

Artificial Intelligence

Manuel Pagliuca

December 13, 2021

Contents

1	Motivation	5
2	Neural networks	6
2.1	Biological background	6
2.2	Reasons for studying the biological background	7
2.3	Neurons	7
2.4	Computer vs Human brain	8
2.5	Threshold Logic Units	10
2.5.1	Conjunction example	12
2.5.2	Implication example	12
2.5.3	Multiple inputs example	12
2.6	Geometric interpretation	13
2.6.1	Linear separability	17
2.6.2	Convex hull	17
2.7	Solution of the bi-implication problem	17
2.8	Arbitrary boolean functions	18
2.9	Training TLUs	19
2.9.1	Negation example	20
2.9.2	Delta rule (Widrow-Hoff)	22
2.9.3	Convergence theorem	24
2.10	Artificial neural network	25
2.11	General structure of the neuron	26
2.12	Type of artificial neural network	27
2.13	Configuration of a neural network	29
2.14	Multi-layer Perceptrons	31
2.15	Regression	38
2.15.1	Linear regression	38
2.15.2	Polynomial regression	39
2.15.3	Multilinear regression	40
2.15.4	Logistic regression	41
2.15.5	Two-class problems	43
2.16	Training multi-layer perceptrons	44

2.16.1	Gradient descend	44
2.16.2	Variant of Gradient Descend	49
2.17	Number of hidden neurons	50
2.18	Cross validation	51
2.18.1	Sensitivity analysis	52
2.19	Deep learning	52
2.20	Radial Basis Function Networks	55
2.21	Training RBFN	59
2.21.1	C-means clustering	63
2.22	Learning Vector Quantization	65
2.23	Self-organizing maps	69
2.24	Hopfield Networks	72
2.25	Boltzman machines	74
2.25.1	Allenamento	75
2.26	Restricted Boltzman Machines	76
2.27	Recurrent Networks	77
3	Fuzzy Systems	78
3.1	Motivation	78
3.2	Imprecision and uncertainty	79
3.2.1	Sorites Paradox	79
3.2.2	Uncertainty	79
3.2.3	Difference between imprecision and uncertainty	79
3.3	Principle of incompatibility	80
3.4	Fuzzy sets	80
3.4.1	Membership functions	80
3.4.2	Linguistic variables and values	81
3.4.3	Semantics of fuzzy systems	81
3.4.4	Definition of a set	83
3.4.5	Support and Core	86
3.4.6	Height of a fuzzy set	86
3.4.7	Convex fuzzy sets	86
3.4.8	Fuzz numbers	87
3.5	Fuzzy logics	87
3.5.1	Set operators	87
3.5.2	Aristotelian logic	87
3.5.3	Classical logic	88
3.5.4	Inference rules	88
3.5.5	Boolean algebra	88
3.5.6	n -valued logics	89
3.5.7	Fuzzy set theory	89
3.5.8	Fuzzy implications	93
3.6	Theory of Fuzzy Logic	94
3.6.1	The extension principle	94
3.6.2	Some relevant fuzzy sets	95
3.6.3	Quantitative fuzzy variables	96

3.6.4	Set Representations of a fuzzy set	96
3.6.5	Fuzzy Relations	97
3.7	Fuzzy binary relations	99
3.7.1	Binary relations on a single stet	100
3.7.2	Properties of Crisp Relations	100
3.7.3	Properties of fuzzy relations	101
3.8	Fuzzy controller	101
3.8.1	Table-based Control	101
3.8.2	Problems of classical approach	102
3.8.3	Fuzzy approach	103
3.8.4	Fuzzy relation	104
3.8.5	Fuzzy control rules	104
3.8.6	Approximate reasoning with fuzzy rules	104
3.8.7	Architecture of a fuzzy controller	104
3.8.8	Defuzzification	105
3.8.9	Mamdani controller	106
3.8.10	Takagi-Sugeno controller	106
3.8.11	Similarity-baseed reasoning (controller)	107
3.9	Fuzzy data analysis	107
3.9.1	Fuzzy clustering	108
3.9.2	Issues with fuzzy clustering	109
3.9.3	Variants	109
3.9.4	Random sets	111
3.10	Neuro-Fuzzy Systems	112
3.10.1	The algorithm	113
4	Evolutionary Computing	114
4.1	Motivation	114
4.1.1	Principles of organismic evolution	114
4.2	Ingredients of an evolutionary algorithm	115
4.3	Formal definition	116
4.4	Meta-heuristics	118
4.4.1	Local Search Methods	118
4.4.2	Tabu search	123
4.4.3	Memetic algorithms	124
4.4.4	Differential evolution	124
4.4.5	Scattered search	124
4.4.6	Cultural algorithm	124
4.5	Elements of evolutionary algorithms	125
4.5.1	Encoding	125
4.5.2	Fitness	125
4.5.3	Selection	126
4.5.4	Genetic operators	128
4.5.5	Adaptation strategies	130
4.6	Meta heuristics	131
4.6.1	Swarm and population based optimization	131

4.7	Theoretical fundations	132
4.8	Genetic programming	133
4.8.1	Initialization	134
4.8.2	Genetic operators	135
4.8.3	Introns	135
4.9	Evolutionary strategies	135
4.9.1	Selection strategies	136
4.9.2	Adaptation	136
4.9.3	Global adaptation	136
4.9.4	Local adaptation	137
4.10	Multi-criteria Optimization	137
4.10.1	Pareto-Optimal solutions	137
4.10.2	Pareto-Optimal solutions with EAs	138

1 Motivation

Artificial neural networks are information processing systems, whose structure and operation principles are inspired by the nervous system and the brain of animals and humans.

Artificial neural networks are studied for various reasons.

- Extracting knowledge from data (*phenomena, events, processes, operating environment,...*).
- Understanding the phenomena that I'm observing, extracting the knowledge about that phenomena.
- Automated construction of computational paradigms for problem solving.

The ultimate goal is to be able to have a model which allow to describe the phenomenon that I'm observing and using this model I would like to solve the specific application problem.

There is an huge variety of application, this is the reason which AI now is so popular, and we have so many interested and expertise in this field.

The next evolution of the economy is based on availability of solution which are extracted from the data by analyzing the data reasoning, the only problems is that our capability of reasoning is limited by the fact that our brain is not able to analyze huge quantity of data, with automated system we can exploit this and make more comprehensive models that describe the process of systems we are observing and develop more accurate solutions.

Basically, artificial intelligence is mimicking the nature. We want to take the data analyze the data make a model, and for doing this we use techniques, we want to enrich the analysis we are doing through sensors from environments, we can use other specific sets of techniques.

We have the need of putting together some ideas and observation, we can define rules for reasoning, we extract the knowledge, but we can also build knowledge through reasoning. Basically, what we want to do is to try to replicate how the living beings observe and operate in the environment, how express, interacts.

Not only but we want to observe how the individual evolves, population evolves, in order to understand the trend of the environment. In this broad variety of approaches is to define some techniques which mainly are in two big categories, from the point of view of intelligence :

- **Symbolic approach**
- **Sub-symbolic approach**, which are sub-symbolic techniques for analyze the environment and the systems.

During this course we are focusing on sub-symbolic reasoning techniques:

- **Neural networks**, which is a simulation of a living brain.
- **Fuzzy systems**, which embed the definition of quantities, which are fuzzy (not defined in a very crisp way).
- **Evolutionary computing**, which is a set of techniques which mimics the natural evolution of the species so we can try to optimize the solution by using the basic concept of the natural evolution of the species.

2 Neural networks

2.1 Biological background

Basically, the aim of the artificial neural networks in *mimicking* the behavior of our brain starts from the fields of neural biology and neural physiology. What this discipline tries to do is to analyze the behavior of the biological neurons cells and understand how they behave for retrieving the information from the outer world.

The neurons use sensors and special cells to connect to the external world and get information. As humans or other biological species, we have five senses, taking for example the eyes, they are receptors which are able to see what around us is, they are able to reconstruct the scene we got around us.

Basically, with artificial neural networks we are trying to replicate what **biology** does for us, in neurobiology we build a model of what is happening in the neurons, this model describes how the neurons are interacting together to extract the knowledge, to build memory, construct reasoning, take decisions.

In computer science we want to build this model in computer just for trying to do something similar, we want to mimic the behavior of the natural brain in order to try to replicate in an artificial environment the same operation.

We are building the *model* for neurons in the computer, in this way we are able to use these artificial models to *learn* the environment and to solve practical problem (predict possible behaviors and solve optimization problem, like we do naturally). We also have some basis in other disciplines like physics and chemistry since we may want to use the neural networks, also for describing physical phenomena, not only to create reasoning in our mind but we want to create models of phenomena, this can be used for various application. We can create an abstract model, and instead of observing the real world we can observe how the model behave in some conditions.

Neural networks and in general artificial intelligence, is a discipline of computer science and engineering, then we use some inspirations from other discipline but the core of the theoretical aspect of the foundation which defines the artificial neural networks this is an area which is been to computer science and engineering.

2.2 Reasons for studying the biological background

The reasons for studying the BB (*biological background*) in computer science are two:

- The **first reason** for studying the neural networks is the fact that they are a very appealing model, since they work in **parallel** means that they have an intrinsically extremely high *parallel process capability*. This is why computer scientist are so attracted to this topic. Our brain in many cases find the solution immediately, this is fascinating, this happens because we are exploiting the parallel capability of our neural networks.
- The **second reason** why we are studying this technology is the fact that there is really a huge amount of *practical application* in broad variety of area : *industrial manufactory, products medicine, finances, economy, social networks, data analysis, ...*

2.3 Neurons

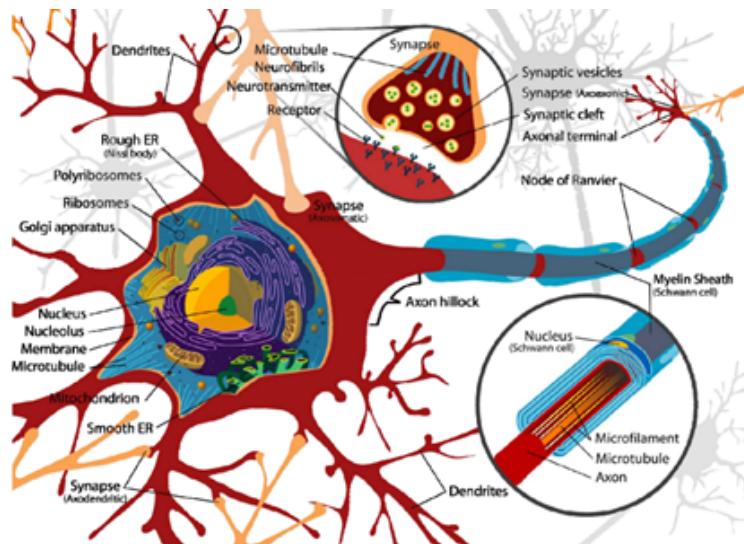


Figure 2.1: Neuron

The *central* part of the neuron is the part, which is managing the entire cell, this core the nucleus is listening what is happening around the neuron, when I have a **solicitation** from the external (other neurons or special cells related to the five senses).

When there is sufficient stimulation coming from this cells the nucleus is solicited and at a given point the nucleus realizes that the solicitation is so high that he has to take an **action**, the action is to send a *signal* around the *axon* (the

long extension covered in blue), this will lead to have a *polarization* of an *electric signal* which is flowing along this connection and this will reach other neurons that are connected to the end of the axon to the synapses.

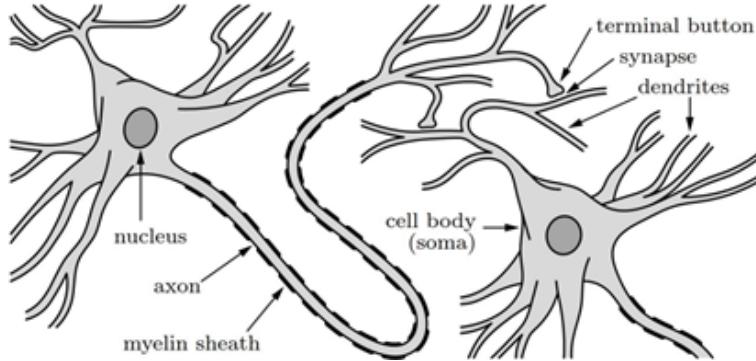


Figure 2.2: Connection between two neurons

The **axon** is covered by an appropriate *protein* (covered in blue) which protect the axon itself, and make sure that the polarization that occur on the nucleus is transmitted along the axon so that this signal is able to reach the synapse.

Basically, what we see is that the **neuron is exciting the nucleus**, which is generating the excitation along the axon, then it will reach the terminal synapses.

Each synapse is connected to the synapses of *another* neuron, so that the signals which are generated by the nucleus and sent to the axon and then to the synapsis terminal at the terminal the synapses are releasing some chemicals called **neurotransmitters**.

These *substances* are exciting the synapses of the connected neuron, bu stimulating the synapses this excitation is propagated from the **dendrites** to the nucleus of the other neuron, when the other cell is *sufficiently excited* by from the amount of stimulation generated by the neurotransmitter the nucleus will generate again a new stimulation which is going along the axon and reach another neuron (and so on...).

The axon is *depolarized* if there is enough excitatory input, basically this is how the signals are propagated through the brain.

Due to the fact that each neuron can stimulate each other neurons connected to them, this will create the **parallel processing** in our brain, so that each component will take care of analyzing each part of the information and this is a consequence to derive a part of the total computation.

2.4 Computer vs Human brain

When we look to the differences between a computer and the human brain what we notice, is that in the computer we got processors composed by many

transistors, the human brain counterpart the got 10^{11} *neurons*, the number of the latter overcomes the number of the cores in a processor.

The neurons aren't able to process the same *complex operation* of a core, but they are still so many that they can overcome the limit of the complexity of the individual operation with the fact that they are working significantly in parallel.

	Personal computer	Human brain
Processing units	1 CPU, 2–10 cores 10^{10} transistors 1–2 graphics cards/GPUs, 10^3 cores/shaders 10^{10} transistors	10^{11} neurons
Storage capacity	10^{10} bytes main memory (RAM) 10^{12} bytes external memory	10^{11} neurons 10^{14} synapses
Processing speed	10^{-9} seconds 10^9 operations per second	$>10^{-3}$ seconds < 1000 per second
Bandwidth	10^{12} bits/second	10^{14} bits/second
Neural updates	10 ⁶ per second	10^{14} per second

Figure 2.3: PC vs Human brain

The **storage** of the neurons has an immense storage potential than any other memory capacity. The **processing speed** seems *much higher* on the computer, but the parallel operation that you can do in parallel in the human brain in the end *overcome* the processing speed of the hardware of a computer.

Basically, what we can observe is that the biological neural networks are able to outperform significantly the processor that we have nowadays, there are some researches carried out which are trying to create processors which are replicating in hardware the operation of the biological neural networks but still the capabilities of this systems are *far* from the capability of human brain.

It may *seem* in some system that they are working faster than humans, it may seem that they are running better, but actually what you have is that you have the ability in the computer to run the algorithms and the explore the possible (deterministic) moves in a fast way but needs some background knowledge in order to that in a very short time.

Advantages of biological neural network:

- High processing speed due to massive parallelism.
- Even if we have a significant amount of the biological neural network damaged, the system is considered **fault tolerance**, it remains functional even if larger parts of the network get *damaged* (maybe some functions

will be disabled). The other cells are able to overcome the death cells, this thanks to the elasticity of the neurons which are able to overcome possible damages in the structures.

- If more neurons are failing, the brain will *degrade the performances* in a *graceful way*, it will not just stop working, will work a bit less not with the same performance and the same functionality but with reduced function. Only when a *really massive* number of neurons is dying at that point a function is not working anymore.
- They are well suited for inductive learning (*learning from examples, generalization from instances*).

What we are trying to do with artificial neural networks is to capture the parallel operation of the brain. The ability to extract the knowledge from the data, we want to replicate these capabilities.

There are some problems due to the ANN (artificial neural networks), if we kill part of the architecture which is replicating the BNN (biological neural networks), the ANN is **not automated to survives**, we have to assure some physical redundancy, this is a problem of the architecture we are using to execute the computation of the ANN.

What we want to do in our model is to construct a *set of abstract models of the neurons* that we call **artificial neurons**, and we like to connect them together to replicate the structure that we have in the natural brain this is why we have the ANN, a connection of neurons that is trying to mimic the behavior of the brain.

The complexity of the brain is so high that is actually difficult to replicate everything in ANN, what we are doing is to replicate a specific function of the brain which is able to solve a specific application problem that we have.

2.5 Threshold Logic Units

This is the first *abstract model* for an artificial neuron of the brain. A **threshold logic unit** (TLU) is a *processing unit* (neuron) with several inputs. It can solve a very simple set of problems.

We have a **core** (the neuron) with several *inputs* that are reaching the neuron, and we have the *output* which is delivered to the subsequent neurons which are connected with it. A TLU is a processing unit in which the output is governed by a threshold θ , if it has a *sufficient excitation* from the inputs, then the TLU became **active** (value 1) and generates the output y .

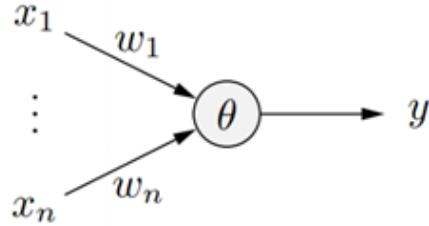


Figure 2.4: Threshold logic unit

We have n inputs identified by x_1, \dots, x_n the unit is generating only one output y each input is **not** delivered directly to the core of the TLU, but each input is **weighted**, some are more relevant, and others are less relevant (exactly how we are doing when we consider the data from the external world).

$$y = \begin{cases} 1, & \text{if } \sum_{i=1}^n w_i x_i \geq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Figure 2.5: TLU conditions

This means that our problem is depending on the most important information, this is what happens in the real world.

In the ANN we are replicating the relevance of the individual input, we can control how much each input is actually important to influence the generation of the output, to do this we use a weight w for each of the input so that the TLU will not see exactly the input but will see a weighted input, this will allow the modulation of each input by giving to each of them the appropriate importance during the generation of the final output.

When the TLU is solicited enough it will generate an output y that will be delivered to the *terminal synapse* to another neuron.

- If the **excitation is enough**, mathematically means that the weighted summation is greater than a thresh θ , we generate 1 (the neuron is active).
- If the **excitation is not enough** for overcoming the threshold θ , we generate 0.

This model that tries to represent what is happening in the BNN from two scientists, also called the **McCulloch-Pitts neuron**.

2.5.1 Conjunction example

The result is equal to 1 only when the two outputs are equal to 1. There is not a standard way to select the threshold θ , we have to choose that in base of the function that we want to implement. In this case I have to select a θ which is greater than the biggest weight.

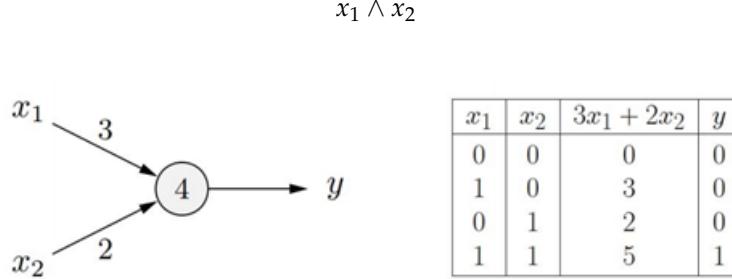


Figure 2.6: Conjunction TLU

2.5.2 Implication example

I can choose the value of the threshold in order to have the proper function, in this case applies on the same ways. *How can I choose the interconnection weights?*

The problem is that there are no general rules, I choose the weights according to the intrinsic relevance of each input variable.

How can I do that when we have a high number of inputs? we will see that there is a procedure for doing that.

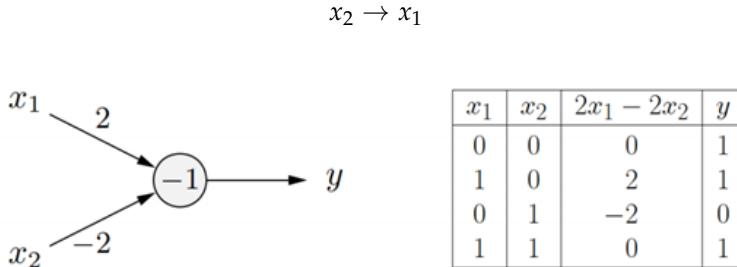


Figure 2.7: Implication TLU

2.5.3 Multiple inputs example

In this case we can see that we have three possible inputs, we can discriminate the inputs in excitatory input and inhibitory input. The first tries to contributes to the final computation of the neuron in such a way that the results will be greater than the threshold, the other neuron does the opposite.

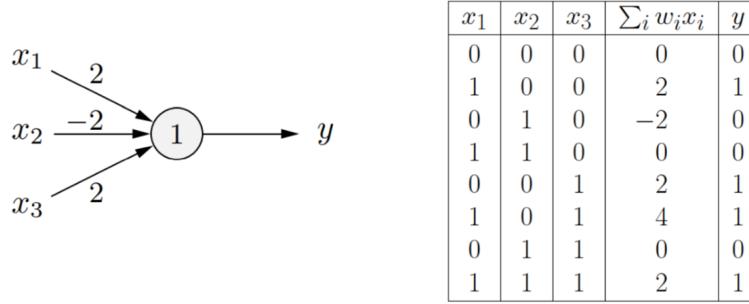


Figure 2.8: Three-input TLU

2.6 Geometric interpretation

The geometric interpretation is significantly helpful to derive a *method* to configure the threshold and the weights starting from the data. We will consider a single and simple TLU, we will try to understand how we can interpret the behavior of the TLU in a geometrical way.

You know that is possible to represent a straight line on a plane in any of the following forms :

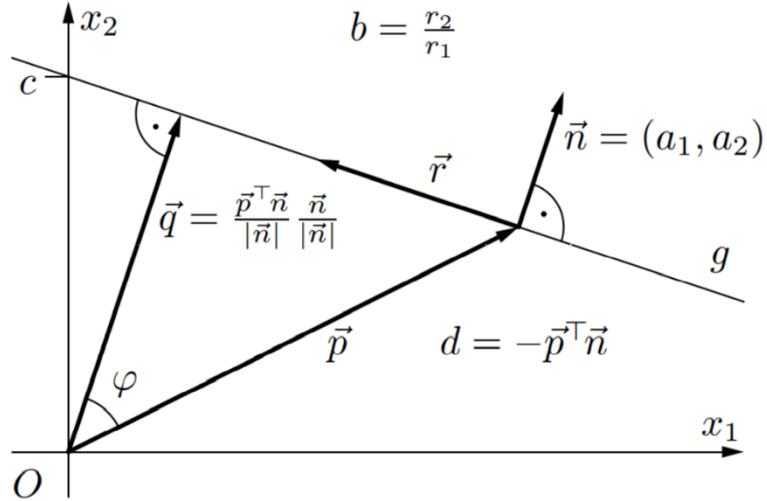
Explicit Form:	$g \equiv x_2 = bx_1 + c$
Implicit Form:	$g \equiv a_1x_1 + a_2x_2 + d = 0$
Point-Direction Form:	$g \equiv \vec{x} = \vec{p} + k\vec{r}$
Normal Form:	$g \equiv (\vec{x} - \vec{p})^\top \vec{n} = 0$

Figure 2.9: Different form for representing a straight line

If i **implicit form**, where you have a weighted combination of the two variables plus a possible threshold, a vector representation in point-direction form and normal form. Any of this is fine to represent a straight line in the plane, we are considering just two variables x_1 and x_2 , and use one of the many representations.

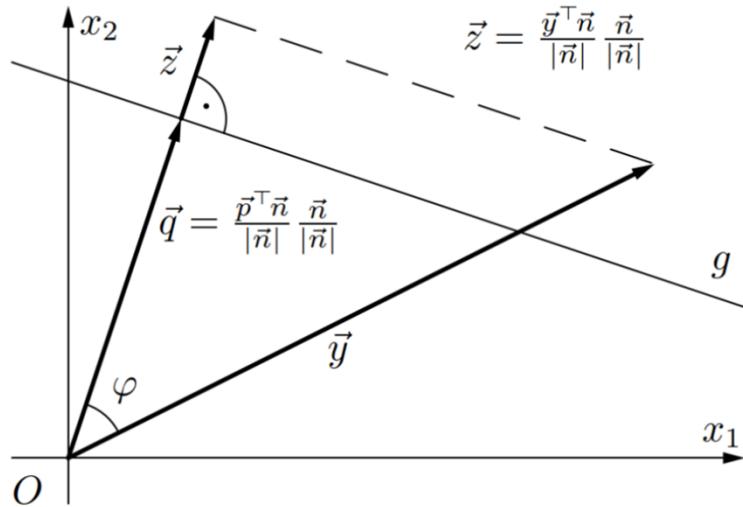
-
- | | |
|-------------|---|
| b : | Gradient of the line |
| c : | Section of the x_2 axis (intercept) |
| \vec{p} : | Vector of a point of the line (base vector) |
| \vec{r} : | Direction vector of the line |
| \vec{n} : | Normal vector of the line |
-

Figure 2.10: Implicit form representation legend



In the case of the *explicit/implicit form* the \vec{b} is the **inclination** (or *gradient*) of the straight line in respect of the horizontal axis, and c is the intercept of the vertical axis.

If we look to the normal \vec{n} we want to pick the vector which is *orthogonal* to the straight line. The straight line is represented by all the points which starts from the origin and has a quantity in the direction of the normal. The vector \vec{p} identifies a point in the example, we consider that point belonging to our straight line, the distance of the straight line respect to the origin O is given by $|\vec{q}|$ which is the projection of \vec{p} on straight line g .

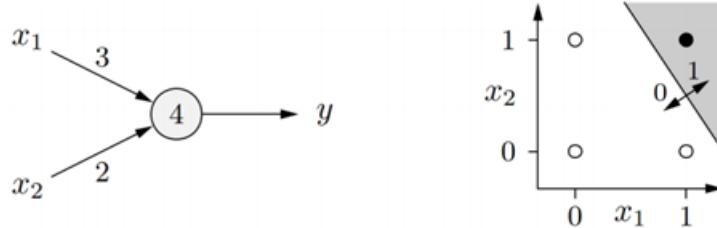


Considering this other graphical representation it is possible to know on which side a points lands. If we take the vector representing \vec{y} a point, and i compute the projection in the direction of the normal, the projection of this is the vector \vec{z} .

To understand on which side the points are, i just need to see if the vector \vec{z} (which is the projection of our point) is shorter or longer than the point i observe on the straight line, which is pointed by \vec{q} . This means that all points (expressed by a vector) which have a **module** higher then the projected point onto the straight line (\vec{q}), are part of the plane above the straight line (they will *satisfies* the solution), viceversa, they will be below the plane if the module will be shorter then \vec{q} (they won't satisfy the condition).

Basically the straight line which defines the behavior of my TLU, splits the plane in two parts (since i'm considering only x_1 and x_2). All the points which distance is greater then

TLU for $x_1 \wedge x_2$



TLU for $x_2 \rightarrow x_1$

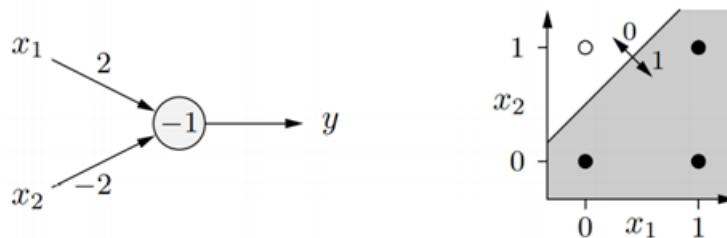


Figure 2.11: Solution of the *conjunction* and *bi-implication*

Now Let's go back to the two examples that we have seen before, starting with the conjunction, in this means that I have three points in which the output has to be 0, and one point where the result is 1. This means that if I represent the straight line with $x_1 = 3$, $x_2 = 3$ and the intercept equal to 4 i will draw a line which is actually separating in a clear way the element in which the output is equal to 1 from the values where the output is equal to 0.

For the case of three variables, this is more complex, we need to generalize this idea moving from a plane to a three-dimensional space. Think about the three axis and look at the combination of the dots, you need to set a plane, so you have to separate one set of inputs for which the output has to be equal to 1 from the set of inputs combinations where the output is equal to 0.

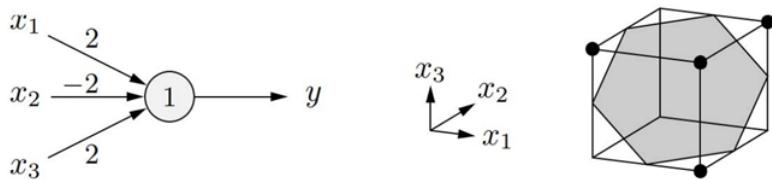


Figure 2.12: Solution of the complex TLU

Basically, if you represent the plane using these information's you will have this kind of plane (that looks like a hexagon) dividing the solution from the 0 values.

So how we choose the threshold and interconnection weights? I have to look at geometrical distribution of the points for the combination of the input which the output must be equal to 1 and the one where the output must be equal to 0, I have to take a straight line or plane and I need to position the divisor in a way that I clearly separate the two groups. If I can do that, that set of values (inputs and threshold) are the one that I need to use in my ANN.



Figure 2.13: The bi-implication problem

Let's analyze another example, the bi-implication problem, this problem results in this kind of distribution in the plane, as you can see it is not possible to separate these points with a straight line. As a consequence, we don't have any TLU for solving this problem, even if this is a simple problem. We have to introduce some notions for understanding the reasons behind this problem.

2.6.1 Linear separability

Two sets of points in the **Euclidean Space**, consider x_1, x_2, \dots, x_n as the variable for the TLU, since we have n possible inputs you are analyzing a problem represented in n dimensional Euclidean space.

If you have in this Euclidean space two sets of points they are **linearly separable**, if and only if there exists at least one point, line, plane or hyperplane, such that all points of the first set lie on one side and all points of the other set lie on the other side of this point, line, plane or hyperplane.

The point sets can be separated by a linear decision function.

If you are in a mono dimensional space you have one input only, your Euclidean Space is a line, you have one point on this straight line which is separating your space the line in two parts, the output is one, for the other point the output is 0. If you have a plane, we have two input variables, we have to find a plane that separates the two sets of points.

2.6.2 Convex hull

A set of points in the Euclidean Space is called **convex** if it is non-empty and connected and for every pair of points in it every point on the straight-line segment connecting the points of the pair is also in the set.

A **convex hull** of a set of points X in a Euclidean Space is the smallest convex set of points that contains X . Alternatively, the convex hull of a set of points X is the intersection of all convex sets that contain X .

2.7 Solution of the bi-implication problem

Two sets of points in Euclidean Space are linearly separable if and only if their convex hulls are disjoint. In the bi-implication problem, the convex hulls are the diagonal line segments.

They share their intersection point and this means that they are not disjoint, therefore the double implication is not linearly separable.

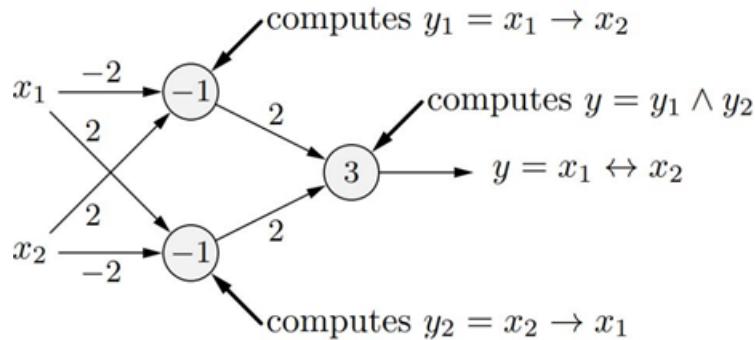


Figure 2.14: TLUs to the bi-implication problem

What we can do is putting together more neurons to try to address more complex problem where one neuron can't solve. We are creating a network of TLUs, we are splitting the problem in two sub problems.

The problem of implication can be solved with a single TLU, we create a more complex structure in which are able to solve the problem.

Let's see what happens geometrically, the two points a and c , are going to be separated from the first two TLUs, I set one of the TLU so that all points which are below the straight line represented by g_2 gives and output equal to 1, and then I set a second straight line represented by g_1 where all points are above over the straight line the values are equal to 1, and then I merge this information.

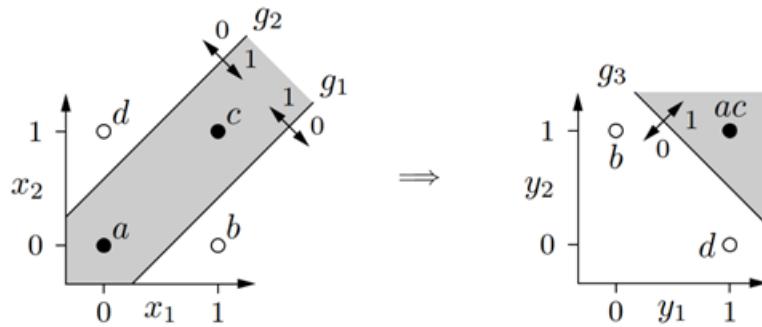


Figure 2.15: Solution of the bi-implication problem

I combine these information from the previous group in the third neuron so that I transform the representation of the information, I can merge the information of the points a and c , so that both the output of the two TLU is equal to 1 is represented by one points (in this case they are the same) and then I have the two other points b and d which are in both the parts above and below the line, for them the solicitation is not enough to generate an output equal to 1, as a consequence I obtained a linear separability.

I transformed the stripe in a semi plane with the third TLU, which identifies the linear separability, I don't have that propriety only with the first TLU but also with the second, both TLUs allow me to partition the first space in three parts.

2.8 Arbitrary boolean functions

I can work with any arbitrary **Boolean function**; in this example I have a Boolean function of three variables. I have these values of the value y according to the formula.

x_1	x_2	x_3	y	C_j
0	0	0	0	
1	0	0	1	$x_1 \wedge \bar{x}_2 \wedge \bar{x}_3$
0	1	0	0	
1	1	0	0	
0	0	1	0	
1	0	1	0	
0	1	1	1	$\bar{x}_1 \wedge x_2 \wedge x_3$
1	1	1	1	$x_1 \wedge x_2 \wedge x_3$

Figure 2.16: Table of values

What we can do is to build a network of TLUs that allow me to compute each of the component for which the output has to be 1 and then with a conjunction unit I put together all the possible value with a or for merging the individual TLUs.

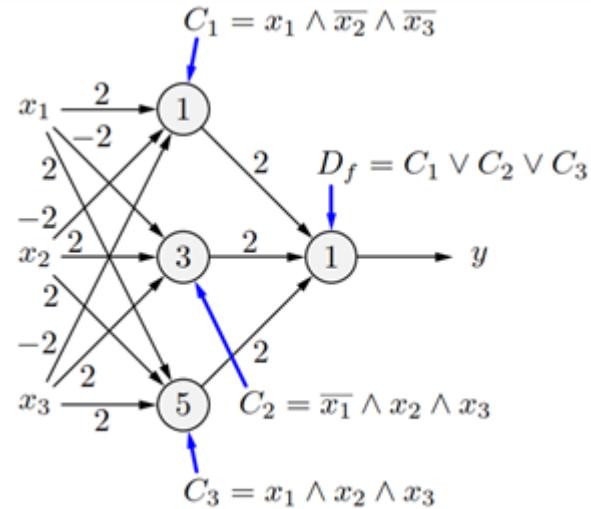


Figure 2.17: Network of TLUs solving the boolean function

2.9 Training TLUs

The geometric interpretation that we have seen before about the operation of TLUs gives us the understanding of how we can place the various parameters

for our networks in order to construct TLUs with 2 and 3 inputs.

But this makes sense when we are using 2 or 3 input, but this is something that is not feasible when we are having more than three inputs, also this is not an automated method.

We want an automatic method for visualizing the space and points in the space, especially when we have more than 3 inputs.

What we want to do is to have an automatic way which adjusts the weights and the threshold of the network so that I can reach the desired solution (if the two sets are linearly separable).

The **automatic training** of TLUs consist in the fact that we can start from a random value of the weights and threshold, and then when we want to configure a TLU we need to evaluate the error that we are generating at the output (of the TLU) in respect to the input pattern that we have presented.

Basically, we have chosen randomly the threshold, the TLU generates an output, we evaluate the error respect to the desired output, and then we try to adjust the weights and the threshold to reduce the error.

We repeat this operation for all inputs until the error is really reduced or vanished.

1. Start with random values for weights and threshold.
2. Determine the error from the output of the TLU.
3. Consider the error as a function of the weights and the threshold $e = e(w_1, \dots, w_n, \theta)$.
4. Adapt weights and threshold so that the error becomes smaller.
5. Iterate adaptation until the error vanishes.

2.9.1 Negation example

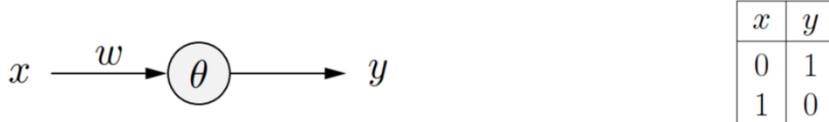


Figure 2.18: TLU that perform negation

$$\neg x$$

In this case we have two really simple parameters, the weights and the threshold. Let's represent the error for all possible weights and all possible threshold at least for a subset that we are interested in analyzing. Let's consider the $x = 0$, in this case the desired output is 1, if we take $w = 2$ and we multiply it by 0 the weighted input will be 0.

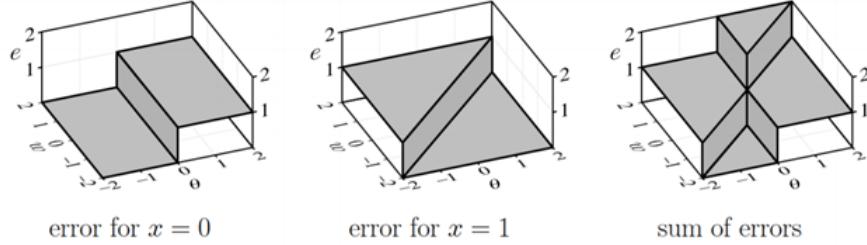


Figure 2.19: Diagrams of the errors expressed in terms of θ and w

Let's consider the first diagram on the left, with the various values of θ and w , and with an input value of $x = 0$. Let's see the error that we have with the various possible combinations, we see that for any values of the weight where the threshold is less than zero, the **error** is zero. Viceversa, for any value of the weight for a θ greater or equal to zero, the error is one (a *plateau*).

Now let's consider $x = 1$, we are on the central diagram, in the left triangular part of the domain there will be the error of 1 defined $\forall w$.

Let's try to sum the errors and see what we obtain: we get an intersection for the error equal to one, and also an intersection for the error equal to *two* (which is strange, since the error in this kind of example is always 1 since the y output is binary, but there we are talking about a sum).

The only part where the error is equal to zero is a small triangle on the bottom. The problem with this definition of the error is that is not suited for define an algorithm which allow me to find the zero error.

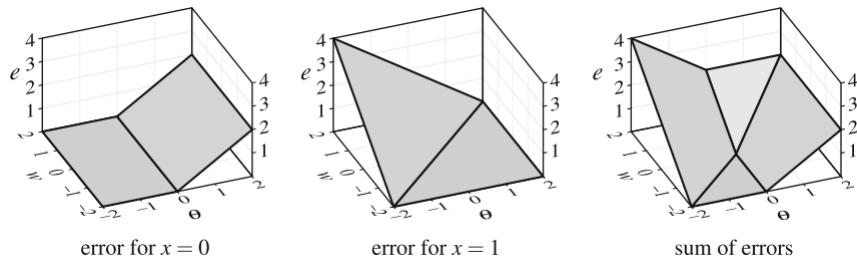


Figure 2.20: Output error modified as function of weight and threshold

What i can do is modify the description of the error and consider that for each value of the input i will not have just the error which is 1,0 or 2, but i will consider a covering and continuos **surface area** (such that is possible to increase progressively the error).

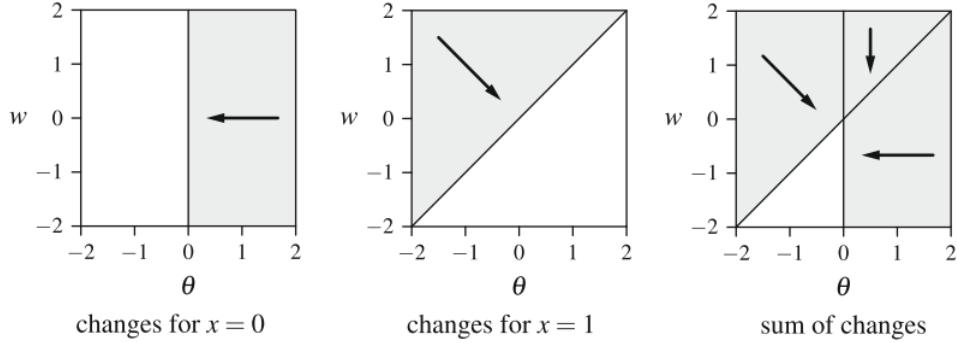


Figure 2.21: Top view output error modified as function of weight and threshold

So I can start from a random point and iteratively adapt parameters according to the direction corresponding to the current point, I will stop if the error vanishes.

There are two ways of training a neural network:

- **online learning**, we receive the learning pattern at time from the external environment, we compute the parameter corrections for this learning pattern (like we saw in the negation example) and in the end we apply the parameter corrections. In case of the *negation*, first we adapt the weight and the threshold according to the left diagram, then we adapt them according to the middle diagram, then we adapt them again according to the left diagram and so forth until the error vanishes.
- **batch learning**, consists in not applying the changes immediately after every training example, but aggregating them over all training examples. Only at the end of a (learning/training) epoch, that is, after all training examples have been traversed, the aggregated changes are applied. Then the training examples are traversed again and at the end the weight and the threshold are adapted and so forth until the error vanishes.

2.9.2 Delta rule (Widrow-Hoff)

Given:

- $\vec{x} = (x_1, \dots, x_n)^T$ be an input vector of TLUs.
- o is the desired output for this input vector.
- y the actual output of the TLU.
- η as learning rate.

If $y \neq o$, then, in order to reduce the error, the threshold θ and the weight vector $\vec{w} = (w_1, \dots, w_n)$ are adapted as follow:

$$\theta^{(new)} = \theta^{(old)} + \Delta\theta, \text{ with } \Delta\theta = -\eta(o - y)$$

$$w_i^{(new)} = w_i^{(old)} + \Delta w_i, \text{ with } \Delta w_i = \eta(o - y)x_i \\ \forall i \in \{1, \dots, n\}$$

The first equation is correction (delta) of the threshold and the second is the correction of the weight. The variation is given by an expression which is proportional to the difference between the actual and expected output (which is the error).

The correction for the weight is proportional not only to the error but also to the input. Essentially, if i have an actual output smaller than the desired one, i take the input that are bigger, and i try to push them to contribute more to the excitation of *threshold logic function* so that the output will be higher and as consequence the error will be lower.

The η controls the speed of updates, the bigger is the bigger will be the update.

Epoch	x	o	\mathbf{xw}	y	e	$\Delta\theta$	Δw	θ	w
1	0	1	-1.5	0	1	-1	0	0.5	2
	1	0	1.5	1	-1	1	-1	1.5	1
2	0	1	-1.5	0	1	-1	0	0.5	1
	1	0	0.5	1	-1	1	-1	1.5	0
3	0	1	-1.5	0	1	-1	0	0.5	0
	1	0	0.5	0	0	0	0	0.5	0
4	0	1	-0.5	0	1	-1	0	-0.5	0
	1	0	0.5	1	-1	1	-1	0.5	-1
5	0	1	-0.5	0	1	-1	0	-0.5	-1
	1	0	-0.5	0	0	0	0	-0.5	-1
6	0	1	0.5	1	0	0	0	-0.5	-1
	1	0	-0.5	0	0	0	0	-0.5	-1

Figure 2.22: Online training with $\theta = \frac{3}{2}, w = 2, \eta = 1$

Epoch	x	o	\mathbf{xw}	y	e	$\Delta\theta$	Δw	θ	w
1	0	1	-1.5	0	1	-1	0	1.5	2
	1	0	0.5	1	-1	1	-1		1
2	0	1	-1.5	0	1	-1	0	0.5	1
	1	0	-0.5	0	0	0	0		
3	0	1	-0.5	0	1	-1	0	0.5	0
	1	0	0.5	1	-1	1	-1		
4	0	1	-0.5	0	1	-1	0	-0.5	0
	1	0	-0.5	0	0	0	0		
5	0	1	0.5	1	0	0	0	0.5	-1
	1	0	0.5	1	-1	1	-1		
6	0	1	-0.5	0	1	-1	0	-0.5	-1
	1	0	-1.5	0	0	0	0		
7	0	1	0.5	1	0	0	0	-0.5	-1
	1	0	-0.5	0	0	0	0		

Figure 2.23: Batch training with $\theta = \frac{3}{2}, w = 2, \eta = 1$

Basically what happens in the online training, if i start in the first move i may go here, and so on, i'll move one step at the time since the amount of change is control by the learning rate.

in this case i move in one direction, since i apply together the two errors correction, so i'm moving in combination of the two changes of the secondo batch, then again i'm applying the two changes and i move. I do less move respect to the other since i combine the two correction for the expression that i shown.

I can try to do the same for the more complex case of the TLU, with two inputs. here i have the conjunction with the truth table describing the implementation. What i'm doing basically i have found the straight line that gives the satisfied and unsatisfied area for determining the solution.

2.9.3 Convergence theorem

If we consider a set of training patterns: $L = \{(\vec{x}, o_1), \dots, (\vec{x}_m, o_m)\}$, each consisting of an input vector $\vec{x}_i \in \mathbb{R}$ and a desired output $o_i \in \{0, 1\}$.

Furthermore, let's consider $L_0 = \{(\vec{x}, o) \in L | o = 0\}$ and $L_1 = \{(\vec{x}, o) \in L | o = 1\}$. If i can show that L_0 and L_1 are **linearly separable**, then is possible to prove that we have a $\vec{w} \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$ such that:

$$\forall(\vec{x}, 0) \in L_0 : \vec{w}^T \vec{x} < \theta$$

and

$$\forall(\vec{x}, 1) \in L_1 : \vec{w}^T \vec{x} \geq \theta$$

This means that it is possible to divide the two sets, and then the online or batch training procedure is able to terminate. So if the two sets are linearly separable they will give us a final value in a **finite time**. The final error will be zero.

The problem is that if we are not able to perform the linear separation of L_0 and L_1 , the algorithm is **not able to terminate**. The algorithm will oscillate around and we will not be able to find a solution with the zero error.

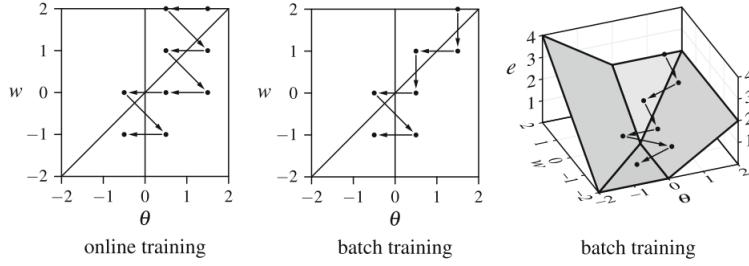


Figure 2.24: Example of online and batch training (also with summed errors)

The correction of the parameters maybe slowed down due to the fact that we have 0 and 1 to the possible values for our outputs. This implies that if we have an output for our threshold logic function equal to 1, we can find an adjustment and apply it. If we have an input which is 0 whichever weight we have, we may not be able to compute a correction that in the end will reduce the final error.

So when we apply an input which is one we can change the weight for finding a better value for the configuration, if we apply the change on an input which is zero the possible changes actually vanishes since we are multiplying by zero, we are not observing a variation.

To avoid to waste too much time on this issue, we can change the representation of the values, for *true* we are considering 1 and for *false* we are considering -1.

What we can point out, is that a *single* TLU, is able to point out any linearly separable function. We have just to apply the delta rule which is easy and fast and guarantee to find a solution, if one exists.

The problem is given by a more complex structure, like a network of TLUs, in this case we cannot apply directly the delta rule, because we have no clue about which is the output. We need something more complex for managing that.

2.10 Artificial neural network

An ANN in general has a very simple definition since it is a **directed** graph $G = (U, C)$ composed by nodes and edges, in the graph we have processing nodes that can elaborate the information incoming in, some arcs that are

bringing in the network some external information, and others that extract the computation performed by our network (which is a structure that mimic our brain).

The connections are the axon-synaptic connections which connect the nucleus of the neurons. In a general neural network this is the very abstract definition, in practice in order to use this kind of structure we will need to have more regular more focused organization of neurons connections.

The set of vertices U is partitioned into:

- U_{in} is the set of input neurons.
- U_{out} is the set of output neurons, whi are the one delivering the result of the computation to the outside world.
- U_{hidden} is the set of hidden neurons, they don't have any direct connection to the external world.

2.11 General structure of the neuron

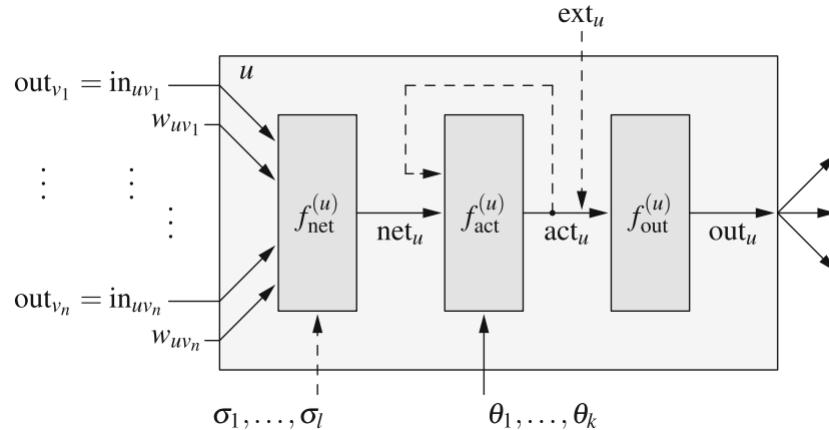


Figure 2.25: Internal representation of a neuron

We see that we have inputs that are coming in the network and usually this inputs are manipulated in order to understand which is the global stimulus that the neuron receive from the external world.

If an incoming signal from a previous neuron has relevance, we want to multiply the amount o signal incoming in by an appropriate constant in order to have a stronger weight for a subsequent processing in this neuron.

In this example we associate the weight w_{uv_1} , which is the weight for the neuron u respect to the incoming neuron v_1 .

These inputs are processed by three stages in which we perform specific operations, we want to see how much the neuron is actually solicited. For these reason we have three stage functions:

- The $f_{net}^{(u)}$, takes the input with the *relative relevance* of the various inputs to generate the **global solicitation** for our neuron. This is a general definition, a specific definition changes in base of the network we are considering.
- The $f_{act}^{(u)}$, the activation functions analyze the network input from the previous function and generate the **excitation status** of our neuron. This will tell us if the neuron is sufficiently excited.
- The $f_{out}^{(u)}$, which takes the excitation status of the neuron and elaborate the final status of the neuron to deliver to the subsequent neurons.

In general we have also an external variable ext_u which tell us how much the excitation should be increased from a stimulus coming directly from the *external world* (not from previous neuron, from the external world).

These are the various inputs that any neuron can have, we may have both the input coming from other neurons and the input coming from the external world, or we can have for a subset of neurons only the stimulus coming from the external world and no other inputs.

We have input neurons coming which are weighted properly, these values are used with the *network input function* which is generating the global excitation status coming from the other neurons, this one will generate the *activation status* of the neuron (how much it is stimulated by the external neurons).

In the end we have the out function which evaluates the final status of the neuron, which is delivered to the other connected neuron.

2.12 Type of artificial neural network

In the practice we won't use the general structure introduced before, we will have different type of ANN:

- **feed forward network** are ANN that doesn't contains any cycles, the acronym is FNN.
- **recurrent network** are ANN that contains cycles (backward connections), the acronym is RNN.

The operation of a general ANN:

1. **input phase**, where the external input are acquired by input neurons in the network.
2. **work phase**, the external input are disconnect (by freezing the input neurons), we move the excitation generated by the input neurons to the connected neurons, we generate the stimuli for the connected neuron, we compute the output status of them and propagate the output to the neurons which are connected to this one.

During the *working phase* if the input of neurons are steady (input stimuli are not changing), the computation of that neuron doesn't change (every stage function generate the same values, FNN).

This is not the case when we have a RNN, since if one of the outputs connected to a neuron is connected to an input will repeat the computation and this maybe change the output status (due to *recomputation*). I always say "*may*" since the actual change will depend specifically by the function that i'm using. The **recomputation** of a neuron output occurs if any of its input changes, we can't loop forever and not evolving. This means that the working phase continues until the external outputs are steady, or a maximum number of recomputation iterations is reached. The **temporal order** of computation depends by the specific NN.

Feed-forward neural network

1. Computation proceeds from input neurons progressively toward output neurons by following the topological order of the neuron in the network.
2. The external inputs are frozen.
3. Input neuron compute their outputs which are maintained steady and forwarded to the connected neurons.
4. Neurons connected to preceding neurons with steady outputs generate their respective outputs and propagated forward to the subsequent neurons, until the external outputs are generated.

Recurrent neural network

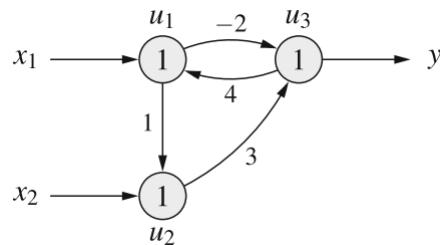


Figure 2.26: A simple recurrent neural network

In the neurons there is the threshold which is considered for the operation, and we have some interconnection weights which define the weights to be applied to the output of a neuron when his output is delivered as input to a subsequent neuron.

	u_1	u_2	u_3	
input phase	1	0	0	
work phase	1	0	0	$\text{net}_{u_3} = -2$
	0	0	0	$\text{net}_{u_1} = 0$
	0	0	0	$\text{net}_{u_2} = 0$
	0	0	0	$\text{net}_{u_3} = 0$
	0	0	0	$\text{net}_{u_1} = 0$

Figure 2.27: Dataset for the previous RNN, for the u_3, u_2, u_1 order

In this case we choice to start from the input 1,0,0 and with an updating order of neurons that is $u_3, u_1, u_2, u_3, u_1, u_2, u_3, \dots$. Each bold number in working phase entry of the table is an output of the *ordered* neuron (0 if greater less than θ or viceversa 1). In this case we can reach a steady state, and exit the RNN.

	u_1	u_2	u_3	
input phase	1	0	0	
work phase	1	0	0	$\text{net}_{u_3} = -2 < 1$
	1	1	0	$\text{net}_{u_2} = 1 \geq 1$
	0	1	0	$\text{net}_{u_1} = 0 < 1$
	0	1	1	$\text{net}_{u_3} = 3 \geq 1$
	0	0	1	$\text{net}_{u_2} = 0 < 1$
	1	0	1	$\text{net}_{u_1} = 4 \geq 1$
	1	0	0	$\text{net}_{u_3} = -2 < 1$

Figure 2.28: Different dataset of the same RNN with u_3, u_2, u_1 order

Now let's consider a different ordering $u_3, u_2, u_1, u_3, u_2, u_1, u_3, \dots$, in this case it is not possible to reach a steady state inside the RNN (**no stable state, oscillation of output**).

2.13 Configuration of a neural network

The details of the configuration strictly depends from the structure opf the network, in general we have two categories of learning procedure (or training procedure):

- **fixed learning task.**
- **free learning task**

Fixed learning task

Considering a NN of n input neurons $U_{in} = \{u_1, \dots, u_n\}$ and m output neurons $U_{out} = \{v_1, \dots, v_m\}$. A fixed learning task is a set of training patterns $l = (\vec{i}^{(l)}, \vec{o}^{(l)})$, each consisting of:

- an input vector $\vec{i}^{(l)} = (ext_{u_1}^{(l)}, \dots, ext_{u_n}^{(l)})$
- an output vector $\vec{o}^{(l)} = (o_{v_1}^{(l)}, \dots, o_{v_m}^{(l)})$.

At the end of the configuration the network is able to generate the desired output that correspond to the output vector. Since the examples are composed by an input vector and the expected output that we want to get, this is called **supervised learning**.

A fixed learning task is solved when for all training patterns $l \in L_{fixed}$ the neural network computes, from the external inputs contained in the input vector $\vec{i}^{(l)}$ of a training pattern l , the outputs contained in the corresponding output vector $\vec{o}^{(l)}$.

The error of a fixed learning task

The error says how well a neural network solves a given fixed learning task. Essentially it is the difference between desired and actual outputs.

$$e = \sum_{l \in L_{fixed}} e^{(l)} = \sum_{v \in U_{out}} e_v = \sum_{l \in L_{fixed}} \sum_{v \in U_{out}} e_v^{(l)}$$

$$e_v^{(l)} = (o_v^{(l)} - out_v^{(l)})^2$$

In order to have a number which tell us the final quality what we need to do we have to consider the module of the error, what we are doing is consider the difference between the actual and the desired output. But in this simple difference could give us a final number which is not reflecting the total accuracy of the NN due to the sign.

A solution to this is to use a squared value, in this way we can avoid the negative sign.

Free learning task

A free learning task is the complementary approach of the fixed learning task, in which the desired behavior is not defined a priori by the pair of vectors for the input and the output.

$$n \text{ input neurons } U_{in} = \{u_1, \dots, u_n\}$$

$$\text{one input vector } \vec{i}^{(l)} = (ext_{u_1}^{(l)}, \dots, ext_{u_n}^{(l)})$$

We only have a set of input vectors which are presented to the network and the learning algorithm will lead to an output vector which is similar to the given input.

Preprocessing

In order to be sure that the learning is working properly, we have to be sure that one of the input is dominating the operation of the network, as consequence we have to do some preprocessing on the data in order to give the same relevance to all neurons.

For each component of the input vector we need to compress the representation in the same range, this is called **normalization** (according to the average of the input value for each component, we need to be sure that the variance σ_k is normalized as well).

$$\mu_k = \frac{1}{|L|} \sum_{l \in L} ext_{u_k}^{(l)}$$

$$\sigma_k = \sqrt{\frac{1}{|L|-1} \sum_{i \in L} (ext_{u_k}^{(l)} - \mu_k)^2}$$

This is the standard normalization of the deviation σ_k , we have also another possibility which is called **unbiased standard deviation** (often preferred by statistician since is not polarized on the two direction).

If we apply the normalization this is the external stimuli that we will have after the normalization, basically all the stimuli are recomputed with respect to the average value and they are normalized in dimension dividing them by the standard deviation.

$$\sigma_{u_k}^{(l)(new)} = \frac{ext_{u_k}^{(l)(old)} - \mu_k}{\sigma_k}$$

There is a problem that we have to seriously consider when we are looking to our examples, we need to have *sufficiently descriptive examples*, which have to be well distributed over the domain that i'm considering (not focussing only on a portion of the domain).

So far we have been dealing with integer and real numbers, we want to use symbols for represent a group of examples that are similar together in a concise way.

For doing that we need to associate a group of identifiers to a group of examples. For being able to represent this symbols we consider the **1-in-N encoding**, if we need to represent n symbols we have a string of 0 and 1 and only one bit the one that corresponds to the symbol we want to represent will be 1 (the others will be 0).

2.14 Multi-layer Perceptrons

A **multi-layer perceptrons** it is essentially a feed-forward neural network in which is present a strictly layered structure. They neurons are connected in groups, each groups receives the input only from the previous layer or from the external input. The output is delivered by the subsequent group, there are different kinds of layers:

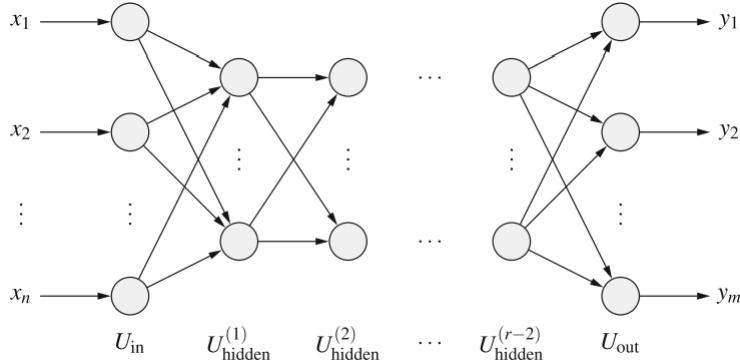


Figure 2.29: General structure of r -layered perceptron

- **input layer**
- **hidden layer**
- **output layer**

Each layers is connected, in multi-layer perceptrons the neurons are strictly connected between layers, jump are not allowed (like in the feed-forward network).

The network input function of each *hidden neuron* and of each *output neuron* is the **weighted sum** of its inputs:

$$f_{net}^{(u)}(\vec{w}_u, \vec{i}_u) = \vec{w}_u \cdot \vec{i}_u = \sum_{v \in pred(u)} w_{uv} out_v$$

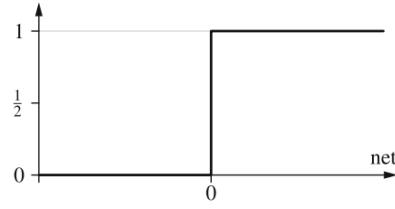
The activation function of each hidden neuron is a so-called **sigmoid function**, that is, a monotonically non-decreasing function with:

$$f : \mathbb{R} \rightarrow [0, 1] \text{ with } \lim_{x \rightarrow -\infty} f(x) = 0 \text{ and } \lim_{x \rightarrow \infty} f(x) = 1$$

It has a shape that allows a minimum and a maximum value (shaped), such that the possible *net status* of the neurons are inside the continuos domain between 0 and 1 of the function (squishification of the net status). If i have enough excitation in the network status the sigmoid function will give a positive excitation of the neuron, viceversa, it will point out a lower value in the function.

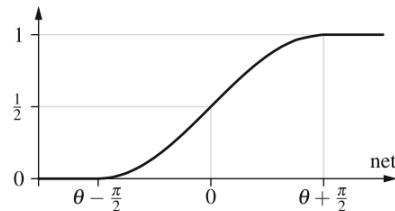
(Heaviside or unit) step function:

$$f_{\text{act}}(\text{net}, \theta) = \begin{cases} 1 & \text{if net} \geq \theta, \\ 0 & \text{otherwise.} \end{cases}$$



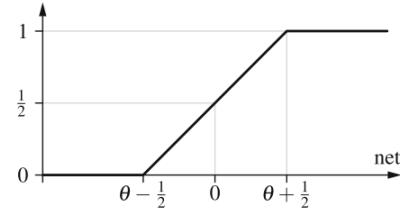
sine up to saturation:

$$f_{\text{act}}(\text{net}, \theta) = \begin{cases} 1 & \text{if net} > \theta + \frac{\pi}{2}, \\ 0 & \text{if net} < \theta - \frac{\pi}{2}, \\ \frac{\sin(\text{net}-\theta)+1}{2} & \text{otherwise.} \end{cases}$$



semi-linear function:

$$f_{\text{act}}(\text{net}, \theta) = \begin{cases} 1 & \text{if net} > \theta + \frac{1}{2}, \\ 0 & \text{if net} < \theta - \frac{1}{2}, \\ (\text{net}-\theta) + \frac{1}{2} & \text{otherwise.} \end{cases}$$



logistic function:

$$f_{\text{act}}(\text{net}, \theta) = \frac{1}{1+e^{-(\text{net}-\theta)}}$$

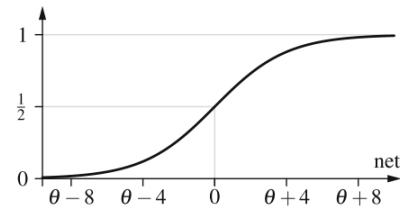


Figure 2.30: Some unipolar sigmoid activation functions

In the contemporary NN the sigmoid function are rarely used, actually the **ReLU** (*Rectified Linear Unit*) function is much more used since it is much easier to train (more similar to the behavior of the biological neurons).

rectified maximum/ramp function:

$$f_{\text{act}}(\text{net}, \theta) = \max\{0, \text{net} - \theta\}$$

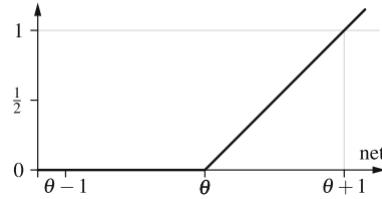


Figure 2.31: ReLU function

The activation function of each output neuron can be a sigmoid or can be a **linear** function (depends on what we want to achieve). There are many types of sigmoid function for representing the general behavior, the most simple is the **step function**.

I have a 0 value until my network excitation reach the value θ , then the output of excitation jump to the strongest status (the value of minimum and maximum activation are generally 0 and 1).

If you prefer a smoother approach you can use the *sin* approach were the derivative is progressively approaching the maximum value.

A **logistic** function is a function with the same shape but that doesn't reach a saturation at the extremes of the range. In this way it is possible to consider the inverse of the function, otherwise it is not possible.

We may have some problems during the learning operation when using the *step function*, since when we have 0 we won't have any effect on the output of the neuron, even if we change the activation the output will be 0. As solution it is possible to use **bipolar** (opposite to the unipolar, which extends only in one pole) sigmoid functions, which changes the ranging to $[-1, +1]$, like the *hyperbolic tangent*.

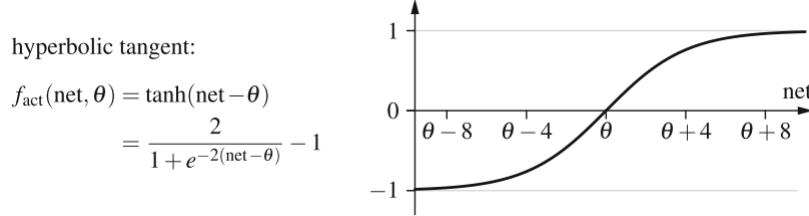


Figure 2.32: Bipolar sigmoid activation function (*hyperbolic tangent*)

I can describe the group of connections between two *consecutive* layers of a multi-layer perception by using a $n \times m$ matrix. Which is the collection of the connection weights between the layers:

$$\mathbf{W} = \begin{pmatrix} w_{u_1v_1} & w_{u_1v_2} & \dots & w_{u_1v_m} \\ w_{u_2v_1} & w_{u_2v_2} & \dots & w_{u_2v_m} \\ \vdots & \vdots & & \vdots \\ w_{u_nv_1} & w_{u_nv_2} & \dots & w_{u_nv_m} \end{pmatrix}$$

Figure 2.33: Matrix describing connection weights between two layers

As a consequence the computation can be described in a very simple way, by using the vectors:

$$\vec{\text{net}}_{U_2} = \mathbf{W} \vec{\text{in}}_{U_2} = \mathbf{W} \vec{\text{out}}_{U_1}$$

You just take the vector of the input neurons of the second layer (which are the output of the preceding layer) and multiply by the weight in between the layers.

Then you apply the sigmoid function σ for squish the values between $[0, 1]$, you can also use a **bias** value for each independent values by using a vector \vec{b} . The bias tells you *how high* the threshold has to be in order for the neuron to fire.

$$\sigma(\vec{W} \vec{in}_{(U_2)} + \vec{b})$$

Let's consider an example with the **bi-implication** three-layer perceptron:

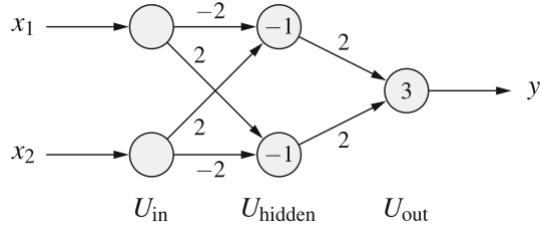


Figure 2.34: Three-layer perceptron for bi-implication

$$\mathbf{W}_1 = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix} \quad \text{and} \quad \mathbf{W}_2 = (2 \ 2)$$

Figure 2.35: Weight matrices for the input and hidden layers

The only difference respect the TLUs network of bi-implication is that each neuron can be an input,hidden and output neuron. In this configuration we are dividing everything in layer of neurons, so we have to **add** a input layer for that (in respect to using TLU).

So far we have seen boolean function, but what we want to look now is how we can extend the dimensions of a feed-forward network in order to consider not only the boolean function of the TLUs but we want to consider in a multi-layer perceptron any type of real number (extend the capability of describing the world).

We can see that the multi-layer perceptron, with any input real-number can be used for function approximation.

Function approximation

Consider the function in the diagram, it is very simple and continuos, and let's consider the points x_1, \dots, x_4 where we want to compute the value of the function and we want to find a method to approximate for any other value x a value in that function.

What i can do is to consider the **midpoint Riemann integration**.

- Approximate a given function by a step function.

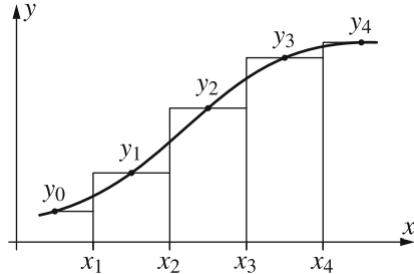


Figure 2.36: Approximating a continuous function with step functions

- Construct a neural network that computes the step function.
- Error is measured as the area between the functions.

It is possible to prove that if you take any Riemann integrable function, it can be approximated with arbitrary accuracy by a *four-layer perceptron*.

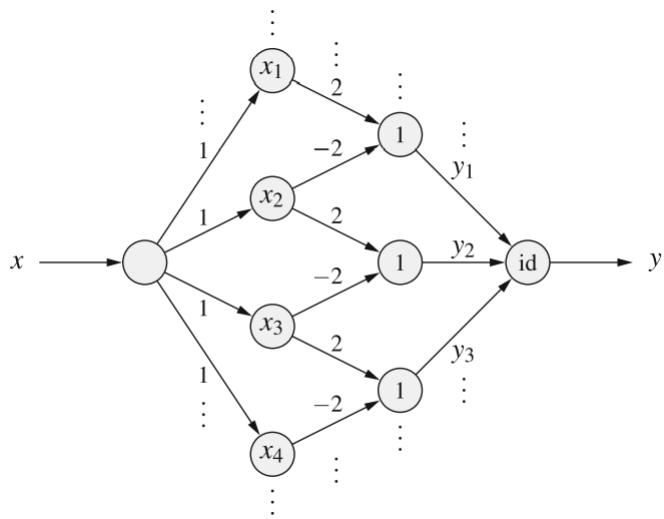


Figure 2.37: NN that computes the step function for integration

The input x is taken by an input neuron, then we have the first layer composed by the subdivision of the domain in the points x_1, x_2, x_3, x_4 .

In the second hidden layer, we create one neuron for each step, which receives input from the two neurons in the first hidden layer that refer to the values x_i and x_{i+1} marking the border of the step. Only one neuron in the second layer can be active, and it will represent the step where the input values lies (this through the **combined excitation** of the neurons in the first hidden layer).

The connection from the neurons of the second layer to the output neuron are weighted with the function values of the stair steps that are represented by the neurons.

Since only one neuron can be active on the second hidden layer, the output neuron receives as input the height of the stair step , in which the input value lies, as result it computes the correct sampling of the function saw in figure 2.36.

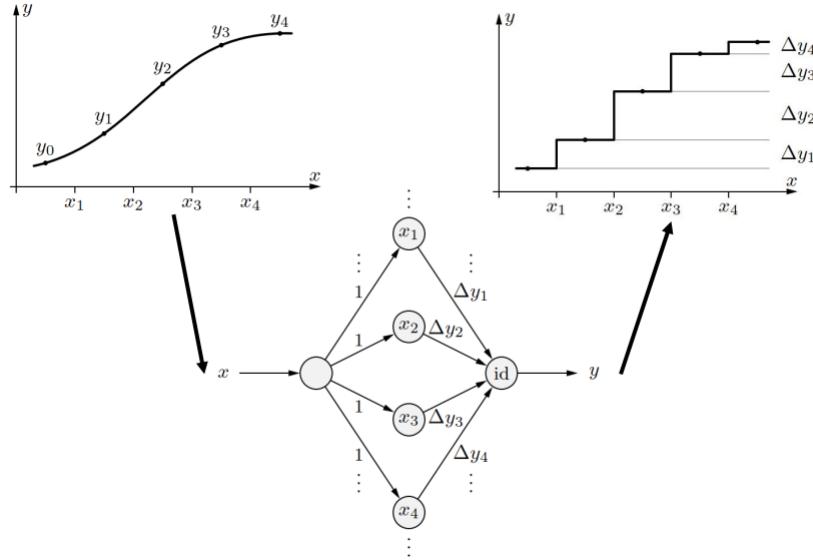
The multi-layer perceptrons can be considered an **universal approximators** of the Riemann integration technique, with the maximum desired error (i can choose the maximum error i want to get for my function by choosing the number of neurons for get that accuracy).

Delta approximation approach

We can actually simplify the error by considering a trick instead of using the combined excitation.

I can split the domain of the function for my network in parts like before, then instead of defining the value of the function for each interval i just use iteratively the Δ variation respect to the value considered.

Figure 2.38



The structure of the network (figure 2.38) will be simpler because i will distribute the input of the hidden neurons, that will tell me if the value x is before x_i or greater than x_i (i don't care where).

For example, if the value is between x_1 and x_2 the first neuron will generate Δy_1 and this will be delivered in output.

If i have a value between x_2 and x_3 i will still have x_1 that will be excited enough and will generate the *contribution* for the output (x_1 and x_2 will be both excited, their contribution will be summed up and then delivered to the output).

This will make simpler the structure of the network.

2.15 Regression

Let's find a way for approximate a function without using the basic calculus approach. In order to do this we have to understand the concept of **regression**. We saw that for training an ANN we need to minimize the error function, which is computed by considering the squared difference between desired output and actual output.

The regression is a technique used in statistical analysis for extrapolating a straight line that better approximate the existing relationship inside a dataset.

Formally, considering the dataset $G = \{(w_0, y_0), \dots, (w_n, y_n)\}$ let's imagine it exists a functional relationship between the input vector w_i and the abscissa y , then the regression will help us to find the parameters of that function. Different kind of functions will give us different kind of regressions.

2.15.1 Linear regression

If we expect that our quantities x and y exhibits a linear dependence, then we have to identify the parameters a and b that extrapolate the straight line $y = g(x) = a + bx$. In general, won't be possible to find a straight line that traverse all the points of the dataset. What we will do is finding a straight line which deviates the least possible, so that minimize the error as follows:

$$F(a, b) = \sum_{i=1}^n (g(x_i) - y_i)^2 = \sum_{i=1}^n (a + bx_i - y_i)^2$$

The necessary condition to find the minimum is given us by mathematical, if i take the partial derivatives between the two parameters a and b (which are the variables in the formula) setting this partial derivatives of the error to 0 it will give to me the condition that allow us to find the minimum (*Fermat's theorem*).

$$\frac{dF}{da} = \sum_{i=1}^n 2(a + bx_i - y_i) = 0$$

$$\frac{dF}{db} = \sum_{i=1}^n 2(a + bx_i - y_i)x_i = 0$$

The *linear algebra* tells me the solution is unique unless i'm so unlucky that all the values x coincides, in that case i have basically one equations but two variable (a and b) and as consequence infinite solutions.

For example, if i need to find the regression of these points:

x	1	2	3	4	5	6	7	8
y	1	3	2	3	4	3	5	6

$$y = \frac{3}{4} + \frac{7}{12}x.$$

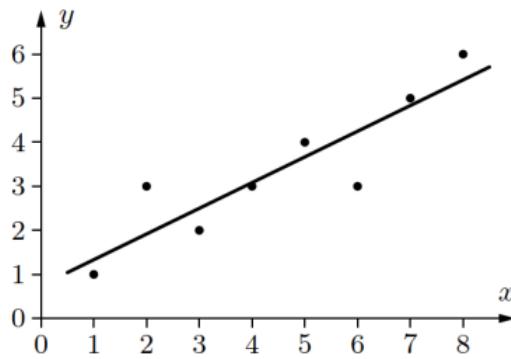


Figure 2.39: Regression line example

2.15.2 Polynomial regression

The previous method is generalizable to polynomial of arbitrary order. The model i have considered is not a good model, is too approximate, i can look if there is instead a very simple straight line a polynomial. Instead of considering a polynomial of grade one, i consider a polynomial of grade m .

$$y = p(x) = a_0 + a_1x + \dots + a_mx^m$$

The minimization fo the error now can be generalized in this way:

$$F(a_1, \dots, a_n) = \sum(p(x_i) - y_i)^2 = \sum(a_0 + a_1x + \dots + a_nx^n - y_i)^2$$

Like in the linear regression, the necessary conditions for the function to be minimized (finding the minimum), are that the partial derivatives respect to the parameters a_i vanishes:

$$\frac{dF}{da_0} = 0, \frac{dF}{da_1} = 0, \dots, \frac{dF}{da_m} = 0$$

- This system can be solved by using the standard methods from linear algebra.
- The solution is unique unless the points lie extactly on a polynomial of lower degree.

2.15.3 Multilinear regression

Until now i just tried to make the approximation more accurate going in the direction of polynomial, *what happens if i have a function in multiple variables?*

$$z = f(x, y) = a + bx + cy$$

If i want to do a multi-linear regression (since i'm applying linear regression to a function of two variables), i have just to apply what i was doing before.

I have to consider the error in the approximated function f , and the actual value given by the samples z_i that i'm collecting (minimizing the sum of squared errors):

$$F(a, b, c) = \sum_{i=1}^n$$

What we need to do? again we need to find when the error is going to 0, this happens when the partial derivatives of the parameters are 0.

$$\frac{dF}{da} = \sum_{i=1}^n 2(a + bx_i + cy_i - z_i) = 0$$

$$\frac{dF}{db} = \sum_{i=1}^n 2(a + bx_i + cy_i - z_i)x_i = 0$$

$$\frac{dF}{dc} = \sum_{i=1}^n 2(a + bx_i + cy_i - z_i)y_i = 0$$

As already said multiple times, the linear algebra will allow me to solve this problem since i have three parameters a, b, c and three equations.

Except the unfortunate case where all points are on a straight line, because in this case i have a plane that can be rotated in any direction and i have *infinite solutions*.

I can consider trying to approximate with a multi-linear regression a function that is defined on m different variables, this is the generalization of what we have seen previously (difficult to draw):

$$y = f(x_1, \dots, x_m) = a_0 + \sum_{k=1}^m a_k x_k$$

I have to minimize the sum of squared errors (vectors for parameters and matrices to define the various values):

$$F(\vec{a}) = (\mathbf{X}\vec{a} - \vec{y})^T (\mathbf{X}\vec{a} - \vec{y})$$

where:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{mn} \end{pmatrix}, \quad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \text{and} \quad \vec{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix}$$

Figure 2.40: Various x values and vector parameters

I need to define the **gradient** which is the generalization of the previous definition defined on vectors and not on the partial derivatives of functions as we seen before. We need to look where the gradient is 0 which will give us the *hyperplane* touching the surface in the minimum.

$$\vec{\nabla}_{\vec{a}} F(\vec{a}) = \vec{\nabla}_{\vec{a}} (\mathbf{X}\vec{a} - \vec{y})^T (\mathbf{X}\vec{a} - \vec{y}) = \vec{0}$$

This comes out as set of regular equations and linear algebra will allow us to solve it as usual unless there is the singular case (if i have all points aligned to a single straight line the plane which is minimizing my error is not defined).

The system of normal equations:

$$\mathbf{X}^T \mathbf{X} \vec{a} = \mathbf{X}^T \vec{y}$$

which has solution unless $\mathbf{X}^T \mathbf{X}$ is a singular:

$$\vec{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y}$$

2.15.4 Logistic regression

In the situation in which the dataset is not approximated with sufficient accuracy from a polynomial function, we could use different kind of functions, for example:

$$y = ax^b$$

We can transform this in a linear equation by applying the operation of logarithm:

$$\ln y = \ln a + b \cdot \ln x$$

In the case of ANN we are interested in particular at the **logistic function**:

$$y = \frac{Y}{1 + e^{a+bx}}$$

If we apply the logarithm of this equation we can derive a different expression of our function y so that the description of this function will be a **linear description**.

Since lot of ANN uses as proper neuron activation function the logistic function, if we would find a way for applying the regression method on that we could determine the parameters of any network.

If we would find a way for applying the method of regression on the neurons, we could determine the parameters of any network at two layers of only one point. The value a of the function represent the threshold of the output neuron, while the value of b represent the weight of the input. We can linearize the logistic function by applying the following transformations (called **logit transformation**):

$$y = \frac{Y}{1 + e^{a+bx}} \longleftrightarrow \frac{1}{y} = \frac{1 + e^{a+bx}}{Y} \longleftrightarrow \frac{Y - y}{y} = e^{a+bx} \longleftrightarrow \ln\left(\frac{Y - y}{y}\right) = a + bx$$

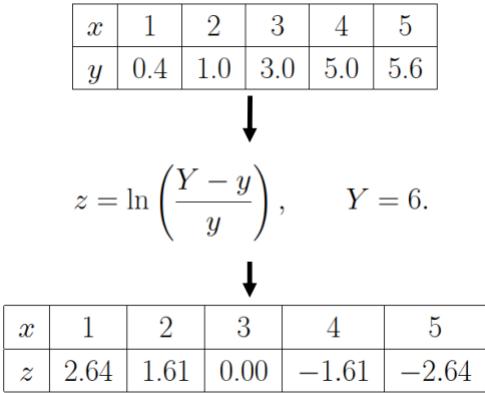


Figure 2.41: Example of logistic regression

The results of the regression line in the example are:

$$z \approx -1.3775x + 4.133, y \approx \frac{6}{1 + e^{-1.3775x+4.133}}$$

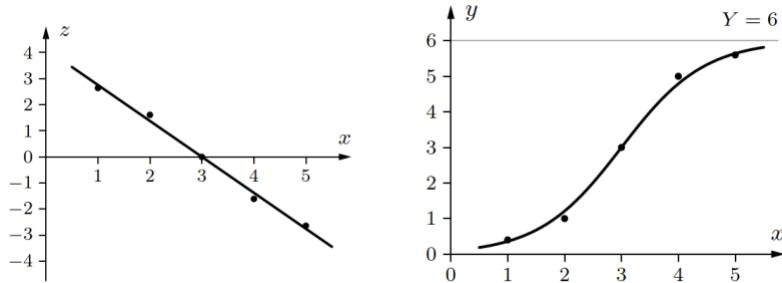


Figure 2.42: Plot of logistic function

How can we understand the values of this non polynomial approximation by transforming it applying the log transformation, so that we can perform the approximation on a function that appear to be linear and then transform it back on the original function.

We can consider a logistic function defined on two variables, and we have this bi-dimensional shape defined on two arguments. we can apply the same principle that we see before, this can be used to solve a problem of separating two groups of examples of different classes (or group).

$$y = \frac{1}{1 + \exp(4 - x_1 - x_2)} = \frac{1}{1 + \exp(4 - (1, 1)^T(x_1, x_2))}$$

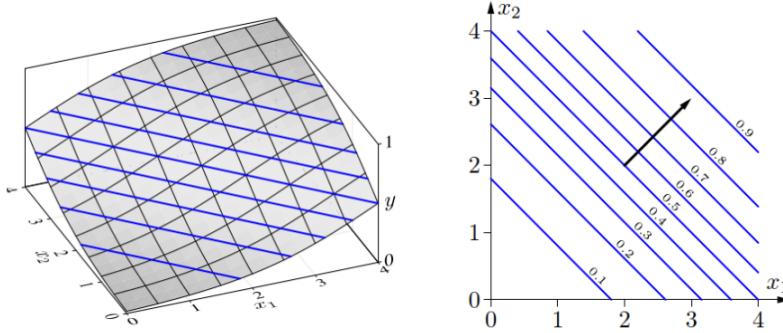


Figure 2.43: Plot of logistic function with two arguments

2.15.5 Two-class problems

Basically the two class problems can be solved by considering the two attributes, which are the name of the class (C is the class of attributes) and a vector random of m dimension.

$$\text{dom}(C) = c_1, c_2$$

We consider the **probability** of belonging to the first class c_1 and the probability to be in the second class c_2 .

$$\begin{aligned} P(C = c_1 | \vec{X} = \vec{x}) &= p(\vec{x}) \\ P(C = c_2 | \vec{X} = \vec{x}) &= 1 - p(\vec{x}) \end{aligned}$$

Given a dataset of points $X = \vec{x}_1, \dots, \vec{x}_n$, each of which belongs to one of two classes c_1 and c_2 .

Basically i want to get a **simple description of the function** $p(\vec{x})$, if i have this function i can apply the **logistic description**:

$$p(\vec{x}) = \frac{1}{1 + e^{a_0 + \vec{a} \cdot \vec{x}}} = \frac{1}{1 + \exp(a_0 + \sum_{i=1}^m a_i x_i)}$$

Then i can apply the **logistic transformation** so that i can obtain the **formal description** of the probability of being on one or the other class:

$$\ln \left(\frac{1 - p(\vec{x})}{p(\vec{x})} \right) = a_0 + \vec{a} \vec{x} = a_0 + \sum_{i=1}^m a_i x_i$$

How can i find the specific value of the probability (and so solving the splitting of the example)? We can consider a specific function called **kernel** that describes how strongly a data point influences the probability estimate for neighboring points.

which give us the correlation in the examples, given some examples how can i say that the example that i observed belongs to the group is influenced by similar examples, or belong to the other group which means that is similar to the examples of the other group.

With a kernel we can describe the similarity of a group of examples, the **Gaussian function** is commonly used. A function like this tell me that if the picked data point is similar to the elements in the middle area, in that case i have a significant value of the function, the more i go far from the values which are similar the less is the value of the function.

$$K(\vec{x}, \vec{y}) = \frac{1}{(2\pi\sigma^2)} \exp \left(-\frac{(\vec{x} - \vec{y})^T (\vec{x} - \vec{y})}{2\sigma^2} \right)$$

If i want to estimate the **probability density** i will use this expression:

$$\hat{f}(\vec{x}) = \frac{1}{n} \sum_{i=1}^n K(\vec{x}, \vec{x}_i)$$

The kernel estimation applied to a two-class problems:

$$\hat{p}(\vec{x}) = \frac{\sum_{i=1}^n c(\vec{x}_i) K(\vec{x}, \vec{x}_i)}{\sum_{i=1}^n K(\vec{x}, \vec{x}_i)}$$

If it is

$$c(\vec{x}) = 1$$

the x_i will belong to class c_1 if $c(\vec{x}) = 0$ otherwise.

2.16 Training multi-layer perceptrons

2.16.1 Gradient descend

3Blue1Brown Conceptually we think about each neuron to be connected to each neuron of the previous layer, and the weighted sum of this neurons expresses the strength of this connections:

$$\sigma(w_1 u_1 + w_2 u_2 + \dots + w_n u_n + b)$$

If we consider all the weights and bias initialized randomly, the MLP network will perform pretty terrible (he is acting randomly).

What you do is defining a **cost function** that can express if the result of the NN is good or bad (for our objective). The greater the function is the greater is the error of the NN respect to desired output.

$$e_v^{(l)} = \left(o_v^{(l)} - out_v^{(l)} \right)^2$$

So then what you do is considering the average cost function over all the samples inside the NN, it is a cost value that express how good or bad the network is.

We want to use this error in such a way that we can perform better during the next execution, let's consider a function $C(w)$ representing just one cost, if we want to find the input that minimize the cost of this function we just have to take the derivative $\frac{dC}{dw}(w) = 0$, but this is not really feasible for complicate functions. You can find different **local minimum** of the function, so what you find isn't the best possible cost function value that you will find. Find the global minimum is pretty hard.

But our function will be much more complicated, since we are dealing with lot of parameters (like the weights and the biases), let's imagine now a function with two variables where the cost function is graphed as a surface above the xy plane. Now, for minimizing the function i should ask my self, in *which direction the $C(x, y)$ function decreases most quickly?*, multi-variable calculus has a concept of **gradient**, which is the direction of the steepest increase $\vec{\nabla}C(x, y)$, and more to that the *length* of that vector expresses how much steep is the increase. The algorithm essentially will be:

- Compute $\vec{\nabla}C$
- Small step in direction $-\vec{\nabla}C$
- Repeat until convergence

The negative gradient of the cost function will tell you how to change all the weights and biases for all the connections, to efficiently decrease the cost (no overshoot). The backpropagation is an algorithm for computing that *crazy* gradient.

The gradient is a vector expressed as n -dimension, where n depends by weights and biases, it is usually a giant number (impossible to visualize). The important thing, is that the magnitude of each component (so the value) will tell you **how sensitive** the cost function output will be respect to the weight and bias (how much it moves respect the previous value).

Lesson notes Now we have to apply this for all the values for all the weights of the network. Essentially when we talk about **training** or the process of learning of a NN it's just about minimizing the cost function (or error function).

More to that, the negative gradient vector will store the relative importance of the weights for each neurons.

What we want to do is to follow the idea of the **gradient descent**, we want to find the *local minimum* of an error function in order to find the proper set of weights.

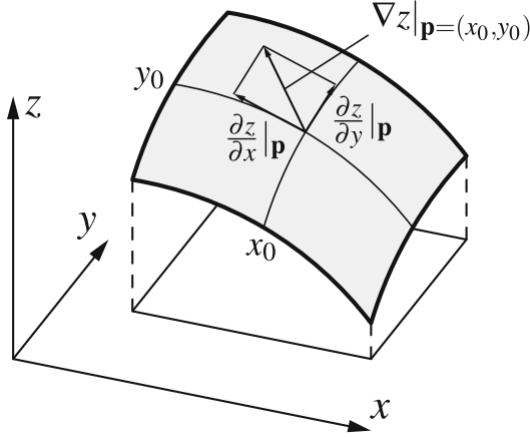


Figure 2.44: Intuitive interpretation of the gradient of a real-valued function $z = f(x, y)$ at a point (x_0, y_0)

Considering this error surface (figure 2.44), if we are in the central point we can look to the derivatives of the error function in the two directions, the **gradient** ∇ is the vector tangent to the surface and will tell us the direction where the surface is more or less increasing.

If we want to find the solution for our learning we need to look for the minimum of this error function, so actually we need to look to the opposite direction to the gradient and we look to reach the minimum.

$$e = \sum_{l \in L_{fixed}} e^{(l)} = \sum_{v \in U_{out}} e_v = \sum_{l \in L_{fixed}} \sum_{v \in U_{out}} e_v^{(l)}$$

So we need to evaluate the gradient to find the direction of the decreasing value, and we need to go in the proper direction in order to reduce the error. In the case of the MLP calculate the gradient means compute the partial derivative of the error function in respect of the threshold and the weights taken as parameters.

Given $\vec{w}_u = (-\theta, w_{u_1}, \dots, w_{u_k})$ as the vector of weights of a single layer extended with the threshold, then compute the gradient as follow:

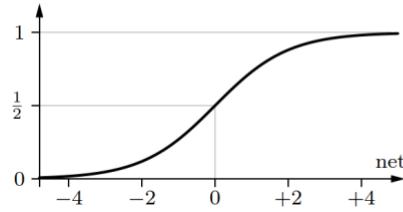
$$\vec{\nabla}_{\vec{w}_u} e = \frac{\partial e}{\partial \vec{w}_u} = \left(-\frac{\partial e}{\partial \theta_u}, \frac{\partial e}{\partial w_{u1}}, \dots, \frac{\partial e}{\partial w_{up_n}} \right)$$

Since the total error e is given by the sum of the individual errors in respect to all neurons and all training patterns l , we get that:

$$\vec{\nabla}_{\vec{w}_u} e = \frac{\partial e}{\partial \vec{w}_u} \sum_{l \in L_{fixed}} e^{(l)} = \sum_{l \in L_{fixed}} \frac{\partial e^{(l)}}{\partial \vec{w}_u}$$

logistic activation function:

$$f_{act}(\text{net}_u^{(l)}) = \frac{1}{1 + e^{-\text{net}_u^{(l)}}}$$



derivative of logistic function:

$$f'_{act}(\text{net}_u^{(l)}) = f_{act}(\text{net}_u^{(l)}) \cdot (1 - f_{act}(\text{net}_u^{(l)}))$$

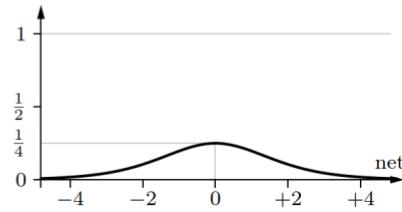


Figure 2.45: Logistic function as activation function

If we have a **logistic function** as $f_{(act)}$ we will have that the changes performed on the vector \vec{w}_u will be proportional to the derivative of the $f_{(act)}$. More near to 0 the value of the function will be, more precipitous will be the learning.

How we can compute the necessary adjustment for each neurons-weight and threshold after we found the error?

The process that allow this is called **error backpropagation**:

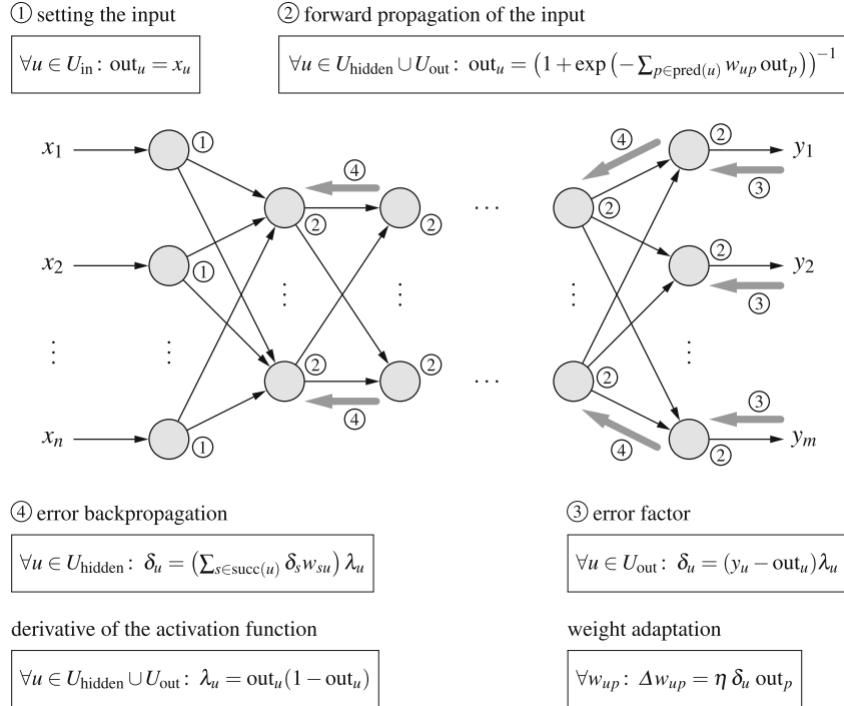


Figure 2.46: Schematized structure of the error backpropagation process

We assume that the activation functions is a *logistic function* for each neuron $u \in U_{(\text{hidden})} \cup U_{(\text{out})}$, except for the input neurons.

1. Apply the input at the input neurons that is returned without modifications at the subsequent first hidden layer.
2. We compute for each neuron of the subsequent layers the weighted sum of the inputs and we apply at the result the logistic function, producing the output that will be propagated in all the network until the terminal neurons.
3. Compute the difference between the desired output and the actual output. since it is possible to invert the logistic function $f_{(\text{act})}$, we can know which was the input (error) that has induced that particular error (δ_u).
4. Now that we have transformed the error of the output variable out_u in the error of the input variable net_u , we can distribute the error (with the correction) in a proportional way back to previous neuron, we back propagate the error until the input neurons.

We have to say that given the shape of the logistic function the error can disappear completely, since the gradient will approximate more the **null vector** the more it will be near the zero.

The weight adaptation is performed by the following formula (this tells me how to perform the correction):

$$\forall w_{up} : \delta w_{up} = \eta \delta_{out_p}$$

If you initialize the *learning rate* η with a too high value instead of descending the curve we risk to jump from a "peak" of the function to another, without ever converging to the minimum. Furthermore, it is not all certain that the minimum reached in this way is the global minimum of the function.

A solution to the problem could consist in *repeating* the learning, initializing the system with a different configuration of weights and threshold, and then choice at the end which configuration result in the better minimum.

Negation example For example let's consider the negation $\neg x$, there is a two-layer perceptron for this function.

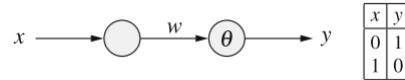


Figure 2.47: Two-layer perceptron with single input and training examples for the negation

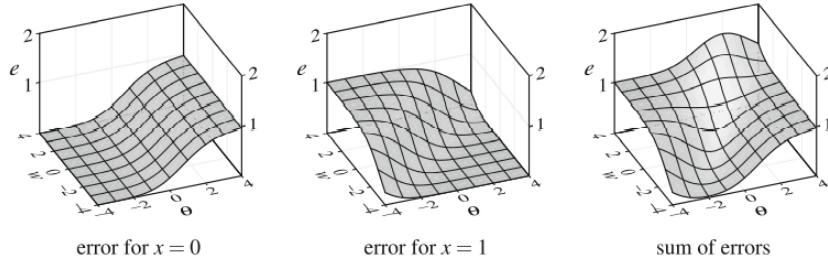


Figure 2.48: Squared errors for computing the negation by using a logic activation function

From a numeric point of view, this procedure won't allow us to reach really the minimum, we will have some residual errors due to the computation of the various parameters (this will change according to the process we are adopting).

2.16.2 Variant of Gradient Descend

The variant aim to change the **learning rate** and the **length of the learning step**.

- *Manhattan training*, taking just the sign of the gradient, very fast but you may not reach the optimum.
- *Flat spot elimination*, it tries to limit the culling of the step length when we get near a function plateau by "lifting" artificially the derivative of the function in that point.
- *Momentum term*, at each successive step we add to the gradient a fraction of the previous with changed weights, so that we can have a sort of memory of how fast it changes respect to the past.
- *Self-adaptive error backpropagation*, it allows to each parameter to have a different learning rate, in this way we can have a grain fine control respect to the characteristic of the single parameter.
- *Resilient error backpropagation*, it's a combination of the Manhattan variation with the Self-adaptive.
- *Quick propagation*, instead of using the gradient to approximate the function with a parable and to jump directly to the peak of this one.
- *Weight decay*, reduces the weights for avoiding to remain trapped in an already saturated region (avoiding too big weights to keep stuck neurons).

2.17 Number of hidden neurons

For a single hidden layer the following rule of thumb is popular (take as granted):

$$\text{Number of hidden neurons} = \left(\frac{\text{number of inputs} + \text{number of outputs}}{2} \right)$$

The problem is that if we have too few neurons in the hidden layer, we will not be able to achieve a good approximation of the function that we need to approximate. We have not enough parameters, this behavior is called **underfitting**.

On the contrary, if we have very large hidden layers we have many neurons, we have many parameters to train and this may force the NN to learn the examples too much, and this will focus the network on these examples only, we lose the ability of generalizing the desired behavior, this is called **overfitting**.

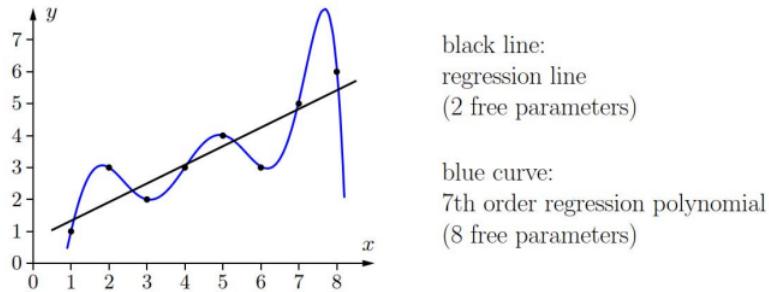


Figure 2.49: The blue curve fits the data points perfectly, not a good model

For example, the NN learned perfectly the blue line to pass exactly in the points, but this is not what we want to have. Now we can't generalize the abstract view of the black line (too many parameters).

What we need to do in order to understand when we have good quality for our learning? Follow, the next subsection.

2.18 Cross validation

In order to understand if we have a good quality for our learning we need to randomly split our data sets in two parts:

- Training set: used for the training.
- Validation set: used for checking the quality of the result.

This will allow us to see (by using the validation set) if the training makes the NN overfitting the configuration or not, if the error that we have is still reasonable low, we have a good quality of the learning (we still can do good generalization), viceversa we have a bad learning quality.

We can do the splitting in two ways:

- **Cross validation:** splits randomly the data in two subsets (for training and validation), we train the MLP with different numbers of hidden neurons on the training data and evaluate them on the validation data. Then you repeat the split of the data and the training-evaluation many times and average the results. At the end you choose the number of hidden neurons with the best average error.
- **N-fold cross validation:** the data is split in n subsets (called *folds*) of about equal size. The relative frequency of the output values in the folds represent as well as possible the relative frequencies of these values in the data set as a whole (**stratification**). Out of these n data subsets, n pairs of training and validation data set are formed by using 1 fold as a validation and $n - 1$ folds for training.

randomly the data in two subsets for training and validation (**cross validation**) and we can repeat this splitting in different phases and at the end we can compare the. or we can split the data set in N-folds (**N-folds validation**) and then use the various n subsets for training and $n - 1$ subsets for validation, in this way we can use most of the example for training instead of validation.

The **stopping criterion** is also very important, we stop the training when the validation error is sufficient low.

2.18.1 Sensitivity analysis

Sensitivity of the NN to the various parameters that we have, the variations that we may have to the various parameters influences the behavior of our NN. We want understand how much the learning applied to the NN is really independent from the behavior of the NN.

This is important to understand how much is generalizable the behavior of our NN.

This is a problem for MLP, when we perform the learning (supervised or unsupervised), basically we have a data set presented to the NN and then the learning algorithm adjust the parameters in ordered to find the desired behavior of the NN.

In our hands we have an algorithm, but we don't have an understanding of the why the parameter are configured in this way. We are trying to give an explanation of why a NN is configured in a specific way, this is actually field of research (we still don't have a rational idea).

For the sensitivity analysis which is part of this understanding, we want to answer this question: *How much the output of each neuron is changing if we change the data set for training* (still describing the desired behavior)? So essentially, which inputs the output(s) react(s) most sensitively, it hints about which inputs aren't needed and may be discarded.

The approach consists in determine the change of output relative to the change of input:

$$\forall u \in U_{in} : s(u) = \frac{1}{|L_{fixed}|} \sum_{l \in L_{fixed}} \sum_{v \in U_{out}} \frac{\partial out_v^{(l)}}{\partial ext_u^{(l)}}$$

2.19 Deep learning

A modern approach to MLP called **Deep learning**, with MLP we can approximate any continuos integrable function, but we have in general a pretty large hidden layer. First problem, *what we can do to simplify the structure?* The second problem, with MLP we have to know the characteristic of the data which are relevant to pick the important data to configure the NN.

It has been experimented that adding 1 or 2 more layers helps to have a concise NN but with the same characteristic of a MLP with a large number of neurons in the hidden layers.

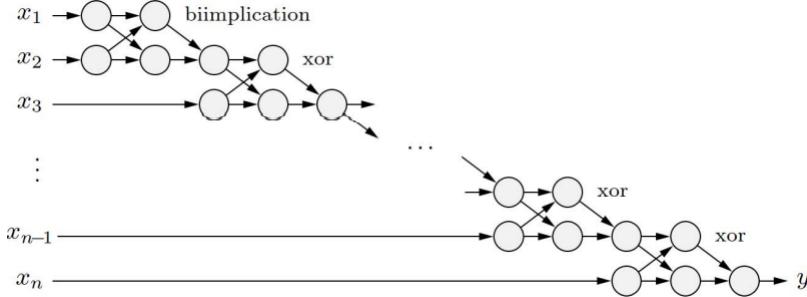


Figure 2.50: n -bit parity function

The deep learning network is basically a MLP, but with a difference: What we want to do it consists in not having any type of knowledge about the problem and try to force the NN to configure itself with a limited and small number of neurons.

For example, let's consider the n -bit parity function, which output 1 if the n -bit word is even, or viceversa 0. If we want to use an MLP with only one hidden layer, which has 2^{n-1} neurons, the number of hidden neurons grows **exponentially** with the number of inputs. This because the function is a disjunction of 2^{n-1} conjunctions. Instead if we use **multiple** hidden layers, the **linear growth** is possible (respect to the input). This finding bring to the developing of *deep learning*, where the "*depth*" is the one given by the longest path which separates the input neurons from the output neurons.

The rational is to allow a greater deepness of the network in change of better performance in construction and computation of the NN. The deep learning bring some downsides:

- *Overfitting*: the increases of neurons given from the extra hidden layers can multiply the parameters in a disproportional way.
- *Vanishing gradient*: during the propagation of the error the gradient will get reduced until the vanishing.

Some solutions to the overfitting issue:

- *Weight decay*: set a maximum limit to the values which can assume weights this for preventing an adjustment too much dependent from the data set.
- *Sparsity constraint*: introducing some limit to the number of neurons of the hidden layers, or limiting the number of the active neurons.
- *Dropout training*: some neurons of the hidden layers are omitted during the evolution of the network.

While the issue of the vanishing gradient is given by the fact that the activation function it's a logistic function, which derivative reach at max the value

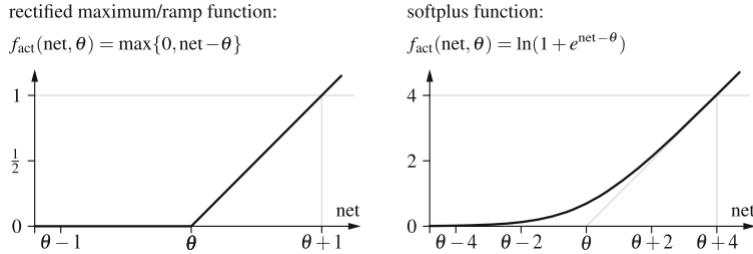


Figure 2.51: Ramp function and softplus function

of $\frac{1}{4}$. As consequence of this the propagation of the error to a previous layer adds a value, often smaller than 1, and in this way reducing the gradient.

A solution consists in editing slightly the activation function in a way that will be always increasing. Some candidates for that are the *ramp function* and the *softplus function*. A totally different approach consist in building the network "layer by layer". A much used technique consists in think at the NN like a stack of *autoencoder*. An autoencoder is an MLP which maps the input inside an approximation, using an hidden layers of less dimensions. The hidden layer works like an *encoder* by encoding the input in a internal representation which will be decoded to the output layer. The autoencoder, has only one layer, it doesn't suffer about the same limitation and can be traversed normally from the backpropagation.

A problem with this approach is that if there are a lot of neurons in hidden layer as much are in the input layer we are going to propagate with less adjustments, and without the autoencoder extracting any knowledge from the data.

There are three possible solutions:

- *Sparse autoencoder*: provide to use a much smaller number of neurons inside the hidden layers, respect to the input layers. The autoencoder will be forced to extract from the input some features instead of just propagating the data.
- *Sparse activation scheme*: in a similar way for avoiding the overfitting, we decide to "shutdown" some neurons during the computation.
- *De-noising autoencoder*: we add randomly some noise to the input.

For obtaining an MLP with multiple layers we combine differente autoencoder. Initially we start by training just one autoencoder. At that time you remove the decoder and you preserve only the internal layer. You use the data processed from this autoencoder for training the second, and so on until you reach a satisfiable number of layers. MLP constructed in this way are really efficient at recognizing with success hand written numbers. If we would want to use similar NN for a more broad class of applications, like, for example, the

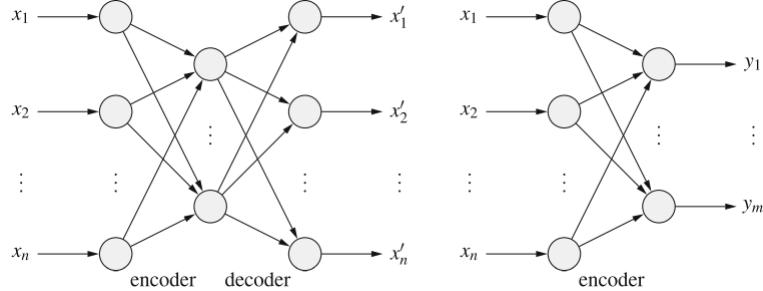


Figure 2.52: Autoencoder and autoencoder with the decoder removed

feature recognized by the internal layers are not located in a specific portion of the image, we would need to use a CNN (*Convolutional Neural Network*, next topic). This architecture is inspired by the working of the human retina, where the neurons used for the perception has a receptive field, or in other words, a limited region where they can answer to stimulus.

This is simulated inside the CNN by connecting the neurons of the first layer just to some neurons of the input. The weights are partitioned in a way that partial networks can be evaluated from different prospective of the image.

During the computation it will proceed by moving the receptive field on the totality of the image. As result you will obtain a convolution of the weight matrix with the input image.

2.20 Radial Basis Function Networks

This is another NN model, we have a special type of structure for the neurons and a special approach for information we are collecting from images (typically).

I have a 3-layered feed forward network, where the activation function of the hidden layer is a *Radial Basis Function* (RBF).

An RBF is a function which try to point out what is relevant in a very specific area of the information we are observing (an image), and tries to diminish what is in the surrounding area. So it will focus on relevant part of the image where we want to understand the contents.

In this kind of NN we have a set of values for various inputs, we represent this with a vector, we have for example a three dimensional space, where each value represent the input. What i'm considering as input of the hidden layer is the difference between this vector and the weight vector, which is the distance between the vectors.

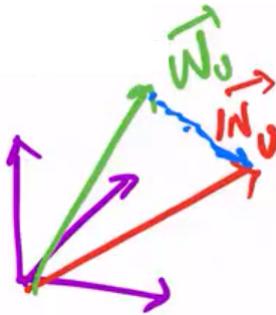


Figure 2.53: Distance between \vec{w} and \vec{in}_u

There are different families of distances, we typically consider the **Minkowski Family**, which are distances defined in general by this formula:

$$d_k(\vec{x}, \vec{y}) = \left(\sum_{i=1}^n |x_i - y_i|^k \right)^{\frac{1}{k}}$$

- $k = 1$: Manhattan distance.
- $k = 2$: Euclidean distance.
- $k \rightarrow \infty$: Maximum distance.

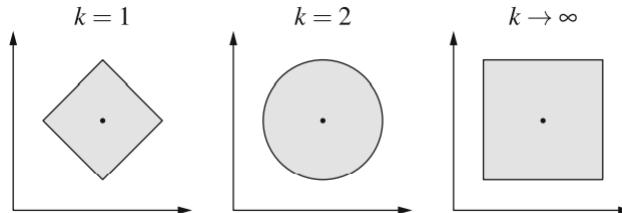


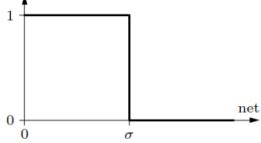
Figure 2.54: Manhattan, Euclidean and Maximum

The RBF has an activation function in the output layer which is **linear**. Differently, the hidden layer has an activation function which is a **radial function**. A radial function decreases from 0 to ∞ in a monotonically way.

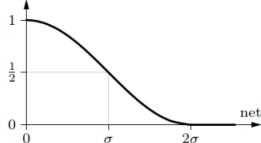
$$f : \mathbb{R}_0^+ \rightarrow [0, 1] \text{ with } f(0) = 1 \text{ and } \lim_{x \rightarrow \infty} f(x) = 0$$

The size of the *catchment region* of the function is defined by the **reference radius** σ , the bigger is the σ the bigger is the area observed by the neuron. The various parameters and the shape of the function determine the width of this radius. The radial activation function are typically shaped in this way:

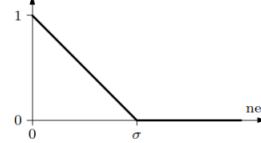
rectangle function:
 $f_{act}(\text{net}, \sigma) = \begin{cases} 0, & \text{if net} > \sigma, \\ 1, & \text{otherwise.} \end{cases}$



cosine until zero:
 $f_{act}(\text{net}, \sigma) = \begin{cases} 0, & \text{if net} > 2\sigma, \\ \frac{\cos(\frac{\pi}{2\sigma}\text{net})+1}{2}, & \text{otherwise.} \end{cases}$



triangle function:
 $f_{act}(\text{net}, \sigma) = \begin{cases} 0, & \text{if net} > \sigma, \\ 1 - \frac{\text{net}}{\sigma}, & \text{otherwise.} \end{cases}$



Gaussian function:
 $f_{act}(\text{net}, \sigma) = e^{-\frac{\text{net}^2}{2\sigma^2}}$

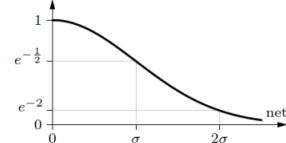


Figure 2.55: Radial activation functions

As example, let's apply an RBFN for simulating the boolean conjunction $x_1 \wedge x_2$. When I'm around the black point, the input is $(1,1)$, what we want to do is creating a circle around that area where the output is 1 (significant information to extract).

Another approach could consist in considering a RBF in the opposite way, where the interesting data is outside the circle.

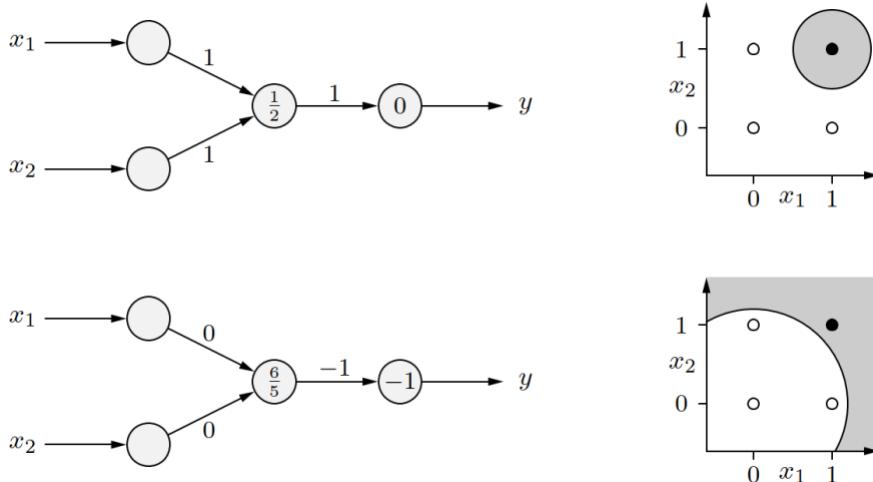


Figure 2.56: RBFN for conjunction

Another example is with the bi-implication $x_1 \longleftrightarrow x_2$:

$$x_1 \leftrightarrow x_2 \equiv (x_1 \wedge x_2) \vee \neg(x_1 \vee x_2)$$

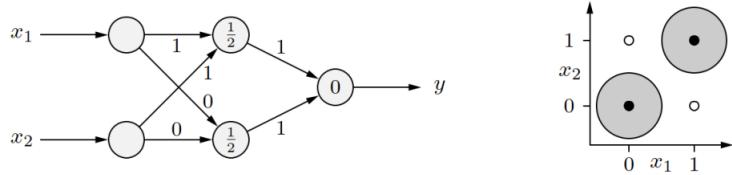
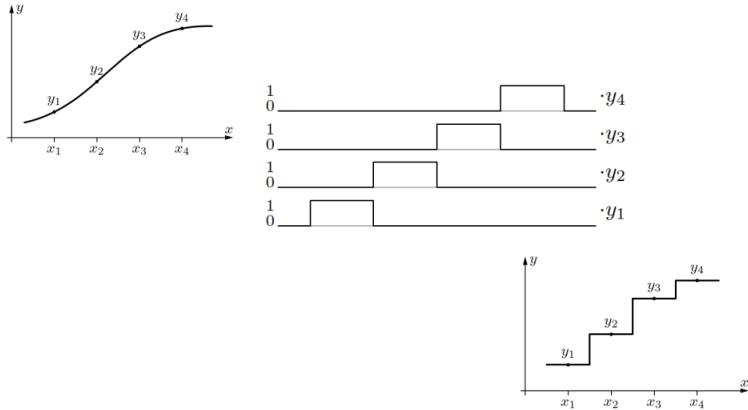


Figure 2.57: RBFN for bi-implication

Function approximation We have two neurons focusing on the relevant part of the domain, the rest will be 0, and by merging this two information you will get the bi-implication. In general an RBFN has the same expressive power of an MLP, and it can be seen as an universal approximator (can approximate any Riemann integrable function) with a small arbitrary error.

The process is the same of the other NN, the function will get approximated with a step function that can be computed easily by a RBF if we define that as the weighted sum of rectangular functions.



Each pulse can be represented by a neuron of a RBFN: The approximation can be enhanced by increasing the number of points where the function is being evaluated. Furthermore, if instead of using the rectangular function, we use the Gaussian we obtain more "soft" transitions, avoiding big jumps (like for the MLP).

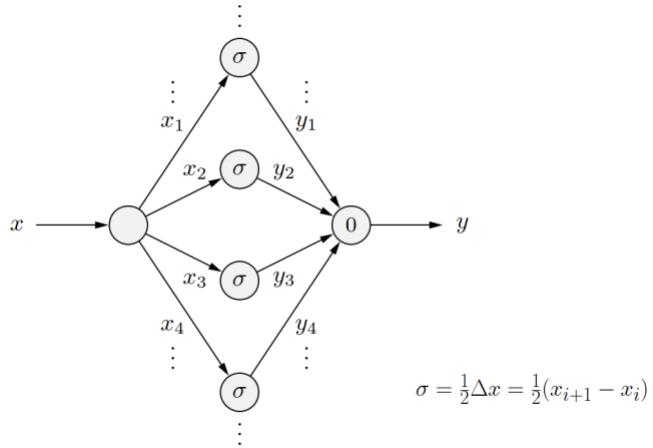


Figure 2.58: RBFN for the approximation, using the step function

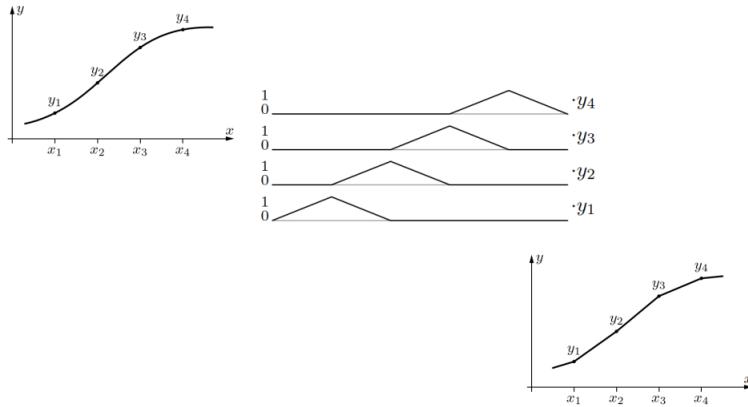


Figure 2.59: RBFN for the approximation, using the Gaussian function

2.21 Training RBFN

We want to apply in principle a sort of backpropagation in order to perform the configuration, like with the FFN. The principle is always the same, having in mind the *minimization* of the final error, but the problem here is *how we can decline this by taking into account the specific structure of this neurons?* With the others ANN the initialization phase was trivial, because you just had to select some random values, instead with RBFN the same approach conduce to suboptimal results.

Let's consider a **simple radial basis function network**, which is a partial view of the general structure of a RBFN, where each training example is asso-

ciated with a proper radial function.

We want to apply a **fixed learning task** $L_{fixed} = \{l_1, \dots, l_m\}$ with m training patterns $l = (\vec{l}^{(l)}, \vec{o}^{(l)})$ composed by input and desired outputs.

We have one hidden neuron $v_k, k = 1, \dots, m$ for each training pattern, let's define the vector of associated weights to the neuron v_k :

$$\forall k \in \{1, \dots, m\} \vec{w}_{v_k} = \vec{l}^k$$

Let's consider the activation function of our neuron equal to the sigma specific to that neuron, the radius can be computed for each of the neuron by considering heuristic approach in which we take the distance between the various input patterns and we take the maximum distance computed between each pair of input vectors (patterns) in our dataset (this value is expressed as d_{max}).

The radius of the neuron σ_k will be choose heuristically if the activation function is Gaussian function:

$$\forall k \in \{1, \dots, m\} : \sigma_k = \frac{d_{max}}{\sqrt{2m}}$$

$$d_{max} = \max_{l_j, l_k \in L_{fixed}} d(\vec{l}^{(l_j)}, \vec{l}^{(l_k)})$$

Where d_{max} is the maximum distance between the input vectors. This choice allow to center the various Gaussians in such a way that them doesn't overlap each other, instead they will distributes in an ordered way respect to the input space. Basically we have a structure where we try to learn with each neuron one of the input patterns, one neuron trying to well understanding a specific input pattern.

I can initialize the connections from the hidden to the output neurons by using weights, this can be done by considering this formula:

$$\forall u : \sum_{k=1}^m w_{u v_m} out_{v_m}^{(l)} - \theta_u = o_u^{(l)}$$

which is the weighted summation of the output of the hidden layer and considering the threshold of the output, which has to be equal to the **desired output**. We know the desired output, we have the threshold and the current output generate by the hidden layer, so the last thing we need is the weights, this can be obtained by multiplying the interconnections weight by the matrix A of the output of the hidden layer (by using $\theta = 0$):

$$\mathbf{A} \cdot \vec{w}_u = \vec{o}_u$$

Where \mathbf{A} is the $m \times m$ matrix that has for components the various output of the neurons in the hidden layer. If the matrix \mathbf{A} has the complete rank, we can invert that and compute the weight vector:

$$\vec{w}_u = \mathbf{A}^{-1} \cdot \vec{o}_u$$

$$\mathbf{A} = \begin{pmatrix} 1 & e^{-2} & e^{-2} & e^{-4} \\ e^{-2} & 1 & e^{-4} & e^{-2} \\ e^{-2} & e^{-4} & 1 & e^{-2} \\ e^{-4} & e^{-2} & e^{-2} & 1 \end{pmatrix} \quad \mathbf{A}^{-1} = \begin{pmatrix} \frac{a}{D} & \frac{b}{D} & \frac{b}{D} & \frac{c}{D} \\ \frac{b}{D} & \frac{a}{D} & \frac{c}{D} & \frac{b}{D} \\ \frac{b}{D} & \frac{c}{D} & \frac{a}{D} & \frac{b}{D} \\ \frac{c}{D} & \frac{b}{D} & \frac{b}{D} & \frac{a}{D} \end{pmatrix}$$

where

$$D = 1 - 4e^{-4} + 6e^{-8} - 4e^{-12} + e^{-16} \approx 0.9287$$

$$a = 1 - 2e^{-4} + e^{-8} \approx 0.9637$$

$$b = -e^{-2} + 2e^{-6} - e^{-10} \approx -0.1304$$

$$c = e^{-4} - 2e^{-8} + e^{-12} \approx 0.0177$$

$$\vec{w}_u = \mathbf{A}^{-1} \cdot \vec{o}_u = \frac{1}{D} \begin{pmatrix} a+c \\ 2b \\ 2b \\ a+c \end{pmatrix} \approx \begin{pmatrix} 1.0567 \\ -0.2809 \\ -0.2809 \\ 1.0567 \end{pmatrix}$$

Figure 2.60: Example of the computing the weight vector in the bi-implication problem

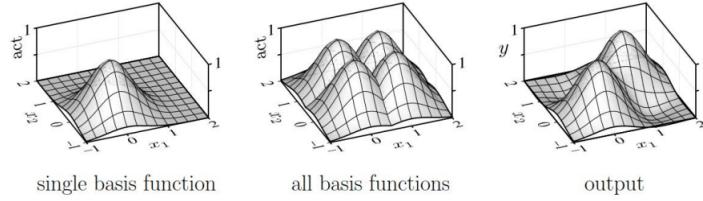


Figure 2.61: RBFN function centered in the examples

This method guarantee a perfect approximation. This means that is not necessary training a simple radial basis function network. In general, if we don't want to have for each pattern a neuron, we will have to select k subsets from the dataset and find, for each subset, a representative che assoceremo to a neuron in the hidden layer.

if we look graphically, for each of the hidden layer we have aRBFN shaped as a Gaussian, which will correspond at each example that the neurons will learn. We will have only two of this Gaussian in the final computation, we will generate an output that will have to be different from 0. This is a RBFN centered in the relevant example that we want to learn.

This is not the general case, if we take a structure like this and we want to create a RBFN having in mind that for each the pattens we have to contribute we need to put a neuron with RBFN as activation function, this implies that if we have many examples that we want to learn we are going to have an incredibile big structure for our network, and this is not smart.

If we are considering the general radial basis function structure in which

we don't want to have a neuron with x RBFN network centered on a specific example to learn, we need to try to limit the number of neurons we need to understand which are the neurons which are relevant and understand where to place it. Basically we need to select among the m patterns that we want to learn a subset of k patterns and then we need to put the neuron centered on this example, then we need to figure out how to build the other outputs by combining this limited number of examples that we learn.

If we want to use this kind of approach we basically need to understand where to place the center, but we need also to understand how large will be the radius in order to be sure that after placing the Gaussians we are able to recreate the examples for which we don't use the Gaussian.

What i can do is splitting the domain of examples, and then for each subdomain find a representative point (not an example) in the subdomain, that will cover the group of the examples (not points) of the subset.

This can be formalized in saying that we select k training patterns as centers, where the connection weights for the hidden neurons are the training pattern, and the connection weights for the output neurons can be obtained by considering this matrix A :

$$A = \begin{pmatrix} 1 & out_{v_1}^{(l_1)} & out_{v_2}^{(l_1)} & \dots & out_{v_k}^{(l_1)} \\ 1 & out_{v_1}^{(l_2)} & out_{v_2}^{(l_2)} & \dots & out_{v_k}^{(l_2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & out_{v_1}^{(l_m)} & out_{v_2}^{(l_m)} & \dots & out_{v_k}^{(l_m)} \end{pmatrix}$$

Figure 2.62: Matrix that expresses the connections weights for output neurons

Now i can apply the invert the matrix for computing the weights since system over-determined (not squared), but i can consider the pseudo inverse matrix:

$$A^+ = (A^T A)^{-1} A^T$$

Now we can apply the actual computation for our vector \vec{w} of the weights, given by:

$$\vec{w}_u = A^+ \vec{o}_u$$

So now, *how we can choose the center for creating the configuration for our RBFN?* As for the simple case we can take all points we just need to determine the radius for one single radius and the output weights (considering that we want to use a better structure, which is practicable).

Options of radial basis function centers:

- All data points as centers, the simple case where only radius and output weights need to be determined. The output values ca be achieved exactly, computing the weights is unfeasible.

- Random subset, this is fast, only radius and output weights need to be determined. The performance depends on the choice of data points.
- Clustering, c-means clustering (the centroid point representative for the examples in the subset).

2.21.1 C-means clustering

This is one of the best options for choosing the RBFN centers:

1. Considering a number of c clusters to be found (input).
2. Initialize the cluster centers **randomly**.
3. Assign initially the data points (the examples) to the nearest center (or *centroid*).
4. Adjusting the position of the center considering the position of the assigned data points (mean vectors of the assigned data points, this is called **cluster center update**).
5. Repeat the step 3 and 4 until clusters do not move anymore (converge).

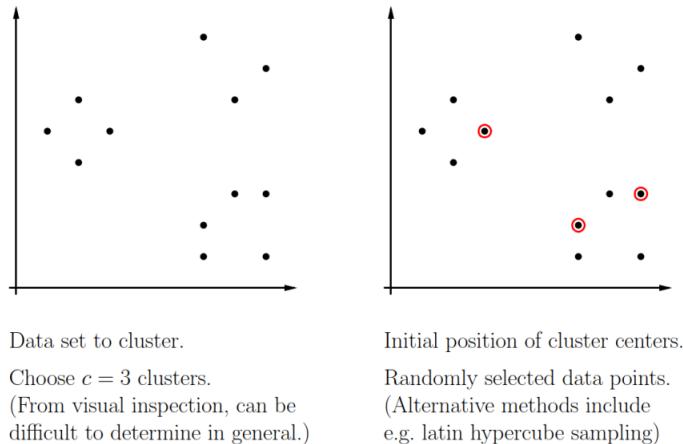


Figure 2.63: Step 1 with $c = 3$, Step 2 randomly choosing the centroids

Then by using the **Delaunay Triangulation** of the centroids, which leads to the **Voronoi Diagram** by taking the line orthogonal to each one of the three edges. This will result in a tessellation of the domain of the function, basically each tassel of the domain is a cluster.

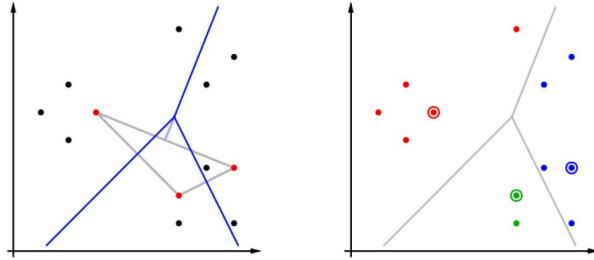


Figure 2.64: Delunay Triangulation (left) and Voronoi Diagram (right)

After that i'm going to search for the new centroid of each cluster, where the sum of the distances is minimal, after finding this element i will consider it as the new center. Then i will apply again the same partition (Delunay-Voronoi) as before.

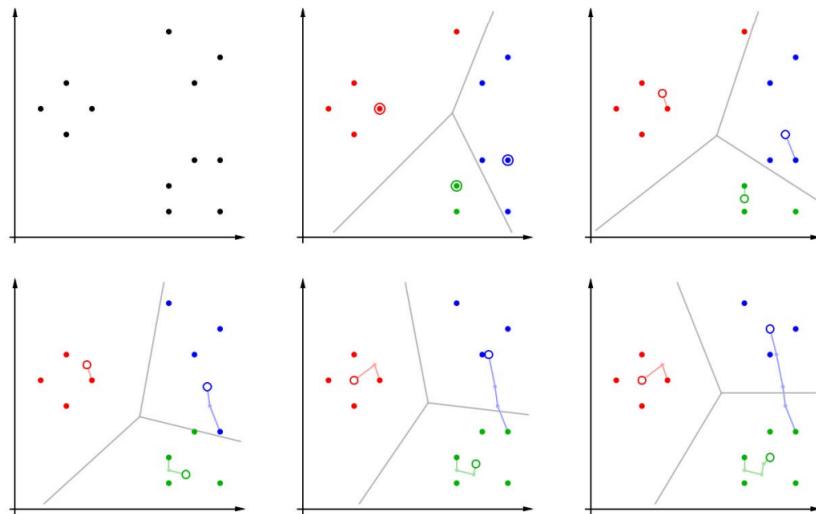


Figure 2.65: Centroid update

Basically what we can do is to apply some rules for updating the weights that are similar to the backpropagation in the MLP; basically you have just to adjust the weights from the hidden to the output neuron we just need to compute the gradient:

and then we need to find the direction of the maximum slope for the gradient and we need to compute the increase of the decrease in order to go in the d

For the hidden layer we need to place the centers, we need to not only find the interconnection weights for the centers but also the size of the radius which allow to combine together the contribution of the various neurons in the

hidden layer for generating the desired output with the requested minimum error.

We have placed the neurons. Now we need to understand how big the radius is big for the neuron. The gradient of the error with respect to the radius :

Update rule for enlarge or reduce radius:
special case of Gaussian activation function:

2.22 Learning Vector Quantization

What happens if we don't have examples with a desired output (only the examples)?
Basically, until now we saw the fixed-learning task, now we want to see what happens with a free-learning task.

Now I'm going for searching outputs that will be similar to the given input.

We still consider the Delaunay and Voronoi approaches. The problem is that what we want to do is a suitable quantization of the domain in which we have our examples in order to understand the group of the examples which are similar. In the previous example we had given a number of clusters, now we want to find the clusters in a given dataset.

Let's consider the input examples (or data points) as empty circles, and the cluster centers (or points) by the full circles.

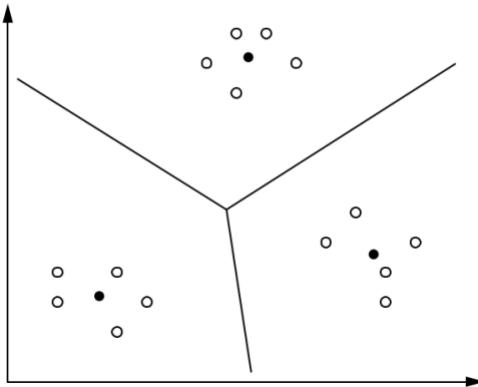


Figure 2.66: Vector quantization

The **Learning Vector Quantization Network** (LVQ/LVQN), is a FFNN with 2-layers that performs this clustering of the examples by similarity, basically from an abstract point of view we see the LVQ as a radial function in which the hidden layer and output layer are collapsed in a single layer.

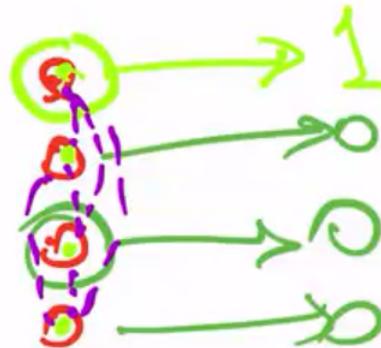
The network input function of each output neuron is the distance between the input vector and the weight vector. The definition of the distance is the

same, we are considering the Mynkowsky family and the euclidean distance for represent the behavior of our function,

The output activation function of each neuron is not a simple function of the activation of the neuron: it considers the activation of **all** output neurons.

$$f_{out}^{(u)}(act_u) = \begin{cases} 1 & \text{if } act_u = \max_{v \in U_{out}} act_v \\ 0 & \text{otherwise} \end{cases}$$

We don't just look locally, for example with 4 neurons we will look to the value of the activation in the activation function of each neuron and we compare the activation value of each one of them. The biggest activation value will lead to an output 1 otherwise 0 (when the activation is smaller). We have one neuron winning over all the other, basically we have a connection between all neurons in the output layer such that each of them is looking to what the other neurons are doing (the activation), this will allow the confrontation.



In the LVQN we can use the same radial functions that are used for any RBFN.

How can we perform the configuration? Basically for each training pattern we look to the *reference vector* which is we find the closest reference vector.

I may have basically two strategy that I can consider in order to perform the adjustment of the weights:

- **Attraction rule**, when the data point and the reference vector have same class:

$$\vec{r}^{(new)} = \vec{r}^{(old)} + \eta(\vec{x} - \vec{r}^{(old)})$$

- **Repulsion rule**, when the data point and the reference vector have different class:

$$\vec{r}^{(new)} = \vec{r}^{(old)} - \eta(\vec{x} - \vec{r}^{(old)})$$

where x is the input, r is the reference vector for the winner neuron, and η is the learning rate. Until now we meant the learning rate as fixed during the

learning, by the way there are some situations where the constant learning rate can bring some problems.

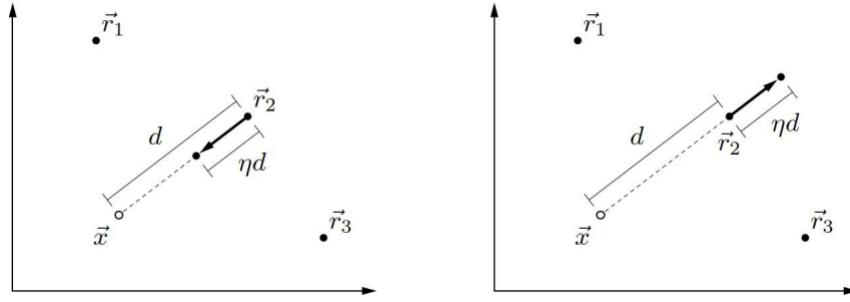


Figure 2.67: Adaptation rules applied on the reference vector

In the case of the attraction rule we move the reference vector ηd towards x which is the same class of the reference vector. Or in the right diagram where we are moving the reference far from the reference x so that that the element moves to another class.

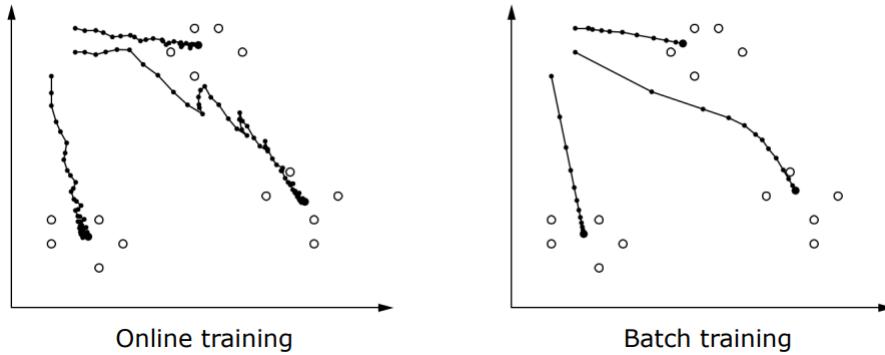


Figure 2.68: Adaptation rules in case of online and batch learning

In the case of the online training we have a step each time according to the various pattern I'm presenting, in the batch training I apply the various connection for the group of examples belonging to the same pattern and perform the update at the end of the entire batch (the solution of the position of the centers is smoother since I am putting together more than one example).

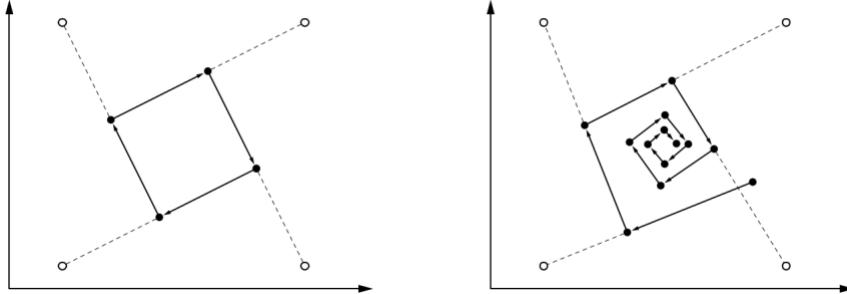


Figure 2.69: Fixed learning rate to time dependent learning rate

The problem i have with this kind of techniques, is that if i consider a fixed learning rate, which is a coefficient which tells me how much i want to change the position by taking into account the similarity.

I may enter in a oscillation (left diagram), in this case what we can do is to dynamically change the learning rate consists by decreasing it as it grows iterations (called **time dependent learning rate**, 2.69 on right).

There are some applications in which i may have some more information about the data tha i have, **classified data**, i may have some more information related to the class to which an example belong and even if i'm not using the supervised learning i can take advantage of this information so that i can update not only the reference vector but i can consider updating also some other reference vector the once which are the nearest the reference vector that i'm considering.

The problem is that this may move the classes very far one from the other, one option to balance this is to consider the *window rule*: a reference vector weill become adapted only if the point p is near the border of the classification, the hyper-surface which separates the contiguous regions of the two classes. The notion of closeness is formalized as :

$$\min \left(\frac{d(\vec{p}, \vec{r}_j)}{d(\vec{x}, \vec{r}_k)}, \frac{d(\vec{x}, \vec{r}_k)}{d(\vec{x}, \vec{r}_j)} \right) > \theta \text{ where } \theta = \frac{1 - \xi}{1 + \xi}$$

where ξ is a parameter specified from the user, which describes the "amplitude" of the window around the border of the classifications. If we assume that the data are randomly chosen from a random set of normal distributions, we could use a *soft* approach, in opposite to a "crisp" division typical of the clustering like the *c-means*. Basically we don't have a RB function which is rather sharp but i'm considering a smooth border between the classes induced by sufficiently large RBF. In this cases i can considering the classing by a mixture og Gaussian which are representing the various classe, with a softer border between the various classes.

In order to do this i have to consider that each element is not interbelonging to one class or another, each element of my data set as a probability of belong-

ing to one or the other class. This is due to the fact that i have a gaussian shape of the radial basis function, In this case what i will do is essentially to evaluate the probability for each example to belong ot each of the classes introduced in my system. Formal definition:

$$f_{\vec{X}}(\vec{x}; \mathbf{C}) = \sum_{y=1}^c f_{\vec{X}, Y}(\vec{x}, y; \mathbf{C}) = \sum_{y=1}^c p_Y(y; \mathbf{C}) \cdot f_{\vec{X}|Y}(\vec{x}|y; \mathbf{C}).$$

\mathbf{C}	is the set of cluster parameters
\vec{X}	is a random vector that has the data space as its domain
Y	is a random variable that has the cluster indices as possible values (i.e., $\text{dom}(\vec{X}) = \mathbb{R}^m$ and $\text{dom}(Y) = \{1, \dots, c\}$)
$p_Y(y; \mathbf{C})$	is the probability that a data point belongs to (is generated by) the y -th component of the mixture
$f_{\vec{X} Y}(\vec{x} y; \mathbf{C})$	is the conditional probability density function of a data point given the cluster (specified by the cluster index y)

Figure 2.70: Probability function

If i take the various examples of my dataset i have a set of clusters in my dataset with the parameter that defines the cluster, and for each element of the dataset i want to understand the prbability of belonging to one class respect to another.

Mathematically this is very complex to be optimized, and as a consequence we try to perform an approximation introducing in an explicit way the position of the various Gaussians with additional random variables which are difficult to compute it in a exact way, we compute a maximum expected value for that, and we try to get the maximum of the likelihood.

Basically what we need to do to perform the maximization of the likelihood, we work in two passes since we have the centers, and the distribution of our variables, what we can do basically we optimize the initial group of centers of the clusters and after this we can compute the maximum likelihood on the data we have. The what we do we do maximization.

1. Computing the center
2. Computing the normal distribution

The final result will be an approximate value of the density function, of belonging to each specified class for each data that we have in the data set. We will have a softer definition of the border but this will bring a much more complex analysis.

2.23 Self-organizing maps

Sometimes are called also *Kohonen maps*, introduce from the Prof. Kohonen. It is a two-layer FFNN similar to the VQLN, instead of having in the output layer

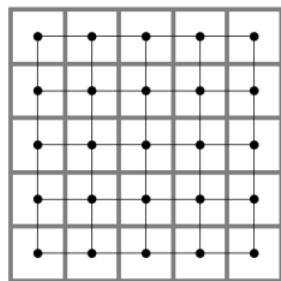
all the other output neurons, the self-organizing map has local connections only among neighboring hidden/output neurons.

We do not have the winner take all over the entire set of output neurons but we work local, this is applied on the neighbor of each neuron. The network input function is the radial function, as before, the

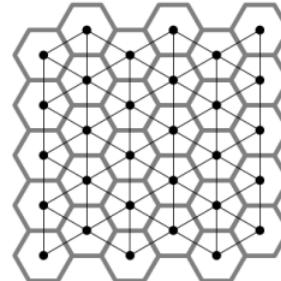


Figure 2.71: Example of neighbor connections for each neuron

By the way in SOM the neighbor are defined in a bi-dimensional space:



quadratic grid



hexagonal grid

Figure 2.72: Neurons expressed as squared or hexagonal grid

the black ones are representing the neurons, where the surrounding square (or hexagon) of a neuron is the neighbor. The grey lines represents the regions assigned to a neuron.

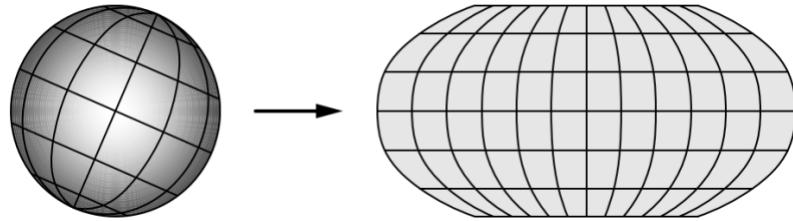


Figure 2.73: The topology is preserved

basically in this kind of structure we want to analyze the data and we want to have a winner for each data, but what we want also is to preserve the neighborhood, so the topology of the structure.

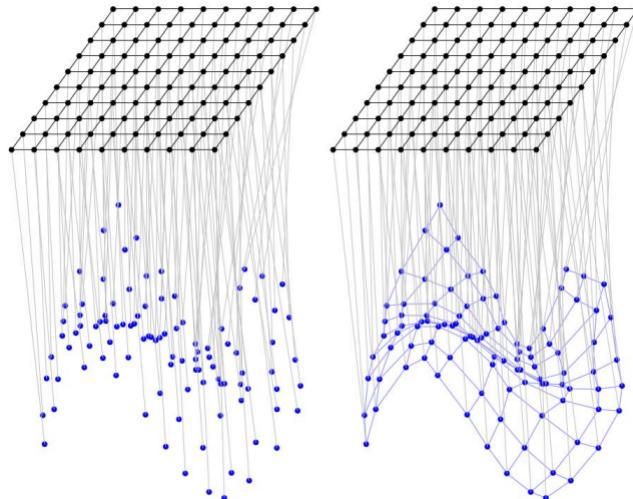


Figure 2.74: The neuron space (2D grid) and the input space (3D grid)

Basically what we want to do in this kind of network is how we associate the element of the dataset with the neuron in my grid preserving the similarity and the concept of neighbouring saw before. This is how we can update parameter of our network, and enforcing the parameters of the winning neuron :

$$\vec{r}_u^{(new)} = \vec{r}_u^{(old)} + \eta(t) f_{nb}(d_{\text{neurons}}(u, u_*), \rho(t)) (\vec{x} - \vec{r}_u^{(old)})$$

in time i will reduce the learning radius , so i can focus the capacity of learning a specific pattern this will allow me to create a much stronger partitioning in classes by similarity. Considering a neighbouring function with is smooth, i may consider a smooth border and an example may excide more t***

The training will work by placing the elements randomly and adjust the interconnection by adjusting the winning neuron***

Example a grid 10 by 10 random associations of the network, each step will associate with the similar example. another example, the problem is always the same the initial configuration is poor and the nn can't flatten the configuration of the grid. Maybe the neighbour radius is too small or the learning rate is too small.

Basically starting from the original 3D we are able to flat in a bidimensional one. This can be used for reduce the bidimensionality of the dataset

2.24 Hopfield Networks

Hopfield network is a NN with a loop in the graph. We have not seen in practice what happens with a loop that feed back the information.

Essentially what we have is a single layer of neurons for which we have the inputs that are giving the incoming data to them, they are operating in such a way that the output of the neuron are feeding the other neurons in the structure.

This means that we have a single layer of neurons that still have an output arc, they can't only send data for feedback. Basically this creates a kind of memory, what is generated in a previous step is reused for the subsequent step (by the way this is the motivation for introducing this kind of network).

The relevance of the output arc is the same of the feedback arc. the network input function is only the weighted summation of the output of all the other neurons.

The we will generate the network neuron of the value, this is the input value of the activation function, the activation function that we are using is essentially the step function.

The output function for this neuron is the identity function. If we want in a concise way represent all the connections we can use a matrix to collect all the weights. The diagonals is composed by 0 because this means that we don't have any contribution to the next neuron. The matrix is symmetric so that the interconnection is passed correctly

simple example In this example we have 1 as interconnection weight, we see the matrix. the problem that we have is when we are updating the neuron? According to how we perform the update the hopfield could converge or not. The computations can oscillate in the case of synchronous update...**+

If we apply the update in an asynchronous way the situation is different, the values are changed one at time as consequence we don't have the cyclic behavior of before. The only problem is that the steady state will be different, we are sure if we use the asynchronous approach that we reach a steady state in the end but the value of the steady state will depend from the order of execution of the computation.

Let's consider a more complex network, if we take 4 or 5 neurons will be more complex. Here is a still readable graph describing the connections

but growing the number of neurons it will be easily unreadable. There is a simplified version of the graph which exploits the characteristic of the network. Every neuron is an output neuron, when a steady configuration is reached is the final output of the neuron every neuron is an output neuron as well. since the matrix of the weights is symmetric, we don't have to create two arcs between the neuron i and j the same***

In order to understand how the computation evolves and how we reach the final steady state we have to introduce the state graph. It will give the values of the activation and transition inside the hopfield network. each node in the graph represent a configuration that we have for each neuron. + is active and - is not active. If we start from this state, if we apply the operation for one of the three neurons what is happening is at the arc going out of the ***

convergence theorem It is actually possible to prove the convergence theorem which tells us that if we active the hopfield network in a sequential asynchronous way a steady state is reached in a finite number of steps. It is possible to prove also that this maximum number is given by n multiplied by 2^n steps.

basically to prove this we can rely on the fact that we can associate to the behaviour of the network an energy which can be defined in this way (not actual physical energy), but if you look to the shape of the function is the typical structure of an energy function: *** It is important to point out that the order of activation of neurons doesn't matter we will reach a steady state.

A steady state is reached when no output is changing, if a neuron is changing the output the previous state of the graph can't be reached. To go back to the previous state we need to increase the energy. Any physical system tries to reduce the total energy in the system. The second problem that forbids the fact of coming back to the previous state is due to the fact that this implies that we may need to change two outputs to go back as consequence even if it is not the complete parallel of the neuron, that we are not following strictly the sequential activation.

To have a better understanding of the energy reasoning we can place the nodes of the state graphs, here is possible to see if we compute that form of energy** not necessary the state graph need to be symmetric, here is an example of another network,

When we are reaching a steady state we have a memory of what we are recording, this is what the hopfield allows us to do. we can configure the network such that the steady state can represent some information that we want to retrieve. if we want to have some specific information that we want to store we have to ** with a single string of bits is possible to show that to achieve this configuration (E is the identity) if we use the weight matrix we are able to present the pattern p and recognize it. what we can state is that if we have the pattern and we want to know if this pattern is one of the considered important

this type of network can do even more, in particular if we present a pattern that is not identical to the original pattern, the network can reconstruct the original pattern (reaching the state of the original pattern). Why this? if we have numbers that we want to recognize, and there is some noise during the connection, we can understand what kind of number is that by similarity.

Solving optimizations problems. we take the function that we want to optimize in our problem ***

2.25 Boltzman machines

Le macchine di Boltzmann sono un modello di network che sono in relazione stretta con gli Hopfield networks. Alcune variazioni speciali di questo modello hanno ottenuto un interesse rilevante recentemente (specialmente nel Deep Learning). Una macchina di Boltzmann standard differisce da un Hopfield Network principalmente nel come gli stati dei neuroni vengono aggiornati (e per il fatto che può contenere neuroni nascosti)..

Come per le reti Hopfield per la risoluzione di problemi di ottimizzazione si affidano ad una **funzione energetica** che assegna un valore numerico ad ogni stato del network. Con l'aiuto di questa funzione energetica viene definita una distribuzione probabilistica su gli stati del network che è basata sulla **distribuzione di Boltzmann**:

$$P(\mathbf{s}) = \frac{1}{c} e^{-\frac{E(\mathbf{s})}{kT}}$$

dove il vettore \mathbf{s} descrive lo stato del sistema, c è la costante di normalizzazione, E è la funzione che restituisce l'energia di uno stato s , T è la temperatura termodinamica del sistema e k è la costante di Boltzmann.

Per le macchine di Boltzmann il prodotto kT è spesso sostituito solo da T , così combinando la temperatura e la costante di Boltzmann in un singolo parametro. Inoltre, lo stato s consiste nel vettore **act** delle attivazioni dei singoli neuroni.

$$E(\mathbf{act}) = -\frac{1}{2} \mathbf{act}^\top \mathbf{W} \mathbf{act} + \boldsymbol{\theta}^\top \mathbf{act}$$

dove \mathbf{W} è la matrice dei pesi connessi e $\boldsymbol{\theta}$ il vettore contenente i valori delle soglie (o threshold).

Ora guardiamo il singolo neurone u e elaboriamo il cambio energetico che viene portato da questo neurone durante il suo cambiamento di stato. Consideriamo la differenza assoluta in energia tra $act_u = 0$ e $act_u = 1$ (mentre tutti gli altri neuroni mantengono la loro attivazione):

$$\Delta E_u = E_{act_u=1} - E_{act_u=0} = \sum_{v \in U - \{u\}} w_{uv} act_v - \theta_u$$

scrivere l'energia in termini di distribuzione di Boltzmann restituisce:

$$P(act_u = 1) = \frac{1}{1 + e^{-\frac{\Delta E_u}{kT}}}$$

cio è, la probabilità di un neurone di essere attivo in una funzione logistica della differenza energetica tra stato attivo ed inattivo. Visto che la differenza energetica è strettamente connessa al network in input del neurone:

$$\Delta E_u = \sum_{v \in U - \{u\}} w_{uv} act_v - \theta_u = net_u - \theta_u$$

questa formula suggerisce una procedura con aggiornamento stocastico per il network.

La procedura di aggiornamento sviluppa come segue:

- Un neurone u di una data macchina di Boltzmann viene scelto e calcola la sua differenza energetica ΔE_u (attraverso la funzione logistica) e con questo la sua probabilità di attivazione.
- Questo aggiornamento viene ripetuto tante volte per ogni neurone scelto casualmente fino alla convergenza del sistema. La convergenza verso uno stato sstabile è garantita dal fatto che la temperatura del sistema non cresce nel tempo, ma diminuisce. Ad un certo punto si raggiungerà uno stato stabile, anche detto **equilibrio termico** (thermal equilibrium) del sistema che rappresenta un minimo (possibilmente locale) della funzione.

Questa procedura si chiama **Markov-Chain Monte Carlo**.

Bisogna notare che una BM potrà calcolare in modo efficace una distribuzione di probabilità se gli esempi forniti sono compatibili con una distribuzione di Boltzmann. Per mitigare questa restrizione si dividono i neuroni di una BM tra neuroni *visibili*, ovvero quelli in grado di ricevere i segnali in input, e neuroni *nascosti*, la cui attivazione non dipende direttamente dal dataset permettendo un adattamento più flessibile ai pattern di allenamento.

2.25.1 Allenamento

L'obiettivo di apprendimento è quello di adattare i pesi ed i threshold in modo che la distribuzione implicita nel dataset sia approssimata dalla distribuzione rappresentata dai neuroni visibili di una BM.

Questo è possibile scegliendo una misura che descriva la differenza tra le due distribuzioni ed utilizzeremo la tecnica del *gradient descent* per minimizzarla. Una delle misure più utilizzate in questo campo è quella di Kullback-Leibler sulla divergenza dell'informazione:

$$KL(p1, p2) = \sum_{\omega \in \Omega} p1(\omega) \ln \frac{p1(\omega)}{p2(\omega)}$$

dove $p1$ si riferisce alla distribuzione del dataset e $p2$ a quella della macchina di Boltzmann. Ogni passo di apprendimento si divide in due fasi:

- *Positive phase*: nella quale i neuroni visibili vengono fissati rispetto ad un dato di input scelto randomicamente e i neuroni nascosti vengono aggiornati fino al raggiungimento di un equilibrio termico.
- *Negative phase*: tutte le unità vengono aggiornate fino al raggiungimento di uno stato stabile.

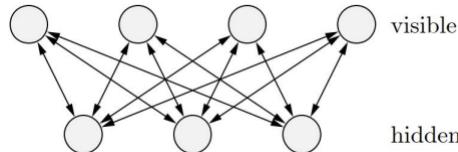
Sia p_u^+ la probabilità che un neurone u sia attivato nella positive phase, e sia p_u^- la probabilità che venga attivato nella fase negativa. Sia anche p_{uv}^+ la probabilità che entrambi i neuroni vengano attivati simultaneamente nella positive phase, e viceversa, p_{uv}^- nella fase negativa. Possiamo definire la regola di aggiornamento dei pesi e delle threshold come:

$$\Delta w_{uv} = \frac{1}{\eta} (p_{uv}^+ - p_{uv}^-) \text{ e } \Delta \theta_u = -\frac{1}{\eta} (p_u^+ - p_u^-)$$

Intuitivamente se lo stesso neurono viene sempre attivato ogniqualvolta viene presentato lo stesso input allora il suo threshold dovrà essere ridotto. Allo stesso modo, se due neuroni vengono spesso attivati assieme allora il peso che corrisponde alla loro connessione verrà aumentato.

2.26 Restricted Boltzman Machines

A special variant of BM was introduced by Harmonium, but nowdays is generally referred to as **restricted Boltzmann machine**. The restriction consists in revoking the condition that the graph underlying the network must be a fully connected graph. Rather one uses a bipartite graph, in which the vertices are split in two, namely the visible and the hidden neurons.



One of the advantages of having a network where there are no connections between neurons of the same group is given by the fact that the process of learning can be accomplished repeating this three steps:

1. Phase I: the input units are fixed respect to a selected random pattern and the hidden are updated in parallel retrieving the so called *positive gradient*.
2. Phase II: obtaining a preprocessed input in the first phase, you invert the parts and you fix the hidden neurons and update the visible neurons, retrieving the *negative gradient*.
3. Phase III: update of the weights and thresholds using the difference between the positive and negative gradient.

The RBM have also been used to build deep networks similar to stacked auto-encoders for the MLP.

2.27 Recurrent Networks

Both HM and BM are examples of recurrent network (RNN) with a very constrained architecture. However, let's analyze what is an RNN without any kind of restriction. As we already know an RNN is an NN with cycles among neurons (on individual neurons or involving groups of neurons).

The output in this kind of network is yield only if it is reached a stable state during the computation.

The configuration of this network can be obtained in two ways:

- by construction if the structure of the computation is known.
- by extending the error backpropagation algorithm in time to deal with recursion to output stability for each input pattern.

RNN can be used to represent differential equations, like:

$$x^{(n)} = f(t, x, x', x'', \dots, x^{(n-1)})$$

which is equivalent to a system of differential equations:

$$\begin{cases} x' = y_1 \\ y'_1 = y_2 \\ \dots \\ y'_{n-2} = y_{n-1} \\ y'_{n-1} = f(t, x, y_1, y_2, \dots, y_{n-1}) \end{cases}$$

using the recursive representation of the equations:

$$x(t_i) = x(t - 1) + \Delta y_1(t_i - 1)$$

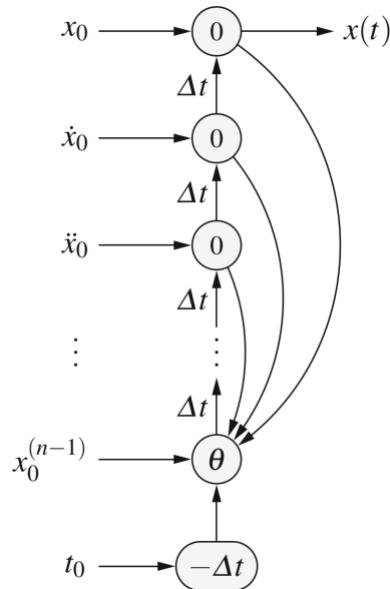
$$y_1(t_i) = y_1(t - 1) + \Delta y_2(t_i - 1)$$

...

$$y_2(t_i) = y_2(t - 1) + \Delta y_3(t_i - 1)$$

$$y_1(t_i) = y_1(t - 1) + f(t_{i-1}, x(t_{i-1}), y_1(t_{i-1}), \dots, y_{n-1}(t_{i-1}))$$

we can represent differential equation using RNN (or viceversa).



we can use the derivative of the function at the previous time slice for computing the following value. This allows to transform the equation in a RNN, creating for each variable a node inside the graph and associating to the connections the value of the differential (like in the image above).

It is possible to generalize this approach with functions with more than one argument thanks to the **vectorial neural network** (VNN). Which are RNN composed by multiple recurrent sub-networks. It can be used to compute vectorial differential equations. The backpropagation is not directly applicable since loops propagate errors in a cyclic manner. The recurrent network must be **unfolded in time** between two training patterns, this any time the RNN traverse a cycle and add a copy of traversed neurons as additional layer.

At this time it will be possible to apply the backpropagation like for the FFN. You compute the adjustments of the weights by backpropagation on the unfolded network, and combine the adjustments of the same weight to generate the value for the recurrent network.

By the way, for calculating the updates of weights and thresholds will be necessary combine the necessary adjustment computer respect to the added neurons.

3 Fuzzy Systems

3.1 Motivation

Humans as imperfect creatures uses imprecise linguistic terms (fat, small, big, tall, etc.). All the complex human reasoning are based on those terms, the clas-

sical mathematics can't manage these terms. Classical logic manages *true* and *false* to each logical proposition or statement. When we describe such formal model, we use a terminolgy which has much more stringent rules than natural language.

We want to introduce these terms in the computer science field, but how? let's talk about **fuzzy logic**.

But first let's introduce some concepts.

3.2 Imprecision and uncertainty

Any notion is said to be **imprecise** when its meaning is not fixed, where the value take places between the conditions *fully*. The *gradualness* is also called *fuzziness*.

A **proposition** is imprecise if it contains gradual predicates, such proposition may be true to a certain degree (*partial truth*). An example of this in the natural language : *very, rather, almost not, etc...*

3.2.1 Sorites Paradox

This is an example of **imprecision**, if a sand dune is small, adding one grain of sand to it leaves it small, a sand dune with a single grain is small, hence all sand dunes are small.

The paradox comes from treatment of "small", *when the dune is not anymore a dune?*

Formal statement:

- Statement $A(n)$: *n grains of sand are a sand dune.*
- $d_n = T(A(n))$ denote "degree of acceptance" for $A(n)$.
- Then $0 = d_0 \leq d_1 \leq \dots \leq d_n \leq \dots \leq 1$ can be seen as truth values of a many valued logic.

3.2.2 Uncertainty

This term refers to situations involving imperfect or unknown information. If we consider the notion "bald", a man without hair on his head is considered bald.

Usually a man is not completely bald, has some hair, where to set baldness/not baldness threshold?

The **fuzzy set theory** does not assume any threshold.

3.2.3 Difference between imprecision and uncertainty

The **imprecision**, refers to "imprecisely" defined concepts neglect of details computing with words ("*today the weather is fine*"). Instead the **uncertainty**, refers to the probability/possibility about unknown or imperfect information ("*this car was probably made in Germany*").

3.3 Principle of incompatibility

As the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance become almost mutually exclusive characteristics.

Fuzzy sets and fuzzy logic are used as mechanism for abstraction of unnecessary or too complex details.

Some fields of fuzzy systems:

- Control Engineering
- Approximate Reasoning
- Data Analysis
- Image Analysis

Advantages brought from the using of fuzzy systems

- Use of imprecise or uncertain information.
- Use of expert knowledge.
- Robust nonlinear control.
- Time to market.
- Marketing aspects.

3.4 Fuzzy sets

A **fuzzy subset** or **fuzzy set** μ of a set X (the universe of discourse) is a mapping $\mu : X \rightarrow [0, 1]$, which assigns to each element $x \in X$ a *degree of membership* $\mu(x)$ to the fuzzy (sub)set μ itself.

This set of all fuzzy subsets (or sets) of X is denoted with $\mathcal{F}(X)$.

3.4.1 Membership functions

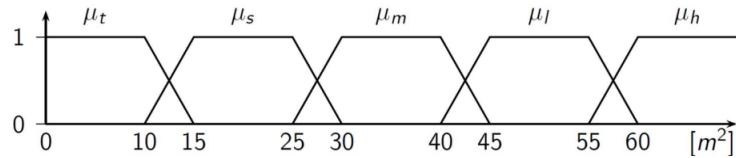
The degree of membership is equal to 1 when the element x fully respects the property that I want to describe with the fuzzy set.

$$\mu_m(u) = 1$$

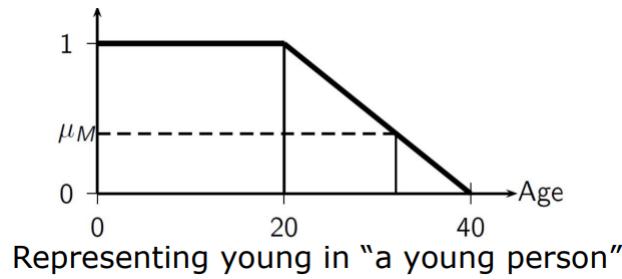
The degree of membership is equal to 0 when it expresses absolutely non-membership.

$$\mu_m(u) = 0$$

Sets can be viewed as special cases of fuzzy sets where only full membership and absolute non-membership are allowed, these sets are called **crisp sets** or



boolean sets. The membership degree between $0 < \mu_M < 1$ represent the **partial membership**.



The membership function attached to a given linguistic description (such as "young") depends on the context. For example, a young retired person is certainly older than a young student. Even idea of young student depends on the user.

The membership degrees are fixed only by convention, the unit interval as range of membership grades is arbitrary. Natural for modeling membership grades of fuzzy sets of real numbers.

There is no precise threshold between young and not young, the fuzzy sets are a natural interface between linguistic and numerical representations.

3.4.2 Linguistic variables and values

The **linguistic variables** represent attributes in fuzzy systems, and they assume **linguistic values**. Linguistic values usually partition the possible values of the linguistic variables subjectively.

All linguistic values have a meaning, not a precise numerical value.

An example of linguistic values:

- Living area of a flat (linguistic variable)
- Tiny, small, medium, large, huge (linguistic values)

Every $x \in A$ has $\mu(x) \in [0, 1]$ to each value.

3.4.3 Semantics of fuzzy systems

Fuzzy systems are relevant in 3 types of information-driven tasks:

1. classification and data analysis
2. decision-making problems
3. approximate reasoning

These tasks exploit three semantics of membership grades:

1. similarity
2. preference
3. possibility

Degree of similarity

$\mu(u)$ is the degree of proximity of u from prototype elements of u . Proximity between pieces of information is modelled.

This view is used in:

- pattern classification
- cluster analysis
- regression

In fuzzy control, similarity degrees are measured between current situation and prototypical ones.

Degree of preference

The μ function represents both:

- set of more or less preferred objects.
- value of a decision variable X

The image $\mu(u)$ represents both:

- intensity of preference in favor of object u .
- feasibility of selecting u as value of X .

Fuzzy sets represent criteria or flexible constraints, this view is used in *fuzzy optimization* (especially in fuzzy linear programming) and in *decision analysis* (decision theory).

The typical applications are in:

- engineering design
- scheduling problems

Degree of possibility

The $\mu(u)$ can be viewed as the *degree of possibility* that the element u is the value of the parameter X , and it is used for quantifying the epistemologic status of an agent.

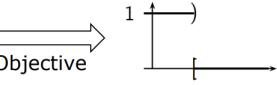
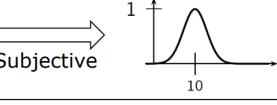
The objective is to distinguish what the agent consider "surprising" from what is "typical".

Supports values are mutually exclusive and membership degrees rank these values by their possibility. This view is used in *expert systems* and *artificial intelligence*.

3.4.4 Definition of a set

By a set we mean every collection made into a whole of definite, distinct objects of our intuition or of our thought. Properties:

- $x \neq \{x\}$
- If $x \in X$ and $X \in Y$, then $X \notin Y$
- The set of all subsets of X is denoted as 2^X
- \emptyset is the empty set

Linguistic description	Model
all numbers smaller than 10 Objective	
all numbers almost equal to 10 Subjective	

A fuzzy set μ of $X \neq \emptyset$ is a function from the reference set X to the unit interval, $X \rightarrow [0, 1]$. $\mathcal{F}(X)$ represents the set of all fuzzy sets of X , $\mathcal{F}(X)^{def} = \{\mu | \mu : X \rightarrow [0, 1]\}$.

Vertical representation

Fuzzy sets are described by their membership function and assigning degree of membership $\mu(x)$ to each element $x \in X$.

$$\mu_{a,b,c,d}(x) = \begin{cases} \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } b \leq x \leq c \\ \frac{x-d}{c-d}, & \text{if } c \leq x < d \\ 0, & \text{if } x < a \text{ or } x > d \end{cases}$$

Horizontal representation

For all membership degrees belonging to chosen subset of $[0, 1]$, human experts lists elements of X that fulfill vague concept of fuzzy set with degree $\geq \alpha$.

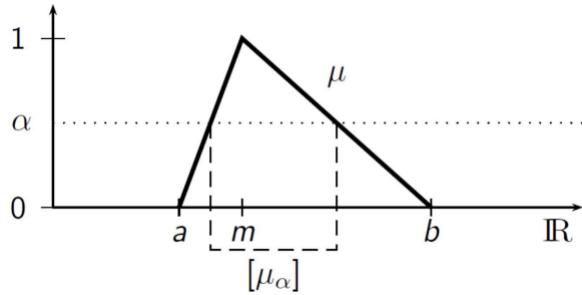
That is the horizontal representation of fuzzy sets by their α -cuts. Let's give the formal definition:

Let $\mu \in \mathcal{X}$ and $\alpha \in [0, 1]$, then the sets

$$[\mu]_\alpha = \{x \in X | \mu(x) \geq \alpha\}, [\mu]_\alpha^* = \{x \in X | \mu(x) > \alpha\}$$

are called **α -cut** and **strict α -cut**.

For example, let's consider the μ be a triangular function on \mathbb{R} .



The α -cut can be constructed in the following steps:

1. Draw an horizontal line parallel to x -axis through point $(0, \alpha)$.
2. Project this section onto x -axis

$$[\mu]_\alpha = \begin{cases} [\alpha + \alpha(m - \alpha), b - \alpha(b - m)] & \text{if } 0 < \alpha \leq 1 \\ \mathbb{R} & \text{otherwise} \end{cases}$$

Properties of α -cuts:

- Any fuzzy set can be described by specifying it's α -cuts.

$$\mu \in \mathcal{F}(X), \alpha \in [0, 1] \text{ and } \beta \in [0, 1]$$

$$\begin{aligned} [\mu]_0 &= X \\ \alpha < \beta \implies [\mu]_\alpha &\supseteq [\mu]_\beta \\ \cap_{\alpha: \alpha < \beta} [\mu]_\alpha &= [\mu]_\beta \end{aligned}$$

- **Representation theorem**

$$\mu \in \mathcal{F}(X)$$

$$[\mu]_0 = \sup_{\alpha \in [0, 1]} \left\{ \min \left(\alpha, \mathcal{X}_{[\mu]_\alpha}(x) \right) \right\}$$

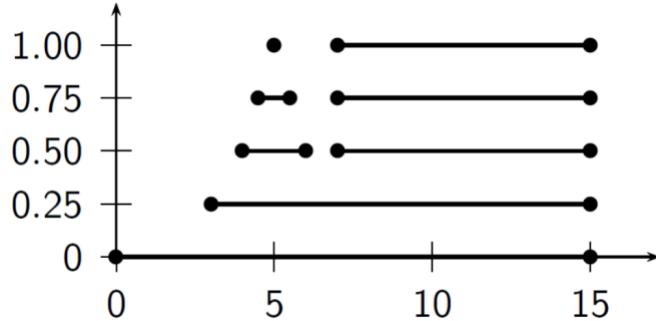
$$\text{where } \mathcal{X}_{[\mu]_\alpha}(x) = \begin{cases} 1, & \text{if } x \in [\mu]_\alpha \\ 0, & \text{otherwise} \end{cases}$$

Fuzzy sets can be obtained as upper envelope of its α -cuts, simply draw the α -cuts parallel to horizontal axis in height of α .

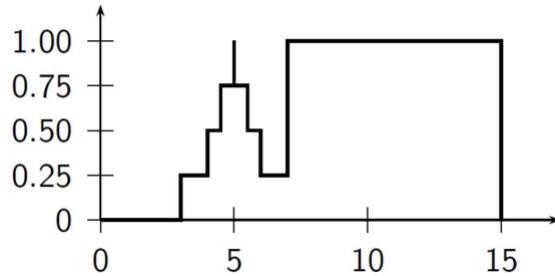
In applications it is recommended to select finite subset $L \subseteq [0, 1]$ of relevant degrees of membership. They must be semantically distinguishable, fix level of fuzzy sets to characterize only for these levels.

This representation of fuzzy sets is used in computers. For example, let's suppose $X = [0, 15]$, an expert choose $L = \{0, 0.25, 0.5, 0.75, 1\}$ and α -cuts:

- $A_0 = [0, 15]$
- $A_{0.25} = [3, 15]$
- $A_{0.5} = [4, 6] \cup [7, 15]$
- $A_{0.75} = [4.5, 5.5] \cup [7, 15]$
- $A_1 = \{5\} \cup [7, 15]$

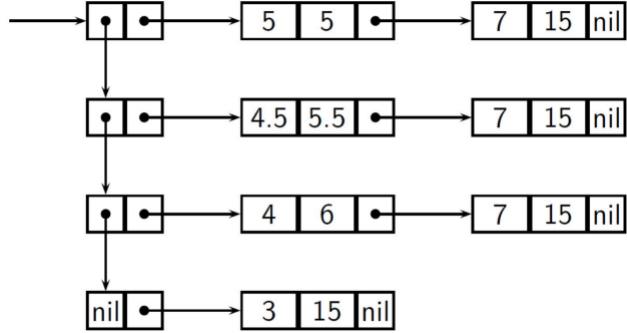


We can see that μ_A is obtained as upper envelope of the family A of sets:



As said the horizontal representation is easier on the computers, this because they are usually stored as chain of singly linked lists (similar to the skip-

list data structure), for each level $\alpha \neq 0$. This data structure is appropriate for arithmetic operators.



3.4.5 Support and Core

The **support** $S(\mu)$ of a fuzzy set $\mu \in \mathcal{F}(X)$ is the crisp set that contains all elements of X that have nonzero membership.

$$S(\mu) = [\mu]_0 = \{x \in X | \mu(x) > 0\}$$

The **core** of $C(\mu)$ of a fuzzy set $\mu \in \mathcal{F}(X)$ is the crisp set that contains all the elements of X that have membership of one.

$$C(\mu) = [\mu]_1 = \{x \in X | \mu(x) = 1\}$$

3.4.6 Height of a fuzzy set

The height $h(\mu)$ of a fuzzy set $\mu \in \mathcal{F}(X)$ is the largest/highest membership grade obtained by any element in that set.

$$h(\mu) = \sup_{x \in X} \{\mu(x)\}$$

A fuzzy set μ is called **normal** if and only if $h(\mu) = 1$, it is called **subnormal**, if and only if $h(\mu) < 1$.

3.4.7 Convex fuzzy sets

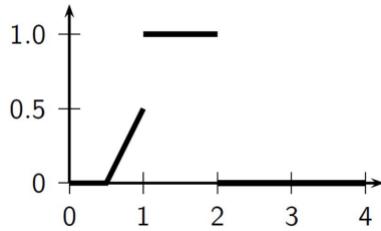
Let X be a vector space, a fuzzy set $\mu \in \mathcal{F}(X)$ is called **fuzzy convex** if its α -cuts are convex for all $\alpha \in (0, 1]$. The membership function of a convex fuzzy set is not a convex function.

The classical definition, the membership functions are actually concave.

3.4.8 Fuzz numbers

A fuzzy set μ is a fuzzy number if and only if μ is normal and $[\mu]_\alpha$ is bounded, close, and convex $\forall \alpha \in (0, 1]$. For example, the term *approximately* x_0 is described by a parametrized class of membership functions.

$$\begin{cases} [1, 2] & \text{if } \alpha \geq 0.5 \\ [0.5 + \alpha, 2] & \text{if } 0 > \alpha < 0.5 \\ \mathbb{R} & \text{if } \alpha = 0 \end{cases}$$



Upper semi-continuous functions are often convenient in applications. In many applications the class of the functions and their exact parameters have a limited influence on the results. In other applications more precise membership degrees are needed.

3.5 Fuzzy logics

An important result of classical logic says that exist and isomorphism between the propositional logic and the theory of finite sets (on a infinite set of variables).

Both these systems can be shown isomorphic to finite boolean algebra.

3.5.1 Set operators

Set operators are defined by using traditional logic operators, let X be the universal set:

$$A \cap B = \{x \in X | x \in A \wedge x \in B\}$$

$$A \cup B = \{x \in X | x \in A \vee x \in B\}$$

$$A^c = \{x \in X | x \notin A\} = \{x \in X | \neg(x \in A)\}$$

$$A \subseteq B \text{ if and only if } (x \in A) \rightarrow (x \in B) \text{ for every } x \in X$$

3.5.2 Aristotelian logic

Aristotelian logic can be seen as formal approach to human reasoning. It's still used today in AI for knowledge representation and reasoning about knowledge.

3.5.3 Classical logic

Classical logic deals with propositions (either true or false). The propositional logic handles combinations of logical variables. The key idea is to express n -ary logic functions with logic primitives.

A set of logic primitives is complete if any logic function can be composed by a finite number of these primitives.

$$\wedge, \vee, \neg, NOR, NAND, |$$

3.5.4 Inference rules

When a variable represented by a logical formula is true for all possible *truth values* it is called **tautology**. Instead, when it is false for all possible truth values it is called **contradiction**.

Various forms of tautologies exist to perform deductive inference, and are called inference rules:

$$(a \wedge (a \rightarrow b)) \rightarrow b \text{ (modus ponens)}$$

$$(\neg b \wedge (a \rightarrow b)) \rightarrow \neg a \text{ (modus tollens)}$$

$$((a \rightarrow b) \wedge (b \rightarrow c)) \rightarrow (a \rightarrow c) \text{ hypothetical syllogism}$$

3.5.5 Boolean algebra

The propositional logic based on finite set of logic variables is isomorphic to **finite set theory**. Both of these systems are isomorphic to a finite Boolean algebra.

A Boolean algebra on a set B is defined as quadruple $\mathcal{B} = (B, +, ., \bar{x})$, where B has at least two elements 0 and 1, $+$ and $.$ are binary operators on B , and \bar{x} is a unary operator on B for which the following properties hold:

(B1) Idempotence	$a + a = a$	$a \cdot a = a$
(B2) Commutativity	$a + b = b + a$	$a \cdot b = b \cdot a$
(B3) Associativity	$(a + b) + c = a + (b + c)$	$(a \cdot b) \cdot c = a \cdot (b \cdot c)$
(B4) Absorption	$a + (a \cdot b) = a$	$a \cdot (a + b) = a$
(B5) Distributivity	$a \cdot (b + c) = (a \cdot b) + (a \cdot c)$	$a + (b \cdot c) = (a + b) \cdot (a + c)$
(B6) Universal Bounds	$a + 0 = a, a + 1 = 1$	$a \cdot 1 = a, a \cdot 0 = 0$
(B7) Complementary	$a + \bar{a} = 1$	$a \cdot \bar{a} = 0$
(B8) Involution	$\bar{\bar{a}} = a$	
(B9) Dualization	$\overline{a + b} = \bar{a} \cdot \bar{b}$	$\overline{a \cdot b} = \bar{a} + \bar{b}$

Every theorem in one theory has a counterpart in each other theory.

Meaning	Set Theory	Boolean Algebra	Prop. Logic
values	2^X	B	$\mathcal{L}(V)$
"meet"/"and"	\cap	\cdot	\wedge
"join"/"or"	\cup	$+$	\vee
"complement"/"not"	c	$\overline{}$	\neg
identity element	X	1	1
zero element	\emptyset	0	0
partial order	\subseteq	\leq	\rightarrow

- The principle of **bivalence**: every proposition is either true or false.
- The principle of **valence**: every proposition has a truth value.

3.5.6 n -valued logics

Various n -valued logics were developed. Usually truth values are assigned by rational number in $[0, 1]$. The key idea is to uniformly divide $[0, 1]$ into n truth values.

The definition: the set T_n of truth values of an n -valued logic is defined as

$$T_n = \left\{ 0 = \frac{0}{n-1}, \frac{1}{n-1}, \dots, \frac{n-1}{n-1}, \frac{n-1}{n-1} = 1 \right\}$$

This values can be interpreted as degree of truth.

The primitives in n -valued logics

Generalization of Lukasiewicz three-valued logic. It uses truth values in T_n and define primitives as follow:

$$\begin{aligned} \neg a &= 1 - a \\ a \wedge b &= \min(a, b) \\ a \vee b &= \max(a, b) \\ a \rightarrow b &= \min(1, 1 + b - a) \\ a \longleftrightarrow b &= 1 - |a - b| \end{aligned}$$

This n -valued logic is denoted by L_n , the sequence (L_2, \dots, L_∞) contains the classical two-valued logic L_2 and an infinite-valued logic L_∞ . The infinite-valued logic L_1 is the logic with all real numbers in $[0, 1]$.

3.5.7 Fuzzy set theory

What does a fuzzy set represent? A logic with values in $[0, 1]$, consider fuzzy proposition A on \mathbb{R} , fuzzy logic offers means to construct such

A is defined by membership function μ_A , truth values $\forall x \in \mathbb{R}$, let $x \in \mathbb{R}$ be a subject/observation, and $\mu_A(x)$ is the degree of truth that x is A .

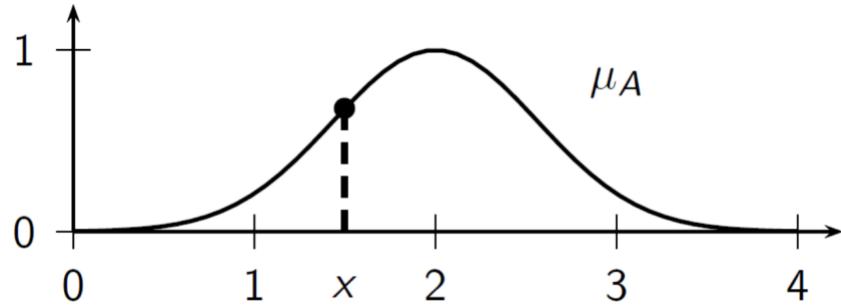


Figure 3.1: Membership function μ_A

Standard fuzzy set operators

We define the following algebraic operations on $\mathcal{F}(X)$

$$(\mu \wedge \mu') =^{\text{def}} \min\{\mu(x), \mu'(x)\} \text{ intersection ("AND")}$$

$$(\mu \vee \mu') =^{\text{def}} \max\{\mu(x), \mu'(x)\} \text{ union ("OR")}$$

$$\neg \mu =^{\text{def}} 1 - \mu(x) \text{ complement ("NOT")}$$

μ is subset of μ' if and only if $\mu \leq \mu'$

Theorem: $(\mathcal{F}(X), \wedge, \vee, \neg)$ is a complete distributive lattice but no Boolean algebra.

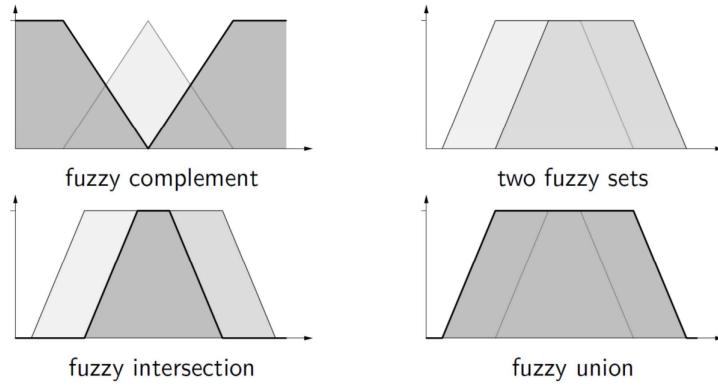


Figure 3.2: Various set operations based on the previous fuzzy definitions

Fuzzy set complement/negation

Let X be a given set and $\mu \in \mathcal{F}(X)$, then the complement $\bar{\mu}$ can be defined pointwise by $\bar{\mu}(x) := 1 - \mu(x)$ where $[0, 1] \rightarrow [0, 1]$ satisfies the following conditions:

$$(0) = 1, (1) = 0$$

and for

$$x, y \in [0, 1], x \leq y \implies x \geq y$$

(where \sim is non-increasing).

A negation is called **strict** if it is also strictly decreasing and continuous. A strict negation is said to be **strong** if it is involutive, too.

standard negation:

$$\sim x = 1 - x$$

threshold negation:

$$\sim_\theta x = \begin{cases} 1 & \text{if } x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Cosine negation:

$$\sim x = \frac{1}{2} (1 + \cos(\pi x))$$

Sugeno negation:

$$\sim x = \frac{1-x}{1+\lambda x}, \quad \lambda > 1$$

Yager negation:

$$\sim x = (1 - x^\lambda)^{\frac{1}{\lambda}}$$

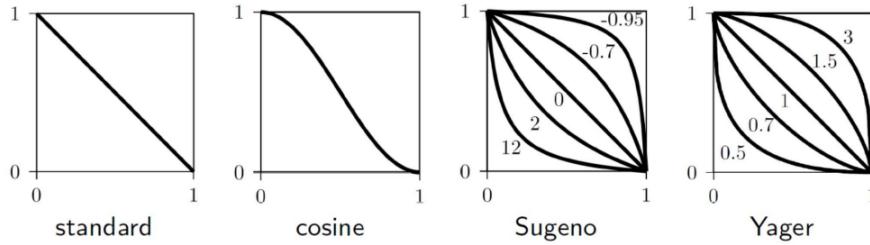


Figure 3.3: Families of negations

Fuzzy set intersection and union

Let A, B be fuzzy subsets of X , $A, B \in \mathcal{F}(X)$. Their intersection and union can be defined pointwise using

$$(A \cup B)(x) = \top(A(x), B(x)) \text{ where } \top : [0, 1]^2 \rightarrow [0, 1]$$

$$(A \cap B)(x) = \perp(A(x), B(x)) \text{ where } \perp : [0, 1]^2 \rightarrow [0, 1]$$

\top is a triangular norm called *t-norm* $\iff \top$ which satisfies the following conditions:

1. $\top(x, 1) = x$
2. $\top(x, y) = \top(y, x)$
3. $\top(x, \top(y, z)) = \top(\top(x, y), z)$
4. $x \leq z \implies \top(x, y) \leq \top(x, z)$

\perp is a triangular conorm called *t-conorm* $\iff \perp$ which satisfies C1 – C4.

1. $\perp(x, 0) = x$
2. $\perp(x, y) = \perp(y, x)$
3. $\perp(x, \perp(y, z)) = \perp(\perp(x, y), z)$
4. $x \leq z \implies \perp(x, y) \leq \perp(x, z)$

Minimum and maximum

The definitions of intersection and union we give in the paragraph "Stanard fuzzy operations" in terms of max and min they satisfy this properties. It is possible also to prove that the operation of minimum is the greatest t-norm and the maximum the smallest t-conorm (they can be easily processed numerically).

Product and probabilistic sum

Also, could be given other definitions of disjunctions and conjunctions, for example in terms of products and probabilistic sum:

$$\top_{prod}(x, y) = x \cdot y$$

$$\perp_{sum}(x, y) = x + y - x \cdot y$$

or, following Łukasiewicz:

$$\top_{Luk}(x, y) = \max\{0, x + y - 1\}$$

$$\perp_{Luk}(x, y) = \min\{1, x + y\}$$

Nilpotent minimum and maximum

$$\begin{aligned}\top_{min_0}(x, y) &= \begin{cases} \min(x, y) & \text{if } x + y > 1 \\ 0 & \text{otherwise} \end{cases} \\ \perp_{max_1}(x, y) &= \begin{cases} \max(x, y) & \text{if } x + y < 1 \\ 1 & \text{otherwise} \end{cases}\end{aligned}$$

Drastic product and sum

$$\begin{aligned}\top_{-1}(x, y) &= \begin{cases} \min(x, y) & \text{if } \max(x + y) = 1 \\ 0 & \text{otherwise} \end{cases} \\ \perp_{-1}(x, y) &= \begin{cases} \max(x, y) & \text{if } \min(x + y) = 0 \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

\top_{-1} is the weakest t-norm, \perp_{-1} is the strongest t-conorm.

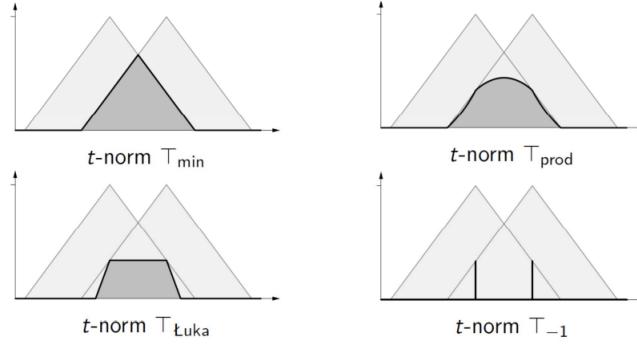


Figure 3.4: Examples of fuzzy intersections

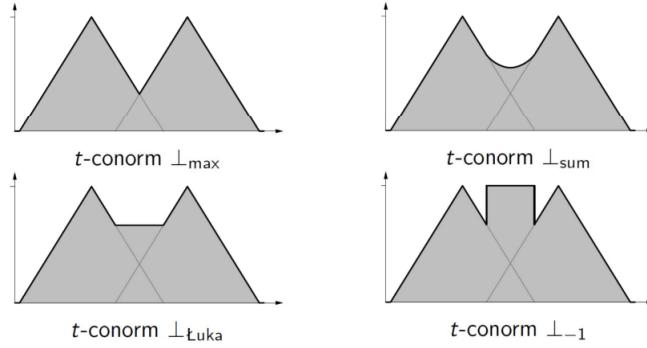


Figure 3.5: Examples of fuzzy unions

3.5.8 Fuzzy implications

A fuzzy set is contained in another fuzzy set if all the elements of the first are contained in the second. By exploiting the isomorphism between the logic and set operators, we will be able to define the concept of subset by starting from the concept of implication:

$$I(a, b) = \neg a \vee b, \forall a, b \in \{0, 1\}$$

In fuzzy logic, disjunction and negation are t-conorm and fuzzy complement, respectively, thus:

$$I(a, b) = \perp(\sim a, b), \forall a, b \in \{0, 1\}$$

Another way in classical logic:

$$I(a, b) = \max\{x \in \{0, 1\} | a \neg x \leq b\}, \forall a, b \in \{0, 1\}$$

in fuzzy logic, conjunction represent t-norm, thus:

$$I(a, b) = \sup\{x \in [0, 1] | \top(a, x) \leq b\}, \forall a, b \in \{0, 1\}$$

S-implications

The implications based on $I(a, b) = \perp(\sim a, b)$ are called *S*-Implications. The symbol *S* is often used to denote the t-conorms.

R-implications

The implications based on $I(a, b) = \sup\{x \in [0, 1] | \top(a, x) \leq b\}$ are called *R*-implication. The symbol *R* represents close connection to residuated semi-group.

QL-implications

Implications based on $\perp(\sim a, \top(a, b))$ are called *QL*-implications, where *QL* stands for quantum logic.

Conclusion about fuzzy logic implications

All *I* come from generalizations of the classical implication. They collapse to the classical implication when truth values are 0 or 1.

Which fuzzy implication should be use for calculating the fuzzy relation R? Since the meaning of *I* is not unique, we must resolve this question. Hence meaningful criteria are needed.

3.6 Theory of Fuzzy Logic

3.6.1 The extension principle

How we extend a function $\phi : X^n \rightarrow Y$ in a fuzzy context in such a way that $\hat{\phi}$ has as domain a tuple of a fuzzy set and as codomain a fuzzy set? A particular case is the one given by the evaluation of the propositions.

Let \mathcal{P} be set of imprecise statements that can be combined by the operators "and" and "or":

$$\text{truth} : \mathcal{P} \rightarrow [0, 1]$$

$\text{truth}(a) = 0$ means that is definitely false

$\text{truth}(a) = 1$ means that is definitely true

Combination of two statements $a, b \in \mathcal{P}$:

$$\text{truth}(a \text{ and } b) = \min\{\text{truth}(a), \text{truth}(b)\}$$

$$\text{truth}(a \text{ or } b) = \max\{\text{truth}(a), \text{truth}(b)\}$$

it is possible also to consider the infinite conjunctions and disjunctions (statements $a_i, i \in I$):

$$\text{truth}(\forall i \in I : a_i) = \inf\{\text{truth}(a_i) | i \in I\}$$

$$\text{truth}(\exists i \in I : a_i) = \sup\{\text{truth}(a_i) | i \in I\}$$

This concept helps to extend $\phi : X^n \rightarrow Y$ to $\hat{\phi} : \mathcal{F}(X)^n \rightarrow \mathcal{F}(Y)$.

Another example for extending the additions is given by the real sum between sets:

$$+ : 2^{\mathbb{R}} \times 2^{\mathbb{R}} \implies 2^{\mathbb{R}}$$

$$(A, B) \rightarrow A + B = \{y | \exists a, b (y = a + b) \wedge (a \in A) \wedge (b \in B)\}$$

The extension to fuzzy set will be defined for μ_1 and μ_2 as:

$$+ : \mathcal{F}(\mathbb{R}) \times \mathcal{F}(\mathbb{R}) \rightarrow \mathcal{F}(\mathbb{R}), (\mu_1, \mu_2) \rightarrow \mu_1 \oplus \mu_2$$

$$\begin{aligned} & \text{truth}(y \in \mu_1 \oplus \mu_2) = \\ & = \sup_{a,b} \{ \text{truth}(y = a + b) \wedge \text{truth}(a \in \mu_1) \wedge \text{truth}(b \in \mu_2) \} = \\ & = \sup_{a,b:y=a+b} \{ \min(\mu_1(a), \mu_2(b)) \} \end{aligned}$$

In general, a function $\phi : X^n \rightarrow Y$ can be extended to $\hat{\phi} : [2^X]^n \rightarrow 2^Y$ on classic sets in this manner:

$$\hat{\phi}(A_1, \dots, A_n) = \{y \in Y | \exists (x_1, \dots, x_n) \in A_1 \times \dots \times A_n : \phi(x_1, \dots, x_n) = y\}$$

Based on this definitoin we can generalize to the fuzzy sets on a domain X :

$$\hat{\phi}(A_1, \dots, A_n) = \sup \{ \min\{\mu_1(x_1), \dots, \mu_n(x_n)\} | (x_1, \dots, x_n) \in X^n \wedge \phi(x_1, \dots, x_n) = y \}$$

assuming that $\sup \emptyset = 0$

3.6.2 Some relevant fuzzy sets

There are various types of fuzzy sets, in our case we give particular relevance to the one defined on \mathbb{R} . This because a fuzzy set on reals got a quantified meaning and can be used for represents fuzzy variables. This variables plays a really important role in lot of applications: fuzzy control, approximated reasoning, optimizations,... Some class of $\mathcal{F}(\mathbb{R})$ that are often referred in literature are the follwing:

- **Normal fuzzy set:**

$$\mathcal{F}_N(\mathbb{R}) = \text{def} \{ \mu \in \mathcal{F}(\mathbb{R}) | \exists x \in \mathbb{R} : \mu(x) = 1 \}$$

- **Upper semi-continuos fuzzy set:**

$$\mathcal{F}_C(\mathbb{R}) = \text{def} \{ \mu \in \mathcal{F}_N(\mathbb{R}) | \forall \alpha \in (0, 1] : [\mu]_\alpha \text{ is compact} \}$$

- **Fuzzy interval:**

$$\mathcal{F}_I(\mathbb{R}) = \text{def} \{ \mu \in \mathcal{F}_N(\mathbb{R}) | \forall a, b, c \in \mathbb{R} : c \in [a, b] \implies \mu(c) \geq \min\{\mu(a), \mu(b)\} \}$$

3.6.3 Quantitative fuzzy variables

Particular interest is held from the fuzzy interval also called *fuzzy numbers* because they allow to define quantitative fuzzy variables. These variables take as value the fuzzy numbers. When the quantity of the fuzzy variables represent linguistic concepts (small, big, ...) we are talking about *linguistic variables*. Each linguistic variable is defined by a quintuple (v, T, X, g, m) , where:

- v is the name of the variable.
- T is the set of linguistic terms of v .
- $X \subseteq \mathbb{R}$ is the base variable.
- g is the syntactic rule, also called grammar (used for generating linguistic terms).
- m is the semantic rule, which assigns meaning $m(t)$ to every $t \in T$, in essence:

$$m : T \rightarrow \mathcal{F}(X)$$

For representing this kind of variables we will need to extend the set and arithmetic operations used for numbers.

3.6.4 Set Representations of a fuzzy set

Definition

A family $(A_\alpha)_{\alpha \in (0,1)}$ of sets is called set representation of $\mu \in \mathcal{F}_N(\mathbb{R})$ if:

- $0 < \alpha < \beta < 1 \implies A_\alpha \subseteq A_\beta \subseteq \mathbb{R}$
- $\mu(t) = \sup\{\alpha \in [0, 1] | t \in A_\alpha\}$

This will assure that a representation for sets is a loyal image of the fuzzy set that represents.

Theorem

Let $\mu \in \mathcal{F}_N(\mathbb{R})$, the family $(A_\alpha)_{\alpha \in (0,1)}$ of sets is a set representation of μ if and only if

$$[\mu]_\underline{\alpha} = \{t \in \mathbb{R} | \mu(t) > \alpha\} \subseteq A_\alpha \subseteq \{t \in \mathbb{R} | \mu(t) \geq \alpha\} = [\mu]_\alpha$$

is valid $\forall \alpha \in (0, 1)$.

If μ_1, \dots, μ_n are normal fuzzy sets on \mathbb{R} and $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ a function $\hat{\phi}$ la sua estensione fuzzy. Then the following holds:

$$\forall \alpha \in [0, 1] : [\hat{\phi}(\mu_1, \dots, \mu_n)]_\underline{\alpha} = \phi([\mu_1]_\underline{\alpha}, \dots, [\mu_n]_\underline{\alpha})$$

$$\forall \alpha \in [0, 1] : [\hat{\phi}(\mu_1, \dots, \mu_n)]_\alpha \subseteq \phi([\mu_1]_\alpha, \dots, [\mu_n]_\alpha)$$

if $(A_\alpha)_{\alpha \in (0,1)}$ is a set representation of μ_i for $1 \leq i \leq n$, then $(\phi((A_1)_\alpha, \dots, (A_n)_\alpha))_{\alpha \in (0,1)}$ is a set representation of $\hat{\phi}(\mu_1, \dots, \mu_n)$.

3.6.5 Fuzzy Relations

A **crisp relation** represents presence or absence of association, interaction or interconnection between elements of ≥ 2 sets. This concept can be generalized to various degrees or strengths of association or interaction between elements.

A **fuzzy relation** generalizes these degrees to membership grades. So, a crisp relation is a restricted case of a fuzzy relation.

Definition of Relation

A relation among crisp sets X_1, \dots, X_n is a subset of $X_1 \times \dots \times X_n$ denoted as $R(X_1, \dots, X_n)$ or $R(X_i | 1 \leq i \leq n)$. So the relation $R \subseteq X_1, \dots, X_n$ is a set too.

The basic concept of sets can be also applied to relations: containment, subset, union, intersection, complement. Each crisp relation can be defined by its characteristic function:

$$R(x_1, \dots, x_n) = \begin{cases} 1, & \text{if and only if } (x_1, \dots, x_n) \in R \\ 0, & \text{otherwise} \end{cases}$$

The membership of (X_1, \dots, X_n) in R means that the elements of the tuple are related to each other. A relation can be written as a set of ordered tuples, this $R(X_1, \dots, X_n)$ represents n -dimensional membership array $R = [r_{i_1}, \dots, r_{i_n}]$. Each element i_1 of R corresponds to exactly one member of X_1 . Each element of i_2 of R corresponds to exactly one member of X_2 , and so on.

If $(x_1, \dots, x_n) \in X_1 \times \dots \times X_n$ corresponds to $r_{i_1}, \dots, r_{i_n} \in R$ then:

$$r_{i_1}, \dots, r_{i_n} = \begin{cases} 1, & \text{if and only if } (x_1, \dots, x_n) \in R \\ 0, & \text{otherwise} \end{cases}$$

The characteristic function of a crisp relation can be generalized to allow tuples to have degrees of membership. Similar to the generalization of the characteristic function of a crisp set. A fuzzy relation is a fuzzy set defined on tuples (x_1, \dots, x_n) that may have varying degrees of membership within the relation.

The membership grade indicates the strength of the present relation between elements of the tuple. The fuzzy relation can also be represented by an n -dimensional membership array.

Subsequences

Consider the Cartesian product of all sets in the family

$$\mathcal{X} = \{X_i | i \in \mathbb{N}_n = \{1, 2, \dots, n\}\}$$

for each sequence (n-tuple)

$$x = (x_1, \dots, x_n) \in \times_{i \in \mathbb{N}_n} X_i$$

and each subsequence (r-tuple, $r \leq n$)

$$y = (y_1, \dots, y_r) \in \times_{j \in J} X_j, J \subseteq \mathbb{N}_n \text{ and } |J| = r$$

y is called subsequence of x if and only if $y_j = x_j, \forall j \in J$. The symbol $y \prec x$ denotes that y is subsequence (precedes) of x .

Projection

Given a relation $R(x_1, \dots, x_n)$. Let $[R \downarrow \mathcal{Y}]$ denote the projection of R on \mathcal{Y} . It disregards all sets in X except those in the family

$$\mathcal{Y} = \{X_j | j \in J \subseteq \mathbb{N}_n\}$$

Then $[R \downarrow \mathcal{Y}]$ is a fuzzy relation whose membership function is defined on the Cartesian product of the sets in:

$$\mathcal{Y}[R \downarrow \mathcal{Y}](y) = \max_{x \succ y} R(x)$$

Under special circumstances, this projection can be generalized by replacing the max operator by another t -conorm.

Cylindric extension

Let \mathcal{X} and \mathcal{Y} denote the same families of sets as used for projection. Let R be a relation defined on Cartesian product of sets in family \mathcal{Y} .

Let $[R \uparrow \mathcal{X} \setminus \mathcal{Y}]$ denote the cylindric extension of R into sets $X_i, (i \in \mathbb{N}_n)$ which are \mathcal{X} but not in \mathcal{Y} . For each x with $x \succ y$

$$[R \uparrow \mathcal{X} \setminus \mathcal{Y}](x) = R(y)$$

The cylindric extension:

- produces the largest fuzzy relation that is compatible with projection.
- is the least specific of all relations compatible with projection.
- guarantees that no information not included in projection is used to determine extended relation.

Cylindric closure

Relations that can be reconstructed from one of their projections by cylindric extension exist. However, they are rather rare and not more common that relation can be exactly reconstructed from several of its projections or by taking set intersection of their cylindric extensions.

The resulting relation is usually called **cylindric closure**. Let the set of projections $\{P_i | i \in I\}$ of a relation on \mathcal{X} be given, then the cylindric closure $cyl\{P_i\}$ is defined for each $x \in \mathcal{X}$ as

$$cyl\{P_i\}(x) = \min_{i \in I} [P_i \uparrow \mathcal{X} \setminus \mathcal{Y}_i](x)$$

where \mathcal{Y}_i denotes the family of sets on which P_i is defined.

3.7 Fuzzy binary relations

Binary relations are significant among n -dimensional relations. They are generalized mathematical functions. On the contrary to functions from X to Y , binary relations $R(X, Y)$ may assign to each element of X two or more elements of Y .

some basic operations on functions, e.g. inverse and composition, are applicable to binary relations as well.

Domain

Given a fuzzy binary relation $R(X, Y)$, its domain $\text{dom } R$ is the fuzzy set on X whose membership function is defined for each $x \in X$ as

$$\text{dom } R(x) = \max_{y \in Y} \{R(x, y)\}$$

in essence each $x \in X$ belongs to the domain of R to a degree equal to the strength of its strongest relation to any $y \in Y$.

Range and Height

The range ran of $R(X, Y)$ is a fuzzy binary relation on Y whose membership function is defined for each $y \in Y$ as

$$\text{ran } R(y) = \max_{x \in X} \{R(x, y)\}$$

The height h of $R(x, y)$ is a number defined by

$$h(R) = \max_{y \in Y} \max_{x \in X} \{R(x, y)\}$$

$h(R)$ is the largest membership grade obtained by any pair $(x, y) \in R$.

Representation and Inverse

Consider the membership matrix

$$R = [r_{xy}] \text{ with } r_{xy} = R(x, y)$$

Its inverse $R^{-1}(Y, X)$ of $R(X, Y)$ is a relation on $Y \times X$ defined by

$$R^{-1}(y, x) = R(x, y), \forall x \in X, \forall y \in Y$$

Then $R^{-1} = [r_{xy}^{-1}]$ representing $R^{-1}(y, x)$ is the transpose of R for $R(X, Y)$

$$(R^{-1})^{-1} = R, \forall R$$

Standard composition

Consider the binary relations $P(X, Y), Q(Y, Z)$ with common set Y . The standard composition of P and Q is defined as

$$(x, z) \in P \circ Q \Leftrightarrow \exists y \in Y : \{(x, y) \in P \wedge (y, z) \in Q\}$$

in the fuzzy case this is generalized by

$$[P \circ Q](x, z) = \sup_{y \in Y} \{ \min\{P(x, y), Q(y, z)\}, \forall x \in X, \forall z \in Z \}$$

If Y is finite, sup operator is replaced by max, then the standard composition is also called max-min composition.

Inverse of standard composition

$$[P(X, Y) \circ Q(Y, Z)]^{-1} = Q^{-1}(Z, Y) \circ P^{-1}(Y, X)$$

Note that the standard composition is not commutative.

Relational Join

Relational Join yields triples (whereas composition returned pairs). For $P(X, Y)$ and $Q(Y, Z)$, the relational join $P \star Q$ is defined by

$$[P \star Q](x, y, z) = \min\{P(x, y), Q(y, z)\}, \forall x \in X, \forall y \in Y, \forall z \in Z$$

3.7.1 Binary relations on a single set

It is also possible to define crisp or fuzzy binary relations among elements of a single set X . Such a binary relation can be denoted by $R(X, X)$ or $R(X^2)$ which is a subset of $X \times X = X^2$.

These relations are often referred to as directed graphs, which is also a representation of them:

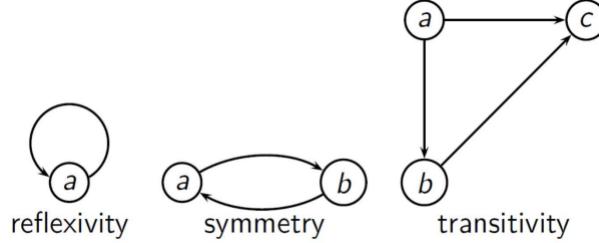
- Each element of X is represented as node.
- Directed connections between nodes indicate pairs of $x \in X$ for which the grade of the membership is nonzero.
- Each connection is labeled by its actual membership grade of the corresponding pair in R .

3.7.2 Properties of Crisp Relations

A crisp relation $R(X, X)$ is called:

- *reflexive* if and only if $\forall x \in X : (x, x) \in R$
- *symmetric* if and only if $\forall x, y \in X : (x, y) \in R \Leftrightarrow (y, x) \in R$

- *transitive* if and only if $(x, z) \in R$ whenever both $(x, y) \in R$ and $(y, z) \in R$ for at least one $y \in X$.



All these properties are preserved under inversion of the relation.

3.7.3 Properties of fuzzy relations

The previous properties can be extended for fuzzy relations, by defining them in terms of the membership function of the relation. A fuzzy relation $R(X, X)$ is called:

- *reflexive* if and only if $\forall x \in X : R(x, x) = 1$
- *symmetric* if and only if $\forall x, y \in X : R(x, y) = R(y, x)$
- *transitive* if it satisfies $R(x, z) \geq \max_{y \in Y} \min\{R(x, y), R(y, z)\}, \forall (x, z) \in X^2$

Note that a fuzzy binary relation that is reflexive, symmetric and transitive is called *fuzzy equivalence relation*.

3.8 Fuzzy controller

An application of particular success of fuzzy system in industry are the so called *fuzzy controller*. They are a special kind of non-linear table-based control method. The concept behind this is the one to define non-linear transitions between different states of the system without specifying a system of differential equation for each variable. So it is possible to define non-linear transition function without specifying each entry individually. The goal is to define a certain behaviour.

3.8.1 Table-based Control

Control systems all share a time-dependent **output variable**:

- Revolutions per minute
- Room temperature

The output is controlled by **control variable**

- Adjustment of current
- Thermostat

Also the *disturbance variables* influences the output

- Load of elevator, ...
- Outside temperature or sunshine through a window, ...

The computation of actual value incorporates both control variable measurements of current output variable ξ and change of output variable $\Delta\xi = \frac{d\xi}{dt}$

If ξ is given in finite time intervals, then set $\Delta\xi(t_{n+1}) = \xi(t_{n+1}) - \xi(t_n)$ in this case measurement of $\Delta\xi$ are not necessary.

Notation

- The input variables ξ_1, \dots, ξ_n
- The control variable η
- The measurements used to determine actual value of η
- η may specify change of η
- $\xi_i, 1 \leq i \leq n$ is value of $X_i, \eta \in Y$
- Contro function φ

$$\begin{aligned}\varphi : X_1 \times \cdots \times X_n &\rightarrow Y \\ (x_1, \dots, x_n) &\rightarrow y\end{aligned}$$

3.8.2 Problems of classical approach

- Often very difficult or even impossible to specify accurate mathematical model.
- Description with differential equations is very complex.
- Profound physical knowledge is needed.
- Exact solution can be very difficult.
- It should be possible to contro a process without a physical-mathematical model.

3.8.3 Fuzzy approach

- Simulate behavior of human who knows how to control.
- That is a *knowledge-based* analysis.
- Directly ask expert to perform analysis.
- Then expert specifies knowledge as *linguistic rules*.

1 - Fuzzy partitioning

- Determine linguistic terms X_1, \dots, X_n and Y is partitioned into fuzzy sets.
- Define p_1 distinct fuzzy sets $\mu_1^{(1)}, \dots, \mu_{p_1}^{(1)} \in \mathcal{F}(X_1)$ on set X_1
- Associate linguistic term with each set.
- Set X_1 corresponds to interval $[a, b]$ of real line, then $\mu_1^{(1)}, \dots, \mu_{p_1}^{(1)} \in \mathcal{F}(X_1)$ are triangular functions

$$\mu_{x_0, \epsilon} : [a, b] \rightarrow [0, 1]$$

$$x \rightarrow 1 - \min\{\epsilon \cdot |x - x_0|, 1\}$$

if $a < x_1 < \dots < x_{p_1} < b$, only $\mu_2^{(1)}, \dots, \mu_{p_1-1}^{(1)}$ are triangular

- Boundaries are treated differently

– left fuzzy set:

$$\begin{aligned} \mu_1^{(1)} &: [a, b] \rightarrow [0, 1] \\ x &\mapsto \begin{cases} 1, & \text{if } x \leq x_1 \\ 1 - \min\{\epsilon \cdot (x - x_1), 1\}, & \text{otherwise} \end{cases} \end{aligned}$$

– right fuzzy set:

$$\begin{aligned} \mu_{p_1}^{(1)} &: [a, b] \rightarrow [0, 1] \\ x &\mapsto \begin{cases} 1, & \text{if } x_{p_1} \leq x \\ 1 - \min\{\epsilon \cdot (x_{p_1} - x), 1\}, & \text{otherwise} \end{cases} \end{aligned}$$

3.8.4 Fuzzy relation

Consider only crisp sets. Solving a control problem means specifying control function

$$\varphi : X \rightarrow Y$$

where φ corresponds to relation

$$R_\varphi = \{(x, \varphi(x)) | x \in X\} \subseteq X \times Y$$

for measured input $x \in X$, the control value

$$\{\varphi(x)\} = \{x\} \circ R_\varphi$$

3.8.5 Fuzzy control rules

If temperature is very high **and** pressure is slightly low, **then** heat change should be slightly negative.

If rate of descent = positive big **and** airspeed = negative big **and** glide slope = positive big, **then** rpm change = positive big **and** elevator angle change = insignificant change.

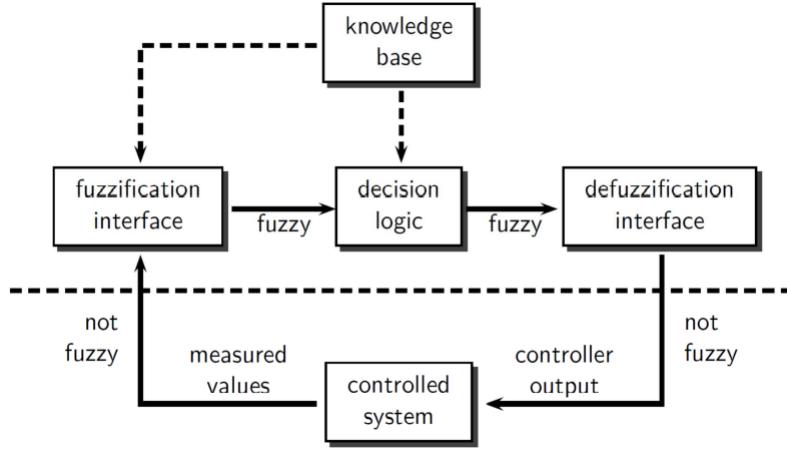
3.8.6 Approximate reasoning with fuzzy rules

Given r **if-then** rules and the fact " X is M' ", we can conclude that " Y is N' ". This kind of reasoning is typically used in fuzzy controllers.

Rule 1:	if X is M_1 , then Y is N_1
Rule 2:	if X is M_2 , then Y is N_2
...	...
Rule r :	if X is M_r , then Y is N_r
Fact:	X is M'
Conclusion:	Y is N'

3.8.7 Architecture of a fuzzy controller

- **Fuzzification interface:** which receives the current input value (and eventually maps it to a suitable domain).
- **Knowledge base:** consists of data base and rule base. Data base contains information about boundaries, possible domain transformations, and fuzzy sets with corresponding linguistic terms. The rule base contains linguistic control rules.
- **Decision logic:** which represents the processing unit. It computes the output from measured input according to knowledge base.

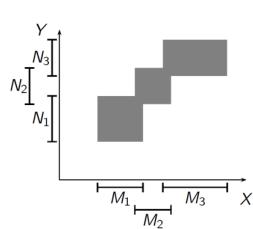


- **Defuzzification interface:** which represents also the processing unit. It determines the crisp output value, and eventually maps it back to appropriate domain (boolean) which are sent as signals to the control of the system.

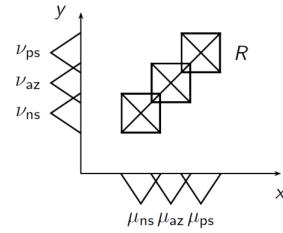
3.8.8 Defuzzification

The mapping of the internal fuzzy signals of the controller to boolean signals which are useful to control the system can be executed in different ways. The most common methods are:

1. Max Criterion Method (MCM)
2. Mean of Maxima (MOM)
3. Center of Gravity (COG)



(a) Rules with crisp values



(b) Rules with fuzzy values

The MCM choices an arbitrary value $y \in Y$ for which it reaches the maximum value of membership. This means that it can be applied to any kind of fuzzy set and to arbitrary Y domain. By the way, it could be hard to aim the

element for which the membership function be maximized. Also, the choice of random values make the behaviour of the controller no deterministic and this can bring to discontinuous actions. The MOM take Y as interval and computes the set Y_{MAX} such that the output in that points is the maximum. The output value will be computed as the average on Y_{MAX} . Like in the previous case, this technique can bring to discontinuous actions. The COG, as the MOM, take Y as interval and returns as output the center of that area. Usually it has a regular behavior, even if the computation is expensive and can bring to counterintuitive results.

3.8.9 Mamdani controller

The first model of fuzzy controller was the so called *Mamdani controller*, it was developed in the 1975 from Mamdani and Assilian. This controller is based on a series of rules like "if X is M_n then Y is N_m " where M_n and N_m are intervals that represents linguistic terms. Even if this rules are in the *if-then* form, they shouldn't be interpreted as logic implications, but as partial definition of a function. Collectively the rules can be represented in the space like the union S of the various intervals:

$$S = \bigcup M_i \times N_i$$

The rules can assume *crisp values* like in the figure 3.6a, or fuzzy values like in the figure 3.6b. In both cases, given an input x_0 the output will be computed as the composition of the singleton of x_0 and the union of the intervals of S , $\{x_0\} \circ S$. This output will be a fuzzy set which represents only a vague and imprecise description of the desired output. For determine the true output value, the preliminary output has to be defuzzified.

For doing this in the Mamdani controller we use the COG method, which allows to find a compromise respect to the single output values of the rules.

3.8.10 Takagi-Sugeno controller

This controller can be viewed as a variant end evolution of the *Mamdani controller*. In the same way the input values are described from the fuzzy sets. By the way, the consecutive of a rule won't be a fuzzy set but a function that got has values the input variables (usually a linear function)

$$R : \text{if } x_1 \text{ is } \mu_1 \text{ and } x_n \text{ is } \mu_n, \text{ then } y = f(x_1, \dots, x_n)$$

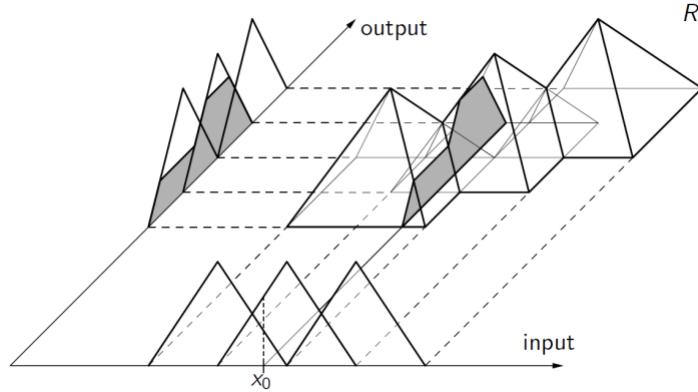


Figure 3.7: Projection of the output respect to the union of the intervals

The idea is that the function is a good function for controlling the described region from the previous rules. For maintaining the readability of the model constructed, occurs avoid overlapping between the various regions described previously from the rules. Since the output computed is already crisp, there is no need for defuzzification.

3.8.11 Similarity-baseed reasoning (controller)

There is a last type of controller which uses the concept of similarity relationship (which is the analogous of the fuzzy sets for the equivalence relationships).

A function $E : X^2 \rightarrow [0, 1]$ is defined as similarity relationship respect to a T-norm if and only if it satisfies the following conditions:

1. $E(x, x) = 1$
2. $E(x, y) = E(y, x)$
3. $\top(E(x, y), E(y, z)) = E(x, z)$

This kind of relationships are used for translating the informations data from the experts, in such a way that the various tuple span all the possible behaviours of the system. From the similarity classes we can extract the rules which are equals to the ones of *Mamdani controller*.

3.9 Fuzzy data analysis

There are two ways for talking about *fuzzy data analysis*. One refers the application of of fuzzy easoning techniques to crisp data (this takes the name of **fuzzy clustering**). Another, instead, concerns the analysis of data presented under the shape of fuzzy set (we will talk about **random set** and **random fuzzy variables**).

3.9.1 Fuzzy clustering

The *fuzzy clustering* is an unsupervised learning procedure which allow to divide the dataset in a way that:

- objects of the same cluster are the most possible similar
- objects from different cluster are the most possible non-similar between them. The similarity relationship is measured in terms of distance function, less is the distance, greater is the probability that 2 elements are from the same cluster.

in the case of the *hard c-means* algorithm:

1. choice a number c of clusters.
2. distribute in a random way the centers.
3. proceeds with the assignment of the nearest points to the centroids of the respective clusters.
4. pass to update the position of the centers through the computation of the center of gravity.
5. repeat the process until the convergence of the position.

The cluster partition is optimal when the sum of the distances between the centers and the elements is minimum. A problem with this approach is that the algorithm could get stucked to the local minimum. For avoiding this problem usually are operated various iterations and then select the best.

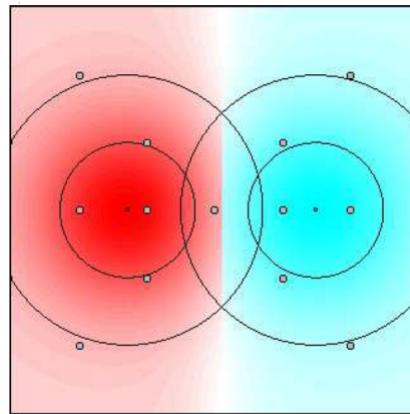


Figure 3.8: Symmetric dataset

A different problem is the one given by the fact that the partition is crisp. Whenever exists an element that is equidistant from two centers like in the

figure 3.8, the assignment to one of the two clusters is puraly arbitrary and doesn't reflects the information given by the data.

The *fuzzy clustering* gives a solution to this problem, introducing a concept of membership which is non-binary but conitnuos in $\in [0, 1]$, this offers the possibility of express the membership of a point to more than one cluster. The results is a partition of the data set to a fuzzy set. We can represent this partition through a matrix which assigns to each component u_{ij} the membership grade of the point x_j to the fuzzy set Γ_i , in symbols $\mu_{ij} = \mu\Gamma_i(x_j)$. There exists two type of fuzzy clustering: the **probabilistic** and the **possibilistic**. The differences plays respect to the imposed conditions of the membership function, in the probabilistic case:

1. $\sum_{j=1}^n u_{ij} > 0, \forall i \in \{1, \dots, c\}$
2. $\sum_{i=1}^c u_{ij} = 1, \forall j \in \{1, \dots, n\}$

The firs conditions indicate that can't exists empty clusters, instead the second means that the membership is worn out from the fuzzy sets that constitutes the partition. In the possiblitisc case we mantain only the first assumption, and we let go the second.

The possiblitisc interpretation is preferable when you have to manage data full of noises or outliers data.

3.9.2 Issues with fuzzy clustering

How can we know if the partition of the cluster executed by our algorithm reflects the implicit informations of the data? What is the correct number of cluster for a dataset? In the case we got a limited number of dimensions, we can represent visually the dataset and having an intuition of how many centers and in which positions locate them. By the way, in general, this is not the case.

Some criteria to research are: a clear separation between clusters, a minimum volume between clusters and the maximum number of points focussed near the centroid of the cluster. In literature are proposed various measured of this kind:

1. **Partition coefficient:** $PC = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2$
2. **Average partition density:** $APD = \frac{1}{c} \sum_{i=1}^c \frac{\sum_{j \in Y_i} u_{ij}}{\sqrt{|\Sigma_i|}}$
3. **Partition entropy:** $PE = \sum_{i=1}^c \sum_{j=1}^n u_{ij} \log u_{ij}$

3.9.3 Variants

The measure of distance more intuitive is the Euclidean, but this one has the inconvenient to allow only spherical clusters. Some variants are proposed for

relaxing the constraints. In the algorithm of *Gustafson-Kessel* the Euclidean distance is substituted with the *Mahalanobis* distance define respect to the cluster Γ_i as:

$$d^2(x_j, C_j) = (x_j - c_i)^\top \sum_i^{-1} (x_j - c_i)$$

where \sum_i is the covariant matrix of the cluster i . This algorithm is favorable in the clustering, used both for the automatic generation of fuzzy rules than for the controllers.

The dimension of the various clusters can change respect to the determinant of the matrix.

- in general the Gustafson-Kessel algorithm extracts more information than the standard algorithm, but it is also more sensible to a correct initialization (trade-off).
- it could be useful to decide on a good initialization to preliminary proceed with some iterations of the standard algorithm.
- since it is present the inverse of a matrix this algorithm is computationally more intensive, and harder to apply on big dataset than the standard one. Restrict the domain to clusters that are distributed along a parabola line respect to the axis reduce the computational cost.

Another approach consist in allow the non-convex clusters. The **shell clustering** algorithms manage to do that, and they are particularly useful for image recognition and analysis.

Name	Prototypes
adaptive fuzzy c-elliptotypes (AFCE)	line segments
fuzzy c-shells	circles
fuzzy c-ellipsoidal shells	ellipses
fuzzy c-quadratic shells (FCQS)	hyperbolas, parabolas
fuzzy c-rectangular shells (FCRS)	rectangles

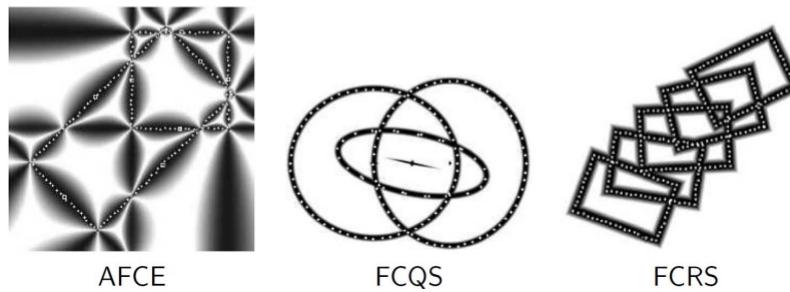


Figure 3.9: Some fuzzy shell clustering algorithms

Another possible approach is the **kernel-based clustering**, which is useful in the case we are managing non-vector data (like sequences, graphs, trees, ...). This method is based on a map:

$$\phi : \mathcal{X} \rightarrow \mathbb{H}$$

where \mathbb{H} is the Hilbert space and \mathcal{X} is the input space. In this way the data are not directly used, but only through their dot product (their existence is guaranteed since they are inside the Hilbert space). It is defined the kernel function k such that:

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}, \forall x, x' \in \mathcal{X} : \langle \phi(x), \phi(x') \rangle = k(x, x')$$

there is no need to know ϕ explicitly. Unlike other algorithms (other fuzzy shell clustering), the kernel methods doesn't extract prototypes from data for single clusters, but it computes the similarity relationship between the various inputs $\in \mathcal{X}$.

The centers $c_i^\phi \in \mathbb{H}$ are linear combinations of transformed data:

$$c_i^\phi = \sum_{r=1}^n a_{ir} \phi(x_r)$$

Some disadvantages are given by the difficulty of an adequate choice of the kernel function and of parameters, and the missing of an explicit representation of the single clusters.

Noise clustering

Last type algorithms are the *noise clustering* algorithms. They adds to c clusters one noise cluster. Which represents the group of all noisy data points or outliers, which are not associative to any other cluster.

- it has not explicitly associated to any prototype.
- It is directly associated to sitance between implicit prototype and data.

The center of the c cluster is choice in such a way to have constant distance δ between all the data points of the dataset.

- all points have same probability of belonging to noise cluster. During the optimization phase the probability is adapted.

3.9.4 Random sets

Until now we have applied fuzzy techniques to crisp data, now we want to extend this techniques in such a way to comprehend fuzzy descriptions. For doing this we have to introduce the concept of *random set*. Standard statistical data analysis is based on random variables $X : \rightarrow U$ A measureable mapping from a probability space to a set U ($U = \mathbb{R}$). A random set is the generalization

of this idea, in the sense that the value of the function won't be anymore an element of the set U , but a subset of U . Given a function $\Gamma : \Omega \rightarrow 2^U$, some useful concepts to define are the one of **upper and lower probability** (expressed as $P^*(A)$) which represent the proportion of elements that the image touches in a certain subsets of U .

$$P^*(A) = P(\{\omega \in \Omega | \Gamma(\omega) \cap A \neq \emptyset\})$$

Example

- $U = \{English, German, French, Spanish\}$
- Ω employees of a working group
- $\Gamma : \Omega \rightarrow 2^U$ collections of language that ω can speak

The other definition to explain is the one of *lower bound of probability*:

$$P_*(A) = P(\{\omega \in \Omega | \Gamma(\omega) \subseteq A \text{ and } \Gamma(\omega) \neq \emptyset\})$$

With these two instruments we can analyze data described in a fuzzy way. We can associate, for each element of the map a probability $E(\Gamma)$, in such a way that:

$$E(\Gamma) = \{E(X) | X(\omega) \in \Gamma(\omega) \text{ where } X \text{ is a random variable such that } E(X), \forall \omega \in \Omega\}$$

We can generalize the approach allowing to the Γ function to map the input in a fuzzy set.

3.10 Neuro-Fuzzy Systems

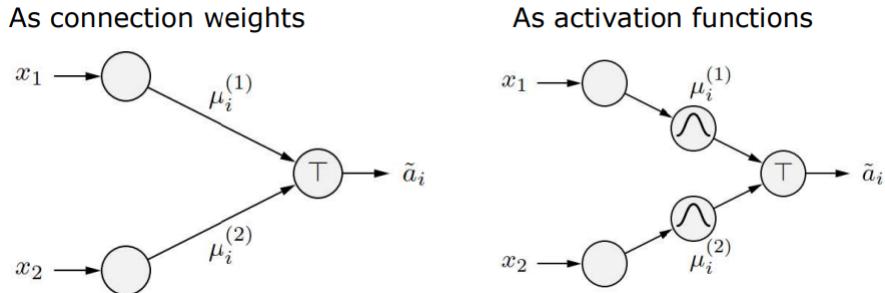
Contrary to the NN, the fuzzy systems manages the high-level reasoning, they do not adapt to the new environment, they use the linguistic informations relative to the domain and they are not based on data.

Neural Networks	Fuzzy Systems
are low-level computational structures	deal with reasoning on a higher level
perform well when enough data are present	use linguistic information from domain experts
can learn	neither learn nor adjust themselves to new environment
are black-boxes for the user	are based on natural language

The fuzzy NN combines the parallel computation and the learning capacity of NN with the high-level representation of fuzzy systems. This allow to have a better interpretation of the internal state of the NN during the computation. There are two modes in which fuzzy systems and neural networks can collaborate:

- *cooperative model*: the two systems works independently. The NN generate some parameters (offline) and then optimize (online) for the fuzzy controller.
- *hybrid model*: fuzzy sets and fuzzy rules are mapped inside a NN. The two data structures are integrated and there is no overhead due to communication. Both offline and online learning are available.

In the hybrid mode, the fuzzy sets that are antecedents to the fuzzy rules, can be modeled both as weight between neurons that as activation function of the neurons. In the first case, the neuron of the first layer represent the rule, in the second the neuron of the first layer represents the rule.



3.10.1 The algorithm

A set of fuzzy rules can be converted in a neural network following this procedure:

1. For each input variable x_i we create a neuron in the input layer.
2. For each output variable y_i we create a neuron in the output layer.
3. For each fuzzy set μ_i^j we create a neuron in the first hidden layer and we connect to the corresponding input neuron x_i .
4. For each fuzzy rule R_i we create a neuron in the second hidden layer and we specify a T-norm for calculating the antecedent of the rule.
5. Connect each neuron to the neurons that represent the fuzzy sets of antecedents relative of the corresponding rule.
6. Now the algorithm diverges in base at which kind of controller you want to use:
 - Mamdani-Assilian controller, it connects each "rule-neuron" to the corresponding output neuron respect to the fuzzy rule domain. As connection weight we can chose the fuzzy set of the corresponding fuzzy rule.

- Takegi-Sugeno-Kang controller, for each "rule-neuron" we create a twin that computes the output function corresponding to the fuzzy rule and then you connect all the input neurons.

The network constructed in this way can be trained with the back-propagation.

4 Evolutionary Computing

An **optimization problem** can be described from a triple Ω, f, \succ where Ω is the search space, f is an evaluation function of form $f : \Omega \rightarrow \mathbb{R}$ and \succ a comparison relationship $\in \{<, >\}$. The set of **global optima** $\mathcal{H} \subseteq \Omega$ is defined as:

$$\mathcal{H} = \{x \in \Omega | \forall x' \in \Omega : f(x) \succeq f(x')\}$$

Given an optimization problem, we want to find an element $x \in \Omega$ which optimized the function f in the whole search space. In literature are present various ways for solving optimization problems:

- Analytic solutions: efficient but rarely applicable.
- Brute-forcing (exhausting exploration): very inefficient, so only usable in small search spaces.
- Random search: always usable, but mostly inefficient.
- Guided search: similar elements in Ω have similar function values.

4.1 Motivation

The Evolutionary Algorithms (EA) are grounded on theory of biological evolution. The fundamental principles are the following:

- Beneficial traits resulting from random variation are favored by natural selection.
- Better chances of procreation and survival of individuals with beneficial traits.

Evolution theory explains diversity and complexity of species and allows unification of all different disciplines in biology.

4.1.1 Principles of organismic evolution

The following are the principles upon which EA are based:

- **Diversity:** All forms of life differ from each other (even from the same species). Different genetic material means **diversity of species**. The currently existing life forms are a tiny fraction of all theoretically possible ones.

- **Variation:** New **variants** are continuously created by mutation and genetic recombination (which is sexual reproduction).
- **Inheritance:** Variations are **heritable** as long as entering the germ line, also they are passed to the next generation.
- **Speciation:** Genetically diverge of individuals and populations.
- **Birth surplus:** More offspring than can ever become mature enough to procreate themselves.
- **Adaptation/Natural selection/Riproduzione differenziale:** On average, hereditary variations of the survivors of a population increase the adaptation to the local environment.
- **Randomness:** Variations are triggered/initiated or caused by random.
- **Gradualism:** Variations happen in comparatively small steps.
- **Evolution:** Adaptation to environment (species evolve in the course of time). The theory of evolution opposes creationism.
- **Discrete Genetic Units:** The store of genetic information in discrete units. Not continuously blend of hereditary traits (otherwise disappearance of differences).
- **Opportunism:** Processes of evolution work exclusively on what is present, the better/optimal solutions are not found if intermediary stages exhibit certain fitness handicaps.
- **Evolution-strategic Principles:** Not only organisms are optimized, but also the mechanisms of evolution: reproduction and mortality rates, life spans, vulnerability to mutations, ...
- **Ecological Niches:** Competitive species can tolerate each other if they occupy different ecological niches (the biospheres). Biological diversity of species is possible in spite of competition and natural selection.
- **Irreversibility:** The course of evolution is irreversible and unrepeatable.
- **Unpredictability:** The course of evolution is neither determined, nor programmed, not predictable.
- **Increasing complexity:** Biological evolution is increasingly more complex.

4.2 Ingredients of an evolutionary algorithm

- *Encoding* for the solution candidates (depends from the issue, there are no general rules).

- A method for creating an *initial population*, this is commonly created as simple random sequences, it depends on the chosen encoding.
- An *evaluation function* to evaluate the individuals. It represents the environment, and often it is the function to optimize.
- The *selection criterias* in relation with the evaluation function, in such a way that it is possible to select the individuals to procreate in the successive generation.
- A set of *genetico operators* to modify chromosomes, the most used operators are two: the **mutation** (randomly modifies the chromosomes) and **crossover** (recombines the chromosome of the parents for the children).
- Various *parameters* (population size, mutation, probability, ...).
- *Termination criterion.*

4.3 Formal definition

For every optimization problem: different representations of solution candidates

- EA separates space Ω , the phenotypes, from representation of the solution candidate in individual, the Genotypes \mathcal{G} .
- Fitness function f is defined on Ω .
- Mutation and Recombination is defined on \mathcal{G} .
- For evaluation, the transformation of genotype represented individual in Ω .
- Decoding function $dec : \mathcal{G} \rightarrow \Omega$ is a transformation of a genotype to the phenotype.

Definiton of individual

An **individual** A is a tuple $(A.G, A.S, A.F)$ containing the solution candidate $(A.G \in \mathcal{G})$, the optional additional information $A.S \in Z$ and the quality assessment $A.F = f(dec(A.G)) \in \mathbb{R}$

The mutation operator

A mutation operator is defined by mapping

$$Mut^\xi : G \times Z \rightarrow G \times Z$$

The recombination operator

The recombination operator with $r \geq 2$ parents and $s \geq 1$ offspring ($r, s \in \mathbb{N}$) is defined by the mapping

$$Rek^\xi : (G \times Z)^r \rightarrow (G \times Z)^s$$

The selection operator

The selection operator allows to choice due to the fitness value between a population of r individuals a number s of individuals that will continue the species. Let $P = \{A_1, \dots, A_r\}$ the population of individuals, then the selection operator will be defined by the mapping

$$Sel^\xi : (\Gamma \times Z \times \mathbb{R})^r \rightarrow (\Gamma \times Z \times \mathbb{R})^s$$

$$\langle A_i \rangle_{1 \leq i \leq r} \rightarrow \langle A^{(IS^\xi(c_1, \dots, c_r)_k)} \rangle_{1 \leq k \leq s} \text{ mit } A^{(i)} = (a_i, b_i, c_i)$$

where the selection has the shape

$$IS^\xi : \mathbb{R}^r \rightarrow \{1, \dots, r\}^s$$

Now we can define an evolutionary algorithm.

Fundamental evolutionary algorithm

A simple EA on an *optimization problem* (ξ, f, \succ) is an 8-tuple

$$(G, dec, Mut, Rek, IS_{Parents}, IS_{Environment}, \mu, \lambda)$$

where

- μ describes the amount of individuals of the parental population.
- λ defines the offspring per generation.

$$Rek : (\mathcal{G} \times \mathcal{Z})^k \rightarrow (\mathcal{G} \times \mathcal{Z})^{k'}$$

$$IS_{Parents} : \mathbb{R}^\mu \rightarrow (1, \dots, \mu)^{\frac{k}{k'} \cdot \lambda} \text{ with } \frac{k}{k'} \cdot \lambda \in \mathbb{N}$$

$$IS_{Environment} : \mathbb{R}^{\mu + \lambda} \rightarrow (1, \dots, \mu + \lambda)^\mu$$

Algorithm 1 General Scheme of an Evolutionary Algorithm

Input: optimization problem $(\xi, f, >)$
 $t \leftarrow 0$
 $\text{pop}(t) \leftarrow \text{create the initial population of size } \mu$
 $\text{evaluate } \text{pop}(t)$
while not termination criterion pop {
 $1 \leftarrow \text{select parents of offsprings with size } \lambda \text{ from } \text{pop}(t)$
 $\text{pop}_2 \leftarrow \text{create offspring by recombination of } \text{pop}_1$
 $\text{pop}_3 \leftarrow \text{mutate individuals in } \text{pop}_2$
 $\text{evaluate } \text{pop}_3$
 $t \leftarrow t + 1$
 $\text{pop}(t) \leftarrow \text{select } \mu \text{ individuals from } \text{pop}_3, \text{pop}(t - 1)$
}
return best individual of $\text{pop}(t)$

There is an internal distinction between EA:

- **Genetic algorithms**, where the encoding is a binary sequence.
- **Evolutionary algorithms** (classical), where the encoding depends from the problem, and so the genetic operators

4.4 Meta-heuristics

A *meta-heuristics* is an algorithmic method for finding approximated solutions for a combinatorial optimization problem. There are defined abstract sequences of steps that can be applied to any problem of this kind. Each single step has to be declined in base to the specificity of the problem.

The need for a meta-heuristic approach borns from the fact that some classes of problems doesn't have an efficient algorithmic solution. The approximation that you will obtain respect to the optimal solution depends from the definition of the problem and from the implementation of the single steps of the meta-heuristic algorithm.

4.4.1 Local Search Methods

Given an optimization problem (Ω, f, \succ) , we want to find an element $x \in \Omega$, which optimizes f (max. or min.). This method searches for the local maximum around the preselected points $\in \Omega$ during the initialization phase.

The assumption states that $f(x_1)$ and $f(x_2)$ differs for slightly similar $x_1, x_2 \in \Omega$, this means that there are no huge jumps in f .

The peculiarity of this evolutionary approach is that itself limits the population to just one individual. This has some consequences, especially for genetic operators: the recombination operator is not necessary anymore, we just limit to the mutation operator.

Algorithm 1 fundamental algorithm of local search

Input: objective function F
Output: solution candidate A

```
1:  $t \leftarrow 0$ 
2:  $A(t) \leftarrow$  create solution candidate
3: evaluate  $A(t)$  by  $F$ 
4: while termination criterion isn't fulfilled {
5:    $B =$  vary  $A(t)$ 
6:   evaluate  $B$  by  $F$ 
7:    $t \leftarrow t + 1$ 
8:   if  $Acc(A(t - 1).F, B.F, t)$  { /* acceptance criterion */
9:      $A(t) \leftarrow B$ 
10:   } else {
11:      $A(t) \leftarrow A(t - 1)$ 
12:   }
13: }
14: return  $A(t)$ 
```

At each step we can decide if continue to mutate the individual or to create a different one (this for avoiding local max/min). The idea is the one to use the *gradient descent* for identifying the max/min point by taking a step in the gradient direction. The width of the step shouldn't be too little because in that case the algorithm converges to slowly, nor too big because there will be oscillations.

Gradient ascent or descent

1. Choose a random starting point $x^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)})$

2. Compute the gradient at the current point x^t

$$\nabla_x f(x^{(t)}) = \left(\frac{\partial}{\partial x_1} f(x^{(t)}), \dots, \frac{\partial}{\partial x_n} f(x^{(t)}) \right)$$

3. Make a small step in the direction of the gradient

$$x^{(t+1)} = x^{(t)} + \eta \nabla_x f(x^{(t)})$$

where η is the step width parameter, the "learning rate".

4. Repeat the steps 2 and 3 until some termination criterion is fulfilled.

It is possible getting stuck in the local maxima

- Due to local gradient information, maybe only local maxima is reachable.
- Problem cannot be remedied in general.
- Chance improvement: multiple execution with different starting points (these can prevent the issues).

Hill Climbing

In the case where f isn't differentiable we can try to determine the direction where f increases by evaluating random points in the vicinity of the current point:

1. Choose a random starting points $x_0 \in \Omega$
 2. Choose a point $x' \in \Omega$ in the "vicinity" of x_t
 - 3.
- $$x_{t+1} = \begin{cases} x' & \text{if } f(x') > f(x_t) \\ x_t & \text{otherwise} \end{cases}$$
4. Repeat steps 2 and 3 until a termination criterion is fulfilled.

Algorithm 2 Acceptance criterion of Hill Climbing

Input: fitness of parents $A.F$, fitness of offspring $B.F$, generation t

Output: true or false

1: **return** $B.F \succ A.F$

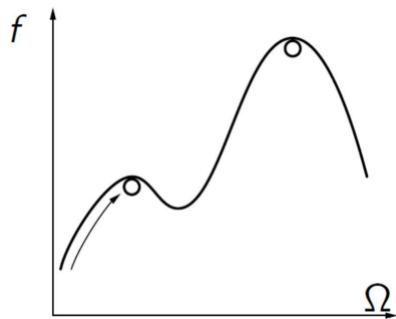
Figure 4.1: Pseudocode of acceptance criterion in fundamental algorithm

Simulated annealing

A generalization of the precedents approaches is the so called *simulated annealing*. The intuition behind this is that moving from the low to the high should be more probable than the contrary.

The guiding principle:

- Random variants of current solution are created.
- Better solutions are always accepted.
- Worse solutions are accepted with certain probability which depends on:
 - quality difference between current and new solution
 - temperature parameter



Algorithm 3 Acceptance criterion of Simulated Annealing

Input: parental fitness $A.F$, fitness of offspring $B.F$, generation t

Output: true oder false

```

1: if  $B.F \succ A.F$  {
2:   return true
3: } else {
4:    $u \leftarrow$  choose randomly from  $U([0, 1])$       /* random number
   between 0 and 1 */
5:   if  $u \leq \exp\left(-\frac{A.F - B.F}{kT_{t-1}}\right)$  {
6:     return true
7:   } else {
8:     return false
9:   }
10: }
```

Figure 4.2: Acceptance criterion of simulated annealing

1. Choose a random starting point $x_0 \in \Omega$
2. Choose a point $x' \in \Omega$ in the "vicinity" of x_t

3. Set

$$x_{t+1} = \begin{cases} x' & \text{if } f(x') \geq f(x_t) \\ x' \text{ with probability } p = e^{-\frac{\delta f}{kT}} & \text{otherwise} \\ x_t & \text{with probability } 1 - p \end{cases}$$

With $\Delta f = f(x') - f(x_t)$ as quality reduction of the solution, with $k = \Delta f_{max}$ estimate of the range of quality values, and T as temperature parameter.

4. Repeat steps 2 and 3 until a termination criterion is fulfilled.

Threshold accepting

Very similar to simulated annealing: worse solutions are sometimes accepted, but with an upper bound on quality degradation.

1. Choose a random starting point $x_0 \in \Omega$
2. Choose a point $x' \in \Omega$ in the "vicinity" of x_t
3. Set

$$x_{t+1} = \begin{cases} x' & \text{if } f(x') \geq f(x_t) - \theta \\ x_t & \text{otherwise} \end{cases}$$

θ threshold for accepting worse solution candidates, when $\theta = 0$ is equivalent to the standard hill climbing.

4. Repeat steps 2 and 3 until a termination criterion is fulfilled

Algorithm 4 Acceptance criterion of Threshold Accepting

Input: parental fitness $A.F$, fitness of offspring $B.F$, generation t

Output: **true** oder **false**

```
1: if  $B.F \succ A.F$  oder  $A.F - B.F \leq \theta$  {  
2:   return true  
3: } else {  
4:   return false  
5: }
```

Gread Deluge Algorithm

Very similar to simulated annealing: worse solutions are sometimes accepted, but absolute lower bound is used

1. Choose a random starting point $x_0 \in \Omega$
2. Choose a point $x' \in \Omega$ "in the vicinity" of x_t
3. Set

$$x_{t+1} = \begin{cases} x' & \text{if } (x') \geq \theta_0 + t \cdot \eta \\ x_t & \text{otherwise} \end{cases}$$

θ_0 lower bound for the quality of the candidate solutions at (initial "water level"))

4. Repeat steps 2 and 3 until a termination criterion is fulfilled

Record-to-Record travel

Similar to great deluge algorithm: rising water level is used, linked to the fitness of the best found individual

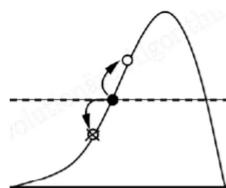
- possible degradation: always seen in relation to the best found individual.
- only if there is an improvement: current individual is important
- similar to threshold accepting: a monotonously increasing sequence of real numbers controls the selection of poor individuals

Algorithm 6 Acceptance criterion of Record-to-Record Travel

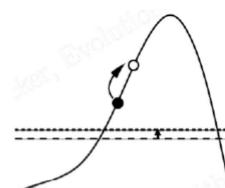
Input: parental fitness $A.F$, fitness of offspring $B.F$, t , best found quality F_{best}

Output: true oder false, F_{best}

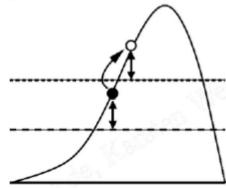
```
1: if  $B.F > F_{\text{best}}$  {
2:    $F_{\text{best}} \leftarrow B.F$ 
3:   return true,  $F_{\text{best}}$ 
4: } else {
5:   if  $B.F - F_{\text{best}} < T_t$  {
6:     return true,  $F_{\text{best}}$ 
7:   }
8: }
9: return false,  $F_{\text{best}}$ 
```



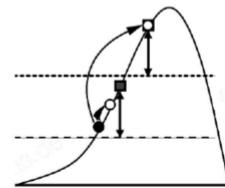
Hill Climbing



Great Deluge Algorithm



Threshold Accepting



Record-to-Record Travel

Figure 4.3: Comparison of local search algorithms

4.4.2 Tabu search

Local search method that considers history when creating new individual. Tabu lists avoid recurrence on previously considered solution candidates. They are implemented as FIFO queue with fixed length, where the entries are complete solution candidate. The mutations are not allowed to create tabu entries. If the new candidate shows up "interesting properties" we can do an exception to the tabu.

4.4.3 Memetic algorithms

This algorithms unified the pros of the *population-based* (slow, but gives more information) approach with the *local seach*. (fast, but sensitive to local minimum). The origin of the name "memes" are elements of the behavior which can be acquired individually. Each new individual will be optimized immediately only in few steps or till local optimum.

This often allow to accelerate the optimization process, but the evolutive dynamic are limited in a critical way: the mutations risks to stuck frequently in local optima, the recombination has a limited starting situation and parts of the space are perhaps unreachable.

4.4.4 Differential evolution

No adaptation of the step width in the set of strategies A.S, we try to use the relationships between individuals in the population as base for the calculation. We introduce a particular genetic operator *DE-operator*. This operator can be seen as a combination of the recombination operator and the mutation operator. In the selection, instead, a descendent can substitute his ancestors if and only if has a better fitness value.

4.4.5 Scattered search

The idea at the base of the *scatter search* algorithms is the one of having a population and to operate a local search around the individuals. Given the values registered from this research, we force the evolution in base of the direction of the maxima registered.

This is a purely deterministic method (different from the precedent methods).

It is an iterative process with two phases:

1. Generating of new individuals and selection of those which guarantee the most possible variety.
2. Recombination of all "pairings" of the chosen individuals, selection of the best individuals and iteration until nothing changes anymore.

4.4.6 Cultural algorithm

Instead of considering only the genetic information, we can also consider the cultural informations, relative to the skills learned by the previous generations. The cultural algorithms tries to take advantage from this generational memory in such a way that individuals are influenced from this last. There are two kinds of cultural knowledge:

- Situational knowledge: relative to generations which are adjacent.
- Normative knowledge: always relevant.

4.5 Elements of evolutionary algorithms

4.5.1 Encoding

The solution of our problem has to be encoded in such a way that is possible to explore the search space of possible solutions through this representation. There is no general receipt: the encoding issue is specific to the problem. By the way there are some generic hints to follow:

1. Represents the similar phenotype with similar genotypes. This assures that mutation of certain genes results in similar genotypes and that radical change allows to evade local minima.
2. The fitness function has to return similar values for similar candidates. This prevents that we choose an encoding which is too much or too little epistatic (in biology an allele of an epistatic gene kills the effects of all possible alleles of another gene. In the context of evolutionary algorithms states the grade of interactions between the chromosomes of the genes).
3. The space Ω has to be close respect to genetic operators. This for obvious reasons, if the search space is not closed respect to the genetic operators, a single modified chromosome couldn't be decoded or interpret.

4.5.2 Fitness

Better individuals should have better chances to create offspring. So we must apply the so called **selective pressure**, which is the strength of preferring good individuals:

- Exploration of the space
 - Deviation of the individuals over Ω as wide as possible.
 - Preferably big chances to find global optimum.
 - Smaller selective pressure.
- Exploitation of good individuals
 - Strive for optimum in the vicinity of good individuals.
 - Convergence to optimum.
 - Higher selective pressure.

There are two metrics for creating selective pressure:

- **Time to takeover:** number of generations until population converges.
- **Selection intensity:** the differential between average quality before and after the selection.

The same selection methods can change in base to the evolutive pressure. One of the most used is the *roulette-wheel selection*. The fitness value is computed relative of each individuals by following:

$$f_{rel}(A_i) = \frac{A_i \cdot F}{\sum_{j=1}^{|P|} A_j \cdot F}$$

The probability of an individuals of being selected for the reproduction will be proportional to his relative fitness values. Some disvantages are:

- the computation of the relative fitness values is expensive and hard to parallelize.
- the individuals with an high fitness value could dominate the selection (no more biodiversity).
- really fast to find optimal locals, but really bad in exploring the space.

The same fitness function can be adapted for avoiding a too fast convergence:

- *linear dynamical scaling*: reduce the relevance of the fitness function subtracting the minimum of the past generations.
- σ -*scaling*: calculated with the formula $f_\sigma(A) = A \cdot F - (\mu_f(t) - \beta \cdot \sigma_f(t))$, where β is a positive parameter.
- *time depending*: the temporal factor is used as exponent rule of selective pressure.
- *Boltzmann-selection*: determines the fitness non directly relative, but through the function $g(x) = \exp^{\frac{f(x)}{kT}}$. T è una variabile that depends from the time and k is the normalizzazion costant.

4.5.3 Selection

There are various strategies available in literature for operating the selection of the individuals which will constitute the genetic pool for the successive generation:

- *Roulette-wheel selection*: Selection of individuals is proportional to the fitness, but random. No guarantee that "fitter" individuals are taken to the next generation, not even for the best individual. the first simple solution is the discretization of the fitness range. (Same approach already saw before for roulette-wheel selection).
- *Rank-based selection*:
 1. Sort individuals in descending order according to their fitness.
 2. Rank is assigne to each individual in the population.

3. Define probability distribution over Rank scale: the lower the rank the lower the probability.
4. Roulette-Wheel selection based on the distribution.

The advantage consist in the avoiding of the dominance issue by decoupling the fitness value and selection probability, and also a regulation of the selective pressure by probability distribution on rank scale. The disadvantage is given by the sorting of the individuals, brings to a complexity $O(n \log n)$.

- *Tournament selection:*

1. Draw k individuals randomly from $P(t)$.
2. Individuals carry out the tournament and best individual wins: tournament winner receives a descendant in the next population.
3. All participants of the tournament are returned to $P(t)$.

The advantage are the same of rank-based selection.

- *Elitist:* Only the expected value model ensures that the best individual enters the next generation. The best individuals of the previous generations constitute the next generation. The chosen elite is not immune from the changes given by the genetic operators. The advantage is that the convergence is achieved rapidly. The disadvantage is that there is the risk of remaining stucked in local optima.
- *Crowding:* The individuals from the next generations should replace the individuals that are more similar to them. The local density in Ω can't grow in an undefined way. This allow a better exploration of the space.

There are some properties that can distinguish the selection methods:

- **Static:** the probabiltiy of selection remains costant.
- **Dynamic:** the probabiltiy of selection can change.
- **Extinguishing:** the probabiltiy of selection could be 0.
- **Preservative:** the probability of selection is always greater than 0.
- **Pure-bred:** the individuals can have descendants only in one generation.
- **Under-bred:** the individuals can have descendants in more than one generation.
- **Right:** all the individuals can reproduce.
- **Left:** the best individuals cannot reproduce.
- **Generational:** the parents can't mutate until their descendants are created.
- **On-the-fly:** the descendent substitutes the parents.

4.5.4 Genetic operators

The genetic operators are applied to a fraction of chosen individuals. There are so generated mutations and recombinations of already existing solutions. The genetic operators are so classified in base to the number of parents:

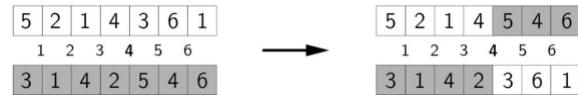
- One-parent operators.
- Two-parent operators.
- Multiple-parent operators.

In the first class we can find the mutation operator, the one which introduce little random changes inside the genome of the solution in which it is applied. Results useful for introducing the biodiversity inside the solution pool and promoting the exploration of the search space. There are various methods for operating the mutation:

- **Standard mutation:** the value of one or more genes is muted.
- **Pair swap:** swap of the positions of two genes.
- **Shift:** right or left shift of a group of genes.
- **Arbitrary permutation:** arbitrary permutation of a group of genes.
- **inversoin:** there is an inversion of the occurrences order of a group of genes.

Instead, an important operator in the two-parent class is the *recombination* or *crossover* operator, given two solutions it creates through a combination of their genetic code the solutions that will constitute the future generation. There are various way to operate this recombination:

- *One-point crossover:* determination of a random position inside the chromosome, and then swap the two sequences on one side of the cutting line.



- *Two-point crossover:* determining of two random cutting points, exchange of the gene sequences between both cutting points.



- **N-point crossover:** this is a generalization of the previous methods. Determining of n random cutting points, alternating exchange/keep of the gene sequences between two following cutting points.
- **Uniform crossover:** On each gene, determine whether to exchange or not (based on a certain parameter of probability).



- **Shuffle crossover:**

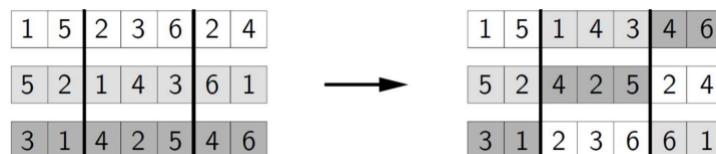
1. Random permutation of the genes.
2. One-point crossover.
3. Unmixing the genes.

- **Uniform order-based crossover:** similar to the uniform crossover, for each gene you choice if keeping it or change it. The spaces are filled in the order of occurrence.



- **Edge recombination crossover:** the chromoseome is represented like a graph. Each gene is a vertex with edges to the neighbours. The edges of the two graphs (chromosome) are mixed. The information relative to the distance is kept.

A case of *multiple-parent* operator is the one of **diagonal crossover**. Which is similar to the n -point crossover, but in this case both parents are partecipating. Given k parents, we choose $k - 1$ points for the crossover and we procede by diagonally shifting the sequences respect the chosen points.



By increasing the number of parents we obtain a optimum grade of exploration of the search space. Some properties which can characterize the crossover operators are:

- **Positional bias:** when the probability that 2 genes are inherited together from the same parent depends from the relative position of the two genes in the chromosome. Undesired since it can make the exact arrangement of the different genes in a chromosome crucial for the success or failure of an evolutionary algorithm.
- **Distributional bias:** when the probability of a certain number of genes are exchanged between parents is not the same for all the possible numbers of genes. This has to be avoided because the partial solutions of different lengths have different probabilities of progress to the next generation. In general, it is less problematic of the positional bias.

For making better the performances of the solutions we have two strategies:

- **Interpolating recombination:** Blend the traits of the parents in such a way that offspring with new traits is created. This creates new alleles, focuses the population on 1 main area. To explore Ω sufficiently at the beginning: using a strong random and diversity-preserving mutation.
- **Extrapolating recombination:** Try to infer information from several individuals, create a prognosis in what direction one can expect fitness improvements.

4.5.5 Adaptation strategies

One last aspect to consider are the **adaptation strategies** which answers to the question: *should we allow that the mutation introduce heavy modification to the phenotype during the optimization?* For answer this question we need a metric for measuring the improvement in fitness between the individual and the muted individual.

Fitness improvement

The fitness improvement of an individual $A \in \mathcal{G}$ to another individual $B \in \mathcal{G}$ is defined as:

$$Improvement(A, B) = \begin{cases} |B.F - A.F| & B.F \succ A.F \\ 0 & \text{otherwise} \end{cases}$$

Relative expected improvement

The relative expected improvement of an operator Mut concerning individual A can be defined as:

$$relEV_{Mut,A} = E(Improvement(A, Mut^\xi(A)))$$

Given this metric it is possible to judge the performance of a particular genetic operator. The quality of a mutation can't, being judge independently of the current fitness level. The operator is never optimal over the complete process of optimization.

Exists various adaptation strategies:

- *Predefined adaptation*: the changes are defined before of the run.
- *Adaptive adaptation*: is defined a metric for establishing which operators use during the run.
- *Self-adaptation*: extra informations are used on the individuals.

4.6 Meta heuristics

4.6.1 Swarm and population based optimization

The **swarm based optimization** and the **population based optimization** are two meta-heuristics used in literature for the development of multi-agent intelligent systems, able to perform a cooperative behaviour. The concept of **swarm intelligence**, used for describing in nature the behavior of some species like bees, ants, ..., means that the capacity of the population of cooperating for the solution of a problem. The idea that single individuals can exchange informations between them without using a central control unit. There are various typology of this kind of heuristics:

- **Particle swarm optimization**: inspired to the biological pattern of food research from birds and fishes. The individuals aggregates informations, creating a set of shared knowledge for the end goal to propose a single solution. Each individual is a possible candidate for being the solution.
- **Ant colony optimization**: inspired to the biological pattern of the ants that search a way that leads to the food. The individuals exchange the informations by editing the environment in such a way that the others could follow the traces. Each individual is a candidate for the solution.

Instead, from the *population based optimization* side we will find the so called *population-based incremental learning*, where the individuals are generated randomly in accord to a probability distribution. In reality, we don't need to conserve the memory of the individuals in an explicit way, it is sufficient conserve the statistics of the population. As recombination operator is used the *uniform crossover*. For the selection, are chosen the individuals that improve the statistics of the population. The mutation, instead, is limited to a simple *bit-flip*.

The distinctive feature is the *learning rate*, the parameter which tunes the possibility of movement of the individuals in the space, this changes in the time and reduces with the number of iteration. This allow initially a great mobility, for establishing then when an optimum is found. Some problems with this strategy:

1. The algorithm can learn some accidental dependency between the chromosome of the individuals and considering the single bits individually isolated from the others.
2. The same statistics representation of the population can be used for different populations.

4.7 Theoretical fundations

Why evolutionary algorithms works? We have to consider the **schemata**, which are binary chromosomes that are only partially specified and encode a particular behavior. Then we investigate how the number of chromosomes matching a schema evolve over several generations. The objective is to give a rough stochastic statement that describes how a genetic algorithm explores the search space.

Definition I

A scheme h is a string of symbols of length L on the alphabet $\{0, 1, *\}$, $h \in \{0, 1, *\}^L$. The character $*$ is a wild card (or "don't care symbol").

Definition II

A chromosome $c \in \{0, 1\}^L$ matches a schema h , if and only if it coincides with h at all positions where h is 0 or 1.

For misuring the effects of the selection occurs computes the fitness of the chromosome that shares a certain schema. Usually as measure is chosen the *mean relative fitness* of chromosome that match the schema h in the population $P(t)$. The individuals that shares the schema in the next generation will be proportional to this value.

Instead, for calculating the influence of the mutation operator we need to measure the probability, that the after the mutation the schema is preserved. We define the concept of **order of a scheme**.

Definition III

The order of a schema h is the number of zeroes and ones in h , that is $ord(h) = \#0 + \#1 = L - \#*$, where $\#$ is the "number of occurrences of".

Now we can calculate the probability that a mutation operator preserves the schema h thanks to the formula:

$$(1 - p_m)^{ord(h)}$$

where the p_m is the "mutation probability". In conclusion, the probaiblity that the crossover preserves the schema will be mesaured respect to the **defining length** (considering the one-point crossover).

Definition IV

The defining length of a schema h is the difference between the position of the last 0/1 and the first 0/1 in h .

Definition V

The probability that the cut point splits a chromosome in such a way that some

of the fixed characters of a schema lie on one side of the cut and some one the other is:

$$\frac{dl(h)}{L - 1}$$

Now let's define important concepts for the following:

Definitino VI

$N(h, t)$ is the expected value of the number of chromosomes that match the schema h during the t -th generation.

Definition VII

$N(h, t + \Delta t_s)$ is the expected value of the number of chromosomes that match the schema h during the t -th generation after selection.

Definition VIII

$N(h, t + \Delta t_s + \Delta t_x)$ is the expected value of the number of chromosomes that match the schema h during the t -th generation after selection and crossover.

Definition IX

$N(h, t + \Delta t_s + \Delta t_x + \Delta t_m) = N(h, t + 1)$ is the expected value of the number of chromosomes that match the schema h during the t -th generation after selection, crossover and mutation. The various Δ will depends from the probabilities of the single operators defined.

Schema theorem

It says that the schemas with a fitness score above the average, a short defining length and a low order, reproduces in an approximately exponential way.

No free lunch theorem

The theorem says that there is no evolutionary algorithm that can be used for all the problems, no other algorithms is superior to the average. The "right" choice and encoding of the "correct" genetic operators depends from our local knowledge relative to the specific problem.

4.8 Genetic programming

Genetic programming creates programs to solve a given application problem. It is a universal way of learning computer programs. GP is based on:

- Describing a solution by connecting certain inputs with a certain output.
- Searching for a matching computer program.
- Representing programs by parse trees.

Essentially the GP it is a family of evolutionary algorithms that allows the automatic creation of programs that can solve problems. For doing that, we need an encoding for representing and manipulate a single program.

For doing this we use *parse trees* (usually), where the internal nodes are the operation and the leaves are variables or constants. F is the set of function symbols and operators, and T is the set of terminal symbols, this two sets are specific for each problem.

Example 1

Example for learning a boolean function:

$$F = \{\text{and}, \text{or}, \text{not}\}$$

$$T = \{x_0, \dots, x_n, 1, 0\}$$

The GP can only solve problems efficiently and effectively, if function and terminal symbol sets are sufficient/complete to ensure an appropriate program can be found. Finding the smallest set is often NP-hard.

Example 2

Can be also helpful for expressions like chromosomes, which are composition of elements from $C = F \cup T$ and brackets.

Algorithm

1. Generate an initialization population of random expressions.
2. Evaluate these expressions by calculating their fitness.
3. Selection with one of the Evolutionary Algorithms strategies.
4. Application of genetic operators, usually only crossover.

4.8.1 Initialization

We need to generate a random population like with the EA. Given the complexity of this structures in the creation process, we need to consider some parameters like the *maximum height* and *maximum number of nodes*. Exists various sub-algorithms that initializes the parse trees:

1. *Grow*: the probability of choosing an internal node or a leaf is distributed in an uniform way at any level of depth. This allow to create "unbalanced" trees.
2. *Full*: the leaves may occur only at maximum height of the tree. This allow the creation of balanced trees.
3. *Ramp-half-and-half*: this approach mix the other two for having a better variation in the exhibited form from the syntax trees.

4.8.2 Genetic operators

The population initialized usually has no good fitness values. The evolutionary progress changes population via genetic operators. The 3 most important operators are:

- Crossover.
- Mutation.
- Clonale reproduction.

In the case of the **crossover** an usual approach is the one of exchanging two sub-expressions: chose two internal nodes and exchange between the two subtrees. In the case of **mutation**, instead, we exchange a subtree of a subexpression with one randomly generated.

4.8.3 Introns

During the evolutionary process the individuals tends to develop large portions of useless code towards the computation aim. A similar concept is given by biology: the **introns** are the portion of DNA that doesn't encode any information at phenotype level (junk DNA, inactive and maybe outdated strips within one gene that serves no function). For avoiding the occurrence of this phenomena, there are some strategies:

- *Breeding recombination*: generates many children using different parameters, with only the best one going on into the next generation.
- *Intelligent recombination*: chooses crossover points selectively.
- *Continuos slight changes*: if the evaluation function can change constraints thus inactive subprograms (introns) might become active again.

4.9 Evolutionary strategies

In an evolutionary strategy we don't only try to optimize the single individual, but the entire evolution process:

- Reproduction
- Mortality rate
- Life span

Which is susceptible to: *mutation, mutation step width, speed of evolution*.

What we do is to consider an optimization problem as a function $f : \mathbb{R} \rightarrow \mathbb{R}$ that we want to minimize (or maximize). The chromosome are represented by an array of real numbers. Then we use only the mutation operator (the crossover is not used), this for moving the chromosome vector inside the search space, this by adding a normal-distributed random vector r .

Elite principle

The selection process is applied to the individuals mutated, only the best individuals enter the next generation (elitism). μ is the number of individuals in the parent generation. λ is the number of offspring individuals that were created by the evolutionary strategy (ES).

4.9.1 Selection strategies

- *Plus-strategy*(+strategy): the selection works on both muted and not-muted individuals $\mu + \lambda$
- *Comma-strategy*(,strategy): it generates offspring of size $\lambda > \mu$, and then chose the individuals that will constitute the next generation, the best μ chromosomes. Chromosome of parent generation are lost.

The advantage of the +strategy:

- Due to strict elite principle, there are only improvements.

The disadvantages of the +strategy:

- Risk of getting stuck in local optima.
- If no improvement over further generations, temporary helpful to switch off to the comma strategy to overcome local minima (increases diversity in population).

The ,strategy could be useful for creating diversity in the population.

4.9.2 Adaptation

Adaptation of random vector's variance.

- Small variance, consist in small changes of chromosomes (local search, exploitation).
- High variance, big changes of chromosomes (global search).

4.9.3 Global adaptation

Choose σ in such a way that the mean convergence rate is optimized. Determining the probabilities for a successful mutation. A successful mutation improves the result.

1/5 success rule

The variance is appropriate when $\frac{1}{5}$ of muted individuals has a better fitness respect to the ones of past generation.

4.9.4 Local adaptation

It is possible to have a more local approach and conservative for each chromosome vector the associated variance as additional information. The chromosomes with "bad" variance will generate "bad" descendants. The chromosome with worse fitness value won't be able to access the following generation and will extinguish.

4.10 Multi-criteria Optimization

There are cases in which optimization problems have different objectives and constraints, possibly in conflicts, each one represented by a fitness function $f_i : \Omega \rightarrow \mathbb{R}$ (for $i = 1, \dots, k$). The most direct approach is the one of combine the various function into one aggregated objective function:

$$f(s) = \sum_{i=1}^k w_i \cdot f_i(x)$$

Each one of the single functions will receive an assigned weight that will reflect the relative relevance respect to the other parameters. The problem is given by finding an adequate distribution of the weights that respects the relevance criteria. If, the objectives are in conflicts will be harder to find a function that will aggregate them together. The problems of this simple approach are:

- Relative significance of criteria needs to be fixed before
- Choice of weights not always simple

The problem of aggregation of preferences:

- The **Arrow's impossibility theorem**: there is no choice function that has all the desired features. In general, it prevents the possibility of the existence of an aggregation function that maximize all the single functions.
- Scaled order of preferences, finding a suitable scaling is more complex than finding suitable weights.

A solution to the Arrow's impossibility theorem is to choose a solution only if it is an *Pareto-optimal*.

4.10.1 Pareto-Optimal solutions

An element $s \in \Omega$ is called Pareto-optimal regarding the objective functions f_i , for $i = 1, \dots, k$, if there is no such element $s' \in \Omega$ for which:

$$\forall i, 1 \leq i \leq k : f_i(s') \geq f_i(s)$$

and

$$\exists, 1 \leq i \leq k : f_i(s') > f_i(s)$$

holds.

No value of an objective function can get better without the value of another function getting worse. The set of Pareto-optimal elements is called **Pareto-front**.

The advantages when searching for Pareto-optimal solutions:

- No need for aggregating objective function.
- No need for choosing weights.
- The search has to be performed only once.
- After this search the solutions are chosen.

4.10.2 Pareto-Optimal solutions with EAs

It is possible to use EAs for finding the most possible number of Pareto-optimal solutions. The objective is to spread the population along the Pareto-Front as widely as possible. The simplest approach consists in using a weighted sum of the objective functions as fitness function.

Vector Evaluated Genetic Algorithm

- Given k criteria with assigned objective functions $f_i, i = 1, \dots, k$.
- $\forall i$: choose $|P|/k$ individuals according to the fitness function f_i .

The advantage is given by the fact that the approach is simple, with low computational effort. The disadvantage is given by the handicap for solutions that satisfy every criterion good, but none perfectly. As consequence the search concentrate on marginal solutions.

It is possible to apply a *selection by dominance* approach, where we do a rating of the individuals of one population and find all the non-dominated solutions of the population. Then we assess the solution candidates to the best rank, and remove them from the population. We repeat this identification and removal of non-dominated solution candidates for other ranks, until population is empty.

Another problem is given by the fact that all individuals of the Pareto-Front are assessed as equally good. This leads to **genetic drift**, Pareto-Front converges at a random point, because of random effects. For remedy to this we can apply niche techniques to be able to decide between individuals with same rank (e.g. **power law sharing**, if more individuals have similar fitness score, the new individual similar to them will have lower fitness score).

Non-Dominated Sorted Genetic Algorithm

1. Generate offsprings by applying genetic operators to the parent population.
2. Select non-dominated individuals from parents+offsprings population: they consistutes the front F_i at the i -th iteration.
3. Remove individuals in F_i from current population.
4. Repeat steps 2 and 3 until the current population is empty.
5. Fill the new generation by taking the individuals from the fronts F_i , from $i = 1$ forwards, until the complete F_i can be included.
6. Apply crowding by similarity to the next front and extract the missing individuals to complete the new generation.

This algorithm lead to a poor approximation of the Pareto-Front.

The strength of the Pareto EA:

- Simple EA.
- Evaluation function is composed by two components: number of individuals dominating this individual, and the distance to the \sqrt{n} -th individual.