Memory in Neural Networks and Glasses
Heerema, M.

Citation for published version (APA):
1 Introduction to neural networks

It is a major event that occurs every four years. The icon associated with it are five rings attached to each other in a specific manner. If, moreover, the fact is added that the origin of this event is to be found in Greece, one will have thought almost certainly of the Olympic Games. Of those persons which recognized the event via these clues it is said that they were able to recall some information that was stored in their memory.

Not everybody is interested in sports and therefore some hear about these facts for the first time. Nevertheless, if those persons were asked, after having read the first paragraph of this introduction, to say something about the Olympic Games they probably could. Those people have learned in the preceding paragraph some information, which they were able to recall in the current paragraph. With this little experiment two processes are illustrated which will be studied in this thesis: learning and recall of information. In the brain many more is to be found and has been found.

Although a research topic for centuries, the brain is up to now for a large part not understood. This fact has focused the attention of more and more people on the brain. Moreover, many human skills tingle our imagination. Among them are the human experience of perception, memory and emotion. As a consequence, brain research is expanding, and nowadays many scientists educated in all kinds of fields try to unravel the secrets of the brain. Of the many questions one can raise, this thesis is restricted to the question how the human brain might process information. This will be done with the limited tools of a theoretical physicist: a pencil, a computer, some paper and a library. Nevertheless, the particular methods, questions and results of a theoretical physicist are hoped to influence, stimulate and enrich the thoughts of researchers from other disciplines.

The diversity of the scientific groups involved in brain research allows the brain to be studied in many different ways. For instance, psychologists investigate the brain as a black box via stimulus-response measurements. Famous is the experiment of Pavlov in which a dog starts to salivate in response to an expected stimulus, i.e., food. A drawback is that the processes which take place in the brain and yield this behaviour remain unclear. This gap is filled by a tremendous amount of diverse studies on the interior of the brain. Examples are the study of the anatomy of the brain via brain imaging methods like functional MRI. Moreover, on a smaller length scale, the electric current flow through membranes of nerve-cells is investigated via patch-clamp techniques. On a molecular level, the method of polymer chain reaction (PCR) has initiated a lot of publications ever since its introduction.

Despite these and many other techniques, it appears to be very difficult to study the influence of the cellular changes at an experimental level. In other words, it is difficult to demonstrate that a change at the cellular level implies a change in behaviour of a human being. This is an important issue since, roughly speaking, the current theories state that this is what learning is about. I will come back to learning in more detail later on, here the message is that a problem is encountered that sounds familiar to a physicist. How to explain a macroscopic phenomenon by means of a microscopic description? In physics this is the domain of statistical physics, commonly applied to the study of gases, liquids or solids. Important in this field is to
ignore those individual properties of the atoms which are considered to be irrelevant for the macroscopic behaviour. Hence, it is advocated that simplified models are sufficient to capture the essentials of a problem. This will also be my attitude when investigating the brain in this thesis. In this context, the nerve-cells, or neurons, play the part of the microscopic entities of statistical physics, while the macroscopic phenomenon associated is with that of a brain process like learning. Simplified models of a part of the brain, commonly referred to as a neural network, will be studied. This will be done both analytically and numerically.

As said before, learning is considered to occur at a cellular level. In particular, it is a change in strength of the synaptical connections between nerve-cells (see below). In neural networks these local changes are modeled via so-called learning rules. The main goal of neural network studies is to find learning rules which are devised in such a manner that the network as a whole operates in some desired fashion. This is a non-trivial requirement. Nevertheless, microscopic learning rules have been found to generate macroscopic behaviour. This fact emphasizes the power and the potential of the use of neural networks models in brain research.

In the spirit of statistical physics, the first neural network models were highly simplified with respect to biology. The current research on neural networks is involved with the addition of more and more biological relevant properties to neural network models. This direction of the ongoing research is reflected in this thesis.

Up to now I have said some words on the relation of neural networks to brain research. However, some neural network models have properties which allow the application in various industrial environments. An example in this context is pattern recognition. Also robotics flourished due to the application of neural networks. Many more can be said about this, but since this thesis is involved with brain research I leave it as a side-remark.

In conclusion, the study of neural networks can be a supplemental tool in the goal to understand the brain. Moreover, it provides a means to investigate the brain not only experimentally, but also analytically and computationally. In physics, the interaction between theory and experiment has proved over and over again to be a rich source of scientific achievements. Hopefully, the theory of neural networks can contribute to such a development in the field of brain research.

In the remainder of this introductory chapter I will first sketch the biochemical organization of the brain, starting with its building block, the nerve cell. Then the neuron-model and the type of neural networks I will use throughout this thesis will be discussed. Also a short description how to look at learning and retrieval in the chosen neural network model will be given. The introductory part is closed with an expose on various contributions to the theory of (attractor) neural networks. Finally, an outline of the research topics of the thesis will be presented.

1.1 Nerve-cells

In this section I will describe in short the structure of a nerve-cell. I do not intend to do this thoroughly, since this topic is worth a book on itself due to its complexity. Rather, I mention those aspects that are necessary for the simplified neuron models I will introduce in the following section.

The nerve-cell, as any other cell, is a complicated thing (see figure 1.1 for a schematic picture). Many types do exist, but basically a nerve-cell consists of a tree like structure, called
the dendrites, which are in contact with the central element of the nerve-cell, the soma. The soma, in turn, is connected to a single offshoot, the axon, which branches at its end.

![Diagram of a nerve-cell](image.jpg)

Figure 1.1: A schematic picture of a nerve-cell, reprinted from [1].

Nerve cells are, mostly, connected in a biochemical way via so-called synaptic junctions. Such a junction is a location where an axon of some nerve-cell, called pre-synaptic nerve-cell, almost touches at a part of the dendrites or the soma of any other nerve-cell, referred to as post-synaptic nerve-cell. To be more precise, on the arrival of a signal at the pre-synaptic side of the cleft between two nerve-cells, a chemical substance is released. These chemical substances, also called neuro-transmitters, reach the opposite side of the cleft by diffusion. The post-synaptic nerve-cell capture the neuro-transmitters via receptors. Then a complex biochemical process leads to either an enhancement or a reduction of the potential at the soma, the post-synaptic potential. In the former case one speaks of an excitatory effect and in the latter of an inhibitory effect.

Also the production of an axonal signal, called an action-potential or a spike, is induced by a complicated biochemical process. In short, a spike is a temporary disturbance of the membrane-potential which propagates through the axon. This process is initiated if via incoming signals a certain critical value of the membrane-potential at the soma is surpassed.

For a more detailed description of all these biochemical processes see, e.g., [2].

1.2 Neuron models

The nerve-cell (neuron), can be regarded as an information processing device, which receives input in the dendrites, integrates this input in the soma and responses to other neurons via its axon. An important issue is how the information on the environment is encoded in neuronal signals. In other words, what kind of signals indicate the recall of the words ‘Olympic Games’?

First of all, one might try to relate the activity of a single neuron to such a recall. It is reported that neurons do exist which respond to specific stimuli, such as neurons that respond only if horizontal (vertical) lines are present in the range of vision. This led to the concept of the ‘grandmother’ cell code, which refers to the famous example of a brain in which the view of a grandmother only triggers a specific group of nerve-cells to emit spikes. Nevertheless, this concept has long been abandoned as the general mechanism in which the brain operates. Nowadays, it is thought that information is stored via ensembles of neurons rather than via single neurons.
A next issue is whether a single spike or a group of spikes between neurons carry information. In favour of the latter case, there is experimental evidence that in sensory neurons the mean firing rate, the time-averaged rate at which a neuron produces signals, is directly related to the stimulus intensity. However, in real life, situations do occur which demand a fast reaction of the brain. Then no time is left for temporal averaging and single spikes should be important.

The issue how to interpret the temporal structure of neural activity is an important issue, which have led to many models, each with its specific solution. Examples are oscillator models like the Kuramoto model [3], the Hodgkin-Huxley model [4] and the integrate-and-fire model [5, 6, 7, 8]. A review can be found in [9, 10].

In this thesis learning and retrieval will be considered to be phenomena generated by the spiking behaviour of a group of neurons. The next section will deal with the way in which neurons are grouped together. Here single neurons will be the topic.

To model the complicated nerve-cells of the preceding section I have chosen the neuron model of McCulloch and Pitts [11], see figure 1.2. In this model, the spiking behaviour of a nerve-cell is described via a two state variable $x$ for the state of a neuron. That is, a neuron $i$ is considered to have a state $x_i = 1$ if it produces one action-potential, or spike, and a state $x_i = 0$ if the neuron is quiescent.

![Figure 1.2: The neuron model of McCulloch and Pitts.](image)

The post-synaptic potential of a neuron $i$ is the result of an integration over all inputs from the pre-synaptic neurons $j$. The integration can be done in various manners. The most simple integration is that of linearly adding the inputs albeit weighted for their influence on the post-synaptic potential. This difference in influence can be due, among other things, to a variable amount of neuro-transmitters that is released in the synaptic cleft or to a variable amount of receptors that is present at the post-synaptic side of the cleft. In formula, the
1.3 Brain models: neural networks

The post-synaptic potential \( h_i(t) \) of neuron \( i \) at time \( t \) as a result of a linear integration reads

\[
h_i(t) = \sum_{j=1}^{N} w_{ij}(t)x_j(t),
\]

where the \( x_j(t) \) are the input signals at time \( t \) and where the \( w_{ij}(t) \) are the weights, also called synaptic efficacies or synaptic strengths at time \( t \). A synaptic weight \( w_{ij} \) takes into account the overall effect of a synaptic connection between a post-synaptic neuron \( i \) and a pre-synaptic neuron \( j \). It may be positive (negative) in case the synaptic connection is excitatory (inhibitory) or zero in case there is no connection between the two neurons.

In case no signal is received by a nerve-cell its soma has a constant resting potential. Input signals lead to an action-potential if the membrane-potential is increased above a certain level. A typical value for this increase is 10 mV [2]. This can be translated into the model saying that in order to generate a spike the post-synaptic potential \( h_i \) has to exceed a constant threshold \( \theta_i \).

The decision to generate a spike has to be given by a dynamical rule. Taking time to be discrete, the output of neuron \( i \) is generated via

\[
x_i(t + \Delta t) = \Theta_H(h_i(t) - \theta_i), \quad (i = 1, \ldots, N),
\]

where \( \Delta t \) is some discrete time step. The symbol \( \Theta_H \) stands for the Heaviside step function, which equals one for positive arguments and vanishes otherwise. Remark that this rule is deterministic although stochastic versions do exist. Moreover, I assume all spikes to contribute as a whole to a post-synaptic potential at a given time. In reality, however, the shape of the synapses and the dendritic tree yield a spike generated by some pre-synaptic neuron at a given time to arrive (via dispersion) in parts at the soma in a few consecutive time-steps and, hence, to contribute in parts to a few post-synaptic potentials consecutive in time.

In this basic model other properties of a nerve-cell can be introduced. For instance, instead of a threshold a refractory function can be defined to model more accurately the reduced excitability of a nerve cell after a spike has been emitted. Another aspect that is neglected in the model of McCulloch and Pitts are axonal delay times and the time propagation time of a signal through the dendritic tree. Inclusion of one or more of these properties yields the spike-response-model [12] or other models [13, 14].
In general, two neurons $A$ and $B$ can be attached to each other in two distinct ways. First, only in one direction, from $A$ to $B$ or, equivalently, from $B$ to $A$ in which case one speaks of a feed-forward neural network of two neurons. The other possibility is that of attaching neuron $A$ to $B$ and neuron $B$ to $A$, the case of a recurrent neural network of two neurons.

In this thesis I will model long-term memory in a simplified manner. First of all, the pathways will be discarded which lead from the sense-organs to the nerve-cells directly involved in the memory process. Instead of a pathway, some input signal to a network of neurons will be defined to represent the information received by the sensory neurons. Moreover, since long-term memory is, primarily, associated with the cortex, a single network of (cortical) neurons will be studied.

Roughly speaking, the cortex is a thin layer with a columnar structure. Within each column neurons are connected densely, i.e., there is a high connectivity, whereas neurons in different columns are connected sparsely. As a consequence, often fully connected recurrent networks are chosen as a model for the memory processes of the cortex. With the words fully connected it is meant that each neuron is attached to all other neurons of the network. To model a column of the cortex in more detail, however, a recurrent network of partially connected neurons has to be used, as I will do below.

A point that still has to be considered is the manner in which the dynamical rule (1.2) is applied to a network of neurons. First of all the rule can be used to update all of the neurons of the network simultaneously. This type of dynamics is called parallel dynamics [15]. The other type of dynamics, sequential dynamics, updates neurons one at a time, either in a fixed order or randomly [16]. Both parallel as sequential updating are approximations to the realistic situation in which nerve-cells produce spikes neither simultaneously nor sequentially. Instead, nerve-cells produce spikes if the post-synaptic potential surpasses the threshold potential, a fact which yield a time-interval between spikes which may vary. Nevertheless, to simplify the problem and to mimic the behaviour of collective oscillations (see section 1.5), I will use parallel dynamics, and, hence, a fixed time-interval.

### 1.4 Learning

In modeling the brain we now have a partially connected recurrent network of McCulloch and Pitts neurons at our disposal. To study this network, however, one first has to know the values of the synaptic connections $w_{ij}$. In a living brain, the efficacy of signal transmission via the synaptic cleft changes in time. This can be due to, among many other factors, a decrease or increase in the release of neuro-transmitters at the pre-synaptic side of a synapse or a decrease or increase in the amount of receptors at the post-synaptic side of a synapse. Mathematically, the change of the strength of the connections can in both cases be written as

$$w_{ij}(t + \Delta t) = w_{ij}(t) + \Delta w_{ij}(t) \tag{1.3}$$

where $\Delta w_{ij}(t)$ is the change in the efficacy at time $t$. In the context of storage of information, an explicit expression for the change $\Delta w_{ij}$ is said to be a learning rule. This brings me to one of the main points of my thesis: how is information learned.

In the forties, the Canadian psychologist Hebb conjectured in his now famous book *The organization of behavior—A neuro-physiological Theory* [17] that the changes of the connections
between the neurons take place according to a ‘neuro-physiological postulate’ that nowadays is referred to as Hebb’s rule: ‘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 so that A’s efficiency, as one of the cells firing B, is increased’. Thus Hebb’s rule is a qualitative statement on the enhancement of synaptic efficiency of signal transmission, but does not state quantitatively, by some mathematical formula, to what extent.

Based on his statement a synapse is defined to be Hebbian if it uses a mechanism to increase the synaptic efficacy that is time-dependent, local and interactive. With interactive I mean that the activities of the pre-synaptic neuron as well as the post-synaptic neuron are to be taken into account.

The statement of Hebb does not imply the conditions for a decrease of the synaptic efficacy. This gap has led to the concept of a generalized Hebbian synaptic mechanism in which both synaptic enhancement as well as depression is present as some function of the activities [18, 19].

The key issue remains to verify the statement of Hebb in an experimental setup. An important phenomenon found experimentally which supports the existence of Hebbian synapses is that of long-term-potentiation (LTP), the long-lasting enhancement of the excitability of a neuron due to repeated stimulation. The first experiments showing LTP were performed with hippocampal neurons [20], but also for neocortical neurons LTP has been shown to exist [21].

The experiments mentioned above were done both in vitro (brain slices) as in vivo. Although there are indications that synapses do exist that behave in a Hebbian way and yield a change in behaviour (learning) [22], a full experimental proof lacks up to now. Moreover, the complexity of the processes involved may be a barrier to high to show the influence of Hebbian synapses on mental processes experimentally.

This emphasizes the importance of theoretical studies on Hebbian synapses. In chapter 2 the statement of Hebb will be rephrased mathematically for the model chosen in this thesis. It will be shown that the statement of Hebb allows for a number of different Hebbian learning rules. Moreover, it will be studied which of the Hebbian learning rules imply storage of information.

### 1.5 Retrieval

In order to study the retrieval of information I have to specify in more detail the question I asked before how the information on the environment is encoded in a neuronal signal. To do so, I now introduce the states $x_i(t)$ of all neurons $i$ ($i = 1, \ldots, N$) of a net, at a given time $t$, as the components of a state-vector $X(t)$ of this neural net, $X(t) := (x_1(t), \ldots, x_N(t))$. Moreover, I define a space, or state-space, as the space spanned by all possible state-vectors $X(t)$. Hence, any state-vector $X(t)$ is represented in state-space as a single point.

Due to the changes in time in the states of the neurons, the state-vector $X(t)$ also changes in time. As a consequence, all state-vectors will follow some trajectory in state-space which can be linked to some mental process like the process of recognition. Therefore, I will proceed by investigating which points sets, or trajectories, in state-space do exist and how they can be interpreted. The dynamics can lead to three type of asymptotic states: i. chaotic, ii. limit-cycles and iii. fixed points. In the cases of chaotic dynamics trajectories wander in an uncorrelated way in state-space, whereas in case of limit-cycles trajectories lead, rapidly, to a
cycle of states. In both situations the dynamics in state-space is sensitive to the initial state-vector, although in the case of limit-cycles this also depends upon the dynamical rule being used. In contrast, point sets which lead to some single state-vector and, moreover, remain stuck at this state, a fixed point, can be rather insensitive to initial state-vectors.

At this point it is important to discuss in more detail the properties related to retrieval of information. From daily experience it is known that objects can be recognized under all kinds of conditions. For instance, this thesis can be viewed and recognized from various distances and from different angles. Also small perturbations as a pencil or a small piece of paper placed on the thesis will not prevent the book to be recognized. In all these situations, the input signal to the brain will be quite different, but the response is identical. This can be translated as a requirement that a neural model, for the retrieval of information, has to be rather insensitive in its response to the input of the neural system. Consequently, trajectories which lead to fixed points have been studied extensively in the context of memory. Also in this thesis fixed points will be the manner in which the environment is encoded in neural signals.

Since a fixed point is some specific network state, a pattern, it is denoted as a vector with components the specific states of the neurons of the net, i.e., \( \xi = (\xi_1, \ldots, \xi_n) \). It has to be emphasized that for the desired behaviour each fixed point has to have a so-called basin of attraction. Roughly speaking, a basin of attraction of a fixed point is defined to be the set of input signals that evolve under the influence of a given dynamical rule to the same fixed point.

In spite of these theoretical arguments, it still is an open discussion whether neural signals are encoded via trajectories which lead to fixed points or via a more complicated dynamics. Namely, there is experimental evidence in favour of the use of fixed points [23] and experimental results do exist that challenge this [24]. Therefore, also trajectories to limit-cycles [25, 26, 27, 28] and chaotic trajectories [29, 30, 31] have been studied.

Given a dynamical rule, e.g., eq. (1.2), it depends upon the synaptic weights \( w_{ij} \) what trajectory in state-space the states of the neurons in a network will follow. In other words, the weights \( w_{ij} \) determine the behaviour of the brain. Hence, the driving force behind the values of the weights, the learning rule, has to be such that fixed points with basins of attractions are created. This will indeed be an important issue in chapters 2 and 4. In chapter 3 the size of the basins of attraction will be the topic.

Neural networks have two time scales, one related to the rate of change of the synaptic efficacies \( w_{ij} \) and one related to the spiking activity of a neuron. The latter time is of the order of milliseconds, the former is less well defined, but can be estimated to lie somewhere between seconds and days: it is a time related to the rate of learning of a brain. Hence, the \( \Delta t \) occurring in equation (1.2) is of the order of milliseconds. The topic of dynamic synapses and neurons is often referred to as double-dynamics and will be studied in chapter 4.

The arrival at a fixed point, using parallel dynamics, can be regarded as a collective oscillation with a frequency given by \( 1/\Delta t \). Taking the characteristic time-step to be one to two milli-seconds, we have in this fashion oscillations of about 500-1000 Hz, which is well above the measured values of 40-50 Hz [32, 33]. The more sophisticated spike-response-model, however, yields results which are in agreement with these experiments [12].

Neural networks based on fixed points are also called attractor neural networks. A good review on this type of networks is presented in [34]. Also books exist which treat neural networks in a more general context [35, 36, 37]. In the next section some major contributions to the study of attractor neural networks will be presented.
1.6 Attractor neural networks

In this section I will go into some detail on the current status of the ongoing theoretical research on attractor neural networks. A development which proved to be very fruitful began when Little pointed out the similarity between recurrent neural networks of McCulloch and Pitts neurons at the one hand and Ising models, models which describe ferromagnetism [38], on the other hand. An Ising system is a lattice model with on each lattice site $i$, a variable $s_i$ which can have the values $s_i = 1$ (spin up) and $s_i = -1$ (spin down). The analogy to a neural network is realized by identifying each spin with a neuron and associating the upward orientation $s_i = 1$ with the active state $x_i = 1$ and the downward orientation $s_i = -1$ with the in-active state $x_i = 0$. This allowed for two representations for the state of a neuron: i. the 'binary' representation, $x_i = 1$ or $x_i = 0$, and ii. the 'bipolar' or 'spin-representation', $s_i = 1$ or $s_i = -1$.

In the spin-representation, the dynamical equation (1.2) can be rewritten as

$$s_i(t + \Delta t) = \text{sgn}(\sum_{j=1}^{N} J_{ij}(t)s_j(t) - T_i(t)), \quad (i = 1, \ldots, N),$$

where the time dependent ‘coupling constants’ $J_{ij}$ are related to the biological weights $w_{ij}$ through $J_{ij} = w_{ij}/2$ and where $s_j = 2x_j - 1$. The time dependent ‘thresholds’ $T_i(t)$ are related to the constant biological thresholds $\theta_i$ according to

$$T_i(t) = \theta_i - \sum_{j=1}^{N} J_{ij}(t), \quad (i = 1, \ldots, N).$$

In the literature the thresholds $T_i(t)$ are usually treated as a constant; most often the constant is taken to vanish [39, 40, 41]. This seemingly innocent fact changes, of course, the dynamics (1.4) of the system in a non-trivial way. As a consequence, the results obtained for, e.g., the adaptation of the coupling constants differ from those obtained when the actual biological dynamics (2.3) is used [see [41] and cf. chapter 2 eqs. (2.41) and (2.42)]. Hence, when modeling adaptation processes of biological neurons with constant thresholds, the use of the binary-representation is obligatory. I therefore proceed in this thesis with the binary-representation, although an identical, but less transparent, reformulation in spin-representation is possible.

The introduction of the spin-representation also allowed the mathematical tools of thermodynamics to be applied to neural networks, in particular the tools developed in the study of spin-glasses. This point will be treated in more detail below and in chapter 5.

A well-known model which incorporates storage of information (via Hebb’s rule) and retrieval (via fixed-points) is that of Hopfield [16]. In this model the synaptic weights $w_{ij}$ between two neurons $i$ and $j$ are given by

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} (2\xi_i^\mu - 1)(2\xi_j^\mu - 1),$$

where $N$ are the number of neurons and $p$ is the number of stored patterns $\xi$. The Hopfield model regarded as a spin-glass system was solved by Amit et. al. using the replica approach [42]. They conclude that the maximal amount of patterns $p_{\text{max}}$ that can be stored in a network of $N$ neurons, defined via the maximal storage capacity $\alpha_c = p_{\text{max}}/N$, is $\alpha_c = 0.138$. 

17
Also within the context of spin-glasses it has been shown by Gardner that the maximal storage capacity cannot exceed the value $\alpha_c = 2$ \cite{43}. This yields a large gap between the patterns that can be stored in the Hopfield model and the model which can store the most patterns. A (mathematical) method to obtain analytical expressions for the weights of a neural network with a storage capacity of $\alpha_c = 1$ is that of the pseudo-inverse \cite{44, 45}. In addition a learning procedure has been found to obtain the values of these connections \cite{46}.

Although often studied for non-biological purposes, feed-forward (layered) neural networks, or in short perceptrons \cite{47}, have been of influence to the current understanding of attractor neural networks. Especially, the perceptron learning rule \cite{48} and the concept of error-back-propagation \cite{49} have been applied to attractor neural networks. For the interested reader a review on perceptron learning is presented in \cite{50}.

In the course of time more and more biological details were taken into account in neural networks with respect to the original simple models. Important to mention is the development of sophisticated neuron models by including all kinds of physiological and anatomical data (see section 1.2 and below).

A first aspect to be considered is that, in general, connections are asymmetric. This is due to the fact that a connection from neuron $A$ to neuron $B$ is independent from the connection of neuron $B$ to neuron $A$ \cite{51, 52, 53, 54}. Another aspect is that the connection of neuron $A$ to neuron $A$ itself, the so-called self-interaction, is found to be absent \cite{39}.

Moreover, both fully connected as partially connected —or diluted— recurrent neural networks have been studied widely. In the case of diluted networks the distinction between dilution before \cite{55, 56} and after \cite{57} learning is made.

A physiological phenomenon is the fact that nerve-cells are either excitatory or inhibitory \cite{58}. In a network environment this is expressed as Dale’s law. In some models the learning rules are formulated such that they comply with Dale’s law \cite{59, 60}. I think of Dale’s law as a constraint on the values of the weights set by nature and not a constraint on the learning rule. Nevertheless, for simplicity I neglect in my studies this phenomenon. As a consequence, the inclusion of Dale’s law may alter some of my results.

A final issue is the fact that, in nature, the connections do have an upper and a lower bound. Although this bound is considered explicitly in some models \cite{61, 62} and is reported to be present implicitly in the model \cite{8}, it is a property often neglected.

This introduction is far too short to give a complete view on the problems encountered and solutions achieved in the field of neural networks. Rather it has to be seen as a crash course in neural networks. To guide the reader to more extensive studies I have added on the appropriate places references to books and review articles.

### 1.7 Outline of the thesis

Now a rough sketch of the theory of neural networks has been presented, I come to the topics that will be treated in this thesis. Although I have abstracted the neural model in many ways by eliminating various biological subtleties, I have tried to model a neural network in a way which, at essential points, is biologically realistic.

The changes of the synapses of a neural network are possibly described by what is called ‘Hebb’s learning rule’, mentioned above already. It is, in fact, a qualitative statement allowing
for 81 interpretations, all of which differ with respect to the question of how the synapses change as a function of dendritic input and axonal output. A more detailed analysis of Hebb's postulate shows that of these 81 possible interpretations only 2 survive, if a simple, biologically plausible, argument is used. This is what the next chapter, chapter 2, starts with.

In the second part of chapter 2, a derivation is given of Hebb's rule. Starting point for this derivation is an expression for the biochemical energy needed to change a synapse. I then derive Hebb's rule by requiring that changes of the synapses take place in an economical way, so that a minimal amount of energy is spent. In this respect, my derivation is 'better', i.e., more realistic, than any other derivation that I encountered in the literature where the Hebb rule is often derived from a 'cost function'. This cost function, in turn, was constructed in such a way that it yields the 'right', i.e., 'guessed to be right' Hebb rule. To the best of my knowledge, a Hebb rule derived from a cost function did not ever lead to the Hebb rule which I derived from a principle of minimal use of energy. Apparently, the rule which I consider to be the actual rule, was difficult to guess!

I named this particular Hebb rule the 'energy saving Hebb rule'. It is a local, mixed Hebbian-Anti-Hebbian learning rule, which is asymmetric with respect to pre- and post-synaptic activity. My mathematically derived energy saving learning rule, as a surprise, encompasses the two rules with which I was left with after my careful inspection, based on biological plausibility, of the 81 different interpretations of Hebb's postulate.

The energy saving learning rule is a quantitative prescription, albeit that there remain two unknowns, not easily accessible within the chosen theoretical framework. They are: an overall factor \( \eta \) (the learning rate), and a constant \( \kappa \) (the margin parameter). Setting apart these facts, my energy saving learning rule can be characterized as quantitative, whereas Hebb's learning rule certainly is qualitative only.

In the third and last part of chapter 2 I derive an analytical formula for the strengths \( w_{ij} \) of the synaptic connections, supposing that the synaptic adaptation takes place according to my energy saving learning rule. The learning procedure used in this derivation is known to yield a result which coincides with the mathematical method of the pseudo-inverse. However, for my biological neural net the original formulation of the well-known method of the pseudo-inverse could not be used. In the appendix of chapter 2 a modification of the method of the pseudo-inverse will be presented which overcomes all problems resulting from biologically induced refinements of the usual theory.

In the following two chapters, chapters 3 and 4, the study of chapter 2 is deepened. Firstly, I take into account that not only ideal, unperturbed, patterns are learned by the neural net: realistic input data is noisy, i.e., the set of input patterns is enlarged. Secondly, I try to get some grip and understanding of the margin parameter \( \kappa \) and the learning rate \( \eta \) occurring in my energy saving learning rule.

To be more specific, in chapter 3 I tackle the non-biological, purely mathematical question to determine the values of the weights of a neural network, independent of any learning rule. This is done by taking into account, explicitly, fixed points as well as basins of attraction, whereas in chapter 2 basins of attraction were introduced only implicitly via the margin-parameter \( \kappa \). The result turns out to be a generalization of the expression for the weights after a learning process with the energy saving learning rule of chapter 2. This is unexpected, because it was found independently of any learning rule, but it certainly is an encouraging surprise.
In chapter 4 I again consider learning in a neural network, just as in chapter 2, but now for noisy instead of fixed patterns. Via a Discrete Time Master Equation, I theoretically follow the process of adaptation of the synapses, using the energy saving learning rule. My final result is an expression for the weights $w_{ij}$, which, for unperturbed input, reduces to the expressions for $w_{ij}$ found in chapter 2, as it should.

For noisy input there is a slight difference with the mathematical result of chapter 2. This is due to the fact that in an actual neural network, the synapses will always change a little bit, so that fixed patterns have a non-fixed representation in the mind. The mathematical approach of chapter 2 was not devised to take into account this subtle, time-dependent effect.

The research of the chapters 2–4 of this thesis, which treat biological neural networks, can be summarized by stating that Hebb's postulate has been derived — for the first time, and half a century after its formulation — from a physical principle involving only economy of consumption of energy. The resulting Hebb rule, which I called 'energy saving rule' is much more precise, and almost quantitative, in contrast to its 1949 predecessor.

In the final chapter 5 I make a switch from neural networks to the related model of a spin-glass. Both neural networks and spin-glasses possess a complex phase-space, which can be studied in many different ways. One of them is the method of damage spreading, in which two identical systems with a different initial spin-configuration, are followed in time. In this manner I show in the specific spherical $p$-spin model the existence of a critical temperature $T_c$, which separates two different dynamical regimes. Moreover, I present an explanation for the observed behaviour of the spin-system. Although this chapter is less or perhaps even totally non-biological, the concept of damage spreading is a useful tool to gain more insight in complex systems, among which biological neural networks.

The remaining chapters are the product of a collaboration with other authors. To express this fact, from now on 'I' will become 'we'.