of the Deity Collected from the Appearances of Nature, Oxford: J. Vincent.

[55] Pincus, M. (.), "A Monte Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems", *Oper. Res.*, 18, pp. 1225–1228.

[56] Rechenberg, I. (1965), "Cybernetic Solution Path of an Experimental Problem", *Roy. Aircr. Establ. Libr. Transl.*, 1122, Farnborough, Hants, UK.

[57] Russel, P. J. (1996), *Genetics*, 4th Ed., Harper Collins College Publishers.

[58] Russell, S. J. and Norvig, P. (1995), *Artificial Intelligence A Modern Approach*, Prentice Hall, New Jersey, U.S.A.

[59] Sarle, W. (1993), *Kangoroos*, article posted on *comp.ai.neural-nets* on the 1st September.

[60] Schwefel, H.-P. (1965), *Kybernetische Evolution als Strategie der Experimentellen Forschung in der Strömungstechnik*, Diploma Thesis, Technical University of Berlin, March.

[61] Schwefel, H.-P. (1981), *Numerical Optimization of Computer Models*, Wiley, Chichester.

[62] Schwefel, H.-P. (1995), *Evolution and Optimum Seeking*, Wiley: New York.

[63] Wolpert, D. H. and Macready, W. G. (1997), "No Free Lunch Theorems for Optimization", *IEEE Trans. on Evolutionary Computation*, **1**(1), pp. 67–82.

[64] Wright, S. (1968–1978), *Evolution and the Genetics of Population*, 4 vols., University of Chicago Press, Chicago, IL.

[65] Zebulum, R. S., Pacheco, M. A., Vellasco, M. M. B. and Zebulum, R. S. (2001), *Evolutionary Electronics: Automatic Design of Electronic Circuits and Systems by Genetic Algorithms*, CRC Press.

NEUROCOMPUTING

"Inside our heads is a magnificent structure that controls our actions and somehow evokes an awareness of the world around ... It is hard to see how an object of such unpromising appearance can achieve the miracles that we know it to be capable of."
(R. Penrose, The Emperor's New Mind, Vintage, 1990; p. 483)

"Of course, <u>something</u> about the tissue in the human brain is necessary for our intelligence, but the physical properties are not sufficient ... Something in the <u>patterning</u> of neural tissue is crucial."
(S. Pinker, How the Mind Works, The Softback Preview, 1998; p. 65)

4.1 INTRODUCTION

How does the brain process information? How is it organized? What are the biological mechanisms involved in brain functioning? These form just a sample of some of the most challenging questions in science. Brains are especially good at performing functions like pattern recognition, (motor) control, perception, flexible inference, intuition, and guessing. But brains are also slow, imprecise, make erroneous generalizations, are prejudiced, and are incapable of explaining their own actions.

Neurocomputing, sometimes called *brain-like computation* or *neurocomputation*, but most often referred to as *artificial neural networks* (ANN)[1], can be defined as information processing systems (computing devices) designed with inspiration taken from the nervous system, more specifically the brain, and with particular emphasis on problem solving. S. Haykin (1999) provides the following definition:

> "A[n artificial] neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use." (Haykin, 1999; p. 2)

Many other definitions are available, such as

> "A[n artificial] neural network is a circuit composed of a very large number of simple processing elements that are neurally based." (Nigrin, 1993; p. 11)
>
> "... neurocomputing is the technological discipline concerned with parallel, distributed, adaptive information processing systems that develop infor-

[1] Although neurocomputing can be viewed as a field of research dedicated to the *design* of brain-like computers, this chapter uses the word neurocomputing as a synonym to artificial neural networks.

mation processing capabilities in response to exposure to an information environment. The primary information processing structures of interest in neurocomputing are neural networks..." (Hecht-Nielsen, 1990, p. 2)

Neurocomputing systems are distinct from what is now known as *computational neuroscience*, which is mainly concerned with the development of biologically-based computational models of the nervous system. Artificial neural networks, on the other hand, take a loose inspiration from the nervous system and emphasize the problem solving capability of the systems developed. However, most books on computational neuroscience not only acknowledge the existence of artificial neural networks, but also use several ideas from them in the proposal of more biologically plausible models. They also discuss the ANN suitability as models of real biological nervous systems.

Neurons are believed to be the basic units used for computation in the brain, and their simplified abstract models are the basic processing units of neurocomputing devices or artificial neural networks. Neurons are connected to other neurons by a small junction called synapse, whose capability of being modulated is believed to be the basis for most of our cognitive abilities, such as perception, thinking, and inferring. Therefore, some essential information about neurons, synapses, and their structural anatomy are relevant for the understanding of how ANNs are designed taking inspiration from biological neural networks.

The discussion to be presented here briefly introduces the main aspects of the nervous system used to devise neurocomputing systems, and then focuses on some of the most commonly used artificial neural networks, namely, single- and multi-layer perceptrons, self-organizing networks, and Hopfield networks. The description of the many algorithms uses a matrix notation particularly suitable for the software implementation of the algorithms. Appendix B.1 provides the necessary background on linear algebra. The biological plausibility of each model is also discussed.

4.2 THE NERVOUS SYSTEM

All multicellular organisms have some kind of nervous system, whose complexity and organization varies according to the animal type. Even relatively simple organisms, such as worms, slugs, and insects, have the ability to learn and store information in their nervous systems. The nervous system is responsible for informing the organism through sensory input with regards to the environment in which it lives and moves, processing the input information, relating it to previous experience, and transforming it into appropriate actions or memories.

The nervous system plays the important role of processing the incoming information (signals) and providing appropriate actions according to these signals. The elementary processing units of the nervous system are the *neurons*, also called *nerve cells*. *Neural networks* are formed by the interconnection of many neurons. Each neuron in the human brain has on the order of hundreds or thousands of connections to other neurons.

Anatomically, the nervous system has two main divisions: *central nervous system* (CNS) and *peripheral nervous system* (PNS), the distinction being their different locations. Vertebrate animals have a *bony spine* (*vertebral column*) and a *skull* (*cranium*) in which the central parts of the nervous system are housed. The peripheral part extends throughout the remainder of the body. The part of the (central) nervous system located in the skull is referred to as the *brain*, and the one found in the spine is called the *spinal cord*. The brain and the spinal cord are continuous through an opening in the base of the skull; both are in contact with other parts of the body through the nerves.

The brain can be further subdivided into three main structures: the *brainstem*, the *cerebellum*, and the *forebrain*, as illustrated in Figure 4.1. The *brainstem* is literally the stalk of the brain through which pass all the nerve fibers relaying input and output signals between the spinal cord and higher brain centers. It also contains the cell bodies of neurons whose axons go out to the periphery to innervate the muscles and glands of the head. The structures within the brainstem are the *midbrain*, *pons*, and the *medulla*. These areas contribute to functions such as breathing, heart rate and blood pressure, vision, and hearing. The cerebellum is located behind the brainstem and is chiefly involved with skeletal muscle functions and helps to maintain posture and balance and provides smooth, directed movements. The *forebrain* is the large part of the brain remaining when the brainstem and cerebellum have been excluded. It consists of a central core, the *diencephalon*, and right and left *cerebral hemispheres* (the *cerebrum*).

The outer portion of the cerebral hemispheres is called *cerebral cortex*. The cortex is involved in several important functions such as thinking, voluntary movements, language, reasoning, and perception. The *thalamus* part of the diencephalon is important for integrating all sensory input (except smell) before it is presented to the cortex. The *hypothalamus*, which lies below the thalamus, is a tiny region responsible for the integration of many basic behavioral patterns, which involve correlation of neural and endocrine functions. Indeed, the hypothalamus appears to be the most important area to regulate the internal environment (homeostasis). It is also one of the brain areas associated with emotions.

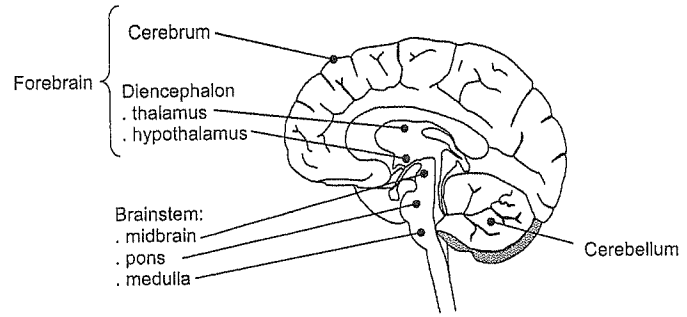


Figure 4.1: Structural divisions of the brain as seen in a midsagittal se

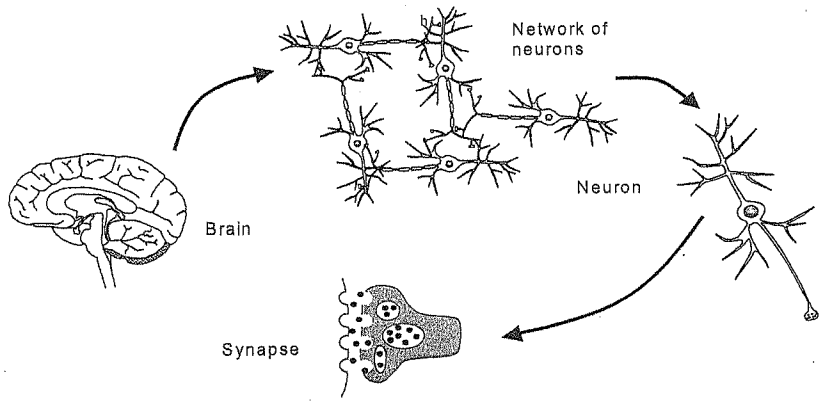


Figure 4.2: Some levels of organization in the nervous system.

4.2.1. Levels of Organization in the Nervous System

What structures really constitute a level of organization in the nervous system is an empirical not an *a priori* matter. We cannot tell, in advance of studying the nervous system, how many levels there are, nor what is the nature of the structural and functional features of any given level (Churchland and Sejnowski, 1992). Therefore, only the structures particularly interesting for the understanding, description, and implementation of artificial neural networks will be described here.

The nervous system can be organized in different levels: molecules, synapses, neurons, networks, layers, maps, and systems (Figure 4.2). An easily recognizable structure in the nervous system is the neuron, which is a cell specialized in signal processing. Depending on environmental conditions, the neurons are capable of generating a signal, more specifically an *electric potential*, that is used to transmit information to other cells to which it is connected. Some processes in the neuron utilize cascades of biochemical reactions that influence information processing in the nervous system. Many neuronal structures can be identified with specific functions. For instance, the *synapses* are important for the understanding of signal processing in the nervous system (Trappenberg, 2002).

Neurons and Synapses

Neurons use a variety of specialized biochemical mechanisms for information processing and transmission. These include *ion channels* that allow a controlled influx and outflux of currents, the generation and propagation of *action potentials*, and the release of *neurotransmitters*. Signal transmission between neurons is the core of the information processing capabilities of the brain. One of the most exciting discoveries in neuroscience was that the effectiveness of the signal

transmission can be modulated in various ways, thus allowing the brain to adapt to different situations. It is believed to be the basis of associations, memories, and many other mental abilities (Trappenberg, 2002). *Synaptic plasticity*, that is the capability of synapses to be modified, is a key ingredient in most models described in this chapter.

Figure 4.3 shows a picture of a schematic generic neuron labeling its most important structural parts. The biological neuron is a single *cell*, thus containing a *cell body* with a *nucleus* (or *soma*) containing DNA, it is filled with fluid and cellular organelles, and is surrounded by a *cell membrane*, just like any other cell in the body. Neurons also have specialized extensions called *neurites*, that can be further distinguished into *dendrites* and *axons*. While the dendrites receive signals from other neurons, the axon propagates the output signal to other neurons.

One peculiarity about neurons is that they are specialized in signal processing utilizing special electrophysical and chemical processes. They can receive and send signals to many other neurons. The neurons that send signals, usually termed *sending* or *presynaptic neurons*, contact the *receiving* or *postsynaptic neurons* in specialized sites named *synapses* either at the cell body or at the dendrites. The synapse is thus the junction between the presynaptic neuron's axon and the postsynaptic neuron's dendrite or cell body.

The general information processing feature of synapses allow them to alter the state of a postsynaptic neuron, thus eventually triggering the generation of an electric pulse, called *action potential*, in the postsynaptic neuron. The action potential is usually initiated at the *axon hillock* and travels all along the axon, which can finally branch and send information to different regions of the nervous system. Therefore, a neuron can be viewed as a device capable of receiving diverse input stimuli from many other neurons and propagating its single output response to many other neurons.

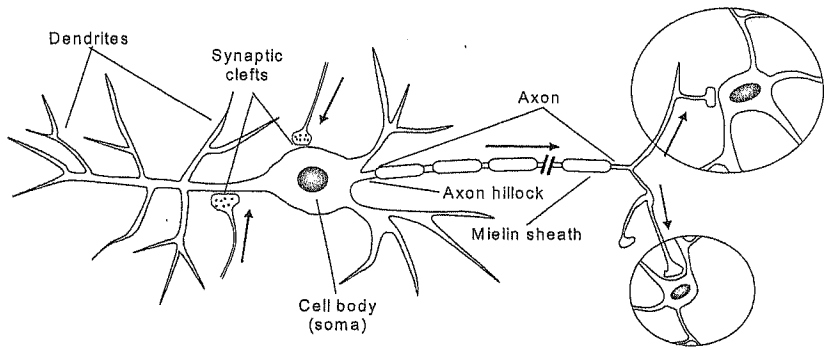


Figure 4.3: Schematic neuron similar in appearance to the pyramidal cells in the brain cortex. The parts outlined are the major structural components of most neurons. The direction of signal propagation between and within neurons is shown by the dark arrows.

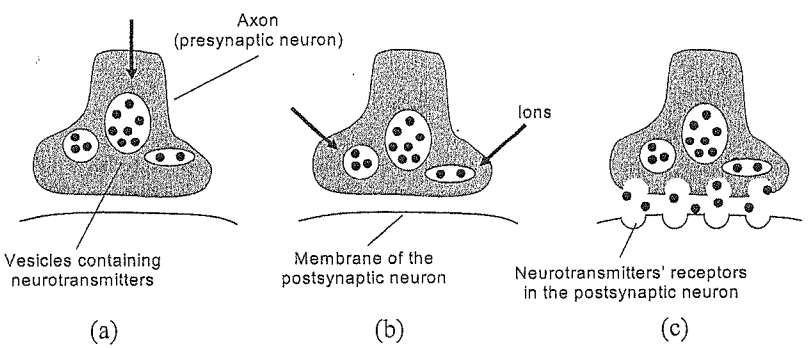


Figure 4.4: Diagram of a chemical synapse. The action potential arrives at the synapse (a) and causes ions to be mobilized in the axon terminal (b), thus causing vesicles to release neurotransmitters into the cleft, which in turn bind with the receptors on the postsynaptic neurons (c).

Various mechanisms exist to transfer information (signals) among neurons. As neurons are cells encapsulated in membranes, little openings in these membranes, called *channels*, allow the transfer of information among neurons. The basic mechanisms of information processing are based on the movement of charged atoms, *ions*, in and out of the channels and within the neuron itself. Neurons live in a liquid environment in the brain containing a certain concentration of ions, and the flow of ions in and out of a neuron through channels is controlled by several factors. A neuron is capable of altering the intrinsic electrical potential, the *membrane potential*, of other neurons, which is given by the difference between the electrical potential within and in the surroundings of the cell.

When an action potential reaches the terminal end of an axon, it mobilizes some ions by opening voltage-sensitive channels that allow the flow of ions into the terminal and possibly the release of some of these stored ions. These ions then promote the release of *neurotransmitters* (chemical substances) into the *synaptic cleft*, which finally diffuse across the cleft and binds with receptors in the postsynaptic neurons. As outcomes, several chemical processes might be initiated in the postsynaptic neuron, or ions may be allowed to flow into it. Figure 4.4 summarizes some of the mechanisms involved in synapse transmission.

As the electrical effects of these ions (action potential) propagate through the dendrites of the receiving neuron and up to the cell body, the process of information can begin again in the postsynaptic neuron. When ions propagate up to the cell body, these signals are *integrated* (summed), and the resulting membrane potential will determine if the neuron is going to *fire*, i.e., to send an output signal to a postsynaptic neuron. This only occurs if the membrane potential of the neuron is greater than the neuron *threshold*. This is because the channels are particularly sensitive to the membrane potential; they will only open when

the potential is sufficiently large (O'Reilly and Munakata, 2000). The action of neural firing is also called *spiking*, *firing a spike*, or the *triggering of an action potential*. The spiking is a very important electrical response of a neuron; once generated it does not change its shape with increasing current. This phenomenon is called the *all-or-none* aspect of the action potential.

Different types of neurotransmitters and their associated ion channels have distinct effects on the state of the postsynaptic neuron. One class of neurotransmitters opens channels that will allow positively charged ions to enter the cell, thus triggering the increase of the membrane potential that drives the postsynaptic neurons towards their excited state. Other neurotransmitters initiate processes that drive the postsynaptic potential towards a resting state; a potential known as the *resting potential*. Therefore, neurotransmitters can promote the initiation of *excitatory* or *inhibitory* processes.

Networks, Layers, and Maps

Neurons can have *forward* and *feedback* connections to other neurons, meaning that they can have either one way or reciprocal connections with other neurons in the nervous system. These interconnected neurons give rise to what is known as *networks of neurons* or *neural networks*. For instance, within a cubic millimeter of cortical tissue, there are approximately 10^5 neurons and about 10^9 synapses, with the vast majority of these synapses arising from cells located within the cortex (Churchland and Sejnowski, 1992). Therefore, the degree of interconnectivity in the nervous system is quite high.

A small number of interconnected neurons (units) can exhibit complex behaviors and information processing capabilities that cannot be observed in single neurons. One important feature of neural networks is the representation of information (knowledge) in a *distributed* way, and the *parallel processing* of this information. No single neuron is responsible for storing "a whole piece" of knowledge; it has to be distributed over several neurons (connections) in the network. Networks with specific architectures and specialized information processing capabilities are incorporated into larger structures capable of performing even more complex information-processing tasks.

Many brain areas display not only networks of neurons but also *laminar organization*. Laminae are *layers of neurons* in register with other layers, and a given lamina conforms to a highly regular pattern of where it projects to and from where it receives projections. For instance, the *superior colliculus* (a particular layered midbrain structure; see Figure 4.1) receives visual inputs in superficial layers, and in deeper layers it receives tactile and auditory input. Neurons in an intermediate layer of the superior colliculus represent information about eye movements (Churchland and Sejnowski, 1992).

One of the most common arrangements of neurons in the vertebrate nervous systems is a layered two-dimensional structure organized with a *topographic* arrangement of unit responses. Perhaps the most well-known example of this is the mammalian cerebral cortex, though others such as the superior colliculus also use maps. The cerebral cortex is the outside surface of the brain. It is a two-

dimensional structure extensively folded with fissures and hills in many larger or more intelligent animals. Two different kinds of cortex exist: an older form with three sub-layers called *paleocortex*, and a newer form that is most prominent in animals with more complex behavior, a structure with six or more sub-layers called *neocortex*.

Studies in human beings by neurosurgeons, neurologists, and neuropathologists have shown that different cortical areas have separate functions. Thus, it is possible to identify visible differences between different regions of the cerebral cortex, each presumably corresponding to a processing module. Figure 4.5 illustrates some specific functional areas in the cerebral cortex of humans. Note that regions associated with different parts of the body can have quite sharp boundaries.

In general, it is known that neocortical neurons are organized into six distinct layers, which can be subdivided into *input*, *hidden*, and *output* layer. The input layer usually receives the sensory input, the output layer sends commands and outputs to other portions of the brain, and the hidden layers receive inputs locally from other cortical layers. This means that hidden layers neither directly receive sensory stimuli nor produce motor and other outputs.

One major principle of organization within many sensory and motor systems is the *topographic map*. For instance, neurons in visual areas of the cortex (in the rear end of the cortex, opposite to the eyes; see Figure 4.5) are arranged topographically, in the sense that adjacent neurons have adjacent visual receptive fields and collectively they constitute a map of the retina. Because neighboring processing units are concerned with similar representations, topographic mapping is an important means whereby the brain manages to save on wiring and also to share wire.

Networking, topographic mapping, and layering are all special cases of a more general principle: the exploitation of geometric and structural properties in information processing design. Evolution has shaped our brains so that these structural organizations became efficient ways for biological systems to assemble in one place information needed to solve complex problems.

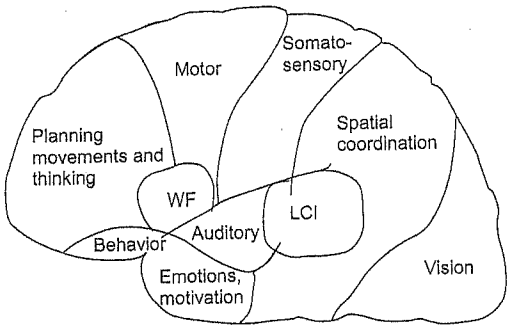


Figure 4.5: Map of specific functional areas in the cerebral cortex. WF: word formation; LCI: Language comprehension and intelligence.

4.2.2. Biological and Physical Basis of Learning and Memory

The nervous system is continuously modifying and updating itself. Virtually all its functions, including perception, motor control, thermoregulation, and reasoning, are modifiable by experience. The topography of the modifications appears not to be final and finished, but an ongoing process with a virtually interminable schedule. Behavioral observations indicate degrees of plasticity in the nervous system: there are fast and easy changes, slower and deeper modifiability, and more permanent but still modifiable changes.

In general, global learning is a function of local changes in neurons. There are many possible ways a neuron could change to embody adaptations. For instance, new dendrites might sprout out, or there might be extension of existing branches, or existing synapses could change, or new synapses might be created. In the other direction, pruning could decrease the dendrites or bits of dendrites, and thus decrease the number of synapses, or the synapses on the remaining branches could be shut down altogether. These are all postsynaptic changes in the dendrites. There could also be changes in the axons; for instance, there might be changes in the membrane, or new branches might be formed, and genes might be induced to produce new neurotransmitters or more of the old ones. Presynaptic changes could include changes in the number of vesicles released per spike and the number of transmitter molecules contained in each vesicle. Finally, the whole cell might die; taking with it all the synapses it formerly supported (Churchland and Sejnowsky, 1992).

This broad range of structural adaptability can be conveniently condensed in the present discussion by referring simply to synapses, since every modification either involves synaptic modification directly or indirectly, or can be reasonably so represented. Learning by means of setting synaptic efficiency is thus the most important mechanism in neural networks, biological and artificial. It depends on both individual neuron-level mechanisms and network-level principles to produce an overall network that behaves appropriately in a given environment.

Two of the primary mechanisms underlying learning in the nervous system are the *long-term potentiation* (LTP) and *long-term depression* (LTD), which refer to the strengthening and weakening of weights in a nontransient form. Potentiation corresponds to an increase in the measured depolarization or excitation delivered by a controlled stimulus onto a receiving neuron, and depression corresponds to a decrease in the measured depolarization. In both cases, the excitation or inhibition of a membrane potential may trigger a complex sequence of events that ultimately result in the modification of synaptic efficiency (strength).

Such as learning, *memory* is an outcome of an adaptive process in synaptic connections. It is caused by changes in the synaptic efficiency of neurons as a result of neural activity. These changes in turn cause new pathways or facilitated pathways to develop for transmission of signals through the neural circuits of the brain. The new or facilitated pathways are called *memory traces*; once established, they can be activated by the thinking mind to reproduce the memories. Actually, one of the outcomes of a learning process can be the creation of a

more permanent synaptic modification scheme, thus resulting in the memorization of an experience.

Memories can be classified in a number of ways. One common classification being into:

- *Short-term memory*: lasts from a few seconds to a few minutes, e.g., one's memory of a phone number.
- *Intermediate long-term memory*: lasts from minutes to weeks, e.g., the name of a nice girl/boy you met in a party.
- *Long-term memory*: lasts for an indefinite period of time, e.g., your home address.

While the two first types of memories do not require many changes in the synapses, long-term memory is believed to require *structural changes* in synapses. These structural changes include an increase in number of vesicle release sites for secretion of neurotransmitter substances, an increase in the number of presynaptic terminals, an increase in the number of transmitter vesicles, and changes in the structures of the dendritic spines.

Therefore, the difference between learning and memory may be sharp and conceptual. Learning can be viewed as the adaptive process that results in the change of synaptic efficiency and structure, while memory is the (long-lasting) result of this adaptive process.

4.3 ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANN) present a number of features and performance characteristics in common with the nervous system:

- The basic information processing occurs in many simple elements called (artificial) *neurons*, *nodes* or *units*.
- These neurons can *receive* and *send stimuli* from and to other neurons and the environment.
- Neurons can be connected to other neurons thus forming networks of neurons or *neural networks*.
- Information (signals) are transmitted between neurons via connection links called *synapses*.
- The efficiency of a synapse, represented by an associated *weight value* or *strength*, corresponds to the information stored in the neuron, thus in the network.
- Knowledge is acquired from the environment by a process known as *learning*, which is basically responsible for *adapting* the connection strengths (weight values) to the environmental stimuli.

One important feature of artificial neural networks is where the knowledge is stored. Basically, what is stored is the connection strengths (synaptic strengths) between unit[illegible]tificial neurons) that allow patterns to be recreated. This feature

has enormous implications, both for processing and learning. The knowledge representation is set up so that the knowledge necessarily influences the course of processing; it becomes a part of the processing itself. If the knowledge is incorporated into the strengths of the connections, then learning becomes a matter of finding the appropriate connection strengths so as to produce satisfactory patterns of activation under some circumstances.

This is an extremely important feature of ANNs, for it opens up the possibility that an information processing mechanism could *learn*, by tuning its connections strengths, to capture the interdependencies between activations presented to it in the course of processing. Another important implication of this type of representation is that *the knowledge is distributed* over the connections among a large number of units. There is no 'special' unit reserved for particular patterns.

An artificial neural network can be characterized by three main features: 1) a set of *artificial neurons*, also termed *nodes*, *units*, or simply *neurons*; 2) the pattern of connectivity among neurons, called the network *architecture* or *structure*; and 3) a method to determine the weight values, called its *training* or *learning* algorithm. Each one of these features will be discussed separately in the following sections.

4.3.1. Artificial Neurons

In the biological neuron, inputs come into the cell primarily through channels located in synapses, allowing ions to flow into and out of the neuron. A membrane potential appears as a result of the integration of the neural inputs, and will then determine whether a given neuron will produce a spike (action potential) or not. This spike causes neurotransmitters to be released at the end of the axon, which then forms synapses with the dendrites of other neurons. The action potential only occurs when the membrane potential is above a critical threshold level. Different inputs can provide different amounts of activation depending on how much neurotransmitter is released by the sender and how many channels in the postsynaptic neuron are opened. Therefore, there are important features of the biological synapses involved in the information processing of neurons.

The net effect of these biological processes is summarized in the computational models discussed here by a *weight* (also called *synaptic strength*, *synaptic efficiency*, *connection strength*, or *weight value*) between two neurons. Furthermore, the modification of one or more of these weight factors will have a major contribution for the neural network learning process. This section reviews three models of neuronal function. The first is the McCulloch-Pitts model in which the neuron is assumed to be computing a logic function. The second is a simple analog *integrate-and-fire* model. And the third is a generic connectionist neuron, which integrates its inputs and generates an output using one particular *activation function*. These neuronal models may be interchangeably called *nodes*, *units*, *artificial neurons*, or simply neurons. It is important to have in mind, though, that the nodes most commonly used in artificial neural networks bear a far resemblance with real neurons.

The McCulloch and Pitts Neuron

W. McCulloch and W. Pitts (1943) wrote a famous and influential paper based on the computations that could be performed by two-state neurons. They did one of the first attempts to understand nervous activity based upon elementary neural computing units, which were highly abstract models of the physiological properties of neurons and their connections. Five physical assumptions were made for their calculus (McCulloch and Pitts, 1943):

1. The behavior of the neuron is a binary process.
2. At any time a number of synapses must be excited in order to activate the neuron.
3. Synaptic delay is the only significant delay that affects the nervous system.
4. The excitation of a certain neuron at a given time can be inhibited by an inhibitory synapse.
5. The neural network has a static structure; that is, a structure that does not change with time.

McCulloch and Pitts conceived the neuronal response as being equivalent to a *proposition* adequate to the neuron's stimulation. Therefore, they studied the behavior of complicated neural networks using a notation of the symbolic *logic of propositions* (see Appendix B.4.2). The 'all-or-none' law of nervous activity was sufficient to ensure that the activity of any neuron could be represented as a proposition.

It is important to note that according to our current knowledge of how the neuron works - based upon electrical and chemical processes - neurons are not realizing any proposition of logic. However, the model of McCulloch and Pitts can be considered as a special case of the most general neuronal model to be discussed in the next sections, and is still sometimes used to study particular classes of nonlinear networks. Additionally, this model has caused a major impact mainly among computer scientists and engineers, encouraging the development of artificial neural networks. It has even been influential in the history of computing (cf. von Neumann, 1982).

The McCulloch and Pitts neuron is binary, i.e., it can assume only one of two states (either '0' or '1'). Each neuron has a fixed *threshold* θ and receives inputs from synapses of identical weight values. The neuronal mode of operation is simple. At each time step t, the neuron responds to its synaptic inputs, which reflect the state of the presynaptic neurons.

If no inhibitory synapse is active, the neuron integrates (sums up) its synaptic inputs, generating the *net input* to the neuron, u, and checks if this sum (u) is greater than or equal to the threshold θ. If it is, then the neuron becomes active, that is, responds with a '1' in its output ($y = 1$); otherwise it remains inactive, that is, responds with a '0' in its output ($y = 0$).

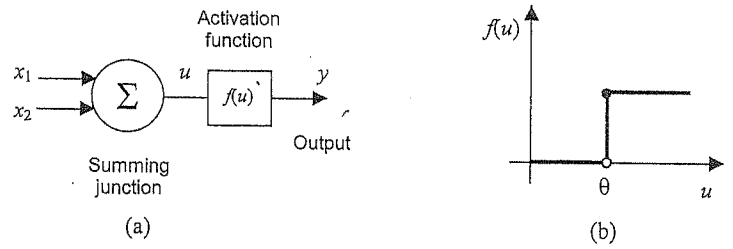


Figure 4.6: The McCulloch and Pitts neuron (a) and its threshold activation function (b).

a	b	a AND b
0	0	0
0	1	0
1	0	0
1	1	1

a	b	a OR b
0	0	0
0	1	1
1	0	1
1	1	1

a	NOT a
0	1
1	0

Figure 4.7: Truth tables for the connectives AND, OR, and NOT.

Although this neuron is quite simple, it already presents some important features in common with most neuron models, namely, the integration of the input stimuli to determine the *net input* u and the presence of an activation function (threshold). Figure 4.6 illustrates the simple McCulloch and Pitts neuron and its activation function.

To illustrate the behavior of this simple unit, assume two excitatory inputs x_1 and x_2 and a threshold $\theta = 1$. In this case, the neuron is going to fire; that is, to produce an output '1', every time x_1 or x_2 has a value '1', thus operating like the logical connective OR (see Figure 4.7). Assume now that the neuron threshold is increased to $\theta = 2$. In this new situation, the neuron is only going to be active (fire) if both x_1 and x_2 have value '1' simultaneously, thus operating like the logical connective AND (see Figure 4.7).

A Basic Integrate-and-Fire Neuron

Assume a noise-free neuron with the net input being a variable of time $u(t)$ corresponding to the membrane potential of the neuron. The main effects of some neuronal channels (in particular the sodium and leakage channels) can be captured by a simple equation of an integrator (Dayan and Abbot, 2001; Trappenberg, 2002):

$$\tau_m \frac{du(t)}{dt} = u_{res} - u(t) + R_m i(t) \qquad (4.1)$$

where τ_m is the membrane time constant of the neuron determined by the average conductance of the channels (among other things); u_{res} is the resting potenti-

al of the neuron; i is the input current given by the sum of the synaptic currents generated by firings of presynaptic neurons; R_m is the resistance of the neuron to the flow of current (ions); and t is the time index.

Equation (4.1) can be very simply understood. The rate of variation of the membrane potential of the neuron is proportional to its current membrane potential, its resting potential, and the potential generated by the incoming signals to the neuron. Note that the last term on the right side of this equation is the Ohm's law ($u = R.i$) for the voltage generated by the incoming currents.

The input current $i(t)$ to the neuron is given by the sum of the incoming synaptic currents depending on the *efficiency* of individual synapses, described by the variable w_j for each synapse j. Therefore, the total input current to the neuron can be written as the sum of the individual synaptic currents multiplied by a weight value

$$i(t) = \sum_j \sum_{t_j^f} w_j f(t - t_j^f) \tag{4.2}$$

where the function $f(\cdot)$ parameterizes the form of the postsynaptic response. This function was termed *activation function* in the McCulloch and Pitts neuron discussed above and this nomenclature will be kept throughout this text. The variable t_j^f denotes the firing time of the presynaptic neuron of synapse j. The firing time of the postsynaptic neuron is defined by the time the membrane potential u reaches a threshold value θ,

$$u(t^f) = \theta \tag{4.3}$$

In contrast to the firing time of the presynaptic neurons, the firing time of the integrate-and-fire neuron has no index. To complete the model, the membrane potential has to be reset to the resting state after the neuron has fired. One form of doing this is by simply resetting the membrane potential to a fixed value u_{res} immediately after a spike.

The Generic Neurocomputing Neuron

The computing element employed in most neural networks is an integrator, such as the McCulloch and Pitts and the integrate-and-fire models of a neuron, and computes based on its connection strengths. Like in the brain, the artificial neuron is an information-processing element that is fundamental to the operation of the neural network. Figure 4.8 illustrates the typical artificial neuron, depicting its most important parts: the synapses, characterized by their *weight values* connecting each input to the neuron; the *summing junction* (*integrator*); and the *activation function*.

Specifically, an input signal x_j at the input of synapse j connected to neuron k is multiplied by the synaptic weight w_{kj}. In this representation, the first subscript of the synaptic weight refers to the neuron, and the second subscript refers to the synapse connected to it. The summing junction adds all input signals weighted by the synaptic weight values plus the neuron's bias b_k; this operation constitutes the dot (or inner) product (see Appendix B.1); that is, a linear combination of the inputs with the weight values, plus the bias b_k. Finally, an activation function is

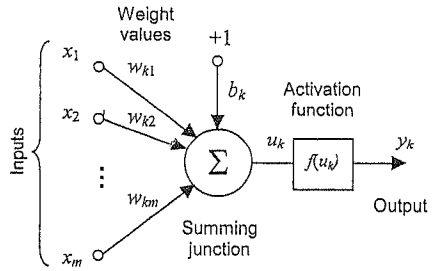


Figure 4.8: Nonlinear model of a neuron.

used to limit the amplitude of the output of the neuron. The activation function is also referred to as a *squashing function* (Rumelhart et al., 1986) for it limits the permissible amplitude range of the output signal to some finite value.

The bias has the effect of increasing or decreasing the net input to the activation function depending on whether it is positive or negative, respectively. In the generic neuron it is usually used in place of a fixed threshold θ for the activation function. For instance, the McCulloch and Pitts neuron (Figure 4.6) will fire when its net input is greater than θ:

$$y = f(u) = \begin{cases} 1 & \text{if } u \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

where $u = x_1 + x_2$ for the neuron of Figure 4.6. It is possible to replace the threshold θ by a bias weight b that will be multiplied by a constant input of value '1':

$$y = f(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $u = x_1 + x_2 - b$. Note however, that while the bias is going to be adjusted during learning, the threshold assumes a fixed value.

It is important to note here that the output of this generic neuron is simply a number, and the presence of discrete action potentials is ignored. As real neurons are limited in dynamic range from zero-output firing rate to a maximum of a few hundred action potentials per second, the use of an activation function could be biologically justified.

Mathematically, the neuron k can be described by a simple equation:

$$y_k = f(u_k) = f\left(\sum_{j=1}^{m} w_{kj} x_j + b_k\right) \tag{4.4}$$

where x_j, $j = 1, \ldots, m$, are the input signals; w_{kj}, $j = 1, \ldots, m$, are the synaptic weights of neuron k; u_k is the net input to the activation function; b_k is the bias of neuron k; $f(\cdot)$ is the activation function; and y_k is the output signal of the neuron.

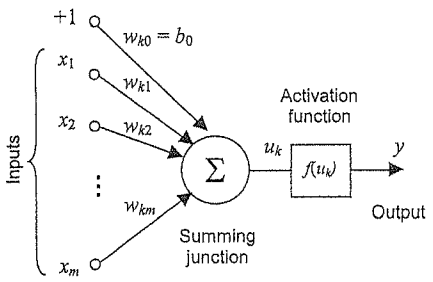


Figure 4.9: Reformulated model of a neuron.

It is possible to simplify the notation of Equation (4.4) so as to account for the presence of the bias by simply defining a constant input signal $x_0 = 1$ connected to the neuron k with associated weight value $w_{k0} = b_k$. The resulting equation becomes,

$$y_k = f(u_k) = f\left(\sum_{j=0}^{m} w_{kj} x_j\right) \tag{4.5}$$

Figure 4.9 illustrates the reformulated model of a neuron.

Types of Activation Function

The activation function, denoted by $f(u_k)$, determines the output of a neuron k in relation to its net input u_k. It can thus assume a number of forms, of which some of the most frequently used are summarized below (see Figure 4.10):

- *Linear function*: relates the net input directly with its output.

$$f(u_k) = u_k \tag{4.6}$$

- *Step or threshold function*: the output is '1' if the net input is greater than or equal to a given threshold θ; otherwise it is either '0' or '−1', depending on the use of a binary {0,1} or a bipolar {−1,1} function. The binary step function is the one originally used in the McCulloch and Pitts neuron:

$$f(u_k) = \begin{cases} 1 & \text{if } u_k \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

- *Signum function*: it is similar to the step function, but has a value of '0' when the net input to the neuron is $u = 0$. It can also be either bipolar or binary; the bipolar case being represented in Equation (4.8).

$$f(u_k) = \begin{cases} 1 & \text{if } u_k > \theta \\ -1 & \text{if } u_k < \theta \end{cases} \tag{4.8}$$

- *Sigmoid function*: it is a strictly increasing function that presents saturation and a graceful balance between linear and nonlinear behavior. This is the most commonly used activation function in the ANN literature. It has an

s-shaped form and can be obtained by several functions, such as the *logistic function*, the *arctangent function*, and the *hyperbolic tangent* function; the difference being that, as u_k ranges from $-\infty$ to $+\infty$, the logistic function ranges from 0 to 1, the arctangent ranges from $-\pi/2$ to $+\pi/2$, and the hyperbolic tangent ranges from −1 to +1:

$$Logistic: f(u_k) = \frac{1}{1+\exp(-u_k)} \tag{4.9}$$

- *Radial basis function*: it is a nonmonotonic function that is symmetric around a base value. This is the main type of function for *radial basis function neural networks*. The expression shown below is that of the Gaussian bell curve:

$$f(u_k) = \exp(-u_k^2) \tag{4.10}$$

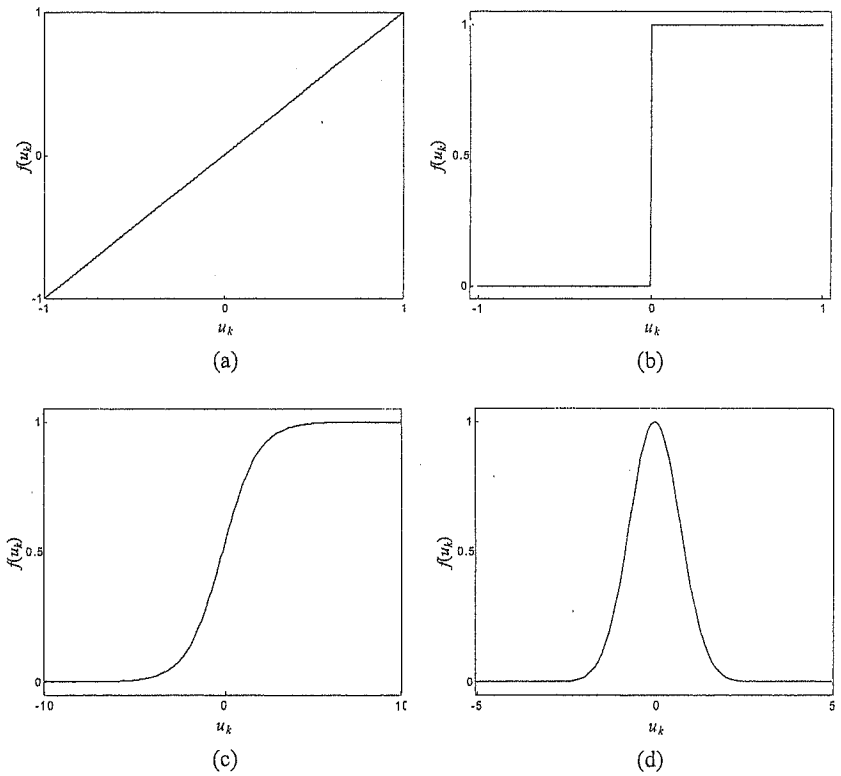


Figure 4.10: Most commonly used types of activation function. (a) Linear activation, $u_k \in [-1,+1]$. (b) Step function with $\theta = 0$, $u_k \in [-1,+1]$. (c) Logistic function, $u_k \in [-10,+10]$. (d) Gaussian bell curve, $u_k \in [-5,+5]$.

It is important to stress that the functions described and illustrated in Figure 4.10 depict only the general shape of those functions. However, it should be clear that these basic shapes can be modified. All these functions can include parameters with which some features of the functions can be changed, such as the *slope* and *offset* of a function. For instance, by multiplying u_k in the logistic function by a value β, it is possible to control the smoothness of the function:

$$f(u_k) = \frac{1}{1+\exp(-\beta.u_k)}$$

In this case, for larger values of β, this function becomes more similar to the threshold function, while for smaller values of β, the logistic function becomes more similar to a linear function.

4.3.2. Network Architectures

One of the basic prerequisites for the emergence of complex behaviors is the interaction of a number of individual agents. In the nervous system, it is known that individual neurons can affect the behavior (firing rate) of others, but, as a single entity, a neuron is meaningless. This might be one reason why evolution drove our brains to a system with such an amazingly complex network of interconnected neurons.

In the human brain not much is known of how neurons are interconnected. Some particular knowledge is available for specific brain portions, but little can be said about the brain as a whole. For instance, it is known that the cortex can be divided into a number of different cortical areas specialized in different kinds of processing; some areas perform pattern recognition, others process spatial information, language processing, and so forth. In addition, it is possible to define anatomic neuronal organizations in the cortex in terms of *layers of neurons*, which are very important for the understanding of the detailed biology of the cortex (Figure 4.5).

In the domain of artificial neural networks, a layer of neurons will refer to functional layers of nodes. As not much is known about the biological layers of neurons, most neurocomputing networks employ some standardized architectures, specially designed for the engineering purposes of solving problems. There are basically three possible layers in a network: an *input layer*, one or more *intermediate* (or *hidden*) *layers*, and an *output layer*. These are so-called because the input layer receives the input stimuli directly from the environment, the output layer places the network output(s) to the environment, and the hidden layer(s) is not in direct contact with the environment.

The way in which neurons are structured (interconnected) in an artificial neural network is intimately related with the learning algorithm that is used for training the network. In addition, this interconnectivity affects the network storage and learning capabilities. In general, it is possible to distinguish three main types of network architectures (Haykin, 1999): *single-layer feedforward networks*, *multi-layer feedforward networks*, and *recurrent networks*.

Single-Layer Feedforward Networks

The simplest case of layered networks consists of an input layer of nodes whose output feeds the output layer. Usually, the input nodes are linear, i.e., they simply propagate the input signals to the output layer of the network. In this case, the input nodes are also called sensory units because they only sense (receive information from) the environment and propagate the received information to the next layer. In contrast, the output units are usually processing elements, such as the neuron depicted in Figure 4.9, with a nonlinear type of activation function. The signal propagation in this network is purely positive or *feedforward*; that is, signals are propagated from the network input to its outputs and never the opposite way (*backward*). This architecture is illustrated in Figure 4.11(a) and the direction of signal propagation in Figure 4.11(b). In order not to overload the picture, very few connection strengths were depicted in Figure 4.11(a), but these might be sufficient to give a general idea of how the weights are assigned in the network.

The neurons presented in Figure 4.11 are of the generic neuron type illustrated in Figure 4.9. (Note that input x_0 is assumed to be fixed in '1' and the weight vector that connects it to all output units w_{i0}, $i = 1, \ldots, o$, corresponds to the bias of each output node $b_i = w_{i0}$, $i = 1, \ldots, o$.) The weight values between the inputs and the output nodes can be written using matrix notation (Appendix B.1) as

$$\mathbf{W} = \begin{bmatrix} w_{10} & w_{11} & \cdots & w_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{o0} & w_{o1} & \cdots & w_{om} \end{bmatrix} \tag{4.11}$$

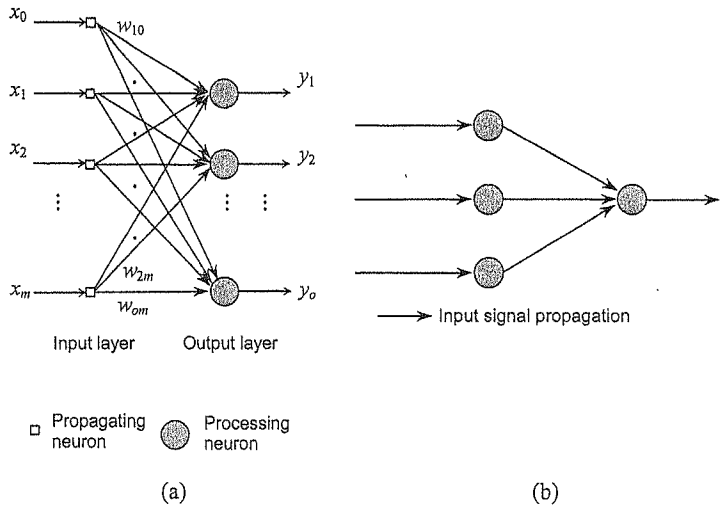


Figure 4.11: *Single-layer feedforward* neural network. (a) Network architecture. (b) Direction of propagation of the input signal.

where the first index i, $i = 1, \ldots, o$, of each element corresponds to the postsynaptic node and the second index j, $j = 0, \ldots, m$, corresponds to the presynaptic node (just remember 'to-from' while reading from left to right). Therefore, each row of matrix $\mathbf{W}$ corresponds to the weight vector of an output unit, and each element of a given column j, $j = 0, \ldots, m$, corresponds to the strength with which an input j is connected to each output unit i. Remember that $j = 0$ corresponds to the unitary input that will multiply the neuron's bias.

The output of each postsynaptic neuron i in Figure 4.11(a), $i = 1, \ldots, o$, is computed by applying an activation function $f(\cdot)$ to its net input, which is given by the linear combination of the network inputs and the weight vector,

$$y_i = f(\mathbf{w}_i.\mathbf{x}) = f(\Sigma_j w_{ij}.x_j), j = 0, \ldots, m \quad (4.12)$$

The output vector of the whole network $\mathbf{y}$ is given by the activation function applied to the product of the weight matrix $\mathbf{W}$ by the input vector $\mathbf{x}$,

$$\mathbf{y} = f(\mathbf{W}.\mathbf{x}) \quad (4.13)$$

Assuming the general case where the weight and input vectors can take any real value, the dimension of the weight matrix and each vector is $\mathbf{W} \in \Re^{o \times (m+1)}$, $\mathbf{w}_i \in \Re^{1 \times (m+1)}$, $i = 1, \ldots, o$, $\mathbf{x} \in \Re^{(m+1) \times 1}$, and $\mathbf{y} \in \Re^{o \times 1}$.

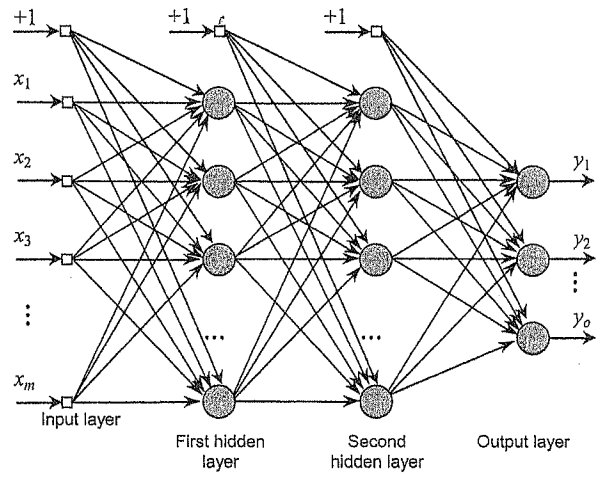
Multi-Layer Feedforward Networks

The second class of feedforward neural networks is known as multi-layer networks. These are distinguished from the single-layered networks by the presence of one or more *intermediate* or *hidden* layers. By adding one or more hidden layers, the nonlinear computational processing and storage capability of the network is increased, for reasons that will become clearer further in the text. The output of each network layer is used as input to the following layer. A multi-layer feedforward network can be thought of as an assembly line: some basic material is introduced into the production line and passed on, the second stage components are assembled, and then the third stage, up to the last or output stage, which delivers the final product.

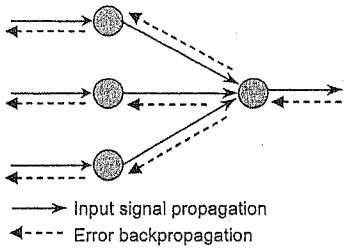
The learning algorithm typically used to train this type of network requires the *backpropagation* of an error signal calculated between the network output and a desired output. This architecture is illustrated in Figure 4.12(a) and the direction of signal propagation is depicted in Figure 4.12(b).

In such networks, there is one weight matrix for each layer, and these are going to be denoted by the letter $\mathbf{W}^k$ with a superscript k indicating the layer. Layers are counted from left to right, and the subscripts of this matrix notation remain the same as previously. Therefore, w_{ij}^k corresponds to the weight value connecting the postsynaptic neuron i to the presynaptic neuron j at layer k.

In the network of Figure 4.12(a), $\mathbf{W}^1$ indicates the matrix of connections between the input layer and the first hidden layer; matrix $\mathbf{W}^2$ contains the connections between the first hidden layer and the second hidden layer; and matrix $\mathbf{W}^3$ contains the connections between the second hidden layer and the output layer.



(a)



(b)

Figure 4.12: *Multi-layer feedforward* neural network. (a) Network architecture (legend as depicted in Figure 4.11). (b) Direction of propagation of the functional and error signals. (The weight values were suppressed from the picture not to overload it.)

In feedforward networks the signals are propagated from the inputs to the network output in a layer-by-layer form (from left to right). Therefore, the network output is given by (in matrix notation):

$$\mathbf{y} = \mathbf{f}^3(\mathbf{W}^3\, \mathbf{f}^2(\mathbf{W}^2\, \mathbf{f}^1(\mathbf{W}^1\mathbf{x}))) \quad (4.14)$$

where $\mathbf{f}^k$ is the vector of activation functions of layer k (note that nodes in a given layer may have different activation functions); $\mathbf{W}^k$ is the weight matrix of layer k and $\mathbf{x}$ is the input vector.

It can be observed from Equation (4.14) that the network output is computed by recursively multiplying the weight matrix of a given layer by the output produced by each previous layer. The expression to calculate the output of each layer is given by Equation (4.12) using the appropriate weight matrix and inputs.

It is important to note that if the intermediate nodes have linear activation functions there is no point in adding more layers to this network, because $f(x) = x$ for linear functions, and thus the network output would be given by

$$\mathbf{y} = \mathbf{f}^3(\mathbf{W}^3\,\mathbf{W}^2\,\mathbf{W}^1\mathbf{x}).$$

and this expression can be reduced to,

$$\mathbf{y} = \mathbf{f}^3(\mathbf{W}\mathbf{x}),$$

where $\mathbf{W} = \mathbf{W}^3\,\mathbf{W}^2\,\mathbf{W}^1$.

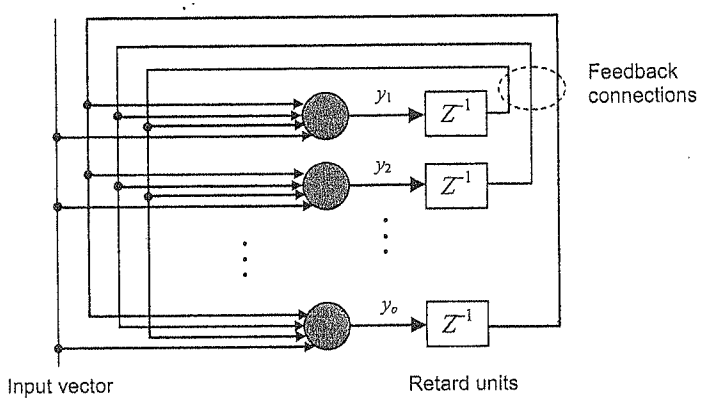
Therefore, if one wants to increase the computational capabilities of a multilayer neural network by adding more layers to the network, nonlinear activation functions have to be used in the hidden layers.

Recurrent Networks

The third class of networks is known as recurrent networks, distinguished from feedforward networks for they have at least one *recurrent* (or *feedback*) loop. For instance, a recurrent network may consist of a single layer of neurons with each neuron feeding its output signal back to the input of other neurons, as illustrated in Figure 4.13. In a feedback arrangement, there is communication between neurons; there can be cross talk and plan revision; intermediate decisions can be taken; and a mutually agreeable solution can be found.

Note that the type of feedback loop illustrated in Figure 4.13 is distinguished from the backpropagation of error signals briefly mentioned in the previous section. The recurrent loop has an impact on the network learning capability and performance because it involves the use of particular branches composed of retard units (Z^{-1}) resulting in a nonlinear dynamic behavior (assuming that the network has nonlinear units).

The network illustrated in Figure 4.13 is fully recurrent, in the sense that all network units have feedback connections linking them to all other units in the network. This network also assumes an input vector $\mathbf{x}$ weighted by the weight vector $\mathbf{w}$ (not shown).



ıre 4.13: Recurrent neural network with no hidden layer.

The network output y_i, $i = 1, \ldots, o$, is a composed function of the weighted inputs at iteration t, plus the weighted outputs in the previous iteration $(t-1)$:

$$y_i(t) = f(\mathbf{w}_i.\mathbf{x}(t) + \mathbf{v}_i.\mathbf{y}(t-1)) = f(\Sigma_j\, w_{ij}.x_j(t) + \Sigma_k\, v_{ik}.y_k(t-1)), \quad j = 1, \ldots, m;\ k = 1, \ldots, o \tag{4.15}$$

where $\mathbf{v}_i$, $i = 1, \ldots, o$ is the vector weighting the retarded outputs fed back into the network ($\mathbf{W} \in \Re^{o \times m}$, $\mathbf{w}_i \in \Re^{1 \times m}$, $i = 1, \ldots, o$, $\mathbf{x} \in \Re^{m \times 1}$, $\mathbf{y} \in \Re^{o \times 1}$, $\mathbf{V} \in \Re^{o \times o}$, and $\mathbf{v}_i \in \Re^{1 \times o}$).

4.3.3. Learning Approaches

It has been discussed that one of the most exciting discoveries in neuroscience was that synaptic efficiency could be modulated to the input stimuli. Also, this is believed to be the basis for learning and memory in the brain. Learning thus involves the adaptation of synaptic strengths to environmental stimuli. Biological neural networks are known not to have their architectures much altered throughout life. New neurons cannot routinely be created to store new knowledge, and the number of neurons is roughly fixed at birth, at least in mammals. The alteration of synaptic strengths may thus be the most relevant factor for learning. This change could be the simple modification in strength of a given synapse, the formation of new synaptic connections or the elimination of preexisting synapses. However, the way learning is accomplished in the biology of the brain is still not much clear.

In the neurocomputing context, *learning* (or *training*) corresponds to the process by which the network's "free parameters" are adapted (adjusted) through a mechanism of presentation of environmental (or input) stimuli. In the standard neural network learning algorithms to be discussed here, these free parameters correspond basically to the connection strengths (weights) of individual neurons. It is important to have in mind though, that more sophisticated learning algorithms are capable of dynamically adjusting several other parameters of an artificial neural network, such as the network architecture and the activation function of individual neurons. The environmental or input stimuli correspond to a set of *input data* (or *patterns*) that is used to *train* the network.

Neural network learning basically implies the following sequence of events:

- Presentation of the input patterns to the network.
- Adaptation of the network free parameters so as to produce an altered pattern of response for the input data.

In most neural network applications the network weights are first adjusted according to a given learning rule or algorithm, and then the network is applied to a new set of input data. In this case there are two steps involved in the use of a neural network: 1) network training, and 2) network application.

With standard learning algorithms a neural network learns through an iterative process of weight adjustment. The type of learning is defined by the way in which the weights are adjusted. The three main learning approaches are: 1) *supervised learning*, 2) *unsupervised learning*, and 3) *reinforcement learning*.

Supervised Learning

This learning strategy embodies the concept of a *supervisor* or *teacher*, who has the knowledge about the environment in which the network is operating. This knowledge is represented in the form of a set of *input-output samples* or *patterns*. Supervised learning is typically used when the class of data is known *a priori* and this can be used as the supervisory mechanism, as illustrated in Figure 4.14. The network free parameters are adjusted through the combination of the input and error signals, where the error signal is the difference between the desired output and the current network output.

To provide the intuition behind supervised learning, consider the following example. Assume you own an industry that produces patterned tiles. In the quality control (QC) part of the industry, an inspection has to be made in order to guarantee the quality in the patterns printed on the tiles. There are a number of tiles whose patterns are assumed to be of good quality, and thus pass the QC. These tiles can be used as input samples to train the artificial neural network to classify good quality tiles. The known or desired outputs are the respective classification of the tiles as good or bad quality tiles.

In the artificial neural network implementation, let neuron j be the only output unit of a feedforward network. Neuron j is stimulated by a signal vector $\mathbf{x}(t)$ produced by one or more hidden layers, that are also stimulated by an input vector. Index t is the discrete time index or, more precisely, the time interval of an iterative process that will be responsible for adjusting the weights of neuron j. The only output signal $y_j(t)$, from neuron j, is compared with a *desired output*, $d_j(t)$. An error signal $e_j(t)$ is produced:

$$e_j(t) = d_j(t) - y_j(t) \tag{4.16}$$

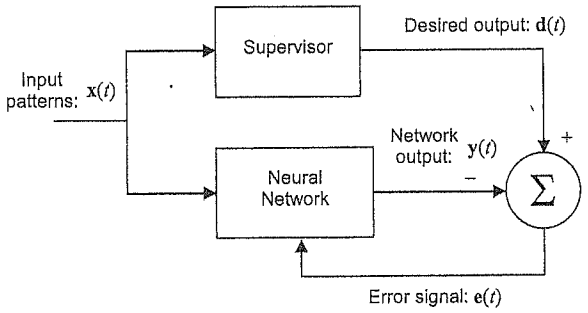


Figure 4.14: In supervised learning, the environment provides input patterns to train the network and a supervisor has the knowledge about the environment in which the network is operating. At each time step t, the network output is compared with the desired output (the error is used to adjust the network response).

The error signal acts as a control mechanism responsible for the application of corrective adjustments in the weights of neuron j. The goal of supervised learning is to make the network output more similar to the desired output at each time step; that is, to *correct the error* between the network output and the desired output. This objective can be achieved by minimizing a *cost function* (also called *performance index*, *error function*, or *objective function*), $\Im(t)$, which represents the instant value of the error measure:

$$\Im(t) = \tfrac{1}{2} e_j^2(t) \tag{4.17}$$

Generalization Capability

Once the supervised learning process is complete, it is possible to present an *unseen* pattern to the network, which will then classify that pattern, with a certain degree of accuracy, into one of the classes used in the training process. This is an important aspect of an artificial neural network, its *generalization capability*. In the present context, *generalization* refers to the performance of the network on (new) patterns that were not used in the network learning process.

This can be illustrated with the help of the following example. Assume we want to use a multi-layer feedforward neural network to approximate the function $sin(x).cos(2x)$ depicted by a solid line in Figure 4.15(a). It is only necessary a few training samples to adjust the weight vectors of a multi-layer feedforward neural network with a single hidden layer composed of five sigmoidal units. The network contains a single input and a single output unit, but these details are not relevant for the present discussion.

It is known that most real-world data contains noise, i.e., irrelevant or meaningless data, or data with disturbances. To investigate the relationship between noisy data and the network capability of approximating these data, some noise was added to the input data by simply adding a uniform distribution of zero mean and variance 0.15 to each input datum. Figure 4.15(a) shows the input data and the curve representing the original noise-free function $sin(x).cos(2x)$. Note that the noisy data generated deviates a little from the function to be approximated.

In this case, if the neural network is not trained sufficiently or if the network architecture is not appropriate, the approximation it will provide for the input data will not be satisfactory; an *underfitting* will occur (Figure 4.15(b)). In the opposite case, when the network is trained until it perfectly approximates the input data, some *overfitting* will occur, meaning that the approximation is too accurate for this noisy data set (Figure 4.15(c)). A better training is the one that establishes a compromise between the approximation accuracy and a good generalization for unseen data (Figure 4.15(d)).

From a biological perspective, generalization is very important for our creation of models of the world. Think of generalization according to the following example. If you just memorize some specific facts about the world instead of trying to extract some simple essential regularity underlying these facts, then you would be in trouble when dealing with novel situations where none of the specifics appear. For instance, if you memorize a situation where you almost

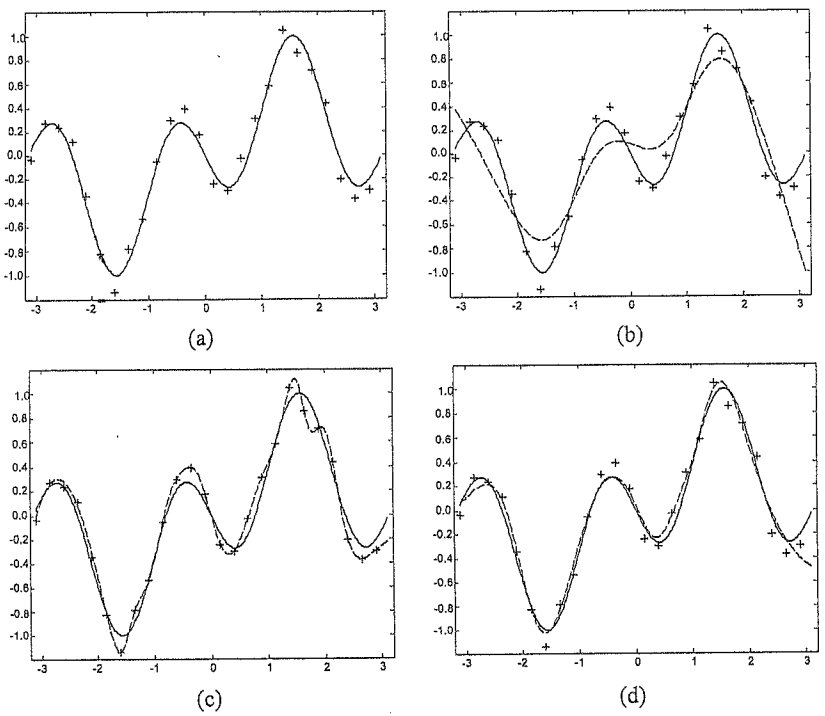


Figure 4.15: Function $sin(x) \times cos(2x)$ with uniformly distributed noise over the interval [-0.15, 0.15]. Legend: + input patterns; — desired output; - - - network output. (a) Training patterns with noise and the original function to be approximated (solid line). (b) Training process interrupted too early, *underfitting* (dashed line). (c) Too much training, *overfitting* (dashed line). (d) Better trade-off between quality of approximation and generalization capability.

drowned in seawater (but you survived) as 'stay away from seawater', then you may be in trouble if you decide to swim in a pool or a river. Of course, if the problem is that you don't know how to swim, then, knowing that seawater in particular is dangerous might not be a sufficiently good 'model' of the world to prevent you from drowning in other types of water. Your internal model has to be general enough so as to suggest 'stay away from water'.

Unsupervised Learning

In the *unsupervised* or *self-organized* learning approach, there is no supervisor to evaluate the network performance in relation to the input data set. The intuitive idea embodied in unsupervised learning is quite simple: given a set of input data, what can you do with it? For instance, if you are given a set of balloons, you could group them by colors, or by shape, or by any other attribute you can qualify.

Note that in unsupervised learning there is no error information being fed back into the network; the classes of the data are *unknown* or *unlabelled* and the presence of the supervisor no longer exists. The network adapts itself to statistical regularities in the input data, developing an ability to create internal representations that encode the features of the input data and thus, generate new classes automatically. Usually, self-organizing algorithms employ a *competitive learning* scheme.

In *competitive learning*, the network output neurons compete with one another to become activated, with a single output neuron being activated at each iteration. This property makes the algorithm appropriate to discover salient statistical features within the data that can then be used to classify a set of input patterns.

Individual neurons learn to specialize on groups (*clusters*) of similar patterns; in effect they become *feature extractors* or *feature detectors* for different classes of input patterns. In its simplest form, a *competitive neural network*, i.e., a neural network trained using a competitive learning scheme, has a single layer of output neurons that is fully connected. There are also lateral connections among neurons, as illustrated in Figure 4.16, capable of inhibiting or stimulating neighbor neurons.

For a neuron i to be the winner, the distance between its corresponding weight vector $\mathbf{w}_i$ and a certain input pattern $\mathbf{x}$ must be the smallest measure (among all the network output units), given a certain metric $\|\cdot\|$, usually taken to be the Euclidean distance. Therefore, the idea is to find the output neuron whose weight vector is most similar (has the shortest distance) to the input pattern presented.

$$i = \arg\min_i \|\mathbf{x} - \mathbf{w}_i\|, \quad \forall i \tag{4.18}$$

If a neuron does not respond to a determined input pattern (i.e., is not the winner), no learning takes place for this neuron. However, if a neuron i wins the competition, then an adjustment $\Delta\mathbf{w}_i$ is applied to the weight vector $\mathbf{w}_i$ associated with the winning neuron i,

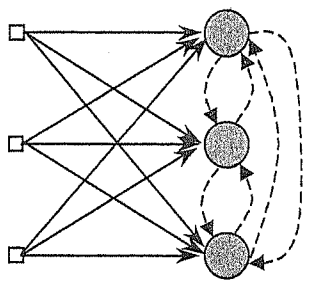


Figure 4.16: Simple competitive network architecture with direct excitatory (feedforward) connections (solid arrows) from the network inputs to the network outputs and inhibitory lateral connections among the output neurons (dashed arrows).

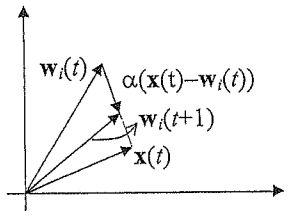


Figure 4.17: Geometric interpretation of the adjustment performed in unsupervised learning. The weight vector of the winning neuron i is moved toward the input pattern $\mathbf{x}$.

$$\Delta \mathbf{w}_i = \begin{cases} \alpha(\mathbf{x} - \mathbf{w}_i) & \text{if } i \text{ wins the competition} \\ 0 & \text{if } i \text{ loses the competition} \end{cases} \tag{4.19}$$

where α is a *learning rate* that controls the step size given by $\mathbf{w}_i$ in the direction of the input vector $\mathbf{x}$. Note that this learning rule works by simply moving the weight vector of the winning neuron in the direction of the input pattern presented, as illustrated in Figure 4.17.

Reinforcement Learning

Reinforcement learning (RL) is distinguished from the other approaches as it relies on learning from direct interaction with the environment, but does not rely on explicit supervision or complete models of the environment. Often, the only information available is a scalar evaluation that indicates how well the artificial neural network is performing. This is based on a framework that defines the interaction between the neural network and its environment in terms of the current values of the network's free parameters (weights), network response (actions), and rewards. Situations are mapped into actions so as to maximize a numerical reward signal (Sutton and Barto, 1998). Figure 4.18 illustrates the network-environment interaction in a reinforcement learning system, highlighting that the network output is fed into the environment that provides it with a reward signal according to how well the network is performing.

In reinforcement learning, the artificial neural network is given a goal to achieve. During learning, the neural network *tries* some actions (i.e., output values) on its environment, then it is *reinforced* or *penalized* by receiving a scalar evaluation (the *reward* or *penalty value*) for its actions. The reinforcement learning algorithm selectively retains the outputs that maximize the received reward over time. The network learns how to achieve its goal by trial-and-error interactions with its environment. At each time step t, the learning system receives some representation of the *state* of the environment $\mathbf{x}(t)$, it provides an output $y(t)$, and one step later it receives a scalar reward $r(t+1)$, and finds itself in a new state $\mathbf{x}(t+1)$. Thus, the two basic concepts behind reinforcement learning are *trial and error search* and *delayed reward.*

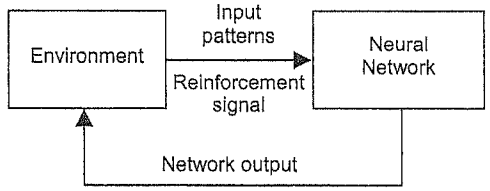


Figure 4.18: In reinforcement learning the network output provides the environment with information about how well the neural network is performing.

The intuitive idea behind reinforcement learning can also be illustrated with a simple example. Assume that you are training in a simulator to become a pilot without any supervisor, and your goal in today's lesson is to land the plane smoothly. Given a certain state of the aircraft, you perform an action that results in the plane crashing on the floor. You thus receive a negative reinforcement (or *punishment*) to that action. If, with the same state of the aircraft as before, you had performed an action resulting in a smooth landing, then you would have received a positive reinforcement (or *reward*). By interacting with the environment and maximizing the amount of reward, you can learn how to smoothly land the plane. We have long been using reinforcement learning techniques to teach animals to play tricks for us, mainly in circuses. For example, consider teaching a dolphin a new trick: you cannot tell it what to do, but you can reward or punish it if it does the right or wrong thing. It has to figure out what it did that made it get the reward or the punishment. This process is also known as the *credit assignment problem.*

4.4 TYPICAL ANNS AND LEARNING ALGORITHMS

The previous section introduced the fundamentals of neurocomputing. It was argued that an artificial neural network can be designed by 1) choosing some abstract models of neurons; 2) defining a network architecture; and 3) choosing an appropriate learning algorithm. In the present section, some of the most commonly used neural networks will be described, focusing on the main three aspects above: type of neuron, network architecture, and learning algorithm. As will be explained throughout this section, the type of learning algorithm is intimately related with the type of application the neural network is going to be used on. In particular, supervised learning algorithms are used for function approximation, pattern classification, control, identification, and other related tasks. On the other hand, unsupervised learning algorithms are employed in data analysis, including clustering and knowledge discovery.

4.4.1. Hebbian Learning

After the 1943 McCulloch and Pitts' paper describing a logical calculus of neuronal activity, N. Wiener published a famous book named *Cybernetics* in 1948, followed by the publication of Hebb's book *The Organization of Behavior*. These were some landmarks in the history of neurocomputing.

In Hebb's book, an explicit statement of a physiological learning rule for synaptic modification was presented for the first time. Hebb proposed that the connectivity of the brain is continually changing as an organism learns distinct functional tasks, and that neural assemblies are created by such changes. Hebb stated that the effectiveness of a variable synapse between two neurons is increased by the repeated activation of one neuron by the other across that synapse. Quoting from Hebb's book *The Organization of Behavior*:

> "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency as one of the cells firing B, is increased." (Hebb, 1949; p. 62)

This postulate requires that change occur in the synaptic strength between cells when the presynaptic and postsynaptic cells are active at the same time. Hebb suggested that this change is the basis for *associative learning* that would result in a lasting modification in the activity pattern of a spatially distributed assemble of neurons.

This rule is often approximated in artificial neural networks by the *generalized Hebb rule*, where changes in connection strengths are given by the product of presynaptic and postsynaptic activity. The main difference being that change is now a result of both stimulatory and inhibitory synapses, not only excitatory synapses. In mathematical terms,

$$\Delta w_{ij}(t) = \alpha\, y_i(t)\, x_j(t) \qquad (4.20)$$

where Δw_{ij} is the change to be applied to neuron i, α is a *learning rate* parameter that determines how much change should be applied, y_i is the output of neuron i, x_j is the input to neuron i (output of neuron j), and t is the time index. Equation (4.20) can be expressed generically as

$$\Delta w_{ij}(t) = g(y_i(t), x_j(t)) \qquad (4.21)$$

where $g(\cdot,\cdot)$ is a function of both pre- and postsynaptic signals. The weight of a given neuron i is thus updated according to the following rule:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) \qquad (4.22)$$

Equation (4.20) clearly emphasizes the correlational or associative nature of Hebb's updating rule. It is known that much of human memory is *associative*, just as the mechanism suggested by Hebb. In an associative memory, an event is linked to another event, so that the presentation of the first event gives rise to the linked event. In the most simplistic version of association, a *stimulus* is linked to a *response*, so that a later presentation of the stimulus evokes the response.

Two important aspects of the Hebbian learning given by Equation (4.20) must be emphasized. First, it is a general updating rule to be used with different types of neurons. In the most standard case, the Hebbian network is but a single neuron with linear output, but more complex structures could be used. Second, this learning rule is unsupervised; that is, no information about the desired behavior is accounted for. However, the Hebb rule can also be employed when the target output for each input pattern is known. This means that the Hebb rule can also be used in supervised learning. In such cases, the modified Hebb rule has the current output substituted by the desired output:

$$\Delta w_{ij}(t) = \alpha\, d_i(t)\, x_j(t) \qquad (4.23)$$

where, d_i is the desired output of neuron i.

Other variations of the Hebb learning rule take into account, for example, the difference between the neuron's output and its desired output. This leads to the Widrow-Hoff and perceptron learning rules to be described later.

Biological Basis of Hebbian Synaptic Modification

Physiological evidence suggests the existence of Hebb synapses at various locations in the brain. Advances in modern neurophysiological techniques have allowed us to see what appears to be Hebbian modification in several parts of the mammalian brain (Anderson, 1995). A part of the cerebral cortex, the *hippocampus*, shows an effect called *long-term potentiation*, in which its neurons can be induced to display long-term, apparently permanent, increases in activity with particular patterns of stimulation.

Kelso et al. (1986) have presented an experimental demonstration of Hebbian modification using a slice of rat hippocampus maintained outside the animal. They showed that when a presynaptic cell is excited and the postsynaptic cells inhibited, little or no change was seen in the efficiency of a synapse. If the postsynaptic cell was excited by raising the membrane potential at the same time the presynaptic cell was active, the excitatory postsynaptic potential from the presynaptic cell was significantly increased in a stable and long-lasting form.

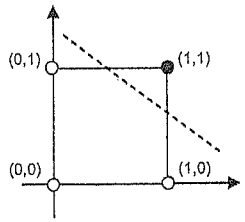
4.4.2. Single-Layer Perceptron

Rosenblatt (1958, 1962) introduced the *perceptron* as the simplest form of a neural network used for the classification of *linearly separable* patterns. Perceptrons constituted the first model of supervised learning, though some perceptrons were self-organized. Before describing the perceptron learning algorithm, the next section discusses in some more detail the problem of linear separability.

Linear Separability

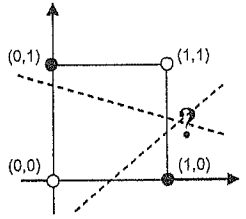
Linear separability can be easily understood with a simple example. Without loss of generality, assume there is a set of input patterns to be classified into a single class. Assume also there is a classifier system that has to respond TRUE if a given input pattern is a member of a certain class, and FALSE if it is not. A TRUE response is represented by an output '1' and a FALSE response by an output '0' of the classifier.

Inputs		Output
x_1	x_2	x_1 AND x_2
0	0	0
0	1	0
1	0	0
1	1	1



(a)

Inputs		Output
x_1	x_2	x_1 XOR x_2
0	0	0
0	1	1
1	0	1
1	1	0



(b)

Figure 4.19: Examples of linear and nonlinear separability. (a) The logic function AND is linearly separable. (b) The logic function XOR is nonlinearly separable.

As one of two responses is required, there is a *decision boundary* that separates one response from the other. Depending on the number m of variables of the input patterns, the decision boundary can have any shape in the space of dimension m. If $m = 2$ and the classifier can be set up so that there is a line dividing all input patterns that produce output '1' from those patterns that produce output '0', then the problem is *linearly separable*; else it is *nonlinearly separable*. This holds true for a space of any dimension, but for higher dimensions, a plane or hyper-plane should exist so as to divide the space into a '1'-class and a '0'-class region.

Figure 4.19 illustrates one linearly separable function and one nonlinearly separable function. Note that, in Figure 4.19(a), several lines could be drawn so as to separate the data into class '1' or '0'. The regions that separate the classes to which each of the input patterns belong to are often called *decision regions*, and the (hyper-)surface that defines these regions are called *decision surfaces*. Therefore, a classification problem can be viewed as the problem of finding decision surfaces that correctly classify the input data. The difficulty lying is that it is not always trivial to automatically define such surfaces.

Simple Perceptron for Pattern Classification

Basically, the perceptron consists of a single layer of neurons with adjustable synaptic weights and biases. Under suitable conditions, in particular if the training patterns belong to linearly separable classes, the iterative procedure of adaptation for the perceptron can be proved to converge to the correct weight set. These weights are such that the perceptron algorithm converges and positions the decision surfaces in the form of (hyper-)planes between the classes. This proof of convergence is known as the *perceptron convergence theorem*.

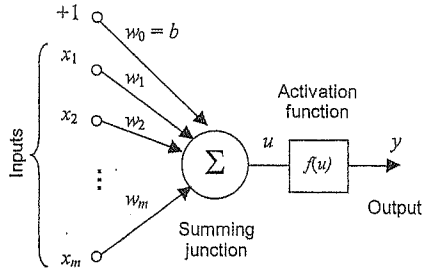


Figure 4.20: The simplest perceptron to perform pattern classification.

The simple perceptron has a single layer of feedforward neurons as illustrated in Figure 4.11(a). The neurons in this network are similar to those of McCulloch and Pitts with a signum or threshold activation function, but include a bias. The key contribution of Rosenblatt and his collaborators was to introduce a learning rule to train the perceptrons to solve pattern recognition and classification problems.

Their algorithm works as follows. For each training pattern $\mathbf{x}_i$, the network output response y_i is calculated. Then, the network determines if an error e_i occurred for this pattern by comparing the calculated output y_i with a *desired value* d_i for that pattern, $e_i = d_i - y_i$. Note that the desired output value for each training pattern is known (supervised learning). The weight vector connecting the inputs (presynaptic neurons) to each output (postsynaptic neuron) and the neuron's bias are updated according to the following rules:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha\, e_i\, \mathbf{x} \tag{4.24}$$

$$b(t+1) = b(t) + \alpha\, e \tag{4.25}$$

where $\mathbf{w} \in \Re^{1\times m}$, $\mathbf{x} \in \Re^{1\times m}$, and $b \in \Re^{1\times 1}$.

Consider now the simplest perceptron case with a single neuron, as illustrated in Figure 4.20. The goal of this network, more specifically neuron, is to classify a number of input patterns as belonging or not belonging to a particular class. Assume that the input data set is given by the N pairs of samples $\{\mathbf{x}_1,d_1\}$, $\{\mathbf{x}_2,d_2\}$, ... , $\{\mathbf{x}_N,d_N\}$, where $\mathbf{x}_j$ is the input vector j and d_j its corresponding *target* or *desired* output. In this case, the target value is '1' if the pattern belongs to the class, and '−1' (or '0') otherwise.

Let $\mathbf{X} \in \Re^{m\times N}$, be the matrix of N input patterns of dimension m each (the patterns are placed column wise in matrix $\mathbf{X}$), and $\mathbf{d} \in \Re^{1\times N}$ be the vector of desired outputs. The learning algorithm for this simple perceptron network is presented in Algorithm 4.1, where $f(\cdot)$ is the signum or the threshold function. The stopping criterion for this algorithm is either a fixed number of iteration steps (`max_it`) or the sum E of the squared errors e_i^2, $i = 1, \ldots, N$, for each input pattern being equal to zero. The algorithm returns the weight vector $\mathbf{w}$ as output.

```
procedure [w] = perceptron(max_it, α, X, d)
    initialize w      //set it to zero or small random values
    initialize b      //set it to zero or small random value
    t ← 1; E ← 1
    while t < max_it & E > 0 do,
        E ← 0
        for i from 1 to N do, //for each training pattern
            y_i ← f(w x_i + b)        //network output for x_i
            e_i ← d_i − y_i           //determine the error for x_i
            w ← w + α e_i x_i         //update the weight vector
            b ← b + α e_i             //update the bias term
            E ← E + e_i^2             //accumulate the error
        end for
        t ← t + 1
    end while
end procedure
```

Algorithm 4.1: Simple perceptron learning algorithm. The function $f(\cdot)$ is the signum (or the threshold) function, and the desired output is '1' if a pattern belongs to the class and '–1' (or '0') if it does not belong to the class.

Assuming the input patterns are linearly separable, the perceptron will be capable of solving the problem after a finite number of iteration steps and, thus, the error should be used as the stopping criterion. The parameter α is a *learning rate* that determines the step size of the adaptation in the weight values and bias term. Note, from Algorithm 4.1, that the perceptron learning rule only updates a given weight when the desired response is different from the actual response of the neuron.

Multiple Output Perceptron for Pattern Classification

Note that the perceptron updating rule uses the error-correction learning of most supervised learning techniques, as discussed in Section 4.3.3. This learning rule, proposed to update the weight vector of a single neuron, can be easily extended to deal with networks with more than one output neuron, such as the network presented in Figure 4.11(a). In this case, for each input pattern $\mathbf{x}_i$ the vector of network outputs is given by Equation (4.13) explicitly including the bias vector $\mathbf{b}$ as follows:

$$\mathbf{y} = f(\mathbf{W}\mathbf{x}_i + \mathbf{b}) \tag{4.26}$$

where $\mathbf{W} \in \Re^{o \times m}$, $\mathbf{x}_i \in \Re^{m \times 1}$, $i = 1, \ldots, N$, $\mathbf{y} \in \Re^{o \times 1}$, and $\mathbf{b} \in \Re^{o \times 1}$. The function $f(\cdot)$ is the signum or the threshold function.

Let the matrix of desired outputs be $\mathbf{D} \in \Re^{o \times N}$, where each column of $\mathbf{D}$ corresponds to the desired output for one of the input patterns. Therefore, the error signal for each input pattern $\mathbf{x}_i$ is calculated by simply subtracting the vectors $\mathbf{d}_i$ and $\mathbf{y}_i$ that are of same dimension: $\mathbf{e}_i = \mathbf{d}_i - \mathbf{y}_i$, where $\mathbf{e}_i, \mathbf{d}_i, \mathbf{y}_i \in \Re^{o \times 1}$. The algorithm to [illegible]nt the perceptron with multiple output neurons is presented in


```
procedure [W] = perceptron(max_it, α, X, D)
    initialize W      //set it to zero or small random values
    initialize b      //set it to zero or small random values
    t ← 1; E ← 1
    while t < max_it & E > 0 do,
        E ← 0
        for i from 1 to N do,  //for each training pattern
            y_i ← f(W x_i + b)        //network outputs for x_i
            e_i ← d_i − y_i           //determine the error for x_i
            W ← W + α e_i x_i^T       //update the weight matrix
            b ← b + α e_i             //update the bias vector
            E ← E + sum(e_ji^2)       //j = 1, ..., o

            ve_i ← e_i
        end for
        t ← t + 1
    end while
end procedure
```

Algorithm 4.2: Learning algorithm for the perceptron with multiple outputs. The function $f(\cdot)$ is the signum (or the threshold) function, and the desired output is '1' if a pattern belongs to the class and '–1' (or '0') if it does not belong to the class. Note that $\mathbf{e}_i$, $\mathbf{d}_i$, $\mathbf{y}_i$, and $\mathbf{b}$ are now vectors and $\mathbf{W}$ is a matrix. e_{ji} corresponds to the error of neuron j when presented with input pattern i.

Algorithm 4.2. As with the simple perceptron, in the perceptron with multiple output neurons the network will converge if the training patterns are linearly separable. In this case, the error for all patterns and all outputs will be zero at the end of the learning phase.

Examples of Application

To illustrate the applicability of the perceptron network, consider the two problems below. The first example - simple classification problem - illustrates the potentiality of the perceptron to represent Boolean functions, and the second example - character recognition - illustrates its capability to recognize a simple set of binary characters.

Simple Classification Problem

Consider the problem of using the simple perceptron with a single neuron to represent the AND function. The training data and its graphical interpretation are presented in Figure 4.21. The training patters to be used as inputs to Algorithm 4.1 are, in matrix notation:

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \quad \mathbf{d} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$$

Inputs		Output
x_1	x_2	x_1 AND x_2
0	0	0
0	1	0
1	0	0
1	1	1

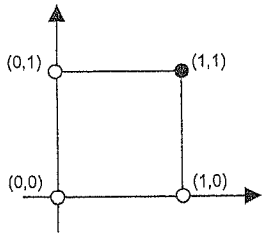


Figure 4.21: The AND function and its graphical representation.

The boundary between the values of x_1 and x_2 for which the network provides a response '0' (not belonging to the class) and the values for which the network responds '1' (belonging to the class) is the separating line given by

$$w_1 x_1 + w_2 x_2 + b = 0 \tag{4.27}$$

Assuming the activation function of the network is the threshold with $\theta = 0$, the requirement for a positive response from the output unit is that the net input it receives be greater than or equal to zero, that is, $w_1 x_1 + w_2 x_2 + b \geq 0$. During training, the values of w_1, w_2, and b are determined so that the network presents the correct response for all training data.

For simplicity, the learning rate is set to 1, $\alpha = 1$, and the initial values for the weights and biases are taken to be zero, $\mathbf{w} = [0\ 0]$ and $b = 0$. By running Algorithm 4.1 with **X**, **d**, and the other parameters given above, the following values are obtained for the perceptron network with a single output.

$$w_1 = 2;\ w_2 = 1;\ b = -3 \tag{4.28}$$

By replacing these values in the line equation for this neuron we obtain

$$2x_1 + 1x_2 - 3 = 0 \tag{4.29}$$

For each pair of training data, the following net input u_i, $i = 1, \ldots, N$, is calculated: $\mathbf{u} = [-3, -2, -1, 0]$. Therefore, the network response, $\mathbf{y} = f(\mathbf{u})$, for each input pattern is $\mathbf{y} = [0, 0, 0, 1]$, realizing the AND function as desired.

The decision line between the two classes can be determined by isolating the variable x_2 as a function of the other variable in Equation (4.27)

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2} = -2x_1 + 3 \tag{4.30}$$

Note that this line passes exactly on top of class '1', as illustrated in Figure 4.22.

An alternative form of initializing the weights for the perceptron network is by choosing random values within a given domain. For instance, a uniform or normal distribution with zero mean and predefined variance (usually less than or equal to one) could be used. Assume now that the following initial values for **w** and b were chosen using a uniform distribution of zero mean and variance one:

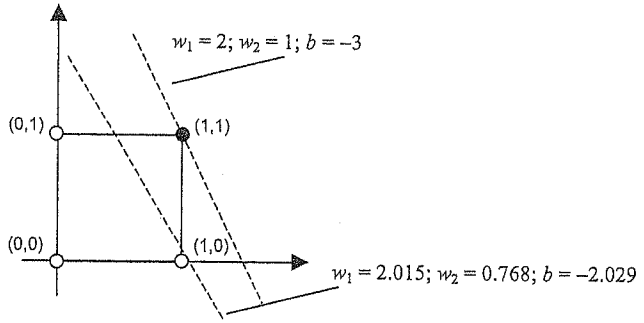


Figure 4.22: Decision lines for the simple perceptron network realizing the logic function AND.

$$w_1 = 0.015;\ w_2 = 0.768;\ b = 0.971$$

After training the network, the following values for the weights were determined:

$$w_1 = 2.015;\ w_2 = 0.768;\ b = -2.029$$

These values define the decision line below, as depicted in Figure 4.22.

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2} = -2.624\, x_1 + 2.642$$

Two important issues can be noticed here. First, the network response is dependent upon the initial values chosen for its weights and biases. Second, there is not a single correct response for a given training data set. These and other important issues on neural network training will be discussed later.

Character Recognition

As a second example of how to use the perceptron for pattern classification, consider the eight input patterns presented in Figure 4.23. Each input vector is a 120-tuple representing a letter expressed as a pattern on a 12 × 10 grid of pixels. Each character is assumed to have its own class. One output unit is assigned to each character/class, thus there are eight categories to which each character could be assigned. Matrix $\mathbf{X} \in \Re^{120\times 8}$, where each column of **X** corresponds to an input pattern, and matrix $\mathbf{D} \in \Re^{8\times 8}$ is a diagonal matrix, meaning that input pattern number 1 (character '0') should activate only output 1, input pattern number 2 (character '1') should activate only output 2, and so on.

Figure 4.23: Training patterns for the perceptron.

```
procedure [y] = run_perceptron(W,b,Z)
   for i from 1 to N do,   //for each training pattern
       y_i ← f(Wz_i + b)   //network outputs for z_i
   end for
end procedure
```

Algorithm 4.3: Algorithm used to run the trained single-layer perceptron network.

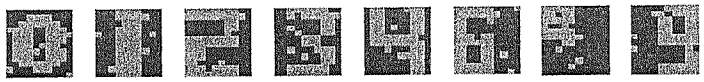


Figure 4.24: Noisy patterns used to test the generalization capability of the single-layer perceptron network trained to classify the patterns presented in Figure 4.23.

After training, the network correctly classifies each of the training patterns. Algorithm 4.3 can be used to run the network in order to evaluate if it is correctly classifying the training patterns. The weight matrix **W** and the bias vector **b** are those obtained after training. Matrix **Z** contains the patterns to be recognized; these can be the original training data **X**, the training data added with noise, or completely new unseen data. Note that, in this case, the network is first trained and then applied to classify novel data; that is, its generalization capability for unseen data can be evaluated.

To test how this network generalizes, some random *noise* can be inserted into the training patterns by simply changing a value '0' by '1' or a value '1' by '0', respectively, with a given probability. Figure 4.24 illustrates the training patterns with 5% noise.

4.4.3. ADALINE, the LMS Algorithm, and Error Surfaces

Almost at the same time Rosenblatt introduced the perceptron learning rule, B. Widrow and his student M. Hoff (Widrow and Hoff, 1960) developed the Widrow-Hoff learning rule, also called the *least mean squared (LMS) algorithm* or *delta rule*. They introduced the ADALINE (ADAptive LInear NEuron) network that is very similar to the perceptron, except that its activation function is linear instead of threshold. Although both networks, ADALINE and perceptron, suffer from only being capable of solving linearly separable problems, the LMS algorithm is more powerful than the perceptron learning rule and has found many more practical uses than the perceptron.

As the ADALINE employs neurons with a linear activation function, the neuron's output is equal to its net input. The goal of the learning algorithm is to minimize the *error* between the network output and the desired output. This allows the network to perform a continuous learning even after a given input pattern has been learnt. One advantage of the LMS algorithm over the perceptron learning rule is that the resultant network after training is not too sensitive to noise. It [illegible] observed in Figure 4.22, that the decision surfaces generated

by the perceptron algorithm usually lie very close to some input data. This is very much the case for the perceptron algorithm, but it is not the case for the LMS algorithm, making the latter more robust to noise. In addition, the LMS algorithm has been broadly used in signal processing applications.

LMS Algorithm (Delta Rule)

The LMS algorithm or delta rule is another example of supervised learning in which the learning rule is provided with a set of inputs and corresponding desired outputs $\{\mathbf{x}_1,\mathbf{d}_1\}$, $\{\mathbf{x}_2,\mathbf{d}_2\}$, ..., $\{\mathbf{x}_N,\mathbf{d}_N\}$, where $\mathbf{x}_j$ is the input vector j and $\mathbf{d}_j$ its corresponding *target* or *desired* output vector.

The delta rule adjusts the neural network weights and biases so as to minimize the difference (error) between the network output and the desired output over all training patterns. This is accomplished by reducing the error for each pattern, one at a time, though weight corrections can also be accumulated over a number of training patterns.

The *sum-squared error* (SSE) for a particular training pattern is given by

$$\Im = \sum_{i=1}^{o} e_i^2 = \sum_{i=1}^{o} (d_i - y_i)^2 \tag{4.31}$$

The gradient of $\Im$, also called the *performance index* or *cost function*, is a vector consisting of the partial derivatives of $\Im$ with respect to each of the weights. This vector gives the direction of most rapid increase of $\Im$; thus, its opposite direction is the one of most rapid decrease in the error value. Therefore, the error can be reduced most rapidly by adjusting the weight w_{IJ} in the following manner

$$w_{IJ} = w_{IJ} - \alpha \frac{\partial \Im}{\partial w_{IJ}} \tag{4.32}$$

where w_{IJ} is the weight from the J-th presynaptic neuron to the I-th postsynaptic neuron, and α is a learning rate.

It is necessary now to explicitly determine the gradient of the error with respect to the arbitrary weight w_{IJ}. As the weight w_{IJ} only influences the output (postsynaptic) unit I, the gradient of the error is

$$\frac{\partial \Im}{\partial w_{IJ}} = \frac{\partial}{\partial w_{IJ}} \sum_{i=1}^{o} (d_i - y_i)^2 = \frac{\partial}{\partial w_{IJ}} (d_I - y_I)^2$$

It is known that

$$y_I = f(\mathbf{w}_I.\mathbf{x}) = f(\Sigma_j w_{Ij}.x_j)$$

Then, we obtain

$$\frac{\partial \Im}{\partial w_{IJ}} = -2(d_I - y_I)\frac{\partial y_I}{\partial w_{IJ}} = -2(d_I - y_I)x_J$$

Therefore, similarly to the updating rule of Hebb's network, the LMS also assumes an updating value Δw_{IJ} to be added to the weight w_{IJ}. The delta rule for

updating the weight from the J-th presynaptic neuron to the I-th postsynaptic neuron is given by

$$\Delta w_{IJ} = \alpha\,(d_I - y_I)\,x_J \tag{4.33}$$

$$w_{IJ} = w_{IJ} + 2\alpha\,(d_I - y_I)\,x_J \tag{4.34}$$

The following equation for updating the bias b_I can be obtained by calculating the gradient of the error in relation to the bias b_I

$$b_I = b_I + 2\alpha\,(d_I - y_I) \tag{4.35}$$

For each input pattern $\mathbf{x}_i$, the LMS algorithm can be conveniently written in matrix notation as

$$\mathbf{W} = \mathbf{W} + \alpha\,\mathbf{e}_i\,\mathbf{x}_i^T$$

$$\mathbf{b} = \mathbf{b} + \alpha\,\mathbf{e}_i$$

where $\mathbf{W} \in \Re^{o\times m}$, $\mathbf{x}_i \in \Re^{m\times 1}$, $i = 1, \ldots, N$, $\mathbf{e}_i \in \Re^{o\times 1}$, and $\mathbf{b} \in \Re^{o\times 1}$. (Note that 2α was replaced by α.)

The beauty of this algorithm is that at each iteration it calculates an approximation to the gradient vector by simply multiplying the error by the input vector. And this approximation to the gradient can be used in a steepest descent-like algorithm with fixed learning rate.

Error Surfaces

Assume a neural network with n weights. This set of weights can be viewed as a point in an n-dimensional space, called *weight space*. If the neural network is used to classify a set of patterns, for each of these patterns the network will generate an error signal. This means that every set of weights (and biases) has an associated scalar error value; if the weights are changed, a new error value is determined. The error values for every set of weights define a surface in the weight space, called the *error surface* (Anderson, 1995; Hagan et al., 1996).

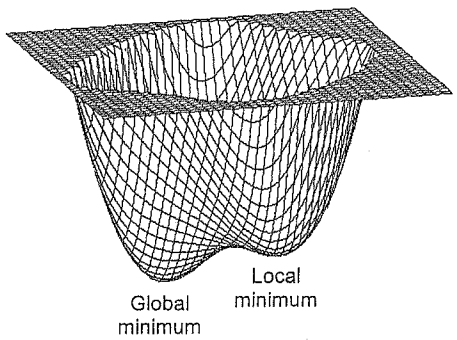


Figure 4.25: An error surface in weight space. The total error is a function of the values of the weights, so every set of weights has an associated error.

The question that arises now is what is the function of learning? The way the LMS algorithm, and the backpropagation algorithm to be explained in the next section, view learning is as the minimization of the error to its smallest value possible. This corresponds to finding an appropriate weight set that leads to the smallest error in the error surface. Figure 4.25 depicts an error surface in weight space. Note that this surface potentially has many *local minima* and one or more *global minimum*. In the example depicted the error surface contains a single local optimum and a single global optimum. Any gradient-like method, such as the LMS and the backpropagation algorithm, can only converge to local optima solutions. Therefore, the choice of an appropriate initial weight set for the network is crucial.

4.4.4. Multi-Layer Perceptron

The multi-layer perceptron is a kind of multi-layer feedforward network such as the one illustrated in Figure 4.12. Typically, the network consists of a set of input units that constitute the input layer, one or more hidden layers, and an output layer. The input signal propagates through the network in a forward direction, layer by layer. This network is a generalization of the single-layer perceptron discussed previously.

In the late 1960's, M. Minsky and S. Papert (1969) released a book called *Perceptrons* demonstrating the limitations of single-layered feedforward networks, namely, the incapability of solving nonlinearly separable problems. This caused a significant impact on the interest in neural network research during the 1970s. Both, Rosenblatt and Widrow were aware of the limitations of the perceptron network and proposed multi-layer networks to overcome this limitation. However, they were unable to generalize their algorithms to train these more powerful networks.

It has already been discussed that a multi-layer network with linear nodes is equivalent to a single-layered network with linear units. Therefore, one necessary condition for the multi-layer network to be more powerful than the single-layer networks is the use of nonlinear activation functions. More specifically, multi-layer networks typically employ a sigmoidal activation function (Figure 4.10).

Multi-layer perceptrons (MLP) have been applied successfully to a number of problems by training them with a supervised learning algorithm known as the *error backpropagation* or simply *backpropagation*. This algorithm is also based on an error-correction learning rule and can be viewed as a generalization of the LMS algorithm or delta rule.

The backpropagation algorithm became very popular with the publication of the *Parallel Distributed Processing* volumes by Rumelhart and collaborators (Rumelhart et al., 1986; McClelland et al., 1986). These books were also crucial for the re-emergence of interest for the research on neural networks. Basically, the error backpropagation learning consists of two passes of computation: a *forward* and a *backward* pass (see Figure 4.12(b)). In the forward pass, an input pattern is presented to the network and its effects are propagated through the

network until they produce the network output(s). In the backward pass, the synaptic weights, so far kept fixed, are updated in accordance with an error correction rule.

The Backpropagation Learning Algorithm

To derive the backpropagation learning algorithm, consider the following notation:

i, j	Indices referring to neurons in the network
t	Iteration counter
N	Number of training patterns
M	Number of layers in the network
$y_j(t)$	Output signal of neuron j at iteration t
$e_j(t)$	Error signal of output unit j at iteration t
$w_{ji}(t)$	Synaptic weight connecting the output of unit j to the input of unit i at iteration t
$u_j(t)$	Net input of unit j at iteration t
$f_j(\cdot)$	Activation function of unit j
$\mathbf{X}$	Matrix of input (training) patterns
$\mathbf{D}$	Matrix of desired outputs
$x_i(t)$	i-th element of the input vector at iteration t
$d_j(t)$	j-th element of the desired output vector at iteration t
α	Learning rate

The description to be presented here follows that of Hagan et al. (1996) and de Castro (1998). For multi-layer networks, the output of a given layer is the input to the next layer,

$$\mathbf{y}^{m+1} = \mathbf{f}^{m+1}(\mathbf{W}^{m+1}\mathbf{y}^m + \mathbf{b}^{m+1}),\ m = 0, 1, \ldots, M-1 \tag{4.36}$$

where M is the number of layers in the network, and the superindex m refers to the layer taken into account (e.g., $m = 0$: input layer, $m = 1$: first hidden layer, ... , $m = M - 1$: output layer). The nodes in the input layer ($m = 0$) receive the input patterns

$$\mathbf{y}^0 = \mathbf{x} \tag{4.37}$$

that represent the initial condition for Equation (4.36). The outputs in the output layer are the network outputs:

$$\mathbf{y} = \mathbf{y}^M \tag{4.38}$$

Equation (4.14) demonstrates that the network output can be a function of only the input vector $\mathbf{x}$ and the weight matrices $\mathbf{W}^m$. If we explicitly consider the bias terms $\mathbf{b}^m$, Equation (4.13) becomes Equation (4.39) for the network of Figure 4.26:

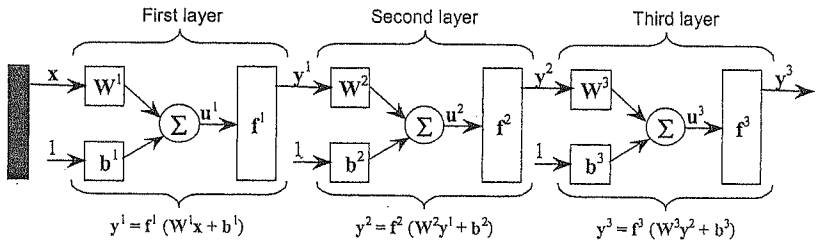


Figure 4.26: An MLP network with three layers. (Matrix notation.)

$$\mathbf{y} = \mathbf{y}^3 = \mathbf{f}^3(\mathbf{W}^3\mathbf{f}^2(\mathbf{W}^2\mathbf{f}^1(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) + \mathbf{b}^2) + \mathbf{b}^3) \tag{4.39}$$

The backpropagation algorithm to MLP networks is a generalization of the LMS algorithm that uses as a performance index the *mean squared error* (MSE). A set of input patterns and desired outputs is provided, as follows:

$$\{(\mathbf{x}_1,\mathbf{d}_1), (\mathbf{x}_2,\mathbf{d}_2), \ldots, (\mathbf{x}_N,\mathbf{d}_N)\}$$

where $\mathbf{x}_i$ is the i-th input to the network and $\mathbf{d}_i$ is the corresponding desired output, $i = 1,\ldots,N$. From Equation (4.39) it is possible to observe that if ω is the vector of network parameters (weights and biases), then the network output can be given as a function of ω and x:

$$y^M = f(\omega, x)$$

After each training pattern is presented to the network, the network output is compared with the desired output. The algorithm must adjust the vector of parameters so as to minimize the mathematical expectation of the mean squared error:

$$\Im(\omega) = E(e(\omega)^2) = E((d - y(\omega))^2) \tag{4.40}$$

If the network has multiple outputs, Equation (4.40) becomes

$$\Im(\omega) = E(\mathbf{e}(\omega)^T\mathbf{e}(\omega)) = E(\ (\mathbf{d} - \mathbf{y}(\omega))^T\ (\mathbf{d} - \mathbf{y}(\omega))\ )$$

Similarly to the LMS algorithm, the mean squared error can be approximated by the following expression:

$$\hat{\Im}(\omega) = \mathbf{e}(t)^T\mathbf{e}(t) = (\mathbf{d}(t) - \mathbf{y}(t))^T\ (\mathbf{d}(t) - \mathbf{y}(t))$$

where the expectation of the mean squared error was substituted by the error at iteration t. In order not to overload the notation, assume $\hat{\Im}(\omega) = \Im(\omega)$.

The updating rule known as steepest descent to minimize the squared error is given by

$$w_{ij}^m(t+1) = w_{ij}^m(t) - \alpha\frac{\partial\Im(t)}{\partial w_{ij}^m} \tag{4.41}$$

$$b_i^m(t+1) = b_i^m(t) - \alpha\frac{\partial\Im(t)}{\partial b_i^m} \tag{4.42}$$

where α is the learning rate.

The most elaborate part is the determination of the partial derivatives that will produce the components of the gradient vector. To determine these derivatives it will be necessary to apply the chain rule a number of times.

The Chain Rule

For a multi-layer network, the error is not a direct function of the weights in the hidden layers, reason why the calculus of these derivatives is not straightforward.

As the error is an indirect function of the weights in the hidden layers, the chain rule must be used to determine the derivatives. The chain rule will be used to determine the derivatives in Equation (4.41) and Equation (4.42).

$$\frac{\partial \Im}{\partial w_{ij}^m} = \frac{\partial \Im}{\partial u_i^m} \times \frac{\partial u_i^m}{\partial w_{ij}^m} \tag{4.43}$$

$$\frac{\partial \Im}{\partial b_i^m} = \frac{\partial \Im}{\partial u_i^m} \times \frac{\partial u_i^m}{\partial b_i^m} \tag{4.44}$$

The second term of both equations above can be easily determined for the net input of layer m as an explicit function of the weights and biases in this layer:

$$u_i^m = \sum_{j=1}^{S^{m-1}} w_{ij}^m y_j^{m-1} + b_i^m \tag{4.45}$$

where S^m is the number of neurons in layer m.

Therefore,

$$\frac{\partial u_i^m}{\partial w_{ij}^m} = y_j^{m-1}, \frac{\partial u_i^m}{\partial b_i^m} = 1 \tag{4.46}$$

Now define the *sensitivity* (δ) of $\Im$ to changes in the i-th element of the net input in layer m as

$$\delta_i^m \equiv \frac{\partial \Im}{\partial u_i^m} \tag{4.47}$$

Equation (4.43) and Equation (4.44) can now be simplified by

$$\frac{\partial \Im}{\partial w_{ij}^m} = \delta_i^m y_j^{m-1} \tag{4.48}$$

$$\frac{\partial \Im}{\partial b_i^m} = \delta_i^m \tag{4.49}$$

Thus, Equation (4.41) and Equation (4.42) become

$$w_{ij}^m(t+1) = w_{ij}^m(t) - \alpha \delta_i^m y_j^{m-1} \tag{4.50}$$

$$b_i^m(t+1) = b_i^m(n) - \alpha \delta_i^m \tag{4.51}$$

In matrix notation these equations become

$$\mathbf{W}^m(t+1) = \mathbf{W}^m(t) - \alpha \boldsymbol{\delta}^m (\mathbf{y}^{m-1})^T \tag{4.52}$$

$$\mathbf{b}^m(t+1) = \mathbf{b}^m(t) - \alpha \boldsymbol{\delta}^m \tag{4.53}$$

where

$$\boldsymbol{\delta}^m \equiv \frac{\partial \Im}{\partial \mathbf{u}^m} = \begin{bmatrix} \frac{\partial \Im}{\partial u_1^m} \\ \frac{\partial \Im}{\partial u_2^m} \\ \vdots \\ \frac{\partial \Im}{\partial u_{S^m}^m} \end{bmatrix} \tag{4.54}$$

Backpropagating the Sensitivities

It is now necessary to calculate the sensitivity δ^m that requires another application of the chain rule. This process gives rise to the term *backpropagation* because it describes a recurrence relationship in which the sensitivity of layer m is determined from the sensitivity of layer $m+1$.

To derive the recurrence relationship for the sensitivities, we will use the following Jacobian matrix

$$\frac{\partial \mathbf{u}^{m+1}}{\partial \mathbf{u}^m} = \begin{bmatrix} \frac{\partial u_1^{m+1}}{\partial u_1^m} & \frac{\partial u_1^{m+1}}{\partial u_2^m} & \cdots & \frac{\partial u_1^{m+1}}{\partial u_{S^m}^m} \\ \frac{\partial u_2^{m+1}}{\partial u_1^m} & \frac{\partial u_2^{m+1}}{\partial u_2^m} & \cdots & \frac{\partial u_2^{m+1}}{\partial u_{S^m}^m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u_{S^{m+1}}^{m+1}}{\partial u_1^m} & \frac{\partial u_{S^{m+1}}^{m+1}}{\partial u_2^m} & \cdots & \frac{\partial u_{S^{m+1}}^{m+1}}{\partial u_{S^m}^m} \end{bmatrix} \tag{4.55}$$

Next, we want to find an expression for this matrix. Consider the element i,j of this matrix

$$\frac{\partial u_i^{m+1}}{\partial u_j^m} = w_{ij}^{m+1} \frac{\partial y_j^m}{\partial u_j^m} = w_{ij}^{m+1} \frac{\partial f^m(u_j^m)}{\partial u_j^m} = w_{ij}^{m+1} \dot{f}^m(u_j^m) \tag{4.56}$$

where

$$\dot{f}^m(u_j^m) = \frac{\partial f^m(u_j^m)}{\partial u_j^m} \tag{4.57}$$

Therefore, the Jacobian matrix can be written as

$$\frac{\partial \mathbf{u}^{m+1}}{\partial \mathbf{u}^{m}} = \mathbf{W}^{m+1}\dot{\mathbf{F}}^{m}(\mathbf{u}^{m}) \tag{4.58}$$

where

$$\dot{\mathbf{F}}^{m}(\mathbf{u}^{m}) = \begin{bmatrix} \dot{f}^{m}(u_1^m) & 0 & \cdots & 0 \\ 0 & \dot{f}^{m}(u_2^m) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dot{f}^{m}(u_{s^m}^m) \end{bmatrix} \tag{4.59}$$

It is now possible to write the recurrence relationship for the sensitivity using the chain rule in matrix form

$$\begin{aligned} \boldsymbol{\delta}^{m} &= \frac{\partial \Im}{\partial \mathbf{u}^{m}} = \left(\frac{\partial \mathbf{u}^{m+1}}{\partial \mathbf{u}^{m}}\right)^{T} \frac{\partial \Im}{\partial \mathbf{u}^{m+1}} = \dot{\mathbf{F}}^{m}(\mathbf{u}^{m})(\mathbf{W}^{m+1})^{T} \frac{\partial \Im}{\partial \mathbf{u}^{m+1}} \\ &= \dot{\mathbf{F}}^{m}(\mathbf{u}^{m})(\mathbf{W}^{m+1})^{T}\boldsymbol{\delta}^{m+1} \end{aligned} \tag{4.60}$$

Note that the sensitivities are (back)propagated from the last to the first layer

$$\boldsymbol{\delta}^{M} \rightarrow \boldsymbol{\delta}^{M-1} \rightarrow \ldots \rightarrow \boldsymbol{\delta}^{2} \rightarrow \boldsymbol{\delta}^{1} \tag{4.61}$$

There is a last step to be executed to complete the backpropagation algorithm. We need the starting point, $\boldsymbol{\delta}^{M}$, for the recurrence relation of Equation (4.60).

$$\delta_i^{M} = \frac{\partial \Im}{\partial u_i^{M}} = \frac{\partial (\mathbf{d}-\mathbf{y})^{T}(\mathbf{d}-\mathbf{y})}{\partial u_i^{M}} = \frac{\partial \sum_{j=1}^{S^M}(d_j - y_j)^2}{\partial u_i^{M}} = -2(d_i - y_i)\frac{\partial y_i}{\partial u_i^{M}} \tag{4.62}$$

Now, since

$$\frac{\partial y_i}{\partial u_i^{M}} = \frac{\partial y_i^{M}}{\partial u_i^{M}} = \frac{\partial f^{M}(u_j^{M})}{\partial u_i^{M}} = \dot{f}^{M}(u_j^{M}) \tag{4.63}$$

It is possible to write

$$\delta_i^{M} = -2(d_i - y_i)\dot{f}^{M}(u_j^{M}) \tag{4.64}$$

Or equivalently in matrix notation

$$\boldsymbol{\delta}^{M} = -2\dot{\mathbf{F}}^{M}(\mathbf{u}^{M})(\mathbf{d}-\mathbf{y}) \tag{4.65}$$

Figure 4.27 provides an adaptation, using matrix notation, of the result presented by Narendra and Parthasarathy (1990) and describes a flow graph for the backpropagation algorithm.

Algorithm 4.4 presents the pseudocode for the standard backpropagation algorithm. The value of α should be small, $\alpha \in (0,1]$. As this algorithm updates the weight matrices of the network after the presentation of each input pattern, this could cause a certain bias toward the order in which the training patterns are presented. In order to avoid this, the patterns are presented to the network in a random or described in Algorithm 4.4.

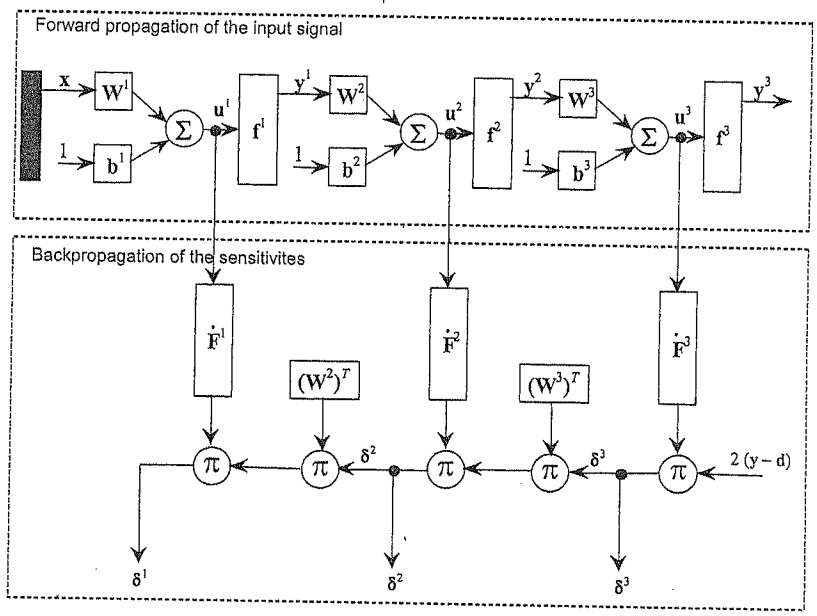


Figure 4.27: Architectural graph of an MLP network with three layers representing the forward propagation of the functional signals and the backward propagation of the sensitivities.

```
procedure [W] = backprop(max_it,min_err,α,X,D)
   for m from 1 to M do,
      initialize W^m //small random values
      initialize b^m //small random values
   end for
   t ← 1
   while t < max_it & MSE > min_err do,
      vet_permut ← randperm(N)   //permutations of N
      for j from 1 to N do,      //for all input patterns
         //select the index i of pattern x_i to be presented
         i ← vet_permut(j)  //present patterns randomly
         //forward propagation of the functional signal
         y^0 ← x_i                                  //Equation (4.37)
         for m from 0 to M - 1 do,
            y_i^{m+1} ← f^{m+1}(W^{m+1} y_i^m + b^{m+1})   //Equation (4.36)
         end for
         //backpropagation of sensitivities
         δ_i^M ← -2Ḟ^M(u_i^M)(d_i - y_i)            //Equation (4.65)
         for m from M - 1 down to 1 do,
            δ_i^m ← Ḟ^m(u_i^m)(W^{m+1})^T δ_i^{m+1},   //Equation (4.60)
         end for
         //update weights and biases
         for m from 1 to M do,
            W^m ← W^m - αδ_i^m (y_i^{m-1})^T,       //Equation (4.52)
            b^m ← b^m - αδ_i^m,                     //Equation (4.53)
         end for
         //calculate the error for pattern i
         E_i ← e_i^T e_i = (d_i - y_i)^T (d_i - y_i)
      end for
      MSE ← 1/N.sum(E_i)                            //Mean Square Error
      t ← t + 1
   end while
end procedure
```

Algorithm 4.4: Learning algorithm for the MLP network trained via the backpropagation algorithm.

Universal Function Approximation

An MLP network can be seen as a generic tool to perform a *nonlinear input-output mappings*. More specifically, let m be the number of inputs to the network and o the number of outputs. The input-output relationship of the network defines a mapping from an input m-dimensional Euclidean space into an output o-dimensional Euclidean space, which is infinitely continuously differentiable.

Cybenko (1989) was the first researcher to rigorously demonstrate that an MLP neural network with a single hidden layer is sufficient to uniformly approximate any continuous function that fits a unit hypercube.

The *universal function approximation theorem* is as follows:

Theorem: Let $f(\cdot)$ be a nonconstant continuous, limited, and monotonically increasing function. Let I_m be a unit hypercube of dimension m, $(0,1)^m$. The space of continuous functions in I_m is denoted by $C(I_m)$. Thus, given any function $g \in C(I_m)$ and $\varepsilon > 0$, there is an integer M and sets of real-valued constants α_i and w_{ij}, where $i = 1, \ldots, o$ and $j = 1, \ldots, m$, it is possible to define

$$F(x_1, x_2, \ldots, x_m) = \sum_{i=1}^{o} \alpha_i f\left(\sum_{j=1}^{m} w_{ij} x_j - w_{i0} \right) \tag{4.66}$$

as an approximation to the function $g(.)$ such that

$|F(x_1, \ldots, x_m) - g(x_1, \ldots, x_m)| < \varepsilon$ for all $\{x_1, \ldots, x_m\} \in I_m$.

Proof: see Cybenko (1989).

This theorem is directly applicable to the MLP networks. First, note that the sigmoidal function used in the MLP networks is continuous, nonconstant, limited, and monotonically increasing; satisfying the constraints imposed to $f(\cdot)$. Then, note that Equation (4.66) represents the outputs of an MLP network, as illustrated in Figure 4.28.

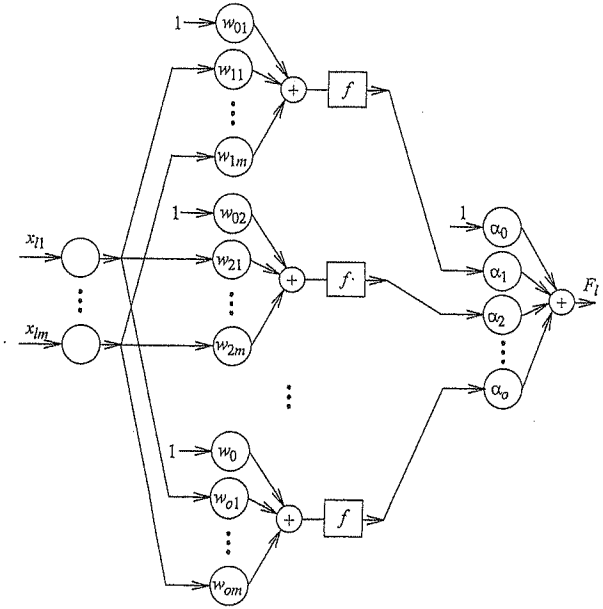


Figure 4.28: An MLP network as a universal function approximator. The network parameters compose Equation (4.66). (Courtesy of © Fernando J. Von Zuben.)

In summary, the theorem states that an MLP network with a single hidden layer is capable of uniform approximation, given an appropriate training data set to represent the function. However, the theorem does not say anything regarding the number of units in the hidden layer necessary to perform an approximation with the precision ε.

Some Practical Aspects

There are a several aspects of the MLP network training that deserve some comments. Some of these aspects are also valid for the other networks presented and to be presented in this chapter.

- *Network architecture*: the number of network inputs is usually defined by the training data, while the number of outputs and hidden units are design issues. For instance, if a data set to be classified has a single class, one or two output units can be used. In the case of a single sigmoidal output in the range $(-1,1)$, a network output less than '0' can be considered as not belonging to the class, and an output greater than or equal to zero can be considered as belonging to the class. If two output units are used, then one of them corresponds to belonging to the class, while the other corresponds to not belonging to the class. Finally, the number of hidden units is chosen so that the network can appropriately classify or approximate the training set. By increasing the number of hidden units, one increases the network mapping capability. It is important to have in mind that an excessive number of hidden units may result in overfitting, while a small number may lead to underfitting.
- *Generalization*: we have already discussed that too much training may result in overfitting, while too little training may result in underfitting. Overfitting is only possible if there are enough hidden units to promote an increasingly better representation of the data set. On the other side, if the network is not big enough to perform an adequate input-output mapping, underfitting may occur.
- *Convergence*: the MLP network is usually trained until a minimum value for the error is achieved; for instance, stop training if the MSE < 0.001. Another choice for the *stopping criterion* is to stop training if the estimated gradient vector at iteration t has a small norm; for instance, stop training if $\|\partial\Im/\partial w\| < 0.001$. A small value for the norm of the gradient value indicates that the network is close to a local minimum.
- *Epoch*: after all the training patterns have been presented to the network and the weights adjusted, an *epoch* is said to have been completed. This terminology is used in all artificial neural network learning algorithms, including the perceptron and self-organizing maps.
- *Data normalization*: when the activation functions of the neurons have a well-defined range, such as all sigmoidal functions, it is important to normalize the data such that they belong to a range equal to or smaller than the range of the activation functions. This helps to avoid the saturation of the neurons in the network. Normalizing the training data is also important when their attributes range over different scales. For example, a data set about cars may include variables such as color (symbolic or discrete attribute), cost (real-valued attribute), type (off road, van, etc.), and so on. Each of these attributes assumes a value on a different range and, thus, influences the network with different degrees. By normalizing all attributes to the same range, we reduce the importance of the differences in scale.

- *Initialization of the weight vectors and biases*: most network models, mainly multi-layer feedforward networks, are sensitive to the initial values chosen for the network weights. This sensitivity can be expressed in several ways, such as guarantee of convergence, convergence speed, and convergence to local optima solutions. Usually, small random values are chosen to initialize the network weights.
- *Learning rate*: the learning rate α plays a very important role in neural network training because it establishes the size of the adaptation step to be performed on a given weight. Too small learning rates may result in a very long learning time, while too large learning rates may result in instability and nonconvergence.

Biological Plausibility of Backpropagation

The backpropagation of error signals is probably the most problematic feature in biological terms. Despite the apparent simplicity and elegance of the backpropagation algorithm, it seems quite implausible that something like the equations described above are computed in the brain (Anderson, 1995; O'Reilly and Munakata, 2000; Trappenberg, 2002).

In order to accomplish this, some form of information exchange between postsynaptic and presynaptic neurons should be possible. For instance, there would be the requirement, in biological terms, that the sensitivity values were propagated backwards from the dendrite of the postsynaptic neuron, across the synapse, into the axon terminal of the presynaptic neuron, down the axon of this neuron, and then integrated and multiplied by both the strength of that synapse and some kind of derivative, and then propagated back out its dendrites, and so on. If a neuron were to gather the backpropagated errors from all the other nodes to which it projects, some synchronization issues would arise, and it would also affect the true parallel processing in the brain.

However, it is possible to rewrite the backpropagation equations so that the error backpropagation between neurons takes place using standard activation signals. This approach has the advantage that it ties into the psychological interpretation of the teaching signal (desired output) as an actual state of experience that reflects something like an outcome or corrected response. The result is a very powerful learning algorithm that need not ignore issues of biological plausibility (Hinton and McClelland, 1987; O'Reilly, 1996).