

The Biological Paradigm

1.1 Neural computation

Research in the field of neural networks has been attracting increasing attention in recent years. Since 1943, when Warren McCulloch and Walter Pitts presented the first model of artificial neurons, new and more sophisticated proposals have been made from decade to decade. Mathematical analysis has solved some of the mysteries posed by the new models but has left many questions open for future investigations. Needless to say, the study of neurons, their interconnections, and their role as the brain's elementary building blocks is one of the most dynamic and important research fields in modern biology. We can illustrate the relevance of this endeavor by pointing out that between 1901 and 1991 approximately ten percent of the Nobel Prizes for Physiology and Medicine were awarded to scientists who contributed to the understanding of the brain. It is not an exaggeration to say that we have learned more about the nervous system in the last fifty years than ever before.

In this book we deal with *artificial neural networks*, and therefore the first question to be clarified is their relation to the biological paradigm. What do we abstract from real neurons for our models? What is the link between neurons and artificial computing units? This chapter gives a preliminary answer to these important questions.

1.1.1 Natural and artificial neural networks

Artificial neural networks are an attempt at modeling the information processing capabilities of nervous systems. Thus, first of all, we need to consider the essential properties of biological neural networks from the viewpoint of information processing. This will allow us to design abstract models of artificial neural networks, which can then be simulated and analyzed.

Although the models which have been proposed to explain the structure of the brain and the nervous systems of some animals are different in many

respects, there is a general consensus that the essence of the operation of neural ensembles is “control through communication” [72]. Animal nervous systems are composed of thousands or millions of interconnected cells. Each one of them is a very complex arrangement which deals with incoming signals in many different ways. However, neurons are rather slow when compared to electronic logic gates. These can achieve switching times of a few nanoseconds, whereas neurons need several milliseconds to react to a stimulus. Nevertheless the brain is capable of solving problems which no digital computer can yet efficiently deal with.

Massive and hierarchical networking of the brain seems to be the fundamental precondition for the emergence of consciousness and complex behavior [202]. So far, however, biologists and neurologists have concentrated their research on uncovering the properties of individual neurons. Today, the mechanisms for the production and transport of signals from one neuron to the other are well-understood physiological phenomena, but how these individual systems cooperate to form complex and massively parallel systems capable of incredible information processing feats has not yet been completely elucidated. Mathematics, physics, and computer science can provide invaluable help in the study of these complex systems. It is not surprising that the study of the brain has become one of the most interdisciplinary areas of scientific research in recent years.

However, we should be careful with the metaphors and paradigms commonly introduced when dealing with the nervous system. It seems to be a constant in the history of science that the brain has always been compared to the most complicated contemporary artifact produced by human industry [297]. In ancient times the brain was compared to a pneumatic machine, in the Renaissance to a clockwork, and at the end of the last century to the telephone network. There are some today who consider computers the paradigm par excellence of a nervous system. It is rather paradoxical that when John von Neumann wrote his classical description of future universal computers, he tried to choose terms that would describe computers in terms of brains, not brains in terms of computers.

The nervous system of an animal is an information processing totality. The sensory inputs, i.e., signals from the environment, are coded and processed to evoke the appropriate response. Biological neural networks are just one of many possible solutions to the problem of processing information. The main difference between neural networks and conventional computer systems is the massive parallelism and redundancy which they exploit in order to deal with the unreliability of the individual computing units. Moreover, biological neural networks are self-organizing systems and each individual neuron is also a delicate self-organizing structure capable of processing information in many different ways.

In this book we study the information processing capabilities of complex hierarchical networks of simple computing units. We deal with systems whose structure is only partially predetermined. Some parameters modify the ca-

pabilities of the network and it is our task to find the best combination for the solution of a given problem. The adjustment of the parameters will be done through a *learning algorithm*, i.e., not through explicit programming but through an automatic adaptive method.

A cursory review of the relevant literature on artificial neural networks leaves the impression of a chaotic mixture of very different network topologies and learning algorithms. Commercial neural network simulators sometimes offer several dozens of possible models. The large number of proposals has led to a situation in which each single model appears as part of a big puzzle whereas the bigger picture is absent. Consequently, in the following chapters we try to solve this puzzle by systematically introducing and discussing each of the neural network models in relation to the others.

Our approach consists of stating and answering the following questions: what information processing capabilities emerge in hierarchical systems of primitive computing units? What can be computed with these networks? How can these networks determine their structure in a self-organizing manner?

We start by considering biological systems. Artificial neural networks have aroused so much interest in recent years, not only because they exhibit interesting properties, but also because they try to mirror the kind of information processing capabilities of nervous systems. Since information processing consists of transforming signals, we deal with the biological mechanisms for their generation and transmission in this chapter. We discuss those biological processes by which neurons produce signals, and absorb and modify them in order to retransmit the result. In this way biological neural networks give us a clue regarding the properties which would be interesting to include in our artificial networks.

1.1.2 Models of computation

Artificial neural networks can be considered as just another approach to the problem of computation. The first formal definitions of computability were proposed in the 1930s and '40s and at least five different alternatives were studied at the time. The computer era was started, not with one single approach, but with a contest of alternative computing models. We all know that the von Neumann computer emerged as the undisputed winner in this confrontation, but its triumph did not lead to the dismissal of the other computing models. Figure 1.1 shows the five principal contenders:

The mathematical model

Mathematicians avoided dealing with the problem of a function's computability until the beginning of this century. This happened not just because existence theorems were considered sufficient to deal with functions, but mainly because nobody had come up with a satisfactory definition of *computability*, certainly a relative concept which depends on the specific tools that can be

used. The general solution for algebraic equations of degree five, for example, cannot be formulated using only algebraic functions, yet this can be done if a more general class of functions is allowed as computational primitives. The squaring of the circle, to give another example, is impossible using ruler and compass, but it has a trivial real solution.

If we want to talk about computability we must therefore specify which tools are available. We can start with the idea that some primitive functions and composition rules are “obviously” computable. All other functions which can be expressed in terms of these primitives and composition rules are then also computable.

David Hilbert, the famous German mathematician, was the first to state the conjecture that a certain class of functions contains all intuitively computable functions. Hilbert was referring to the primitive recursive functions, the class of functions which can be constructed from the zero and successor function using composition, projection, and a deterministic number of iterations (primitive recursion). However, in 1928, Wilhelm Ackermann was able to find a computable function which is not primitive recursive. This led to the definition of the general recursive functions [154]. In this formalism, a new composition rule has to be introduced, the so-called μ operator, which is equivalent to an indeterminate recursion or a lookup in an infinite table. At the same time Alonzo Church and collaborators developed the lambda calculus, another alternative to the mathematical definition of the computability concept [380]. In 1936, Church and Kleene were able to show that the general recursive functions can be expressed in the formalism of the lambda calculus. This led to the *Church thesis* that computable functions are the general recursive functions. David Deutsch has recently added that this thesis should be considered to be a statement about the physical world and be given the same status as a physical principle. He thus speaks of a “Church principle” [109].

The logic-operational model (Turing machines)

In his classical paper “On Computable Numbers with an Application to the Entscheidungsproblem” Alan Turing introduced another kind of computing model. The advantage of his approach is that it consists in an operational, mechanical model of computability. A Turing machine is composed of an infinite tape, in which symbols can be stored and read again. A read-write head can move to the left or to the right according to its internal state, which is updated at each step. The *Turing thesis* states that computable functions are those which can be computed with this kind of device. It was formulated concurrently with the Church thesis and Turing was able to show almost immediately that they are equivalent [435]. The Turing approach made clear for the first time what “programming” means, curiously enough at a time when no computer had yet been built.

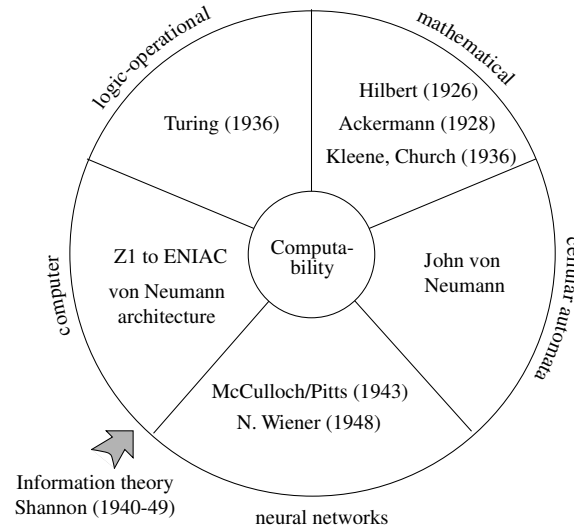


Fig. 1.1. Five models of computation

The computer model

The first electronic computing devices were developed in the 1930s and '40s. Since then, “computation-with-the-computer” has been regarded as computability itself. However the first engineers developing computers were for the most part unaware of Turing’s or Church’s research. Konrad Zuse, for example, developed in Berlin between 1938 and 1944 the computing machines Z1 and Z3 which were programmable but not universal, because they could not reach the whole space of the computable functions. Zuse’s machines were able to process a sequence of instructions but could not iterate. Other computers of the time, like the Mark I built at Harvard, could iterate a constant number of times but were incapable of executing open-ended iterations (WHILE loops). Therefore the Mark I could compute the primitive but not the general recursive functions. Also the ENIAC, which is usually hailed as the world’s first electronic computer, was incapable of dealing with open-ended loops, since iterations were determined by specific connections between modules of the machine. It seems that the first universal computer was the Mark I built in Manchester [96, 375]. This machine was able to cover all computable functions by making use of conditional branching and self-modifying programs, which is one possible way of implementing indexed addressing [268].

Cellular automata

The history of the development of the first mechanical and electronic computing devices shows how difficult it was to reach a consensus on the architecture of universal computers. Aspects such as the economy or the dependability of the building blocks played a role in the discussion, but the main problem was the definition of the minimal architecture needed for universality. In machines like the Mark I and the ENIAC there was no clear separation between memory and processor, and both functional elements were intertwined. Some machines still worked with base 10 and not 2, some were sequential and others parallel.

John von Neumann, who played a major role in defining the architecture of sequential machines, analyzed at that time a new computational model which he called *cellular automata*. Such automata operate in a “computing space” in which all data can be processed simultaneously. The main problem for cellular automata is communication and coordination between all the computing cells. This can be guaranteed through certain algorithms and conventions. It is not difficult to show that all computable functions, in the sense of Turing, can also be computed with cellular automata, even of the one-dimensional type, possessing only a few states. Turing himself considered this kind of computing model at one point in his career [192].

Cellular automata as computing model resemble massively parallel multi-processor systems of the kind that has attracted considerable interest recently.

The biological model (neural networks)

The explanation of important aspects of the physiology of neurons set the stage for the formulation of artificial neural network models which do not operate sequentially, as Turing machines do. Neural networks have a hierarchical multilayered structure which sets them apart from cellular automata, so that information is transmitted not only to the immediate neighbors but also to more distant units. In artificial neural networks one can connect each unit to any other. In contrast to conventional computers, no program is handed over to the hardware – such a program has to be created, that is, the free parameters of the network have to be found adaptively.

Although neural networks and cellular automata are potentially more efficient than conventional computers in certain application areas, at the time of their conception they were not yet ready to take center stage. The necessary theory for harnessing the dynamics of complex parallel systems is still being developed right before our eyes. In the meantime, conventional computer technology has made great strides.

There is no better illustration for the simultaneous and related emergence of these various computability models than the life and work of John von Neumann himself. He participated in the definition and development of at least three of these models: in the architecture of sequential computers [417],

the theory of cellular automata and the first neural network models. He also collaborated with Church and Turing in Princeton [192].

Artificial neural networks have, as initial motivation, the structure of biological systems, and constitute an alternative computability paradigm. For that reason we will review some aspects of the way in which biological systems perform information processing. The fascination which still pervades this research field has much to do with the points of contact with the surprisingly elegant methods used by neurons in order to process information at the cellular level. Several million years of evolution have led to very sophisticated solutions to the problem of dealing with an uncertain environment. In this chapter we will discuss some elements of these strategies in order to determine what features we want to adopt in our abstract models of neural networks.

1.1.3 Elements of a computing model

What are the *elementary components* of any conceivable computing model? In the theory of general recursive functions, for example, it is possible to reduce any computable function to some composition rules and a small set of primitive functions. For a universal computer, we ask about the existence of a minimal and sufficient instruction set. For an arbitrary computing model the following metaphoric expression has been proposed:

$$\textit{computation} = \textit{storage} + \textit{transmission} + \textit{processing}.$$

The mechanical computation of a function presupposes that these three elements are present, that is, that data can be stored, communicated to the functional units of the model and transformed. It is implicitly assumed that a certain coding of the data has been agreed upon. Coding plays an important role in information processing because, as Claude Shannon showed in 1948, when noise is present information can still be transmitted without loss, if the right code with the right amount of redundancy is chosen.

Modern computers transform storage of information into a form of information transmission. Static memory chips store a bit as a circulating current until the bit is read. Turing machines store information in an infinite tape, whereas transmission is performed by the read-write head. Cellular automata store information in each cell, which at the same time is a small processor.

1.2 Networks of neurons

In biological neural networks information is stored at the contact points between different neurons, the so-called *synapses*. Later we will discuss what role these elements play for the storage, transmission, and processing of information. Other forms of storage are also known, because neurons are themselves

complex systems of self-organizing signaling. In the next few pages we cannot do justice to all this complexity, but we analyze the most salient features and, with the metaphoric expression given above in mind, we will ask: how do neurons compute?

1.2.1 Structure of the neurons

Nervous systems possess global architectures of variable complexity, but all are composed of similar building blocks, the neural cells or neurons. They can perform different functions, which in turn leads to a very variable morphology. If we analyze the human cortex under a microscope, we can find several different types of neurons. Figure 1.2 shows a diagram of a portion of the cortex. Although the neurons have very different forms, it is possible to recognize a hierarchical structure of six different layers. Each one has specific functional characteristics. Sensory signals, for example, are transmitted directly to the fourth layer and from there processing is taken over by other layers.

Fig. 1.2. A view of the human cortex [from Lassen et al. 1988]

Neurons receive signals and produce a response. The general structure of a generic neuron is shown in Figure 1.3¹. The branches to the left are the transmission channels for incoming information and are called *dendrites*. Dendrites receive the signals at the contact regions with other cells, the synapses

¹ Some animals have neurons with a very different morphology. In insects, for example, the dendrites go directly into the axon and the cell body is located far from them. The way these neurons work is nevertheless very similar to the description in this chapter.

mentioned already. Organelles in the body of the cell produce all necessary chemicals for the continuous working of the neuron. The mitochondria, visible in Figure 1.3, can be thought of as part of the energy supply of the cell, since they produce chemicals which are consumed by other cell structures. The output signals are transmitted by the *axon*, of which each cell has at most one. Some cells do not have an axon, because their task is only to set some cells in contact with others (in the retina, for example).

Fig. 1.3. A typical motor neuron [from Stevens 1988]

These four elements, dendrites, synapses, cell body, and axon, are the minimal structure we will adopt from the biological model. Artificial neurons for computing will have input channels, a cell body and an output channel. Synapses will be simulated by contact points between the cell body and input or output connections; a *weight* will be associated with these points.

1.2.2 Transmission of information

The fundamental problem of any information processing system is the transmission of information, as data storage can be transformed into a recurrent transmission of information between two points [177].

Biologists have known for more than 100 years that neurons transmit information using electrical signals. Because we are dealing with biological structures, this cannot be done by simple electronic transport as in metallic cables. Evolution arrived at another solution involving ions and semipermeable membranes.

Our body consists mainly of water, 55% of which is contained within the cells and 45% forming its environment. The cells preserve their identity and biological components by enclosing the protoplasm in a membrane made of

a double layer of molecules that form a diffusion barrier. Some salts, present in our body, dissolve in the intracellular and extracellular fluid and dissociate into negative and positive ions. Sodium chloride, for example, dissociates into positive sodium ions (Na^+) and negative chlorine ions (Cl^-). Other positive ions present in the interior or exterior of the cells are potassium (K^+) and calcium (Ca^{2+}). The membranes of the cells exhibit different degrees of permeability for each one of these ions. The permeability is determined by the number and size of pores in the membrane, the so-called *ionic channels*. These are macromolecules with forms and charges which allow only certain ions to go from one side of the cell membrane to the other. Channels are selectively permeable to sodium, potassium or calcium ions. The specific permeability of the membrane leads to different distributions of ions in the interior and the exterior of the cells and this, in turn, to the interior of neurons being negatively charged with respect to the extracellular fluid.

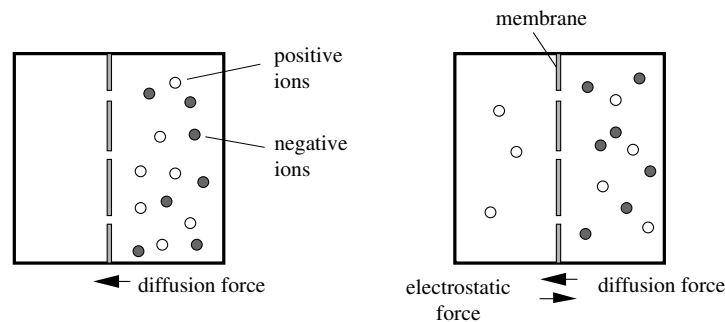


Fig. 1.4. Diffusion of ions through a membrane

Figure 1.4 illustrates this phenomenon. A box is divided into two parts separated by a membrane permeable only to positive ions. Initially the same number of positive and negative ions is located in the right side of the box. Later, some positive ions move from the right to the left through the pores in the membrane. This occurs because atoms and molecules have a thermodynamical tendency to distribute homogeneously in space by the process called diffusion. The process continues until the electrostatic repulsion from the positive ions on the left side balances the diffusion potential. A potential difference, called the *reversal potential*, is established and the system behaves like a small electric battery. In a cell, if the initial concentration of potassium ions in its interior is greater than in its exterior, positive potassium ions will diffuse through the open potassium-selective channels. If these are the only ionic channels, negative ions cannot disperse through the membrane. The interior of the cell becomes negatively charged with respect to the exterior, creating a potential difference between both sides of the membrane. This balances the

diffusion potential, and, at some point, the net flow of potassium ions through the membrane falls to zero. The system reaches a steady state. The potential difference E for one kind of ion is given by the Nernst formula

$$E = k(\ln(c_o) - \ln(c_i))$$

where c_i is the concentration inside the cell, c_o the concentration in the extracellular fluid and k is a proportionality constant [295]. For potassium ions the equilibrium potential is -80 mV.

Because there are several different concentrations of ions inside and outside of the cell, the question is, what is the potential difference which is finally reached. The exact potential in the interior of the cell depends on the mixture of concentrations. A typical cell's potential is -70 mV, which is produced mainly by the ion concentrations shown in Figure 1.5 (A^- designates negatively charged biomolecules). The two main ions in the cell are sodium and potassium. Equilibrium potential for sodium lies around 58 mV. The cell reaches a potential between -80 mV and 58 mV. The cell's equilibrium potential is nearer to the value induced by potassium, because the permeability of the membrane to potassium is greater than to sodium. There is a net outflow of potassium ions at this potential and a net inflow of sodium ions. However, the sodium ions are less mobile because fewer open channels are available. In the steady state the cell membrane experiences two currents of ions trying to reach their individual equilibrium potential. An ion pump guarantees that the concentration of ions does not change with time.

intracellular fluid (concentration in mM)		extracellular fluid (concentration in mM)	
K ⁺	125	K ⁺	5
Na ⁺	12	Na ⁺	120
Cl ⁻	5	Cl ⁻	125
A ⁻	108	A ⁻	0

Fig. 1.5. Ion concentrations inside and outside a cell

The British scientists Alan Hodgkin and Andrew Huxley were able to show that it is possible to build an electric model of the cell membrane based on very simple assumptions. The membrane behaves as a capacitor made of two isolated layers of lipids. It can be charged with positive or negative ions. The different concentrations of several classes of ions in the interior and exterior of the cell provide an energy source capable of negatively polarizing the interior of the cell. Figure 1.6 shows a diagram of the model proposed by Hodgkin and

Huxley. The specific permeability of the membrane for each class of ion can be modeled like a conductance (the reciprocal of resistance).

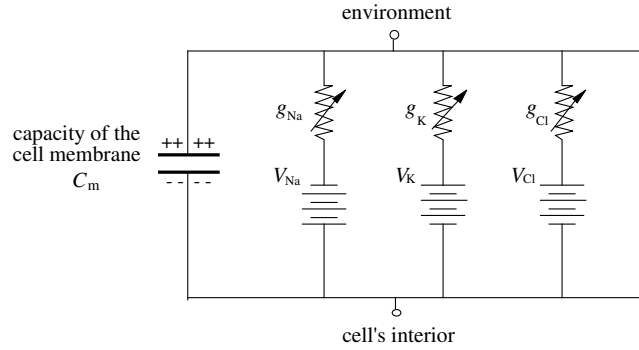


Fig. 1.6. The Hodgkin–Huxley model of a cell membrane

The electric model is a simplification, because there are other classes of ions and electrically charged proteins present in the cell. In the model, three ions compete to create a potential difference between the interior and exterior of the cell. The conductances g_{Na} , g_K , and g_L reflect the permeability of the membrane to sodium, potassium, and leakages, i.e., the number of open channels of each class. A signal can be produced by modifying the polarity of the cell through changes in the conductances g_{Na} and g_K . By making g_{Na} larger and the mobility of sodium ions greater than the mobility of potassium ions, the polarity of the cell changes from -70 mV to a positive value, nearer to the 58 mV at which sodium ions reach equilibrium. If the conductance g_K then becomes larger and g_{Na} falls back to its original value, the interior of the cell becomes negative again, overshooting in fact by going below -70 mV. To generate a signal, a mechanism for depolarizing and polarizing the cell in a controlled way is necessary.

The conductance and resistance of a cell membrane in relation to the different classes of ions depends on its permeability. This can be controlled by opening or closing excitable ionic channels. In addition to the static ionic channels already mentioned, there is another class which can be electrically controlled. These channels react to a depolarization of the cell membrane. When this happens, that is, when the potential of the interior of the cell in relation to the exterior reaches a threshold, the sodium-selective channels open automatically and positive sodium ions flow into the cell making its interior positive. This in turn leads to the opening of the potassium-selective channels and positive potassium ions flow to the exterior of the cell, restoring the original negative polarization.

Figure 1.7 shows a diagram of an electrically controlled sodium-selective channel which lets only sodium ions flow across. This effect is produced by the

small aperture in the middle of the channel which is negatively charged (at time $t = 1$). If the interior of the cell becomes positive relative to the exterior, some negative charges are displaced in the channel and this produces the opening of a gate ($t = 2$). Sodium ions flow through the channel and into the cell. After a short time the second gate is closed and the ionic channel is sealed ($t = 3$). The opening of the channel corresponds to a change of membrane conductivity as explained above.

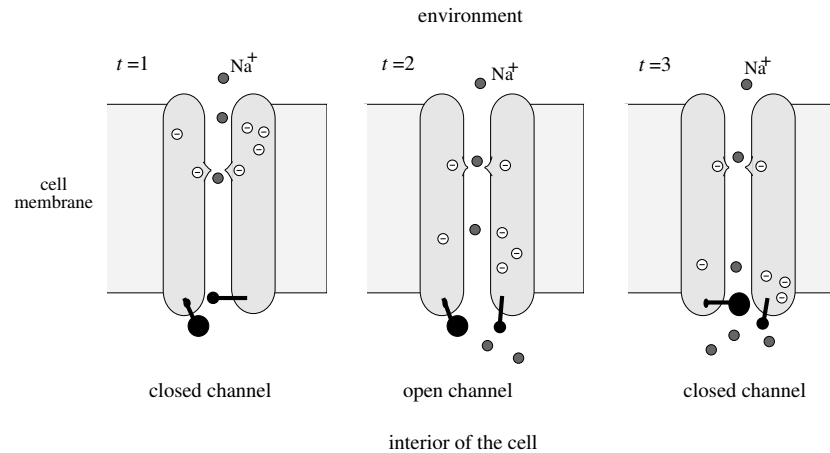


Fig. 1.7. Electrically controlled ionic channels

Static and electrically controlled ionic channels are not only found in neurons. As in any electrical system there are charge losses which have to be continuously balanced. A sodium ion pump (Figure 1.8) transports the excess of sodium ions out of the cell and, at the same time, potassium ions into its interior. The ion pump consumes adenosine triphosphate (ATP), a substance produced by the mitochondria, helping to stabilize the polarization potential of -70 mV. The ion pump is an example of a self-regulating system, because it is accelerated or decelerated by the differences in ion concentrations on both sides of the membrane. Ion pumps are constantly active and account for a considerable part of the energy requirements of the nervous system.

Neural signals are produced and transmitted at the cell membrane. The signals are represented by depolarization waves traveling through the axons in a self-regenerating manner. Figure 1.9 shows the form of such a depolarization wave, called an *action potential*. The x -dimension is shown horizontally and the diagram shows the instantaneous potential in each segment of the axon.

An action potential is produced by an initial depolarization of the cell membrane. The potential increases from -70 mV up to $+40$ mV. After some time the membrane potential becomes negative again but it overshoots, going

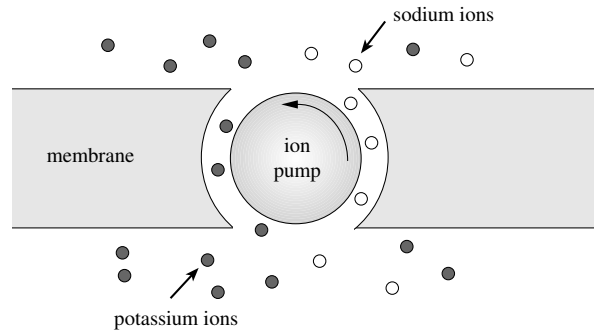


Fig. 1.8. Sodium and potassium ion pump

as low as -80 mV. The cell recovers gradually and the cell membrane returns to the initial potential. The switching time of the neurons is determined, as in any resistor-capacitor configuration, by the RC constant. In neurons, 2.4 milliseconds is a typical value for this constant.

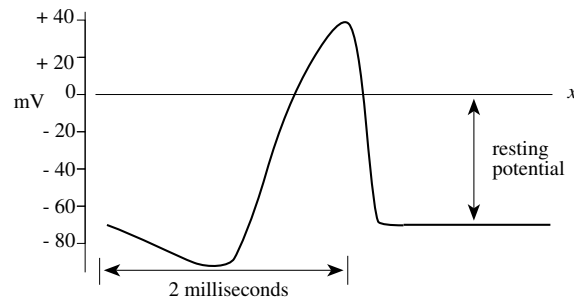


Fig. 1.9. Typical form of the action potential

Figure 1.10 shows an action potential traveling through an axon. A local perturbation, produced by the signals arriving at the dendrites, leads to the opening of the sodium-selective channels in a certain region of the cell membrane. The membrane is thus depolarized and positive sodium ions flow into the cell. After a short delay, the outward flow of potassium ions compensates the depolarization of the membrane. Both perturbations – the opening of the sodium and potassium-selective channels – are transmitted through the axon like falling dominos. In the entire process only local energy is consumed, that is, only the energy stored in the polarized membrane itself. The action potential is thus a wave of Na^+ permeability increase followed by a wave of K^+ permeability increase. It is easy to see that charged particles only move a short

distance in the direction of the perturbation, only as much as is necessary to perturb the next channels and bring the next “domino” to fall.

Figure 1.10 also shows how impulse trains are produced in the cells. After a signal is produced a new one follows. Each neural signal is an all-or-nothing self-propagating regenerative event as each signal has the same form and amplitude. At this level we can safely speak about digital transmission of information.

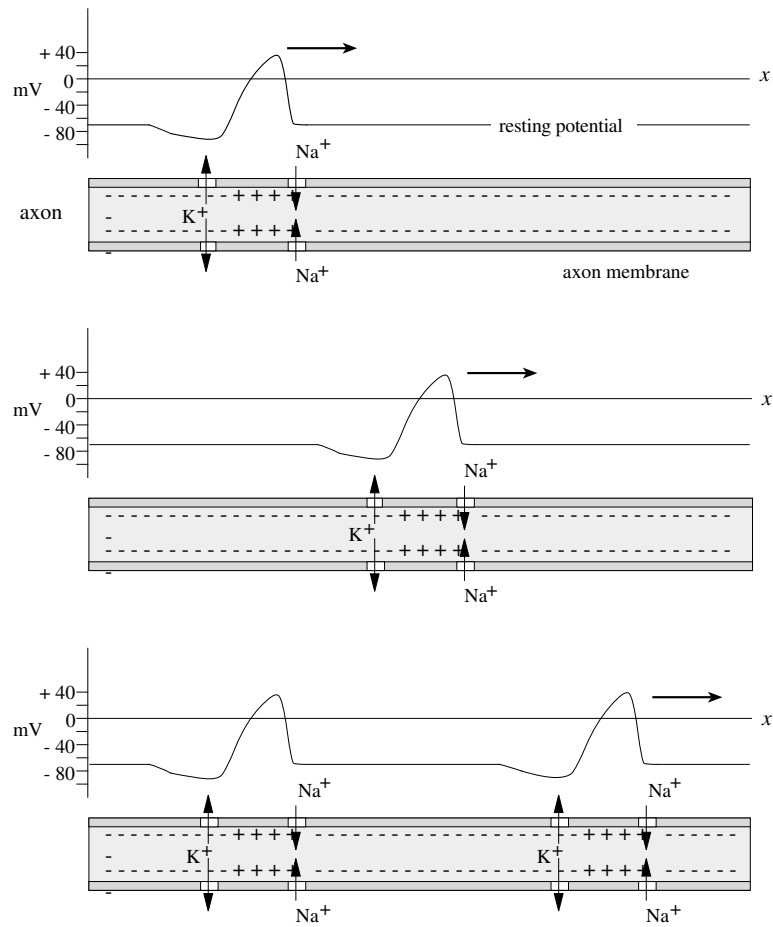


Fig. 1.10. Transmission of an action potential [Stevens 1988]

With this picture of the way an action potential is generated in mind, it is easy to understand the celebrated Hodgkin–Huxley differential equation which

describes the instantaneous variation of the cell's potential V as a function of the conductances of sodium, potassium and leakages (g_{Na} , g_K , g_L) and of the equilibrium potentials for all three groups of ions called V_{Na} , V_K and V_L with respect to the current potential:

$$\frac{dV}{dt} = \frac{1}{C_m}(I - g_{Na}(V - V_{Na}) - g_K(V - V_K) - g_L(V - V_L)). \quad (1.1)$$

In this equation C_m is the capacitance of the cell membrane. The terms $V - V_{Na}$, $V - V_K$, $V - V_L$ are the electromotive forces acting on the ions. Any variation of the conductances translates into a corresponding variation of the cell's potential V . The variations of g_{Na} and g_K are given by differential equations which describe their oscillations. The conductance of the leakages, g_L , can be taken as a constant.

A neuron codes its level of activity by adjusting the frequency of the generated impulses. This frequency is greater for a greater stimulus. In some cells the mapping from stimulus to frequency is linear in a certain interval [72]. This means that information is transmitted from cell to cell using what engineers call frequency modulation. This form of transmission helps to increase the accuracy of the signal and to minimize the energy consumption of the cells.

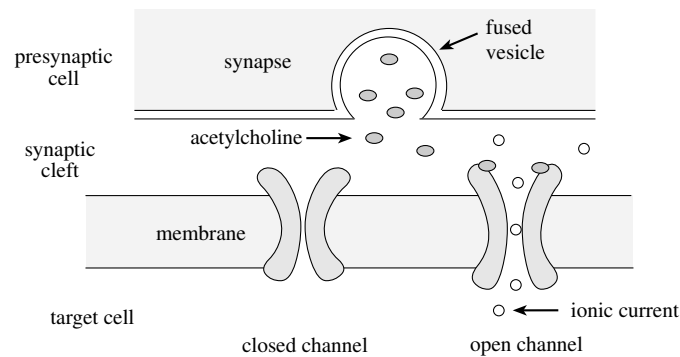
1.2.3 Information processing at the neurons and synapses

Neurons transmit information using action potentials. The processing of this information involves a combination of electrical and chemical processes, regulated for the most part at the interface between neurons, the synapses.

Neurons transmit information not only by electrical perturbations. Although electrical synapses are also known, most synapses make use of chemical signaling. Figure 1.11 is a classical diagram of a typical synapse. The synapse appears as a thickening of the axon. The small vacuoles in the interior, the synaptic vesicles, contain chemical transmitters. The small gap between a synapse and the cell to which it is attached is known as the synaptic gap.

When an electric impulse arrives at a synapse, the synaptic vesicles fuse with the cell membrane (Figure 1.12). The transmitters flow into the synaptic gap and some attach themselves to the ionic channels, as in our example. If the transmitter is of the right kind, the ionic channels are opened and more ions can now flow from the exterior to the interior of the cell. The cell's potential is altered in this way. If the potential in the interior of the cell is increased, this helps prepare an action potential and the synapse causes an excitation of the cell. If negative ions are transported into the cell, the probability of starting an action potential is decreased for some time and we are dealing with an inhibitory synapse.

Synapses determine a direction for the transmission of information. Signals flow from one cell to the other in a well-defined manner. This will be expressed in artificial neural networks models by embedding the computing elements in a

Fig. 1.11. Transversal view of a synapse [from Stevens 1988]**Fig. 1.12.** Chemical signaling at the synapse

directed graph. A well-defined direction of information flow is a basic element in every computing model, and is implemented in digital systems by using diodes and directional amplifiers.

The interplay between electrical transmission of information in the cell and chemical transmission between cells is the basis for neural information processing. Cells process information by integrating incoming signals and by reacting to inhibition. The flow of transmitters from an excitatory synapse leads to a depolarization of the attached cell. The depolarization must exceed a threshold, that is, enough ionic channels have to be opened in order to produce an action potential. This can be achieved by several pulses arriving simultaneously or within a short time interval at the cell. If the quantity of transmitters reaches a certain level and enough ionic channels are triggered,

the cell reaches its activation threshold. At this moment an action potential is generated at the axon of this cell.

In most neurons, action potentials are produced at the so-called axon hillock, the part of the axon nearest to the cell body. In this region of the cell, the number of ionic channels is larger and the cell's threshold lower [427]. The dendrites collect the electrical signals which are then transmitted electrotonically, that is through the cytoplasm [420]. The transmission of information at the dendrites makes use of additional electrical effects. Streams of ions are collected at the dendrites and brought to the axon hillock. There is spatial summation of information when signals coming from different dendrites are collected, and temporal summation when signals arriving consecutively are combined to produce a single reaction. In some neurons not only the axon hillock but also the dendrites can produce action potentials. In this case information processing at the cell is more complex than in the standard case.

It can be shown that digital signals combined in an excitatory or inhibitory way can be used to implement any desired logical function (Chap. 2). The number of computing units required can be reduced if the information is not only transmitted but also weighted. This can be achieved by multiplying the signal by a constant. Such is the kind of processing we find at the synapses. Each signal is an all-or-none event but the number of ionic channels triggered by the signal is different from synapse to synapse. It can happen that a single synapse can push a cell to fire an action potential, but other synapses can achieve this only by simultaneously exciting the cell. With each synapse i ($1 \leq i \leq n$) we can therefore associate a numerical weight w_i . If all synapses are activated at the same time, the information which will be transmitted is $w_1 + w_2 + \dots + w_n$. If this value is greater than the cell's threshold, the cell will fire a pulse.

It follows from this description that neurons process information at the membrane. The membrane regulates both transmission and processing of information. Summation of signals and comparison with a threshold is a combined effect of the membrane and the cytoplasm. If a pulse is generated, it is transmitted and the synapses set some transmitter molecules free. From this description an *abstract neuron* [72] can be modeled which contains dendrites, a cell body and an axon. The same three elements will be present in our artificial computing units.

1.2.4 Storage of information – learning

In neural networks information is stored at the synapses. Some other forms of information storage may be present, but they are either still unknown or not very well understood.

A synapse's efficiency in eliciting the depolarization of the contacted cell can be increased if more ionic channels are opened. In recent years NMDA receptors have been studied because they exhibit some properties which could help explain some forms of learning in neurons [72].

NMDA receptors are ionic channels permeable for different kinds of molecules, like sodium, calcium, or potassium ions. These channels are blocked by a magnesium ion in such a way that the permeability for sodium and calcium is low. If the cell is brought up to a certain excitation level, the ionic channels lose the magnesium ion and become unblocked. The permeability for Ca^{2+} ions increases immediately. Through the flow of calcium ions a chain of reactions is started which produces a durable change of the threshold level of the cell [420, 360]. Figure 1.13 shows a diagram of this process.

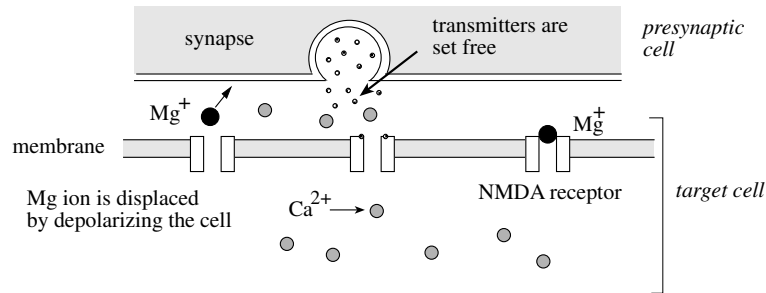


Fig. 1.13. Unblocking of an NMDA receptor

NMDA receptors are just one of the mechanisms used by neurons to increase their plasticity, i.e., their adaptability to changing circumstances. Through the modification of the membrane's permeability a cell can be trained to fire more often by setting a lower firing threshold. NMDA receptors also offer an explanation for the observed phenomenon that cells which are not stimulated to fire tend to set a higher firing threshold. The stored information must be refreshed periodically in order to maintain the optimal permeability of the cell membrane.

This kind of information storage is also used in artificial neural networks. Synaptic efficiency can be modeled as a property of the edges of the network. The networks of neurons are thus connected through edges with different transmission efficiencies. Information flowing through the edges is multiplied by a constant which reflects their efficiency. One of the most popular learning algorithms for artificial neural networks is *Hebbian learning*. The efficiency of synapses is increased any time the two cells which are connected through this synapse fire simultaneously and is decreased when the firing states of the two cells are uncorrelated. The NMDA receptors act as coincidence detectors of presynaptic and postsynaptic activity, which in turn leads to greater synaptic efficiency.

1.2.5 The neuron – a self-organizing system

The short review of the properties of biological neurons in the previous sections is necessarily incomplete and can offer only a rough description of the mechanisms and processes by which neurons deal with information. Nerve cells are very complex self-organizing systems which have evolved in the course of millions of years. How were these exquisitely fine-tuned information processing organs developed? Where do we find the evolutionary origin of consciousness?

The information processing capabilities of neurons depend essentially on the characteristics of the cell membrane. Ionic channels appeared very early in evolution to allow unicellular organisms to get some kind of feedback from the environment. Consider the case of a paramecium, a protozoan with cilia, which are hairlike processes which provide it with locomotion. A paramecium has a membrane cell with ionic channels and its normal state is one in which the interior of the cell is negative with respect to the exterior. In this state the cilia around the membrane beat rhythmically and propel the paramecium forward. If an obstacle is encountered, some ionic channels sensitive to contact open, let ions into the cell, and depolarize it. The depolarization of the cell leads in turn to a reversing of the beating direction of the cilia and the paramecium swims backward for a short time. After the cytoplasm returns to its normal state, the paramecium swims forward, changing its direction of movement. If the paramecium is touched from behind, the opening of ionic channels leads to a forward acceleration of the protozoan. In each case, the paramecium escapes its enemies [190].

From these humble origins, ionic channels in neurons have been perfected over millions of years of evolution. In the protoplasm of the cell, ionic channels are produced and replaced continually. They attach themselves to those regions of the neurons where they are needed and can move laterally in the membrane, like icebergs in the sea. The regions of increased neural sensitivity to the production of action potentials are thus changing continuously according to experience. The electrical properties of the cell membrane are not totally predetermined. They are also a result of the process by which action potentials are generated.

Consider also the interior of the neurons. The number of biochemical reaction chains and the complexity of the mechanical processes occurring in the neuron at any given time have led some authors to look for its *control system*. Stuart Hameroff, for example, has proposed that the cytoskeleton of neurons does not just perform a static mechanical function, but in some way provides the cell with feedback control. It is well known that the proteins that form the microtubules in axons coordinate to move synaptic vesicles and other materials from the cell body to the synapses. This is accomplished through a coordinated movement of the proteins, configured like a cellular automaton [173, 174].

Consequently, transmission, storage, and processing of information are performed by neurons exploiting many effects and mechanisms which we still do

not understand fully. Each individual neuron is as complex or more complex than any of our computers. For this reason, we will call the elementary components of artificial neural networks simply “computing units” and not neurons. In the mid-1980s, the PDP (*Parallel Distributed Processing*) group already agreed to this convention at the insistence of Francis Crick [95].

1.3 Artificial neural networks

The discussion in the last section is only an example of how important it is to define the primitive functions and composition rules of the computational model. If we are computing with a conventional von Neumann processor, a minimal set of machine instructions is needed in order to implement all computable functions. In the case of artificial neural networks, the primitive functions are located in the nodes of the network and the composition rules are contained implicitly in the interconnection pattern of the nodes, in the synchrony or asynchrony of the transmission of information, and in the presence or absence of cycles.

1.3.1 Networks of primitive functions

Figure 1.14 shows the structure of an abstract neuron with n inputs. Each input channel i can transmit a real value x_i . The *primitive function* f computed in the body of the abstract neuron can be selected arbitrarily. Usually the input channels have an associated weight, which means that the incoming information x_i is multiplied by the corresponding weight w_i . The transmitted information is integrated at the neuron (usually just by adding the different signals) and the primitive function is then evaluated.

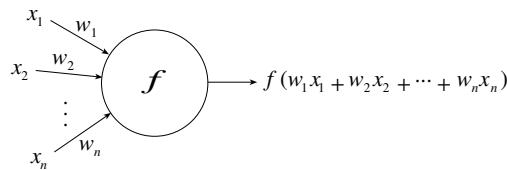


Fig. 1.14. An abstract neuron

If we conceive of each node in an artificial neural network as a primitive function capable of transforming its input in a precisely defined output, then artificial neural networks are nothing but *networks of primitive functions*. Different models of artificial neural networks differ mainly in the assumptions about the primitive functions used, the interconnection pattern, and the timing of the transmission of information.

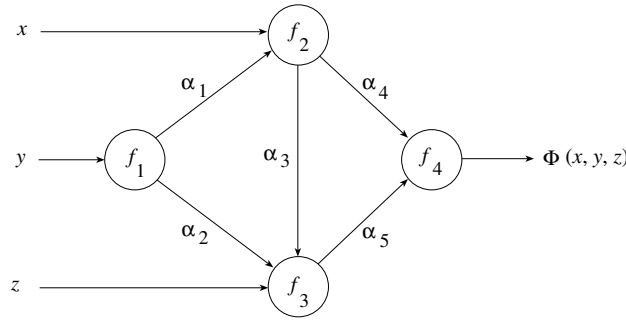


Fig. 1.15. Functional model of an artificial neural network

Typical artificial neural networks have the structure shown in Figure 1.15. The network can be thought of as a function Φ which is evaluated at the point (x, y, z) . The nodes implement the primitive functions f_1, f_2, f_3, f_4 which are combined to produce Φ . The function Φ implemented by a neural network will be called the *network function*. Different selections of the weights $\alpha_1, \alpha_2, \dots, \alpha_5$ produce different network functions. Therefore, tree elements are particularly important in any model of artificial neural networks:

- the structure of the nodes,
- the topology of the network,
- the learning algorithm used to find the weights of the network.

To emphasize our view of neural networks as networks of functions, the next section gives a short preview of some of the topics covered later in the book.

1.3.2 Approximation of functions

An old problem in approximation theory is to reproduce a given function $F : \mathbb{R} \rightarrow \mathbb{R}$ either exactly or approximately by evaluating a given set of primitive functions. A classical example is the approximation of one-dimensional functions using polynomials or Fourier series. The Taylor series for a function F which is being approximated around the point x_0 is

$$F(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \dots + a_n(x - x_0)^n + \dots,$$

whereby the constants a_0, \dots, a_n depend on the function F and its derivatives at x_0 . Figure 1.16 shows how the polynomial approximation can be represented as a network of functions. The primitive functions $z \mapsto 1, z \mapsto z^1, \dots, z \mapsto z^n$ are computed at the nodes. The only free parameters are the constants a_0, \dots, a_n . The output node additively collects all incoming information and produces the value of the evaluated polynomial. The weights of the network can be calculated in this case analytically, just by computing the first $n + 1$

terms of the Taylor series of F . They can also be computed using a learning algorithm, which is the usual case in the field of artificial neural networks.

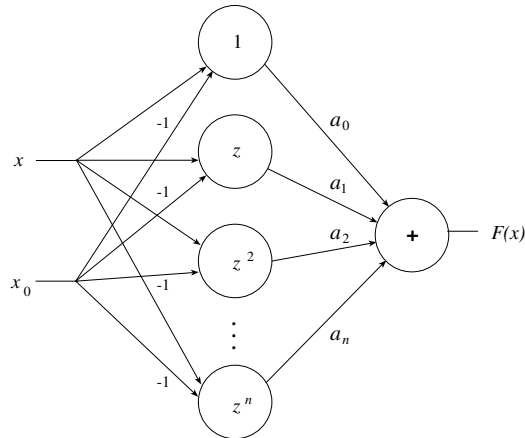


Fig. 1.16. A Taylor network

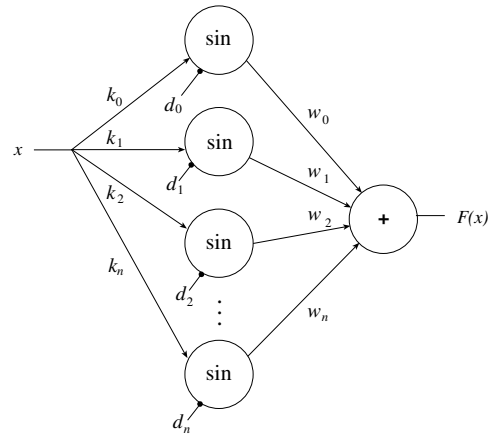


Fig. 1.17. A Fourier network

Figure 1.17 shows how a Fourier series can be implemented as a neural network. If the function F is to be developed as a Fourier series it has the form

$$F(x) = \sum_{i=0}^{\infty} (a_i \cos(ix) + b_i \sin(ix)). \quad (1.2)$$

An artificial neural network with the sine as primitive function can implement a finite number of terms in the above expression. In Figure 1.17 the constants k_0, \dots, k_n determine the wave numbers for the arguments of the sine functions. The constants d_0, \dots, d_n play the role of phase factors (with $d_0 = \pi/2$, for example, we have $\sin(x + d_0) = \cos(x)$) and we do not need to implement the cosine explicitly in the network. The constants w_0, \dots, w_n are the amplitudes of the Fourier terms. The network is indeed more general than the conventional formula because non-integer wave numbers are allowed as are phase factors which are not simple integer multiples of $\pi/2$.

The main difference between Taylor or Fourier series and artificial neural networks is, however, that the function F to be approximated is given not explicitly but implicitly through a set of input-output examples. We know F only at some points but we want to generalize as well as possible. This means that we try to adjust the parameters of the network in an optimal manner to reflect the information known and to extrapolate to new input patterns which will be shown to the network afterwards. This is the task of the *learning algorithm* used to adjust the network's parameters.

These two simple examples show that neural networks can be used as universal function approximators, that is, as computing models capable of approximating a given set of functions (usually the integrable functions). We will come back to this problem in Chap. 10.

1.3.3 Caveat

At this point we must issue a warning to the reader: in the theory of artificial neural networks we do not consider the whole complexity of real biological neurons. We only abstract some general principles and content ourselves with different levels of detail when simulating neural ensembles. The general approach is to conceive each neuron as a primitive function producing numerical results at some points in time. These will be the kinds of model that we will discuss in the first chapters of this book. However we can also think of artificial neurons as computing units which produce pulse trains in the way that biological neurons do. We can then simulate this behavior and look at the output of simple networks. This kind of approach, although more closely related to the biological paradigm, is still a very rough approximation of the biological processes. We will deal with asynchronous and spiking neurons in later chapters.

1.4 Historical and bibliographical remarks

Philosophical reflection on consciousness and the organ in which it could possibly be localized spans a period of more than two thousand years. Greek philosophers were among the first to speculate about the location of the

soul. Several theories were held by the various philosophical schools of ancient times. Galenus, for example, identified nerve impulses with pneumatic pressure signals and conceived the nervous system as a pneumatic machine. Several centuries later Newton speculated that nerves transmitted oscillations of the ether.

Our present knowledge of the structure and physiology of neurons is the result of 100 years of special research in this field. The facts presented in this chapter were discovered between 1850 and 1950, with the exception of the NMDA receptors which were studied mainly in the last decade. The electrical nature of nerve impulses was postulated around 1850 by Emil du Bois-Reymond and Hermann von Helmholtz. The latter was able to measure the velocity of nerve impulses and showed that it was not as fast as was previously thought. Signals can be transmitted in both directions of an axon, but around 1901 Santiago Ramón y Cajal postulated that the specific networking of the nervous cells determines a direction for the transmission of information. This discovery made it clear that the coupling of the neurons constitutes a hierarchical system.

Ramón y Cajal was also the most celebrated advocate of the *neuron theory*. His supporters conceived the brain as a highly differentiated hierarchical organ, while the supporters of the reticular theory thought of the brain as a grid of undifferentiated axons and of dendrites as organs for the nutrition of the cell [357]. Ramón y Cajal perfected Golgi's staining method and published the best diagrams of neurons of his time, so good indeed that they are still in use. The word *neuron* (Greek for *nerve*) was proposed by the Berlin Professor Wilhelm Waldeger after he saw the preparations of Ramón y Cajal [418].

The chemical transmission of information at the synapses was studied from 1920 to 1940. From 1949 to 1956, Hodgkin and Huxley explained the mechanism by which depolarization waves are produced in the cell membrane. By experimenting with the giant axon of the squid they measured and explained the exchange of ions through the cell membrane, which in time led to the now famous Hodgkin–Huxley differential equations. For a mathematical treatment of this system of equations see [97].

The Hodgkin–Huxley model was in some ways one of the first artificial neural models, because the postulated dynamics of the nerve impulses could be simulated with simple electric networks [303]. At the same time the mathematical properties of artificial neural networks were being studied by researchers like Warren McCulloch, Walter Pitts, and John von Neumann. Ever since that time, research in the neurobiological field has progressed in close collaboration with the mathematics and computer science community.

Exercises

1. Express the network function Φ in Figure 1.15 in terms of the primitive functions f_1, \dots, f_4 and of the weights $\alpha_1, \dots, \alpha_5$.

2. Modify the network of Figure 1.17 so that it corresponds to a finite number of addition terms of equation (1.2).
3. Look in a neurobiology book for the full set of differential equations of the Hodgkin–Huxley model. Write a computer program that simulates an action potential.

