more accurate answer. For example, the X-ray-based system errs 10% of the time. If a machine learning system of the same type is used to decide who shall be released from jail – as it often does nowadays – this means that in 10% of the cases:

- either reformed folks, who can become productive members of the society, unnecessarily remain jailed – which is not good,
- or, which is probably even worse, unreformed criminals are mistakenly released, thus endangering the population.

And when there are no explanations, it is impossible to argue against such decisions.

It is therefore extremely important to be able to provide explanations for AI-based decisions, explanations that would state, in plain natural-language words, why this decision was made. A few AI systems already provide such explanations, but as of now, most machine learning-based system do not provide such an explanation.

2 Why Fuzzy Techniques Seem a Reasonable Approach for Explainable AI

We need explainable AI: reminder. We want to translate numerical results – produced by AI algorithms – into natural language. In other words, we want to have some relation between:

- numerical recommendations and
- natural-language explanations.

In search for the desired translation, a natural idea is to look for known relations between numerical recommendations and natural-language explanations. And such a relation is indeed well-known: it is so-called *fuzzy techniques.*

Before we start analyzing them in detail, let us first briefly overview why these techniques were invented in the first place.

A brief history of fuzzy techniques. Fuzzy techniques first appeared in 1960s, when the appearance of first easy-to-use and not-very-expensive computers led to a boom in computer-based automated control systems. One of the main specialists in optimal control was Professor Lotfi Zadeh (pronounced LOT-fih Zah-DEH), a co-author of then the most popular and most widely used textbook on optimal control. For example, control folks are familiar with the term z-transform; this z stands for Zadeh – Zadeh did not invent this notion, but he made the corresponding mathematical technique widely used in control applications.

Professor Zadeh was born in Baku, Azerbaijan. His father was an Iranian consul.

- Lotfi Zadeh got his Bachelor's degree in Iran.
- Then he moved to the US, where he got his graduate degrees.
- Upon receiving his PhD, he became a professor – first at Columbia University, then at the University of California-Berkeley.

His main interest at that time was in improving control systems – since in many cases, their performance was not as good as the performance of human controllers. For example, a chemical plant controlled by an automatic system was usually not as productive as when controlled by a skilled engineer.

At first, he tried to bridge the gap between human and automatic controllers by *optimizing* controllers – making sure that their control strategies provided the best possible values of the corresponding objective function. However, even when a supposedly optimal control was implemented, the automatic system were still not performing as well as human-controlled ones.

Of course, if the automatic control was based on the fully adequate description of a controlled system, this gap would not be possible: by definition of a maximum, nothing is larger than the maximum value, so there is no way to control better than the optimal control. So, the fact that the automatic controllers did not perform as well as human controllers means that:

- the automatic systems operated based on a not-fully-adequate description of the system, while
- human controllers – who achieved a better performance – clearly had some more adequate description in mind.

In other words, some expert knowledge about the controlled systems and control strategies that was *not* implemented in the computer-based system. So, Zadeh concluded that we need to extract this knowledge and use it in automatic control.

The reason why this knowledge was not used is *not* that the experts do not want to share it. Such cases are rare: sometimes, a company protects its expertise – but even in this case, the most skilled engineers are willing to share their knowledge with other engineers from the same company.

The reason why this knowledge was not implemented in the computer-based computer systems was much more fundamental:

- computers only understand precise language of numbers, but
- a significant part of expert knowledge is described by imprecise ("fuzzy") words from natural language.

Let us use self-driving cars as an example. In this case, we are not talking about a difficult skill of controlling a chemical plant or analyzing X-ray images, we are talking about a skill that the vast majority of US folks have: ability to drive a car. So why not ask people how they drive, and then implement this strategy in an automatic system?

Here is an example:

- You are driving on a freeway with a speed of 60 miles per hour.
- The car in front of you – at 60 feet distance – is traveling at the exact same maximum speed.
- Then the car in front of you brakes a little bit, reducing its speed to 55 miles per hour.

Your reaction: slow down a little bit.

The problem is that a computer does not understand what "a little but" means, it needs to know:

- for how many milliseconds and
- with what exactly pressure (in pounds per square inch)

we need to press the brakes. Hardly any driver can describe his/her driving in these terms.

So, we need to translate expert rules that use imprecise ("fuzzy") words from natural language into a precise numerical control strategy. To take care of this need, Zadeh developed a methodology for such a translation, a methodology that he called *fuzzy*.

In the 1980s-1990s, this methodology had many applications:

- in industrial plants,
- in control of trains, cars, and elevators – it is still used by many automotive companies when designing automatic transmissions,
- in consumer devices such as video cameras, rice cookers, washing machines, etc.

This methodology provides the relation between numerical control and natural-language rules – exactly the relation that we need to implement explainable AI. So, it is reasonable to try to use this methodology when designing explainable AI systems.

3 What Is Fuzzy Methodology

Case study: description. To explain Zadeh's ideas, let us start with a toy example of a simple thermostat that is controlled by turning a knob:

- if we turn it to the right, it warms the room up – and the more we turn it, the more the room warms up;
- if we turn it to the left, it cools the room down – and the more we turn it, the more the room cools down.

We would like to maintain an ideal temperature T_0 – e.g., 25 C. When the temperature T is different, i.e., when the difference $\Delta T = T - T_0$ is different from 0, we need to apply some control, i.e., turn the knob some degrees from its original position. Let us denote the angle between the resulting position of the knob and its original position by u. Then:

- if $u > 0$, we turn the knob u degrees to the right – i.e., in the direction of heating, and
- if $u < 0$, we turn the knob u degrees to the left – i.e., in the direction of cooling.

Commonsense rules. Let us describe natural commonsense rules.

- If the difference ΔT is negligible, then we should not heat or cool the room. In other words, in this case, the angle u should also be negligible.
- If the room is slightly warmer than desired – i.e., if the difference ΔT is small positive – then we should cool the room a little bit. Cooling corresponds to negative angles u, and cooling a little bit corresponds to small values u. So, in this case, the control u should be small negative.
- If the room is slightly colder than desired – i.e., if the difference ΔT is small negative – then we should heat the room a little bit. Heating corresponds to positive angles u, and heating a little bit corresponds to small values u. So, in this case, the control u should be small positive.

We can have many more rules, with ΔT medium, large, etc., but for simplicity, let us consider only these three rules.

Thus, we have the following three rules:

- If ΔT is negligible, then u should be negligible.
- If ΔT is small positive, then u should be small negative.
- If ΔT is small negative, then u should be small positive.

When is a control value reasonable? Based on these rules, when, for a given different ΔT, is u a reasonable control – we will denote it by $R(\Delta T, u)$? When one of these rules has been applied, i.e., when:

(ΔT is negligible and u is negligible) or
(ΔT is small positive and u is small negative) or
(ΔT is small negative and u is small positive).

How to describe this in precise terms? To describe the above statement in precise terms, we need first to understand how to describes statements like "ΔT is negligible" in precise terms.

For some values of ΔT, this is easy. For example:

- the difference $\Delta T = 0$ is clearly definitely negligible, while
- the difference $\Delta T = 5$ – meaning that instead of 25 C, the room temperature is 30 C – this difference is clearly not negligible.

However, when it comes to values in between 0 and 5, such as 1 degree, 2 degrees, 3 degrees, the situation is not so clear. If you ask a person whether he/she considers this difference to be negligible, the answer will probably be not "absolutely negligible" or "absolutely not negligible", but rather something like "somewhat negligible" or "to some extent negligible". How can describe such answers in precise terms?

A natural idea is to ask a person to mark, on some scale – e.g., on a scale from 0 to 5 – to what extent they consider this difference to be negligible.

This sound weird, but this is exactly what students do when they evaluate the instructor. The students do not just have an option of answering "yes" and "no" to questions on how prepared the instructor is for the class – that would correspond to 0 and 5 – they also have an option to select one of the intermediate values.

It does not have to be scale from 0 to 5, we can use a scale from 0 to 10, or any other scale. To be able to compare different results, it makes sense to divide the resulting mark by the largest value. For example, if an expert marked 3 on a scale from 0 to 5, we take a degree $3/5 = 0.6$. These "scaled" values are known as *degrees of confidence.*

If there are several experts, then there is another natural way to assign a degree of confidence to each statement like "the temperature difference of 2 degrees ($\Delta T = 2$) is negligible": polling. If out of 10 folks, 6 think that this value is negligible, then we take the ratio $6/10$ as the desired degree.

In both cases:

- if we are absolutely sure that the statement is true, then its degree is 1, and
- if we are absolutely sure that the statement is false, then its degree is 0.

Comment. We may have a more complex situation, when each expert marks his/her degree of confidence in a statement on a scale. This is an important case, but in these introductory lectures, we will not study this case.

Membership functions and fuzzy sets. For each property P (in our case, the property "to be negligible") and for each possible value of the quantity x (in our case, of the quantity ΔT), we can ask an expert (or experts) and get the degree of confidence that the value x satisfies the property P. This value is usually denoted by $\mu_P(x)$.

The function that assigns, to each value x, this degree $\mu_P(x)$ is known as the *membership function* or, alternatively, a *fuzzy set.*

Comment. μ (pronounced "mu") is the Greek analogue of the letter m – the first letter of the word *membership.*

Need for interpolation/extrapolation. There are infinitely many possible values x – e.g., we can have different in temperatures equal to 1.5 or 0.8 degrees – and we cannot ask infinitely many questions to the expert. So, once we asked finitely many questions and got finitely many values $\mu\left(x^{(1)}\right)$, $\mu\left(x^{(2)}\right)$, $\ldots$, to get the values $\mu(x)$ for all other x, we need to use the procedure that we mentioned earlier – interpolation/extrapolation.

There are many different interpolation/extrapolation techniques. The simplest is when we use linear functions to describe the missing values. This is called *linear interpolation.* Let us describe how it works.

Linear interpolation: formulation of the problem.

- We know that the dependence of a quantity y on a quantity x is linear, i.e., that
$$y(x) = a \cdot x + b \tag{1}$$
for some values a and b.

- We know two cases in which we measured x and y:
  - we know the values x_1 and y_1 for which
  $$a \cdot x_1 + b = y_1, \tag{2}$$
  and
  - we know the values x_2 and y_2 for which
  $$a \cdot x_2 + b = y_2. \tag{3}$$

Based on this information, we need to find the formula for $y(x)$.

Linear interpolation: solution. Subtracting (2) from (3), we conclude that
$$a \cdot (x_2 - x_1) = y_2 - y_1. \tag{4}$$
Dividing both sides by the difference $x_2 - x_1$, we conclude that
$$a = \frac{y_2 - y_1}{x_2 - x_1}. \tag{5}$$

Subtracting (2) from (1), we conclude that
$$a \cdot (x - x_1) = y - y_1, \tag{6}$$
therefore
$$y = y_1 + a \cdot (x - x_1). \tag{7}$$
Substituting the expression (5) for a into this formula, we conclude that
$$y = y_1 + (x - x_1) \cdot \frac{y_2 - y_1}{x_2 - x_1}. \tag{8}$$
This formula is known as *linear interpolation*.

Linear interpolation: example 1. Suppose that we know that:

- for $x_1 = 0$, we have $y_1 = 1$, and
- for $x_2 = 5$, we have $y_2 = 0$.

In this case, we have
$$y = 1 + (x - 0) \cdot \frac{0 - 1}{5 - 0} = 1 + x \cdot \frac{-1}{5} = 1 - \frac{x}{5}. \tag{9}$$

Linear interpolation: example 2. Suppose that we know that:

- for $x_1 = -5$, we have $y_1 = 0$, and
- for $x_2 = 0$, we have $y_2 = 1$.

In this case, we have

$$y = 0 + (x - (-5)) \cdot \frac{1-0}{0-(-5)} = (x+5) \cdot \frac{1}{5} = \frac{x+5}{5} = 1 + \frac{x}{5}. \qquad (10)$$

Triangular membership function. Let us use linear interpolation to find the membership function $\mu_N(x)$ describing what is negligible. Let us consider the simplest situation, when the only thing we know is that:

- the value $x = 0$ is definitely negligible, i.e., $\mu_N(0) = 1$;
- the value $x = 5$ is definitely *not* negligible, i.e., $\mu_N(5) = 0$; and, similarly,
- the value $x = -5$ is definitely *not* negligible, i.e., $\mu_N(-5) = 0$.

Common sense tell us that if a difference is not negligible, then larger differences in temperature are not negligible too, so $\mu_N(x) = 0$ for all $x \geq 5$ and for all $x \leq -5$.

To find the values $\mu_N(x)$ for x between 0 and 5, we can use linear interpolation. In this case:

- for $x_1 = 0$, we have $y_1 = 1$, and
- for $x_2 = 5$, we have $y_2 = 0$.

This is exactly the above Example 1, so we conclude that for $x \in [0, 5]$, we have $\mu_N(x) = 1 - x/5$.

To find the values $\mu_N(x)$ for x between -5 and 0, we can also use linear interpolation. In this case:

- for $x_1 = -5$, we have $y_1 = 0$, and
- for $x_2 = 0$, we have $y_2 = 1$.

This is exactly above Example 2, so we conclude that for $x \in [-5, 0]$, we have $\mu_N(x) = 1 + x/5$.

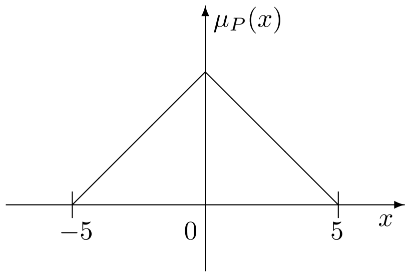
Thus, we have the following expression for the membership function $\mu_N(x)$:

- for $x \leq -5$, we have $\mu_N(x) = 0$;
- for $-5 \leq x \leq 0$, we have $\mu_N(x) = 1 + x/5$;
- for $0 \leq x \leq 5$, we have $\mu_N(x) = 1 - x/5$; and
- for $x \geq 5$, we have $\mu_N(x) = 0$.

Comment. For borderline values $x = -5$, $x = 0$, and $x = 5$, we can use two different formulas, but this does not create any problem, since both formulas lead to the same result. For example, for $x = -5$, we have:

- we have $\mu_N(-5) = 0$ according to the formula from the first bullet, and
- we get the exact same value $\mu_N(-5) = 1 + (-5)/5 = 1 + (-1) = 0$ if we us the formula from the second bullet.

The membership function $\mu_N(x)$ has the following form:



Comment. The graph of this function has the shape of the triangle, so such functions are known as *triangular* membership functions.

Membership functions for "small positive" and "small negative". For small positive, we can assume that:

- the value $x = 0$ is *not* small positive,
- the value $x = 5$ is definitely small positive, and
- the value $x = 10$ is again *not* small positive.

So, we have $\mu_{SP}(0) = 0$, $\mu_{SP}(5) = 1$, and $\mu_{SP}(10) = 0$. If we use linear interpolation to find the values $\mu_{SP}(x)$ for intermediate values x, we get the following expression:

- for $x \le 0$, we have $\mu_{SP}(x) = 0$;
- for $0 \le x \le 5$, we have $\mu_{SP}(x) = 1 + (x - 5)/5 = 1 + x/5 - 1 = x/5$;
- for $5 \le x \le 10$, we have $\mu_{SP}(x) = 1 - (x - 5)/5 = 1 - x/5 + 1 = 2 - x/5$; and
- for $x \ge 10$, we have $\mu_{SP}(x) = 0$.

Similarly, for small negative, we can assume that:

- the value $x = 0$ is *not* small negative,
- the value $x = -5$ is definitely small negative, and

- the value $x = -10$ is again *not* small negative.

So, we have $\mu_{SN}(-10) = 0$, $\mu_{SN}(-5) = 1$, and $\mu_{SN}(0) = 0$. If we use linear interpolation to find the values $\mu_{SN}(x)$ for intermediate values x, we get the following expression:

- for $x \le -10$, we have $\mu_{SN}(x) = 0$;
- for $-10 \le x \le -5$, we have $\mu_{SN}(x) = 1+(x+5)/5 = 1+x/5+1 = 2+x/5$;
- for $-5 \le x \le 0$, we have $\mu_{SN}(x) = 1-(x+5)/5 = 1-x/5-1 = -x/5$; and
- for $x \ge 0$, we have $\mu_{SN}(x) = 0$.

Need for "and"- and "or"-operations. So far, we learned how to describe statements like "ΔT is negligible" and "u is negligible". However, what we need is to find our degree of confidence in more complex statements, e.g., in a statement

"ΔT is negligible *and* u is negligible".

The situation is easy if both statements are either absolutely true or absolutely false: then, we can simply use the usual truth table for "and", acroding to which $0\,\&0 = 0\,\&\,1 = 1\,\&\,1 = 0$ and $1\,\&\,1 = 1$ – remember that here:

- 0 means "false", and
- 1 means "true".

But what shall we do in all other cases?

In principle, we can ask the expert about all possible pairs, but this largely increases the number of questions that we need to ask the expert, and if we have not one but many inputs, this becomes infeasible.

Since we cannot directly elicit the degrees of confidence of such complex statements $A\,\&\,B$ from the expert, we need to to estimate this degree based on the known degrees $a = d(A)$ and $b = d(B)$ of statements A and B. In other words, we need to have an algorithm that:

- given the degrees of confidence a and b of the statements A and B,
- returns an estimate of the degree of confidence in the composite statement $A\,\&\,B$.

We will denote such algorithm by $f_\&(a,b)$. Such algorithms are known as *"and"-operations.*

Similarly, we need to have an algorithm that:

- given the degrees of confidence a and b of the statements A and B,
- returns an estimate of the degree of confidence in the composite statement $A \vee B$.

We will denote such algorithm by $f_\vee(a,b)$. Such algorithms are known as *"or"-operations.*

Comment. For historical reasons:

- "and"-operations are also called *t-norms*, while
- "or"-operations are also called *t-conorms.*

Here, t comes from *triangle.*

Natural properties of "and"-operations.

- The statement $A \,\&\, B$ means the same as $B \,\&\, A$. It is therefore reasonable to require that our estimates for the degree of confidence of these two composite statements be equal, i.e., that $f_\&(a,b) = f_\&(b,a)$. This property is known as *commutativity.*
- Similarly, if we have three statements, their "and"-combination does not depend on the order in which we combine them: $A \,\&\, (B \,\&\, C)$ means the same as $(A \,\&\, B) \,\&\, C$. Thus, we must have $f_\&(a, f_\&(b,c)) = f_\&(f_\&(a,b), c)$. This property is known as *associativity.*
- If our degrees of confidence in A and/or B increases, our degree of confidence in $A \,\&\, B$ should also increase: if $a \le a'$ and $b \le b'$, then we must have $f_\&(a,b) \le f_\&(a',b')$. This property is known as *monotonicity.*
- Finally, for the case when both statements A and B are either absolutely true or absolutely false, we must have the same value as in the usual truth table: $f_\&(0,0) = f_\&(0,1) = f_\&(1,0) = 0$ and $f_\&(1,1) = 1$.

There exist many "and"-operations that satisfy all these properties. The most widely used are $\min(a,b)$ and $a \cdot b$.

Comment. The "and"-operation $f_\&(a,b) = a \cdot b$ is known as *algebraic product.* The reason for this name is as follows:

- in middle school and in high school, in algebra classes, we study only one type of product: the usual product of two numbers;
- however later, at the university level, we learn that there are many other types of products: dot-product, matrix product, vector product (used in physics), etc.

The adjective *algebraic* is added to emphasize that we are *not* dealing with any of these complex products, we are dealing with the simple product of two numbers – as in high school algebra.

Natural properties of "or"-operations. Similar properties hold for "or"-operations:

- The statement $A \vee B$ means the same as $B \vee A$. It is therefore reasonable to require that our estimates for the degree of confidence of these two composite statements be equal, i.e., that $f_\vee(a,b) = f_\vee(b,a)$. This property is known as *commutativity.*

- Similarly, if we have three statements, their "or"-combination does not depend on the order in which we combine them: $A \vee (B \vee C)$ means the same as $(A \vee B) \vee C$. Thus, we must have $f_\vee(a, f_\vee(b,c)) = f_\vee(f_\vee(a,b),c)$. This property is known as *associativity.*

- If our degrees of confidence in A and/or B increases, our degree of confidence in $A \vee B$ should also increase: if $a \le a'$ and $b \le b'$, then we must have $f_\vee(a,b) \le f_\vee(a',b')$. This property is known as *monotonicity.*

- Finally, for the case when both statements A and B are either absolutely true or absolutely false, we must have the same value as in the usual truth table: $f_\vee(0,0) = 0$, $f_\vee(0,1) = f_\vee(1,0) = f_\&(1,1) = 1$.

Note that the last condition is the only one which is different from the properties of an "and"-operation.

There exist many "or"-operations that satisfy all these properties. The most widely used are $\max(a,b)$ and $a + b - a \cdot b$.

Comment. The operation $a+b-a\cdot b$ can be explained if we recall the de Morgan law, according to which $A \vee B$ is equivalent to $\neg(\neg A \,\&\, \neg B)$ Let us explain this law.

For example, you can get a discount on using El Paso buses if you are either a student (A) or a senior person (B). So, a person gets a discount if $A \vee B$ is true.

Thus, *not* getting a discount $\neg(A \vee B)$ means that you are *not* a student *and not* a senior person: $\neg A \,\&\, \neg B$. So, $\neg(A \vee B)$ is equivalent to $\neg A \,\&\, \neg B$. Hence, the same equivalence is true for negations of these two statements: $A \vee B$ is equivalent to $\neg(\neg A \,\&\, \neg B)$.

Negation is naturally described as $1 - a$. So, if we use algebraic product for &, then the de Morgan formula leads to

$$f_\vee(a,b) = 1-(1-a)\cdot(1-b) = 1-(1-a-b+a\cdot b) = 1-1+1+b-a\cdot b = a+b-a\cdot b.$$

Let us bring all this together. Now:

- we know how to describe the original statements like "ΔT is negligible",

- we know how to describe "and"-combinations, and

- we know how to describe "or"-combinations.

We can therefore conclude that the degree $\mu_R(\Delta T, u)$ to which the control u is reasonable for the given value of ΔT can be computed as follows:

$$\mu_R(\Delta T, u) =$$

$$f_\vee(f_\&(\mu_N(\Delta T), \mu_N(u)), f_\&(\mu_{SP}(\Delta T), \mu_{SN}(u)), f_\&(\mu_{SN}(\Delta T), \mu_{SP}(u))).$$

Numerical example. Let us take $\Delta T = 2$ and $u = -4$, and, for simplicity, let us use:

- min as an "and"-operation and
- max as an "or"-operation.

In this case:

- Since the value $\Delta T = 2$ is between 0 and 5, we get
$$\mu_N(\Delta T) = 1 - 2/5 = 0.6.$$
- The value $u = -4$ is between -5 and 0, so we get
$$\mu_N(u) = 1 + (-4)/5 = 0.2.$$

Thus, $f_\&(\mu_N(\Delta T), \mu_N(u)) = \min(0.6, 0.2) = 0.2$.

- Since the value $\Delta T = 2$ is between 0 and 5, we get $\mu_{SP}(\Delta T) = 2/5 = 0.4$.
- The value $u = -4$ is between -5 and 0, we get $\mu_{SN}(u) = -(-2)/5 = 0.4$.

Thus, $f_\&(\mu_{SP}(\Delta T), \mu_{SN}(u)) = \min(0.4, 0.4) = 0.4$.

- Since the value $\Delta T = 2$ is larger than 0, we get $\mu_{SN}(2) = 0$.
- The value $u = -4$ is smaller than 0, so we get $\mu_{SP}(u) = 0$.

Thus, $f_\&(\mu_{SN}(\Delta T), \mu_{SP}(u))) = \min(0, 0)$.

So, $\mu_R(\Delta T, u) = \max(0.2, 0.4, 0) = 0.4$ – this is the degree to which, for the difference in temperatures $\Delta T = 2$, it is reasonable to turn the thermostat's know $u = -4$ degrees, i.e., 4 degrees of the left.

4 Summary of Fuzzy Methodology

Let us summarize what we have learned so far. We will describe the genera steps – and, *in italics*, we illustrate the general description by explaining what we did in our example.

What problem we are solving. We want to come up with an algorithm that, given the inputs $x_1, \ldots, x_n$, generates an output y.

In our example, we have only one input $x_1 = \Delta T$; based on this input, we want to generate the value of $y = u$.

Step 1: eliciting rules from the experts. First, we ask the experts to provide natural-language rules that describe y based on x_i.

In our example, we had three such rules:

- *If ΔT is negligible, then u is negligible.*
- *If ΔT is small positive, then u is small negative.*
- *If ΔT is small negative, then u is small positive.*

Step 2: describing what is reasonable – first, in natural-language terms. Based on the expert rules, we describe what it means for a value y to be reasonable for the given inputs $x_1, \ldots, x_n$. This means that:

- either the first rule is applicable, i.e., its condition(s) is satisfied and its conclusion is satisfied,
- or the second rule is applicable, i.e., its condition(s) is satisfied and its conclusion is satisfied, etc.

In our example, this reasonableness took the following form:

(ΔT is negligible and u is negligible) or
(ΔT is small positive and u is small negative) or
(ΔT is small negative and u is small positive).

Step 3: eliciting membership functions. The natural-language rules provided by experts use imprecise natural-language words.

In our example, we used three imprecise words: negligible, small positive, and small negative.

For each of these words, we ask the expert, for different inputs x, to provide his/her degree of confidence that the given value satisfies the property.

In our example, for negligible (N, for short), we had three values:

- $\mu_N(-5) = 0$, *meaning that the difference in temperatures $\Delta T = -5$ degrees is definitely not negligible;*
- $\mu_N(0) = 1$, *meaning that the difference in temperatures $\Delta T = 0$ degrees is definitely negligible; and*
- $\mu_N(5) = 0$, *meaning that the difference in temperatures $\Delta T = 5$ degrees is definitely not negligible.*

For small positive (SP, for short), we had the following three values:

- $\mu_{SP}(0) = 0$, *meaning that the difference in temperatures $\Delta T = 0$ degrees is definitely not small positive;*

- $\mu_{SP}(5) = 1$, *meaning that the difference in temperatures* $\Delta T = 5$ *degrees is definitely small positive; and*

- $\mu_{SP}(10) = 0$, *meaning that the difference in temperatures* $\Delta T = 10$ *degrees is definitely not small positive.*

For small negative (SN, for short), we had the following three values:

- $\mu_{SN}(-10) = 0$, *meaning that the difference in temperatures* $\Delta T = -10$ *degrees is definitely not small negative;*

- $\mu_{SN}(-5) = 1$, *meaning that the difference in temperatures* $\Delta T = -5$ *degrees is definitely small negative; and*

- $\mu_{SN}(0) = 0$, *meaning that the difference in temperatures* $\Delta T = 0$ *degrees is definitely not small negative.*

We then use interpolation – e.g., linear interpolation – to find the values of all the membership functions for all other values of the inputs.

In our example, we used linear interpolation to come up with expressions for the three membership functions $\mu_N(x)$, $\mu_{SP}(x)$, *and* $\mu_{SN}(x)$.

Step 4. We select an "and"-operation $f_\&(a, b)$, and we use it to combine the degrees to which different conditions are satisfied into a degree to which the rule is applicable.

In our example:

- *We know that the degree to which the statement "*ΔT *is negligible" is satisfied is equal to* $\mu_N(\Delta T)$.

- *We know that the degree to which the statement "u is negligible" is satisfied is equal to* $\mu_N(u)$.

So, we conclude that the degree to which the statement

$$\Delta T \textit{ is negligible and } u \textit{ is negligible}$$

is satisfied is equal to $f_\&(\mu_N(\Delta T), \mu_N(u))$.

Similarly, we conclude that the degree to which the statement

$$\Delta T \textit{ is small positive and } u \textit{ is small negative}$$

(which corresponds to the second rule) is satisfied is equal to $f_\&(\mu_{SP}(\Delta T), \mu_{SN}(u))$, *and the degree to which the statement*

$$\Delta T \textit{ is small negative and } u \textit{ is small positive}$$

(which corresponds to the third rule) is satisfied is equal to $f_\&(\mu_{SN}(\Delta T), \mu_{SP}(u))$.

Step 5. We select an "or"-operation $f_\vee(a, b)$, and we use it to combine the degrees to which each rule is satisfied into a degree to which the value y is reasonable. This way, once we know the inputs, then, for each value y, we know the degree to which this value is reasonable for the given inputs.

In our example, we combined the three degrees

$$f_\&(\mu_N(\Delta T), \mu_N(u)), \quad f_\&(\mu_{SP}(\Delta T), \mu_{SN}(u)). \text{ and } f_\&(\mu_{SN}(\Delta T), \mu_{SP}(u))$$

into a single degree

$$f_\vee(f_\&(\mu_N(\Delta T), \mu_N(u)), f_\&(\mu_{SP}(\Delta T), \mu_{SN}(u)), f_\&(\mu_{SN}(\Delta T), \mu_{SP}(u))).$$

This way, once we know the difference $\Delta T = T - T_0$ between the actual temperature T and the desired temperature T_0, we can compute, for each angle u, the degree to which this angle leads to a reasonable control.

What next. If we simply want to provide advice to a human decision maker, then this is all we need: we tell the decision maker which values are more reasonable and which are less reasonable, and let him/her decide what to do.

However, in many practical situations, we are interested in an automatic control. In this case, we need to transform such "fuzzy" recommendations into an exact control value. This transformation is known as *defuzzification.* We will study it in the next lecture.