Machine learning with neural networks

October 15 (2020)

BERNHARD MEHLIG

Department of Physics
University of Gothenburg
Göteborg, Sweden 2020

PREFACE

These are lecture notes for my course on *Artificial Neural Networks* that I have given at Chalmers and Gothenburg University. This course describes how neural networks are used in machine learning. The lecture notes cover Hopfield nets, perceptrons, deep learning, recurrent networks, reinforcement learning, and other supervised and unsupervised machine-learning algorithms.

When I first developed my lectures, my main source was the book by Hertz, Krogh, and Palmer [1]. Other sources were the book by Haykin [2], as well as the lecture notes of Horner [3]. My main sources for the Chapter on deep learning were the book by Goodfellow, Bengio & Courville [4], and the online-book by Nielsen [5].

I am grateful to Martin Čejka who typed the first version of my hand-written lecture notes and made most of the Figures, and to Erik Werner and Hampus Linander for their interest and their help in preparing Chapter 7. I would like to thank also Johan Fries and Oleksandr Balabanov for implementing the algorithms described in Section 8.3. Johan Fries and Marina Rafajlovic made many exam questions. Finally a large number of students – past and present – pointed out misprints and errors, and suggested improvements. I thank them all.

CONTENTS

1 Introduction

The term *neural networks* refers to networks of neurons in the mammalian brain. Neurons are its fundamental units of computation, they are connected together in networks to process data. This can be a very complex task, and the dynamics of such neural networks in response to external stimuli is therefore often quite intricate. Inputs and outputs of each neuron vary as functions of time, in the form of spike trains. Also the network itself changes over time: we learn and improve our data-processing capacities by establishing reconnections between neurons.

Neural-network algorithms for machine learning are inspired by the architecture and the dynamics of networks of neurons in the brain. The algorithms use neuron models that are highly simplified, compared with real neurons. Nevertheless, the fundamental principle is the same: artificial neural networks learn by reconnection. Such networks can perform a multitude of information-processing tasks. They can learn to recognise structures in a set of training data and, to some extent, generalise what they have learnt (supervised learning). A training set contains a list of input data sets together with a list of corresponding labels or target values that encode the properties of the input data the network is supposed to learn. Artificial neural nets can learn to classify such data very accurately, and can generalise the result to other input-data sets – provided that the new data comes from the same data distribution as the original one. A prime example for a problem of this type is object recognition in images, for instance in the sequence of camera images taken by a self-driving car.

The tools for machine learning with neural nets were developed long ago, during second half of the last century, demonstrating some success in solving problems with neural-net based machine-learning tools. During the past decade, however, machine learning with neural nets has become especially popular, driven in part by the success of neural nets in object recognition. The foremost reason is perhaps that industry is in acute need of such algorithms. An important aspect is also that the algorithms are much easier to train nowadays, because there are much better data sets available – not only larger but also more accurate. Last but not least the computer hardware has improved, so that networks with many layers containing many neurons can be efficiently trained (deep learning).

Another task at which neural nets excel is *machine translation*. These networks are dynamical, or recurrent. They take sentences as inputs. As one feeds word after word, the network outputs the words in the translated sentence. Recurrent nets can be efficiently trained on large training sets of input sentences and their translations. Google translate works in this way. Recurrent nets have also been used with considerable success to predict chaotic dynamics.

In general, neural nets are good at analysing large sets of unlabeled, often high-

dimensional data – where it is difficult to determine *a priori* which questions may be most relevant and rewarding to ask. Unsupervised-learning algorithms allow neural nets to learn without labels. Instead the network organises the unlabeled input data in relevant ways: it can detect familiarity and similarity (clusters) of input patterns and other structures in the input data. Unsupervised-learning algorithms work well when there is redundancy in the input data, and they are particularly useful for high-dimensional data sets, where it is a challenge to detect clusters or other data structures by inspection.

Many problems lie between these two extremes of supervised and unsupervised learning. Consider how an agent may learn to navigate a complex environment, in order to get from one location to another one as quickly as possible, or expending as little energy as possible. Reinforcement learning allows the agent to optimise its behaviour in response to environmental cues in the form of penalties and rewards. The agent learns to act in such a way that it receives positive feedback (reward) more more often than a penalty. Such algorithms are used, for instance, in the software AlphaGo [6] that plays the game of Go.

These different algorithms have much in common, because they share the same building blocks: the neurons are modeled as linear threshold units, so-called McCulloch-Pitts neurons, and the learning rules are similar, based on Hebb's rule. Closely related questions arise also regarding the network dynamics. Some noise (not too much) can improve the performance, and ensures that the long-time dynamics approaches a steady state, making it possible to analyse under which circumstances the algorithms converges, so that they yield a definite learning outcome.

There are many connections to methods used in Mathematical Statistics, such as Markov-chain Monte-Carlo algorithms and simulated annealing. Certain unsupervised learning algorithms are related to principal-component analysis, others to clustering algorithms such as K-means clustering. Supervised learning with deep networks is essentially regression analysis, trying to fit an input-output function to the training data. In other words this is just function fitting – and usually with a very large number of fitting parameters. Recent convolutional neural nets have millions of parameters. To determine so many parameters requires very large and accurate data sets. This makes it very clear that neural nets are not a *solution of everything*. One of the difficult problems is to understand when machine learning with neural nets may be appropriate, and when not. We need a detailed understanding of how the algorithms work, and in particular when and how they fail.

The goal of this book is to explain the fundamental principles that allow neural nets to learn, and to demonstrate how they are implemented from scratch, with all indices in the right places. This understanding and these details are necessary to apply machine-learning methods with success.

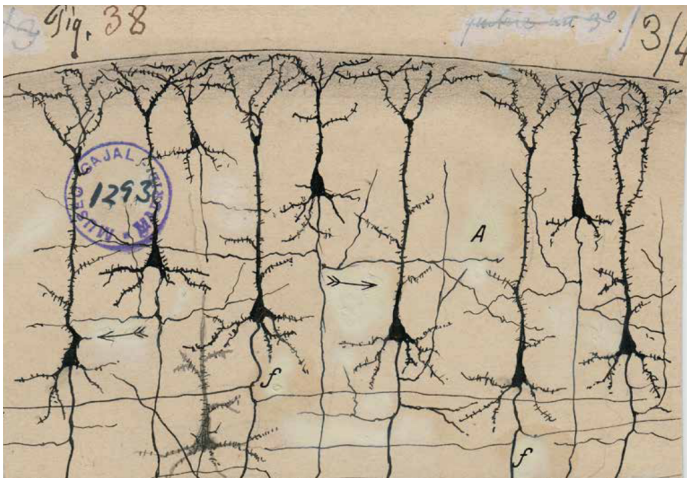


Figure 1.1: Neurons in the cerebral cortex (outer layer of the cerebrum, the largest and best developed part of the mammalian brain). Drawing by Santiago Ramón y Cajal, the Spanish neuroscientist who received the Nobel Prize in Physiology and Medicine in 1906 together with Camillo Golgi 'in recognition of their work on the structure of the nervous system' [7]. Courtesy of the Cajal Institute, 'Cajal Legacy', Spanish National Research Council (CSIC), Madrid, Spain.

1.1 Neural nets

The mammalian brain consists of different regions that perform different tasks. The *cerebral cortex* is the outer layer of the mammalian brain. We can think of it as a thin sheet (about 2 to 5 mm thick) that folds upon itself to form a compact structure with a large surface area. The cortex is the largest and best developed part of the Human brain. It contains large numbers of nerve cells, *neurons*. The Human cerebral cortex contains about 10^{10} neurons. They are linked together by nerve strands (*axons*) that branch and end in *synapses*. These synapses are the connections to other neurons. The synapses connect to *dendrites*, branches extending from the neural cell body that are designed to receive input from other neurons in the form of electrical signals. A neuron in the Human brain may have thousands of synaptic connections with other neurons. The resulting network of connected neurons in the cerebral cortex is responsible for processing of visual, audio, and sensory data.

Figure 1.1 shows neurons in the cerebral cortex. This drawing was made by Santiago Ramón y Cajal more than 100 years ago. By microscope he studied the structure of the neural network in the brain, and he documented his observations by ink-on-paper drawings like the one in Figure 1.1. One can distinguish the cell bodies of the neural cells, their axons (*f*), and their dendrites. In region *A*, the axon of one neuron connects to the dendrites of another neuron, forming a neural network. A

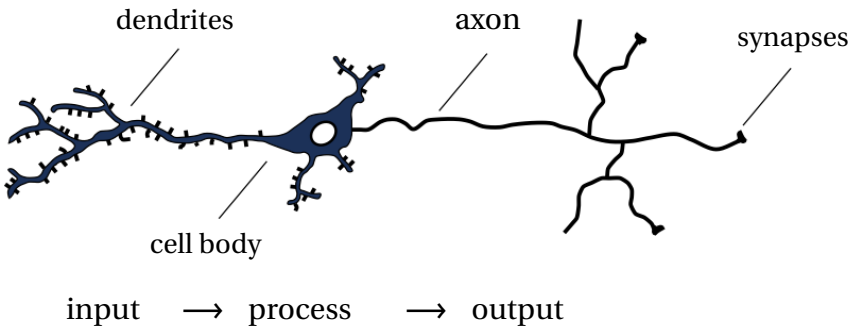


Figure 1.2: Schematic image of a neuron. Dendrites receive input in the form of electrical signals, via synapses. The signals are processed in the cell body of the neuron. The cell nucleus is shown in white. The output travels from the neural cell body along the axon which connect by synaptic couplings to other neurons.

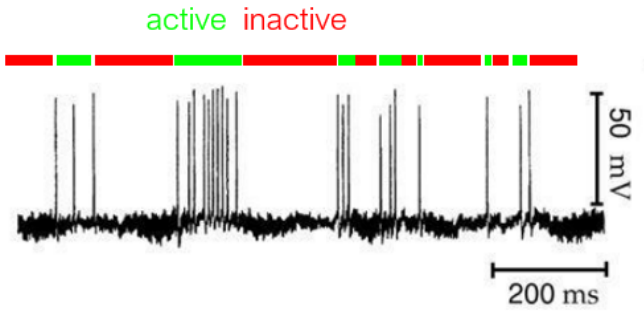


Figure 1.3: Spike train in electrosensory pyramidal neuron in fish (*eigenmannia*). The time series is from Ref. [8]. It is reproduced by permission of the publisher.

more schematic image of a neuron is drawn in Figure 1.2. Information is processed from left to right. On the left are the dendrites that receive signals and connect to the cell body of the neuron where the signal is processed. The right part of the Figure shows the axon, through which the output is sent to the dendrites of other neurons.

Information is transmitted as an electrical signal. Figure 1.3 shows an example of the time series of the electric potential for a pyramidal neuron in fish [8]. The time series consists of an intermittent series of electrical-potential spikes. Quiescent periods without spikes occur when the neuron is *inactive*, during spike-rich periods the neuron is *active*.

1.2 McCulloch-Pitts neurons

In artificial networks, the ways in which information is processed and signals are transferred are highly simplified. The model we use nowadays for the computational

unit, the artificial neuron, goes back to McCulloch and Pitts [9]. Rosenblatt [10, 11] described how to connect such units in artificial neural networks to process information. He referred to these networks as *perceptrons*.

In its simplest form, the model for the artificial neuron has only two states, *active* or *inactive*. The model works as a linear threshold unit: it sums the weighted input signals and computes an output. If the output exceeds a given threshold then the state of the neuron is said to be *active*, otherwise *inactive*. The model is illustrated in Figure 1.4. Neurons usually perform repeated computations, and one divides up time into discrete time steps $t = 0, 1, 2, 3, \ldots$. The state of neuron number j at time step t is denoted by

$$s_j(t) = \begin{cases} -1 & \text{inactive}, \\ 1 & \text{active}, \end{cases} \tag{1.1}$$

Given the signals $s_j(t)$, neuron number i computes

$$s_i(t+1) = \operatorname{sgn}\Big(\underbrace{\sum_{j=1}^{N} w_{ij} s_j(t) - \theta_i}_{=b_i(t)} \Big). \tag{1.2}$$

Here $\operatorname{sgn}(b)$ is the signum function (Figure 1.4):

$$\operatorname{sgn}(b) = \begin{cases} -1, & b < 0, \\ +1, & b \geq 0. \end{cases} \tag{1.3}$$

Strictly speaking, the signum function is not defined at $b = 0$. To avoid problems in our computer algorithms we define $\operatorname{sgn}(0) = 1$. The argument of the signum function,

$$b_i(t) = \sum_{j=1}^{N} w_{ij} s_j(t) - \theta_i, \tag{1.4}$$

is called the *local field*. We see that the neuron performs a weighted linear average of the inputs $s_j(t)$. The parameters w_{ij} are called *weights*. Here the first index, i, refers to the neuron that does the computation, and j labels all neurons that connect to neuron i. The connection strengths between different pairs of neurons are in general different, reflecting different strengths of the synaptic couplings. When the value of w_{ij} is positive, we say that the coupling is *excitatory*. When w_{ij} is negative, the connection is called *inhibitory*. When $w_{ij} = 0$, there is no connection. The threshold[1] for neuron i is denoted by θ_i.

[1] In this book we use the sign convention in Equation (1.4) where the thresholds θ_i come with a minus sign. Other authors use the update rule $s_i(t+1) = \operatorname{sgn}(\sum_{j=1}^{N} w_{ij} s_j(t) + \theta_i)$.

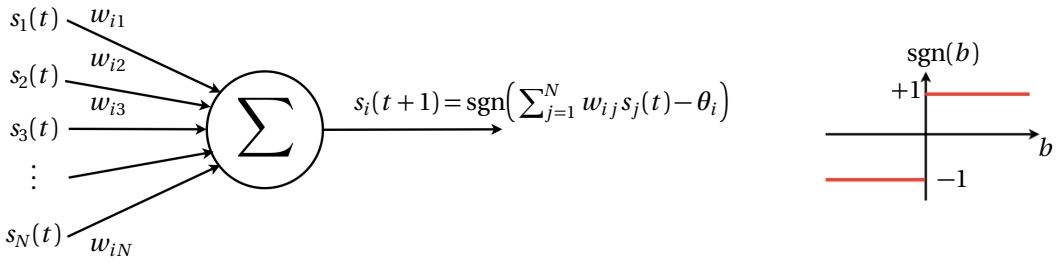


Figure 1.4: Schematic diagram of a McCulloch-Pitts neuron. The index of the neuron is i, it receives inputs from N neurons. The strength of the connection from neuron j to neuron i is denoted by w_{ij}. The *threshold* value for neuron i is denoted by θ_i. The index $t = 0, 1, 2, 3, \ldots$ labels the discrete time sequence of computation steps, and sgn(b) stands for the signum function [Equation (1.3)].

Finally, the computation (1.2) is performed for all neurons i in parallel, and the outputs s_i are the inputs to all neurons at the next time step, therefore the outputs have the time argument $t+1$. These steps are repeated many times, resulting in time series of the activity levels of all neurons in the network.

1.3 Other models for neural computation

The McCulloch-Pitts model is just a caricature of the time series of electrical signals in the cortex. It models the patterns of spiking activity in Figure 1.3 in terms of two states, -1 and $+1$, representing the inactive and active periods shown in the Figure. For many computation tasks this is sufficient, and for our purposes it does not matter that the dynamics of real neurons is quite different in detail. The aim is not to model the neural dynamics in the brain, but to construct computation models inspired by real neural dynamics.

In the course of these lectures it will become apparent that the simplest model described above must be generalised for certain tasks and questions. For example, the jump in the signum function at $b = 0$ may cause large fluctuations in the activity levels of a network of neurons, caused by infinitesimal changes of the local fields across $b = 0$. To avoid this, one allows the neuron to respond continuously to its inputs, replacing Eq. (1.2) by

$$s_i(t+1) = g\left(\sum_j w_{ij} s_j(t) - \theta_i\right). \tag{1.5}$$

Here $g(b)$ is a continuous *activation function*. It could just be a linear function, $g(b) \propto b$. But we shall see that many tasks require non-linear activation functions, such as $\tanh(b)$ (Figure 1.6). When the activation function is continuous, the states

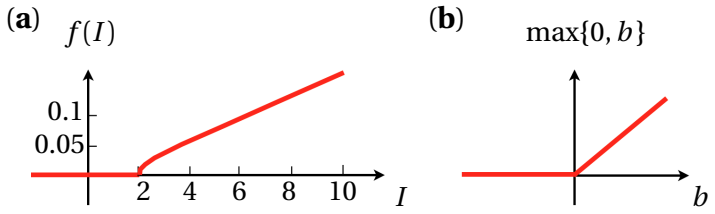


Figure 1.5: (**a**) Firing rate of a *leaky integrate-and-fire* neuron as a function of the electrical current I through the cell membrane, Equation (1.7) for $\tau = 25$ and $U_c/R = 2$ (see text). (**b**) Piecewiwe linear activation function, $g(b) = \max\{0, b\}$.

assume continuous values too, not just the discrete values -1 and $+1$ given in Equation (1.1).

Alternatively one may use a piecewise linear activation function (Figure 1.5). This is motivated in part by the response curve of *leaky integrate-and-fire* neurons. This is a model for the relation between the electrical current I through the cell membrane into the neuron cell, and the membrane potential U. The simplest models for the dynamics of the membrane potential represent the neuron as a capacitor. In the leaky integrate-and-fire neuron, leakage is added by a resistor R in parallel with the capacitor C, so that

$$I = \frac{U}{R} + C\frac{\mathrm{d}U}{\mathrm{d}t}\,. \tag{1.6}$$

For a constant current, the membrane potential grows from zero as a function of time, $U(t) = RI[1-\exp(-t/\tau)]$, where $\tau = RC$ is the time constant of the model. One says that the neuron produces a *spike* when the membrane potential exceeds a critical value, U_c. Immediately after, the membrane potential is set to zero (and begins to grow again). In this model, the *firing rate* $f(I)$ is thus given by t_c^{-1}, where t_c is the solution of $U(t) = U_c$. It follows that the firing rate exhibits a threshold behaviour (the system works like a rectifier):

$$f(I) = \begin{cases} 0 & \text{for} \quad I \le U_c/R\,, \\ \left[\tau \log\left(\frac{RI}{RI-U_c}\right)\right]^{-1} & \text{for} \quad I > U_c/R\,. \end{cases} \tag{1.7}$$

This response curve is illustrated in Figure 1.5 (**a**). The main message is that there is a threshold below which the response is strictly zero (this is not the case for the activation function shown in Figure 1.6). The response function looks qualitatively like the piecewise linear function

$$g(b) = \max\{0, b\}\,, \tag{1.8}$$

shown in panel (**b**). Neurons with this activation function are called *rectified linear units*, and the activation function (1.8) is called the *ReLU* function.

Finally, Equations (1.2) and (1.5) are called *synchronous* update rules, because all neurons are updated in parallel: at time step t all inputs $s_j(t)$ are stored. Then all

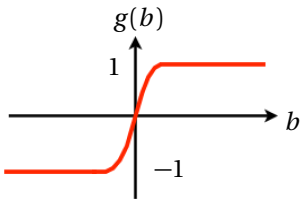


Figure 1.6: Continuous activation function.

neurons i are simultaneously updated using the stored inputs. An alternative is to update only a single neuron per iteration, the one with index m say:

$$s_i(t+1) = \begin{cases} g\left(\sum_j w_{mj} s_j(t) - \theta_m\right) & \text{for } i = m\,, \\ s_i(t) & \text{otherwise}\,. \end{cases} \tag{1.9}$$

This an *asynchronous* update rule. Different schemes for choosing neurons are used in asynchronous updating. One possibility is to arrange the neurons into a two-dimensional array and to update them one by one, in a certain order. In the *typewriter scheme*, for example, one updates the neurons in the top row of the array first, from left to right, then the second row from left to right, and so forth. A second possibility is to choose randomly which neuron to update.

If there are N neurons, then one synchronous step corresponds to N asynchronous steps, on average. This difference in time scales is not the only difference between synchronous and asynchronous updating. The asynchronous dynamics can be shown to converge to a definite state in certain cases, while the synchronous dynamics may fail to do so, resulting in periodic cycles that persist forever.

1.4 Summary

Artificial neural networks use a highly simplified model for the fundamental computation unit, the neuron. In its simplest form, the model is just a binary threshold unit. The units are linked together by weights w_{ij}, and each unit computes a weighted average of its inputs. The network performs these computations in sequence. Usually one considers discrete sequences of computation time steps, $t = 0, 1, 2, 3, \ldots$. Either all neurons are updated simultaneously in one time step (synchronous updating), or only one chosen neuron is updated (asynchronous updating). Most neural-network algorithms are built using the model described in this Chapter.

PART I
HOPFIELD NETS

The Hopfield net [12, 13] is an artificial neural network that can recognise or reconstruct images. Consider for example the binary images of digits in Figure 2.1. The images are stored in the network by assigning the weights w_{ij} in a certain way (called *Hebb's rule*). Then one feeds an image of a distorted version of one of the digits (Figure 2.2) to the network by assigning the initial states of the neurons in the network to the bits in the distorted image. The idea is that the neural-network dynamics converges to the correct undistorted digit. In this way the network can recognise the input as a distorted image of the correct digit (*retrieve* this digit). Hopfield nets recognise patterns with many bits very efficiently, and in the past such networks were used to perform pattern recognition tasks. Today there are more efficient algorithms for this purpose (Chapter 8).

Nevertheless, Hopfield nets exemplify fundamental principles of machine learning with neural networks. For a start, all other neural-network algorithms discussed in these lectures are built from the same building blocks and use learning rules that are closely related to Hebb's rule. Furthermore, generalisations of Hopfield nets (*restricted Boltzmann machines*) can learn the distribution underlying an ensemble of input patterns. This makes it possible to generate image textures and to complete partially obscured images [14]. Also, the dynamics of Hopfield nets is closely related to *Markov-chain Monte-Carlo algorithms* which are much used for a wide range of problems in Physics and Mathematical Statistics. Last but not least, Hopfield nets exemplify the role of stochasticity in neural-network dynamics. A certain degree of noise (not too much) can substantially improve the performance of Hopfield nets. In Engineering problems it is usually better to avoid stochasticity, when it is due to errors in the form of multiplicative or additive noise that diminish the performance of the system. In neural-network dynamics, by contrast, stochasticity is often helpful, as we shall see below. In general it is very challenging to analyse the stochastic dynamics. But for the Hopfield network much is known. The reason is that Hopfield nets are closely related to stochastic systems studied in Physics, namely *random magnets* and *spin glasses*.

2 Deterministic Hopfield nets

2.1 Pattern recognition

As an example for a pattern-recognition task consider p images (*patterns*), each with N bits. The patterns could be the letters in the alphabet, or the digits shown in Figure 2.1. The different patterns are labeled by the index $\mu = 1, \ldots, p$. The bits of pattern μ are denoted by $x_i^{(\mu)}$. The index i labels the bits of a given pattern, it

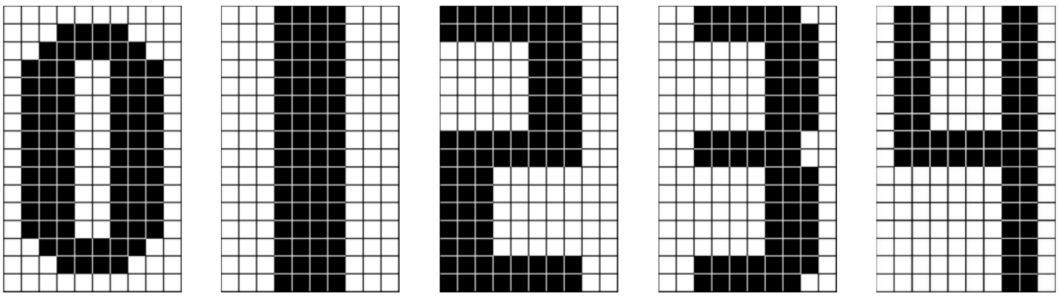


Figure 2.1: Binary representation of the digits 0 to 4. Each pattern has 10×16 pixels. Adapted from Figure 14.17 in Ref. [2]. The slightly peculiar shapes help the Hopfield net to distinguish the patterns [15].

ranges from 1 to N. The bits are *binary*: they can take only the values -1 and $+1$, as illustrated in Figure 2.2. To determine the generic properties of the algorithm, one often turns to *random patterns* where each bit $x_i^{(\mu)}$ is chosen randomly, taking either value with probability $\frac{1}{2}$. It is convenient to gather the bits of a pattern in a column vector

$$\boldsymbol{x}^{(\mu)} = \begin{bmatrix} x_1^{(\mu)} \\ x_2^{(\mu)} \\ \vdots \\ x_N^{(\mu)} \end{bmatrix} . \tag{2.1}$$

In this book, vectors are written in bold math font.

The task of the neural net is to recognise distorted patterns, to determine for instance that the pattern on the right in Figure 2.2 is a distorted version of the digit zero (left pattern in Figure 2.2). In practice one *stores p* patterns in the network and presents it with a distorted version of one of these patterns. The network retrieves the stored pattern that is most similar to the distorted one.

The formulation of the problem requires to define how similar a distorted pattern $\boldsymbol{x}$ is to any of the stored patterns, $\boldsymbol{x}^{(\mu)}$ say. The *Hamming distance* h_μ between $\boldsymbol{x}$ and $\boldsymbol{x}^{(\mu)}$ is defined as

$$h_\mu \equiv \sum_{i=1}^{N} \Big[(1 + x_i^{(\mu)})(1 - x_i) + (1 - x_i^{(\mu)})(1 + x_i) \Big] . \tag{2.2}$$

It equals the number of bits by which the two patterns differ. More generally one can quantify the distance by the mean squared error $\frac{1}{N} \sum_{i=1}^{N} (x_i^{(\mu)} - x_i)^2$. For ± 1 patterns both measures are equivalent, up to the factor N^{-1}. Note that the Hamming distance does not refer to distortions by translations, rotations, or shearing. An improved measure of distance might take the minimum Hamming distance between the

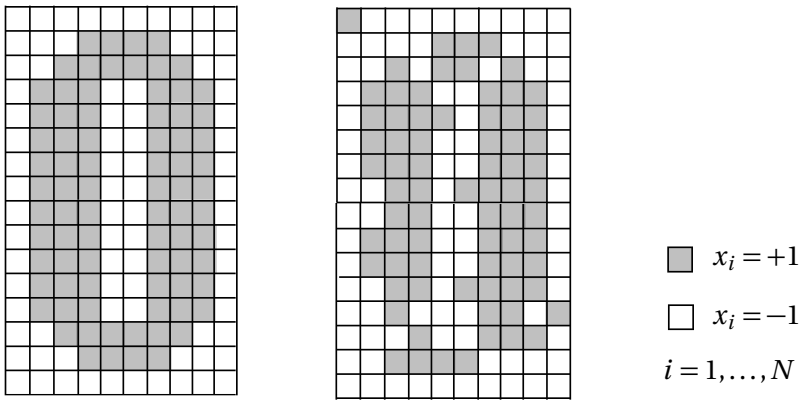


Figure 2.2: Binary image ($N = 160$) of the digit 0, and a distorted version of the same image.

patterns subject to all possible translations, rotations, and so forth. In summary, the task is to find the index ν for which the Hamming distance h_ν is minimal, $h_\nu \leq h_\mu$ for all $\mu = 1, \ldots, p$. How can one solve this task using a neural network? One feeds the distorted pattern $\boldsymbol{x}$ into the network by assigning $s_i(t=0) = x_i$.

Assume that $\boldsymbol{x}$ is a distorted version of $\boldsymbol{x}^{(\nu)}$. The idea is to find a set of weights w_{ij} so that the network dynamics (2.6) converges to the correct stored pattern:

$$s_i(t) \to x_i^{(\nu)} \quad \text{as} \quad t \to \infty \,. \tag{2.3}$$

Which weights to choose depends on the patterns $\boldsymbol{x}^{(\mu)}$, so the weights must be functions of $\boldsymbol{x}^{(\mu)}$. We say that we *store* these patterns in the network by choosing the appropriate weights. If the network converges as in Equation (2.3), the pattern $\boldsymbol{x}^{(\nu)}$ is said to be an *attractor* of the dynamics.

2.2 Hopfield nets

Hopfield nets [12, 13] are networks of McCulloch-Pitts neurons designed to solve the pattern-recognition task described in the previous Section. The space of all possible states

$$\boldsymbol{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} \tag{2.4}$$

of this network is called *state space*. The states are updated either with the synchronous update rule (1.2):

$$s_i(t+1) = \operatorname{sgn}\big[b_i(t)\big] \quad \text{with local field} \quad b_i(t) = \sum_{j=1}^{N} w_{ij} s_j(t) - \theta_i\,, \tag{2.5}$$

or with the asynchronous rule

$$s_i(t+1) = \begin{cases} \operatorname{sgn}\big[b_m(t)\big] & \text{for } i = m\,, \\ s_i(t) & \text{otherwise}\,. \end{cases} \tag{2.6}$$

The argument of the signum function in Equation (2.6) is again the local field, $b_m(t) = \sum_{j=1}^{N} w_{mj} s_j(t) - \theta_m$.

Now we need a strategy for choosing the weights w_{ij}, so that the patterns $\boldsymbol{x}^{(\mu)}$ are attractors. If one feeds a pattern $\boldsymbol{x}$ close to $\boldsymbol{x}^{(\nu)}$ to the network, we want the network to converge to $\boldsymbol{x}^{(\nu)}$

$$\boldsymbol{s}(t=0) = \boldsymbol{x} \approx \boldsymbol{x}^{(\nu)}; \quad \boldsymbol{s}(t) \to \boldsymbol{x}^{(\nu)} \quad \text{as} \quad t \to \infty\,. \tag{2.7}$$

This means that the network succeeds in correcting a small number of errors. If the number of errors is too large, the network may converge to another pattern, or it may converge to some other state that bears no or little relation to the stored patterns. The region in configuration space around pattern $\boldsymbol{x}^{(\nu)}$ in which all patterns $\boldsymbol{x}$ converge to $\boldsymbol{x}^{(\nu)}$ is called the *region of attraction* of $\boldsymbol{x}^{(\nu)}$. The size of the region around $\boldsymbol{x}^{(\nu)}$ depends in an intricate way the ensemble of stored patterns, and there is no general convergence proof. Therefore we try to answer a different question first: if one feeds one of the undistorted patterns $\boldsymbol{x}^{(\nu)}$, does the network recognise that it is one of the stored, undistorted patterns? The network should not make any changes to $\boldsymbol{x}^{(\nu)}$ because all bits are correct:

$$\boldsymbol{s}(t=0) = \boldsymbol{x}^{(\nu)}; \quad \boldsymbol{s}(t) = \boldsymbol{x}^{(\nu)} \quad \text{for all} \quad t = 0, 1, 2, \ldots\,. \tag{2.8}$$

Even this question is in general difficult to answer. We therefore consider a simple limit of the problem first, namely $p = 1$. There is only one pattern to recognize, $\boldsymbol{x}^{(1)}$. A suitable choice of weights w_{ij} is *Hebb's rule*

$$w_{ij} = \frac{1}{N} x_i^{(1)} x_j^{(1)} \quad \text{and} \quad \theta_i = 0\,. \tag{2.9}$$

We say that the pattern $\boldsymbol{x}^{(1)}$ is *stored* in the network by assigning the weights w_{ij} using the rule (2.9). Note that the weights are symmetric, $w_{ij} = w_{ji}$.

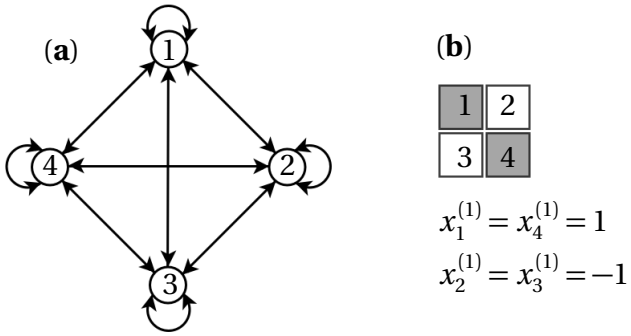


Figure 2.3: Hopfield net with $N = 4$ neurons. (**a**) Network layout. Neurons are represented as ◯. Arrows indicate symmetric connections. (**b**) Pattern $\boldsymbol{x}^{(1)\mathsf{T}} = [1,-1,-1,1]^{\mathsf{T}}$. Here T denotes the transpose of the column vector $\boldsymbol{x}^{(1)}$.

To check that the rule (2.9) does the trick, feed the pattern to the network by assigning $s_j(t=0) = x_j^{(1)}$, and evaluate Equation (2.5):

$$\sum_{j=1}^{N} w_{ij} x_j^{(1)} = \frac{1}{N} \sum_{j=1}^{N} x_i^{(1)} x_j^{(1)} x_j^{(1)} = \frac{1}{N} \sum_{j=1}^{N} x_i^{(1)} . \tag{2.10}$$

The last equality follows because $x_j^{(1)}$ can only take the values ± 1. The sum evaluates to N, so that

$$\operatorname{sgn}\left(\sum_{j=1}^{N} w_{ij} x_j^{(1)} \right) = x_i^{(1)} . \tag{2.11}$$

Recall that $x_i^{(\mu)} = \pm 1$, so that $\operatorname{sgn}(x_i^{(\mu)}) = x_i^{(\mu)}$. Comparing Equation (2.11) with the update rule (2.5) shows that the bits $x_j^{(1)}$ of the pattern $\boldsymbol{x}^{(1)}$ remain unchanged under the update, as required by Eq. (2.8). The network recognises the pattern as a stored one, so Hebb's rule (2.9) does what we asked for.

But does the network correct small errors? In other words, is the pattern $\boldsymbol{x}^{(1)}$ an attractor [Eq. (2.7)]? This question cannot be answered in general. Yet in practice Hopfield nets work often very well! It is a fundamental insight that neural networks may work well although it is impossible to strictly prove that their dynamics converges to the correct solution.

To illustrate the difficulties consider an example, a Hopfield net with $p = 1$ and $N = 4$ (Figure 2.3). Store the pattern $\boldsymbol{x}^{(1)}$ shown in Figure 2.3 by assigning the weights w_{ij} using Hebb's rule (2.9). Now feed a distorted pattern $\boldsymbol{x}$ to the network that has a non-zero distance to $\boldsymbol{x}^{(1)}$:

$$h_1 = \frac{1}{4} \sum_{i=1}^{4} \left(x_i - x_i^{(1)} \right)^2 > 0 . \tag{2.12}$$

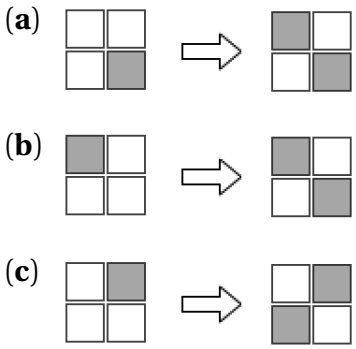


Figure 2.4: Reconstruction of a distorted pattern. Under synchronous updating (2.5) the first two distorted images (**a**) and (**b**) converge to the stored pattern $\boldsymbol{x}^{(1)}$, but pattern (**c**) does not.

The factor $\frac{1}{4}$ takes into account that the patterns take the values ± 1 and not $0/1$ as in Section 2.1. To feed the pattern to the network, one sets $s_i(t=0) = x_i$. Now iterate the dynamics using synchronous updating (2.5). Results for different distorted patterns are shown in Figure 2.4. We see that the first two distorted patterns (distance 1) converge to the stored pattern, cases (**a**) and (**b**). But the third distorted pattern does not [case (**c**)].

To understand this behaviour it is most convenient to analyse the synchronous dynamics using the *weight matrix*

$$\mathbb{W} = \frac{1}{N} \boldsymbol{x}^{(1)} \boldsymbol{x}^{(1)\mathsf{T}} . \tag{2.13}$$

Here $\boldsymbol{x}^{(1)\mathsf{T}}$ denotes the *transpose* of the column vector $\boldsymbol{x}^{(1)}$, so that $\boldsymbol{x}^{(1)\mathsf{T}}$ is a row vector. The standard rules for matrix multiplication apply also to column and row vectors, they are just $N \times 1$ and $1 \times N$ matrices. So the product on the r.h.s. of Equation (2.13) is an $N \times N$ matrix. In the following, matrices with elements A_{ij} or B_{ij} are written as $\mathbb{A}$, $\mathbb{B}$, and so forth. The product in Equation (2.13) is also referred to as an *outer product*. The product

$$\boldsymbol{x}^{(1)\mathsf{T}} \boldsymbol{x}^{(1)} = \sum_{j=1}^{N} [x_j^{(1)}]^2 = N , \tag{2.14}$$

by contrast, is just a number (equal to N). The product (2.14) is called *scalar product*. It also is denoted by $\boldsymbol{x} \cdot \boldsymbol{x} = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{x}$. We use the same notation for multiplying a transposed vector with a matrix from the left: $\boldsymbol{x} \cdot \mathbb{A} = \boldsymbol{x}^{\mathsf{T}} \mathbb{A}$. An excellent source for those not familiar with these terms from Linear Algebra (Figure 2.5) is Chapter 6 of the book by Mathews and Walker [16].

Using Equation (2.14) we see that $\mathbb{W}$ *projects* onto the vector $\boldsymbol{x}^{(1)}$,

$$\mathbb{W} \boldsymbol{x}^{(1)} = \boldsymbol{x}^{(1)} . \tag{2.15}$$

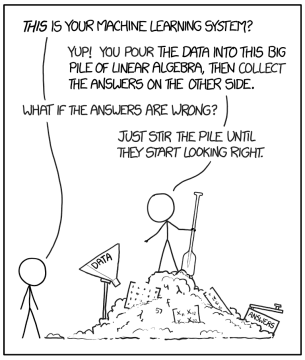


Figure 2.5: Reproduced from xkcd.com/1838 under the creative commons attribution-noncommercial 2.5 license.

In addition, we note that the matrix (2.13) is *idempotent*:

$$\mathbb{W}^n = \mathbb{W} \quad \text{for } n = 1,2,3,\ldots. \tag{2.16}$$

Equations (2.15) and (2.16) imply that the network recognises the pattern $\boldsymbol{x}^{(1)}$ as the stored one. The pattern is not updated [Eq. (2.8)]. This example illustrates the general proof, Equations (2.10) and (2.11).

Now consider the distorted pattern (**a**) in Figure 2.4. We feed this pattern to the network by assigning

$$\boldsymbol{s}(t=0) = \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \end{bmatrix}. \tag{2.17}$$

To compute one step in the synchronous dynamics (2.5) we simply apply $\mathbb{W}$ to $\boldsymbol{s}(t=0)$. This is done in two steps, using the outer-product form (2.13) of the weight matrix. We first multiply $\boldsymbol{s}(t=0)$ with $\boldsymbol{x}^{(1)\mathsf{T}}$ from the left

$$\boldsymbol{x}^{(1)\mathsf{T}}\boldsymbol{s}(t=0) = \begin{bmatrix} 1, & -1, & -1, & 1 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \end{bmatrix} = 2, \tag{2.18}$$

and then we multiply this result with $\boldsymbol{x}^{(1)}$. This gives:

$$\mathbb{W}\boldsymbol{s}(t=0) = \tfrac{1}{2}\boldsymbol{x}^{(1)}. \tag{2.19}$$

The signum of the i-th component of the vector $\mathbb{W}\boldsymbol{s}(t=0)$ yields $s_i(t=1)$:

$$s_i(t=1) = \operatorname{sgn}\Big(\sum_{j=1}^{N} w_{ij}\, s_j(t=0)\Big) = x_i^{(1)}. \tag{2.20}$$

This means that the state of the network converges to the stored pattern, in one synchronous update. Since $\mathbb{W}$ is idempotent, the network stays there: the pattern $\boldsymbol{x}^{(1)}$ is an attractor. Case (**b**) in Figure 2.4 works in a similar way.

Now look at case (**c**), where the network fails to converge to the stored pattern. We feed this pattern to the network by assigning $\boldsymbol{s}(t=0)=[-1,1,-1,-1]^{\mathsf{T}}$. For one iteration of the synchronous dynamics we first evaluate

$$\boldsymbol{x}^{(1)\mathsf{T}}\boldsymbol{s}(0)=\begin{bmatrix}1, & -1, & -1, & 1\end{bmatrix}\begin{bmatrix}-1\\1\\-1\\-1\end{bmatrix}=-2\,. \tag{2.21}$$

It follows that

$$\mathbb{W}\boldsymbol{s}(t=0)=-\tfrac{1}{2}\boldsymbol{x}^{(1)}\,. \tag{2.22}$$

Using the update rule (2.5) we find

$$\boldsymbol{s}(t=1)=-\boldsymbol{x}^{(1)}\,. \tag{2.23}$$

Equation (2.16) implies that

$$\boldsymbol{s}(t)=-\boldsymbol{x}^{(1)}\quad\text{for}\quad t\geq 1\,. \tag{2.24}$$

Thus the network shown in Figure 2.3 has two attractors, the pattern $\boldsymbol{x}^{(1)}$ as well as the *inverted* pattern $-\boldsymbol{x}^{(1)}$. This is a general property of McCulloch-Pitts dynamics with Hebb's rule: if $\boldsymbol{x}^{(1)}$ is an attractor, then the pattern $-\boldsymbol{x}^{(1)}$ is an attractor too. But one ends up in the correct pattern $\boldsymbol{x}^{(1)}$ when more than half of the bits in $\boldsymbol{s}(t=0)$ are correct. In the next Section we discuss what happens when more than one patterns are stored in the Hopfield net.

2.3 The cross-talk term

When there are more than one patterns, the first question is how to generalise Hebb's rule (2.9). A guess is to simply sum Equation (2.9) over the stored patterns:

$$w_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}x_i^{(\mu)}x_j^{(\mu)}.\quad\text{and}\quad\theta_i=0 \tag{2.25}$$

The weights are, in other words, proportional to the second-order statistics of the pattern bits. The prefactor is not important. Here it is chosen in such a way that the large-N analysis simplifies. As for $p=1$ the weight matrix is symmetric, $\mathbb{W}=\mathbb{W}^{\mathsf{T}}$, so

that $w_{ij} = w_{ji}$. The diagonal weights are not zero in general. An alternative version of Hebb's rule [2] defines the diagonal weights to zero:

$$w_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} x_i^{(\mu)} x_j^{(\mu)} \quad \text{for} \quad i \neq j\,, \quad w_{ii} = 0\,, \quad \text{and} \quad \theta_i = 0\,. \tag{2.26}$$

If we store only one pattern, $p = 1$, this modified rule Hebb's rule (2.26) satisfies Equation (2.8). In this Section we use Equation (2.26).

If we assign the weights according to Equation (2.26), does the network recognise distorted patterns? We saw in the previous Section that this question is difficult to answer in general, even for $p = 1$. Therefore we ask, first, whether the network recognises the stored pattern $\boldsymbol{x}^{(\nu)}$. The question is whether

$$\underbrace{\operatorname{sgn}\Big(\frac{1}{N}\sum_{j\neq i}\sum_{\mu} x_i^{(\mu)} x_j^{(\mu)} x_j^{(\nu)}\Big)}_{=b_i^{(\nu)}} \stackrel{?}{=} x_i^{(\nu)}\,. \tag{2.27}$$

To check whether Equation (2.27) holds or not, we repeat the calculation described in the previous Section. As a first step we evaluate the local field

$$b_i^{(\nu)} = \big(1 - \tfrac{1}{N}\big) x_i^{(\nu)} + \frac{1}{N}\sum_{j\neq i}\sum_{\mu\neq\nu} x_i^{(\mu)} x_j^{(\mu)} x_j^{(\nu)}\,. \tag{2.28}$$

Here we have split the sum over the patterns into two contributions. The first term corresponds to $\mu = \nu$, where ν refers to the pattern that was fed to the network, the one that we want the network to recognise. The second term in Equation (2.28) contains the sum over the remaining patterns. For large N we can approximate $(1 - \frac{1}{N}) \approx 1$. It follows that condition (2.27) is satisfied if the second term in (2.28) does not affect the sign of the r.h.s. of this Equation. This second term is called *cross-talk* term.

Whether adding the cross-talk term to $\boldsymbol{x}^{(\nu)}$ affects $\operatorname{sgn}(b_i^{(\nu)})$ depends on the stored patterns. Since the cross-talk term contains a sum over μ we may expect that this term does not matter if p is small enough. If this is true for all i and ν then all p stored patterns are recognised. Furthermore, by analogy with the example described in the previous Section, it is plausible that the stored patterns are then also attractors, so that slightly distorted patterns converge to the correct stored pattern.

For a more quantitative analysis of the effect of the cross-talk term we store patterns with random bits (*random patterns*). Different bits (different values of i and/or μ) are assigned ± 1 independently with equal probability:

$$\text{Prob}(x_i^{(\nu)} = \pm 1) = \tfrac{1}{2}\,. \tag{2.29}$$

This means that different patterns are *uncorrelated* because their *covariance* vanishes:

$$\langle x_i^{(\mu)} x_j^{(\nu)} \rangle = \delta_{ij}\delta_{\mu\nu}. \tag{2.30}$$

Here $\langle \cdots \rangle$ denotes an average over many realisations of random patterns, and δ_{ij} is the *Kronecker delta*, equal to unity if $i = j$ but zero otherwise. Note that it follows from Equation (2.29) that $\langle x_j^{(\mu)} \rangle = 0$.

We now ask: what is the probability that the cross-talk term changes $\mathrm{sgn}(b_i^{(\nu)})$? In other words, what is the probability that the network produces a wrong bit in one asynchronous update, if all bits were initially correct? The magnitude of the cross-talk term does not matter when it has the same sign as $x_i^{(\nu)}$. If it has a different sign, then the cross-talk term may matter. It does if its magnitude is larger than unity (the magnitude of $x_i^{(\nu)}$). To simplify the analysis one wants to avoid having to distinguish between the two cases, whether or not the cross-talk term has the same sign as $x_i^{(\nu)}$. To this end one defines:

$$C_i^{(\nu)} \equiv -x_i^{(\nu)} \underbrace{\frac{1}{N} \sum_{j \neq i} \sum_{\mu \neq \nu} x_i^{(\mu)} x_j^{(\mu)} x_j^{(\nu)}}_{\text{cross-talk term}}. \tag{2.31}$$

If $C_i^{(\nu)} < 0$ then the cross-talk term has same sign as $x_i^{(\nu)}$, so that the cross-talk term does not matter, adding it does not change the sign of $x_i^{(\nu)}$. If $0 < C_i^{(\nu)} < 1$ it does not matter either, only when $C_i^{(\nu)} > 1$. The network produces an error in bit i of pattern νif $C_i^{(\nu)} > 1$.

2.4 One-step error probability

The one-step *error probability* $P_{\text{error}}^{t=1}$ is defined as the probability that an error occurs in one attempt to update a bit, given that initially all bits are correct. Therefore $P_{\text{error}}^{t=1}$ is given by:

$$P_{\text{error}}^{t=1} = \mathrm{Prob}(C_i^{(\nu)} > 1). \tag{2.32}$$

Since patterns and bits are identically distributed, $\mathrm{Prob}(C_i^{(\nu)} > 1)$ does not depend on i or ν. Therefore $P_{\text{error}}^{t=1}$ does not carry any indices.

How does $P_{\text{error}}^{t=1}$ depend on the parameters of the problem, p and N? When both p and N are large we can use the *central-limit theorem* to answer this question. Since different bits/patterns are independent, we can think of $C_i^{(\nu)}$ as a sum of independent random numbers c_m that take the values -1 and $+1$ with equal probabilities,

$$C_i^{(\nu)} = -\frac{1}{N} \sum_{j \neq i} \sum_{\mu \neq \nu} x_i^{(\mu)} x_j^{(\mu)} x_j^{(\nu)} x_i^{(\nu)} = -\frac{1}{N} \sum_{m=1}^{(N-1)(p-1)} c_m. \tag{2.33}$$

There are $M=(N-1)(p-1)$ terms in the sum on the r.h.s. because terms with $\mu=\nu$ are excluded, and also those with $j=i$ [Equation (2.26)]. If we use Equation (2.25) instead, then there is a correction to Equation (2.33) from the diagonal weights. For $p \ll N$ this correction is small.

When p and N are large, then the sum $\sum_m c_m$ contains a large number of independently identically distributed random numbers with mean zero and variance unity. It follows from the central-limit theorem that $\frac{1}{N}\sum_m c_m$ is Gaussian distributed with mean zero, and with variance

$$\sigma_C^2 = \frac{1}{N^2}\Big\langle \Big(\sum_{m=1}^{M} c_m\Big)^2 \Big\rangle = \frac{1}{N^2}\sum_{n=1}^{M}\sum_{m=1}^{M}\langle c_n c_m\rangle . \tag{2.34}$$

Here $\langle\cdots\rangle$ denotes an average over realisations of c_m. Since the random numbers c_m are independent for different indices, $\langle c_n c_m\rangle = \delta_{nm}$. So only the diagonal terms in the double sum contribute, summing up to $M \approx Np$. Therefore

$$\sigma_C^2 \approx \frac{p}{N} . \tag{2.35}$$

One way of showing that $\sum_m c_m$ is approximately Gaussian distributed is to represent it in terms of *Bernoulli trials*. The sum $\sum_{m=1}^{M} c_m$ equals $2k-M$ where k is the number of occurrences $+1$ in the sum. Since the probability of $c_m = \pm 1$ is $\frac{1}{2}$, the probability of drawing k times $+1$ and $M-k$ times -1 is

$$P_{k,M} = \binom{M}{k} \left(\tfrac{1}{2}\right)^k \left(\tfrac{1}{2}\right)^{M-k} . \tag{2.36}$$

Here $\binom{M}{k} = M!/[k!(M-k)!]$ denotes the number of ways in which k occurrences of $+1$ can be distributed over M places. We expect that the quantity $2k-M$ is Gaussian distributed with mean zero and variance M. To demonstrate this, it is convenient to use the variable $z=(2k-M)/\sqrt{M}$, because it should then be Gaussian with mean zero and unit variance. To check whether this is the case, we substitute $k=\frac{M}{2}+\frac{\sqrt{M}}{2}z$ into Equation (2.36) and take the limit of large M using Stirling's approximation

$$n! \approx \mathrm{e}^{n\log n - n + \frac{1}{2}\log 2\pi n} . \tag{2.37}$$

Expanding $P_{k,M}$ to leading order in M^{-1} assuming that z remains of order unity gives $P_{k,M} = \sqrt{2/(\pi M)}\exp(-z^2/2)$. Now one changes variables from k to z. This squeezes local neighbourhoods dk to dz. Conservation of probability implies that $P(z)\mathrm{d}z = P(k)\mathrm{d}k$. It follows that $P(z) = (\sqrt{M}/2)P(k)$, so that $P(z) = (2\pi)^{-1/2}\exp(-z^2/2)$. In other words, the distribution of z is Gaussian with zero mean and unit variance, as we intended to show.

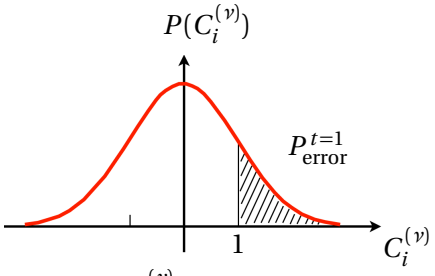


Figure 2.6: Gaussian distribution of $C_i^{(\nu)}$. The hashed area equals the one-step error probability $P_{\text{error}}^{t=1}$.

Returning to the distribution of C [Equation (2.31)], we conclude that it is Gaussian,

$$P(C) = (2\pi\sigma_C^2)^{-1/2} \exp[-C^2/(2\sigma_C^2)], \tag{2.38}$$

with mean zero and variance $\sigma_C^2 \approx \frac{p}{N}$, as illustrated in Figure 2.6. To determine $P_{\text{error}}^{t=1}$ [Equation (2.32)] we must integrate this distribution from 1 to ∞:

$$P_{\text{error}}^{t=1} = \frac{1}{\sqrt{2\pi}\sigma_C} \int_1^\infty \mathrm{d}C \; \mathrm{e}^{-\frac{C^2}{2\sigma_C^2}} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{\frac{N}{2p}}\right)\right]. \tag{2.39}$$

Here erf is the *error function* defined as [17]

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{d}x \; \mathrm{e}^{-x^2}. \tag{2.40}$$

Since erf(z) increases monotonically as z increases we conclude that $P_{\text{error}}^{t=1}$ increases as p increases, or as N decreases. This is expected: it is more difficult for the network to distinguish stored patterns when there are more of them. On the other hand, it is easier to differentiate stored patterns if they have more bits. We also see that the one-step error probability depends on p and N only through the combination

$$\alpha \equiv \frac{p}{N}. \tag{2.41}$$

The parameter α is called the *storage capacity* of the network. Figure 2.7 shows how $P_{\text{error}}^{t=1}$ depends on the storage capacity. For $\alpha = 0.2$ for example. the one-step error probability is slightly larger than 1%.

In the derivation of Equation (2.39) we assumed that the stored patterns are random with independent bits. Realistic patterns are not random. We nevertheless expect that $P_{\text{error}}^{t=1}$ describes the typical one-step error probability of the Hopfield net when p and N are large. However, it is straightforward to construct counter examples. Consider for example *orthogonal patterns*:

$$\boldsymbol{x}^{(\mu)} \cdot \boldsymbol{x}^{(\nu)} = 0 \quad \text{for} \quad \mu \neq \nu. \tag{2.42}$$

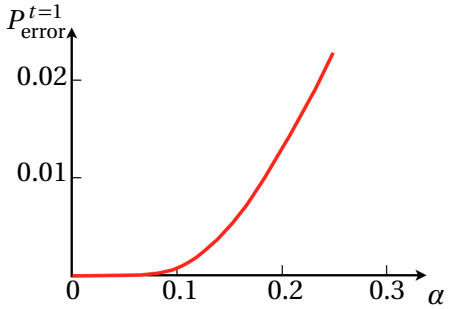


Figure 2.7: Dependence of the one-step error probability on the storage capacity α according to Equation (2.39).

The cross-talk term vanishes in this case, so that $P_{\text{error}}^{t=1} = 0$.

More importantly, the error probability defined in this Section refers only to the initial update, the first iteration. What happens in the next iteration, and after many iterations? Numerical experiments show that the error probability can be much higher in later iterations, because more errors tend to increase the probability of making another error. So the estimate $P_{\text{error}}^{t=1}$ is only a lower bound.

2.5 Energy function

Consider the long-time limit $t \to \infty$. Does the algorithm converge, as required by Equation (2.7)? This is perhaps the most important question in the analysis of neural-net algorithms, because an algorithm that does not converge to a meaningful solution is useless.

The standard way of analysing convergence of neural-net algorithms is to define an *energy function* $H(\boldsymbol{s})$ that has a minimum at the desired solution, $\boldsymbol{s} = \boldsymbol{x}^{(\nu)}$ say. We monitor how the energy function changes as we iterate, and keep track of the smallest values of H encountered, to find the minimum. If we store only one pattern, $p = 1$, then a suitable energy function is

$$H = -\frac{1}{2N}\left(\sum_{i=1}^{N} s_i x_i^{(1)}\right)^2 . \tag{2.43}$$

This function is minimal when $\boldsymbol{s} = \boldsymbol{x}^{(1)}$, *i.e.*, when $s_i = x_i^{(1)}$ for all i. It is customary to insert the factor $1/(2N)$, this does not change the fact that H is minimal at $\boldsymbol{s} = \boldsymbol{x}^{(1)}$.

A crucial point is that the asynchronous McCulloch-Pitts dynamics (2.6) *converges* to the minimum. This follows from the fact that H cannot increase under (2.6). To prove this important property, we begin by evaluating the expression on the r.h.s. of

Equation (2.43):

$$H = -\frac{1}{2N}\Big(\sum_i s_i x_i^{(1)}\Big)\Big(\sum_j s_j x_j^{(1)}\Big) = -\frac{1}{2}\sum_{ij}^{N}\underbrace{\Big(\frac{1}{N}x_i^{(1)}x_j^{(1)}\Big)}_{=w_{ij}} s_i s_j\,. \tag{2.44}$$

Using Hebb's rule (2.9) we find that the energy function (2.43) takes the form

$$H = -\frac{1}{2}\sum_{ij} w_{ij}s_i s_j\,. \tag{2.45}$$

This function has the same form as the energy function (or *Hamiltonian*) for certain physical models of magnetic systems consisting of interacting spins [18], where the interaction energy between spins s_i and s_j is $\frac{1}{2}(w_{ij}+w_{ji})s_i s_j$. Note that Hebb's rule (2.9) yields symmetric weights, $w_{ij} = w_{ji}$, and $w_{ii} > 0$. Setting the diagonal weights to zero does not change the fact that H is minimal at $\boldsymbol{s} = \boldsymbol{x}^{(1)}$ because $s_i^2 = 1$. So the diagonal weights just give a constant contribution to H, independent of $\boldsymbol{s}$.

The second step is to show that H cannot increase under the asynchronous McCulloch-Pitts dynamics (2.6), and we say that the energy function is a *Lyapunov function*, or *loss function*. To demonstrate that the energy function is a Lyapunov function, choose a neuron m and update it according to Equation (2.6). We denote the updated state of neuron m by s'_m:

$$s'_m = \operatorname{sgn}\Big(\sum_j w_{mj}s_j\Big), \tag{2.46}$$

all other neurons remain unchanged. There are two possibilities, either $s'_m = s_m$ or $s'_m = -s_m$. In the first case H remains unchanged, $H' = H$. Here H' refers to the value of the energy function after the update (2.46). The other case is $s'_m = -s_m$. In this case the energy function changes by the amount

$$\begin{aligned} H' - H &= -\frac{1}{2}\sum_{j\neq m}(w_{mj}+w_{jm})(s'_m s_j - s_m s_j) - \frac{1}{2}w_{mm}(s'_m s'_m - s_m s_m) \\ &= \sum_{j\neq m}(w_{mj}+w_{jm})s_m s_j\,. \end{aligned} \tag{2.47}$$

The sum goes over all neurons j that are connected to the neuron m, the one to be updated in Equation (2.46). Now if the weights are symmetric, $H'-H$ equals

$$H' - H = 2\sum_{j\neq m} w_{mj}s_m s_j = 2\sum_j w_{mj}s_m s_j - 2w_{mm}\,. \tag{2.48}$$

Since the sign of $\sum_j w_{mj} s_j$ is that of $s'_m = -s_m$, and since $w_{mm} > 0$ it follows that

$$H' - H < 0\,. \tag{2.49}$$

In other words,the value of H must decrease when the state of neuron m changes, $s'_m \neq s_m$. In summary,[1] H either remains constant under the asynchronous McCulloch-Pits dynamics ($s'_m = s_m$), or its value decreases ($s'_m \neq s_m$). Note that this does not hold for the synchronous dynamics (2.5), see Exercise 2.9. Since the energy function cannot increase under the asynchronous McCulloch-Pits dynamics, it must converge to minima of the energy function. For the energy function (2.43) this implies that the dynamics must either converge to the stored pattern or to its inverse. Both are attractors.

We assumed the thresholds to vanish, but the proof also works when the thresholds are not zero, in this case for the energy function

$$H = -\frac{1}{2}\sum_{ij} w_{ij} s_i s_j + \sum_i \theta_i s_i \tag{2.50}$$

in conjunction with the update rule $s'_m = \mathrm{sgn}(\sum_j w_{mj} s_j - \theta_m)$.

Up to now we considered only one stored pattern, $p = 1$. If we store more than one pattern [Hebb's rule (2.25)], the proof that (2.45) cannot increase under the McCulloch-Pitts dynamics works in the same way because no particular form of the weights w_{ij} was assumed, only that they must be symmetric, and that the diagonal weights must not be negative. In this case it follows that minima of the energy function must correspond to attractors, as illustrated schematically in Figure 2.8. The state space of the network – corresponding to all possible choices of $(s_1, \ldots s_N)$ – is illustrated drawn as a single axis, the x-axis. But when N is large, the state space is really very high dimensional.

However, when $p > 1$ some stored patterns may not be attractors. This follows from our analysis of the cross-talk term in Section 2.2 . If the cross-talk term causes errors for a certain stored pattern that is fed into the network, then this pattern is not located at a minimum of the energy function. To see this, note that Equations (2.25) and (2.45) give

$$H = -\frac{1}{2N}\sum_{\mu=1}^{p}\left(\sum_{i=1}^{N} s_i x_i^{(\mu)}\right)^2 . \tag{2.51}$$

[1] The derivation outlined here did not use the specific form of Hebb's rule (2.9), only that the weights are symmetric, and that $w_{mm} > 0$. It is clear that the argument works also when $w_{mm} = 0$. However, it does not work when $w_{mm} < 0$. In this case it is still true that H assumes a minimum at $\boldsymbol{s} = \boldsymbol{x}^{(1)}$, but H can increase under the update rule, so that convergence is not guaranteed. We therefore require that the diagonal weights are not negative.

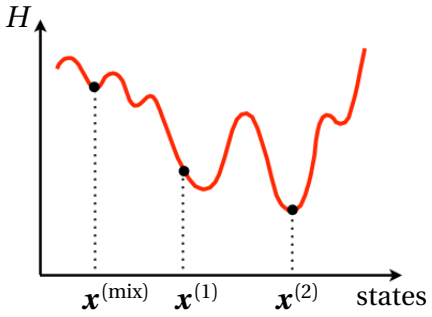


Figure 2.8: Minima in the energy function are attractors in state space. Not all minima correspond to stored patterns, and stored patterns need not correspond to minima.

While the energy function defined in Equation (2.43) has a minimum at $\boldsymbol{x}^{(1)}$, Equation (2.51) need not have a minimum at $\boldsymbol{x}^{(1)}$ (or at any other stored pattern), because a maximal value of $\left(\sum_{i=1}^{N} s_i x_i^{(1)}\right)^2$ may be compensated by terms stemming from other patterns. This happens rarely when p is small (Section 2.2).

Conversely there may be minima that do not correspond to stored patterns. Such states are referred to as *spurious states*. The network may converge to spurious states, this is undesirable but inevitable. It occurs even when there is only one stored pattern, as we saw in Section 2.2: the McCulloch-Pitts dynamics may converge to the inverted pattern. You can see this also in Equations (2.43) and (2.51). If $\boldsymbol{s} = \boldsymbol{x}^{(1)}$ is a local minimum of H, then so is $\boldsymbol{s} = -\boldsymbol{x}^{(1)}$.

2.6 Spurious states*

Stored patterns may be minima of the energy function (attractors), but they need not be. In addition there can be other local minima (spurious states), different from the stored patterns. For example, we saw in the previous Section that $-\boldsymbol{x}^{(1)}$ is a local minimum if $\boldsymbol{x}^{(1)}$ is a local minimum. This follows from the invariance of H under $\boldsymbol{s} \to -\boldsymbol{s}$.

There are other types of spurious states besides inverted patterns. An example are *mixed states, superpositions* of an odd number $2n+1$ of patterns. For $n = 1$, for example, the bits of a mixed state read:

$$x_i^{(\text{mix})} = \text{sgn}(\pm x_i^{(1)} \pm x_i^{(2)} \pm x_i^{(3)}). \tag{2.52}$$

The number of distinct mixed states increases as n increases. There are $2^{2n+1}\binom{p}{2n+1}$ mixed states that are superpositions of $2n+1$ out of p patterns, for $n = 1, 2, \ldots$ (Exercise 2.4).

It is difficult to determine under which circumstances the network dynamics converges to a certain mixed state. But we can at least check whether a mixed state is

$x_j^{(1)}$	$x_j^{(2)}$	$x_j^{(3)}$	$x_j^{(\mathrm{mix})}$	$s_j^{(1)}$	$s_j^{(2)}$	$s_j^{(3)}$
1	1	1	1	1	1	1
1	1	-1	1	1	1	-1
1	-1	1	1	1	-1	1
1	-1	-1	-1	-1	1	1
-1	1	1	1	-1	1	1
-1	1	-1	-1	1	-1	1
-1	-1	1	-1	1	1	-1
-1	-1	-1	-1	1	1	1

Table 2.1: Mixed states. Possible signs of $s_j^{(\mu)} = x_j^{(\mu)} x_j^{(\mathrm{mix})}$ for a given bit j (see text).

recognised by the network (although we do not want this to happen). As an example consider the mixed state

$$x_i^{(\mathrm{mix})} = \mathrm{sgn}(x_i^{(1)} + x_i^{(2)} + x_i^{(3)}). \tag{2.53}$$

To check whether this state is recognised, we must determine whether or not

$$\mathrm{sgn}\left(\frac{1}{N}\sum_{\mu=1}^{p}\sum_{j=1}^{N} x_i^{(\mu)} x_j^{(\mu)} x_j^{(\mathrm{mix})}\right) = x_i^{(\mathrm{mix})}, \tag{2.54}$$

under the update (2.6) using Hebb's rule (2.25). To this end we split the sum in the usual fashion

$$\frac{1}{N}\sum_{\mu=1}^{p}\sum_{j=1}^{N} x_i^{(\mu)} x_j^{(\mu)} x_j^{(\mathrm{mix})} = \sum_{\mu=1}^{3} x_i^{(\mu)} \frac{1}{N}\sum_{j=1}^{N} x_j^{(\mu)} x_j^{(\mathrm{mix})} + \text{cross-talk term}. \tag{2.55}$$

Let us ignore the cross-talk term for the moment and check whether the first term reproduces $x_i^{(\mathrm{mix})}$. To make progress we assume random patterns [Equation (2.29)], and compute the probability that the sum on the r.h.s of Equation (2.55) yields $x_i^{(\mathrm{mix})}$. The sum over j on the r.h.s. of Equation (2.55) is an average of $s_j^{(\mu)} = x_j^{(\mu)} x_j^{(\mathrm{mix})}$. Table 2.1 lists all possible combinations of bits, and the corresponding values of $s_j^{(\mu)}$. We see that on average $\langle s_j^{(\mu)} \rangle = \frac{1}{2}$, so that

$$\frac{1}{N}\sum_{\mu=1}^{p}\sum_{j=1}^{N} x_i^{(\mu)} x_j^{(\mu)} x_j^{(\mathrm{mix})} = \frac{1}{2}\sum_{\mu=1}^{3} x_i^{(\mu)} + \text{cross-talk term}. \tag{2.56}$$

Neglecting the cross-talk term and taking the signum function we see that $\boldsymbol{x}^{(\mathrm{mix})}$ is reproduced. So mixed states such as (2.53) are recognised, at least for small α, and it may happen that the network converges to these states.

Algorithm 1 pattern recognition with deterministic Hopfield net

1: store patterns $\boldsymbol{x}^{(\mu)}$ using Hebb's rule;
2: feed distorted pattern $\boldsymbol{x}$ into network by assigning $\boldsymbol{s}(t=0) \leftarrow \boldsymbol{x}$;
3: **for** $t = 1, \ldots, T$ **do**
4: choose a value of m and update $s_m(t) \leftarrow \operatorname{sgn}\big(\sum_{j=1}^{N} w_{mj} s_j(t-1)\big)$;
5: **end for**
6: read out pattern $\boldsymbol{s}(T)$;

Finally, for large values of p there are local minima of H that are not correlated with any number of the stored patterns $x_j^{(\mu)}$. Such *spin-glass* states are discussed further in the book by Hertz, Krogh and Palmer [1].

2.7 Summary

We analysed how Hopfield nets recognise (or *retrieve*) patterns using Algorithm 1. Hopfield nets are networks of McCulloch-Pitts neurons. Their layout is defined by connection strengths (weights) w_{ij}, chosen according to Hebb's rule. The w_{ij} are symmetric, and the network is in general fully connected. Hebb's rule ensures that stored patterns are recognised, at least most of the time if the number of patterns is not too large. A single-step estimate for the error probability was given in Section 2.2. If one iterates several steps, the error probability is generally much larger, but it is difficult to evaluate. For stochastic Hopfield nets the steady-state error probability can be estimated more easily, because the dynamics converges to a definite steady state.

2.8 Exercises

2.1 Modified Hebb's rule. Show that the modified Hebb's rule (2.26) satisfies Equation (2.8) if we store only one pattern, $p = 1$.

2.2 Orthogonal patterns. Show that the cross-talk term vanishes for orthogonal patterns, so that $P_{\text{error}}^{t=1} = 0$. Show that this works for both versions of Hebb's rule, (2.25) as well as (2.26).

2.3 Cross-talk term. Expression (2.33) for the cross-talk term was derived using modified Hebb's rule, Equation (2.26). How does Equation (2.33) change if you use the rule (2.25) instead? Show that the distribution of $C_i^{(\nu)}$ then acquires a non-zero mean, obtain an estimate for this mean value, and compute the one-step error

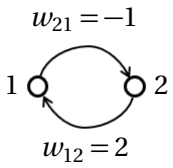


Figure 2.9: Two neurons with asymmetric connections.

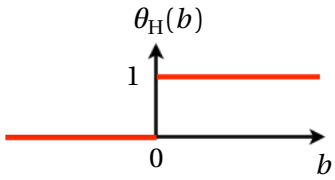


Figure 2.10: Heaviside function (Exercise 2.8).

probability. Show that your result approaches (2.39) for small values of α. Explain why your result is different from (2.39) for large α.

2.4 Mixed states. Explain why there are no mixed states that are superpositions of an even number of stored patterns. Show that there are $2^{2n+1}\binom{p}{2n+1}$ mixed states that are superpositions of $2n+1$ out of p patterns, for $n=1,2,\ldots$.

2.5 One-step error probability for mixed states. Write a computer program implementing the asynchronous deterministic dynamics of a Hopfield net to determine the one-step error probability for the mixed state (2.53). Plot how the one-step error probability depends on α for $N=50$ and $N=100$. Repeat this exercise for mixed patterns that are superpositions of the bits of 5 and 7 patterns.

2.6 Energy function. Figure 2.9 shows a network with two neurons with asymmetric weights, $w_{12}=2$ and $w_{21}=-1$. Show that the energy function $H=-\frac{w_{12}+w_{21}}{2}s_1s_2$ can increase under the asynchronous McCulloch-Pitts rule $s_2'=\mathrm{sgn}(w_{21}s_1)$.

2.7 Higher-order Hopfield nets. Determine under which conditions the energy function $H=-\frac{1}{2}\sum_{ij}w^{(2)}_{ij}s_is_j-\frac{1}{6}\sum_{ijk}w^{(3)}_{ijk}s_is_js_k$ is a Lyapunov function for the modified asynchronous McCulloch-Pitts dynamics: $s_m'=\mathrm{sgn}(b_m)$ for neuron m with $b_m=\partial H/\partial s_m$.

2.8 Hebb's rule and energy function for $0/1$ units. Suppose that the state of a neuron takes the values 0 (inactive) and 1 (active). The corresponding asynchronous update rule is $n_m'=\theta_{\rm H}(\sum_j w_{mj}n_j-\mu_m)$ with threshold μ_m, and where $\theta_{\rm H}(b)$ is the Heaviside function, equal to 0 is $b<0$ and equal to 1 if $b\geq 0$ (Figure 2.10). Write down Hebb's rule for such $0/1$ units and show that if one stores only one pattern, then this pattern is recognised. Show that $H=-\frac{1}{2}\sum_{ij}w_{ij}n_in_j+\sum_i\mu_in_i$ cannot

increase under the asynchronous update rule (it is assumed that the weights are symmetric, and that $w_{ii} \geq 0$). See Ref. [19].

2.9 Energy function and synchronous dynamics. Analyse how the energy function (2.45) changes under the synchronous dynamics (2.5). Show that the energy function can increase, even though the weights are symmetric and the diagonal weights are zero.

2.10 Continuous Hopfield net. Hopfield [20] also analysed a version of his model with continuous-time dynamics. Here we use $\tau \frac{\mathrm{d}}{\mathrm{d}t} n_i = -n_i + g(\sum_j w_{ij} n_j - \theta_i)$ with $g(b) = (1+\mathrm{e}^{-b})^{-1}$ (this dynamical equation is slightly different from the one used by Hopfield [20]). Show that the energy function $E = -\frac{1}{2}\sum_{ij} w_{ij} n_i n_j + \sum_i \theta_i n_i + \sum_i \int_0^{n_i} \mathrm{d}n\, g^{-1}(n)$ cannot increase under the network dynamics if the weights are symmetric. It is not necessary to assume that $w_{ii} \geq 0$.

2.11 Hopfield net with four neurons. The pattern shown in Fig. 2.11 is stored in a Hopfield net using Hebb's rule $w_{ij} = \frac{1}{N} x_i^{(1)} x_j^{(1)}$. There are 2^4 four-bit patterns. Apply each of these to the Hopfield net, and perform one synchronous update. List the patterns you obtain and discuss your results.

2.12 Recognising letters with a Hopfield net. The five patterns in Figure 2.12 each have $N = 32$ bits. Store the patterns $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ in a Hopfield net using Hebb's rule $w_{ij} = \frac{1}{N}\sum_{\mu=1}^{2} x_i^{(\mu)} x_j^{(\mu)}$. Which of the patterns in Figure 2.12 remain unchanged after one synchronous update with $s_i' = \mathrm{sgn}(\sum_{j=1}^{N} w_{ij} s_j)$? *Hint*: read off $\sum_{j=1}^{N} x_j^{(\mu)} x_j^{(\nu)}$ from the Hamming distances between the patterns, and use this quantity to express the local fields $b_i^{(\mu)}$ as linear combinations of $x_i^{(1)}$ and $x_i^{(2)}$.

2.13 Diluted Hopfield net. In the diluted Hopfield net with N neurons, only a fraction $\frac{K}{N} \ll 1$ of the weights w_{ij} is active: $w_{ij} = \frac{K_{ij}}{K}\sum_{\mu=1}^{p} x_i^{(\mu)} x_j^{(\mu)}$ where K_{ij} are random numbers, equal to $K_{ij} = 1$ with probability $\frac{K}{N}$ and zero otherwise. The parameter K determines the average number of connections to neuron i, $\langle \sum_{j=1}^{N} K_{ij} \rangle_{\mathrm{c}} = K$, where $\langle \ldots \rangle_{\mathrm{c}}$ denotes the average over random realisations of K_{ij}. Derive the approx-

Figure 2.11: The pattern $\boldsymbol{x}^{(1)}$ has $N = 4$ bits, $x_1^{(1)} = 1$, and $x_i^{(1)} = -1$ for $i = 2,3,4$. Exercise 2.11.

imate self-consistent equation [1]

$$m_\nu = \mathrm{erf}\Big(\frac{m_\nu}{\sqrt{2p/K}}\Big) \tag{2.57}$$

for the order parameter $m_\nu = \langle \frac{1}{N}\sum_{i=1}^{N} x_i^{(\nu)} \langle s_i \rangle_c \rangle$ in the deterministic limit, assuming that $K \gg 1$ and $1 \ll p \ll N$. Here the outer average is over random patterns. *Hint*: show that the distribution of the cross-talk term $\langle C_i^{(\nu)} \rangle_c$ over input patterns is Gaussian with mean zero. Determine how the variance of $\langle C_i^{(\nu)} \rangle_c$ depends upon K.

2.14 Mixed states. Consider p random patterns $\boldsymbol{x}^{(\mu)}$ ($\mu = 1,\ldots,p$) with N bits $x_i^{(\mu)}$ ($i = 1,\ldots,N$), equal to 1 or -1 with probability $\frac{1}{2}$. Store the patterns in a deterministic Hopfield net using Hebb's rule $w_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} x_i^{(\mu)} x_j^{(\mu)}$. In the limit of $N \gg 1$, $p \gg 1$, $p \ll N$, show that the network recognises bit $x_i^{(\mathrm{mix})}$ of the *mixed state* $\boldsymbol{x}^{(\mathrm{mix})}$ with bits

$$x_i^{(\mathrm{mix})} = \mathrm{sgn}\big(x_i^{(1)} + x_i^{(2)} + x_i^{(3)}\big), \tag{2.58}$$

after a single asynchronous update $s_i \leftarrow \mathrm{sgn}(\sum_{j=1}^{N} w_{ij} s_j)$. Follow the steps outlined below. Firs, feed the mixed state (2.58) to the network. Use the weights w_{ij} you obtained by applying Hebb's rule and express $\sum_{j=1}^{N} w_{ij} x_j^{(\mathrm{mix})}$ in terms of $\langle s_\mu \rangle$, defined by $\langle s_\mu \rangle = \frac{1}{N}\sum_{j=1}^{N} x_j^{(\mu)} x_j^{(\mathrm{mix})}$, for $\mu = 1 \ldots p$. Second, assume that the bits $x_i^{(\mu)}$ are independent random numbers, equal to 1 or -1 with equal probabilities. What is the value of $\langle s_\mu \rangle$ for $\mu = 1$, 2 and 3? What is the value for $\langle s_\mu \rangle$ for $\mu > 3$? Third, rewrite your result as a sum of two terms. The first term is a sum over $\mu = 1,2,3$. The second term is the cross-talk term, a sum over the remaining values of μ. Explain why the cross-talk term can be neglected in the limit stated above. Fourth, combine your results to show that the network recognises the mixed state (2.58).

2.15 XOR function. The Boolean XOR function takes two binary inputs. For the inputs $[-1,-1]$ and $[1,1]$ the function evaluates to -1, for the other two to $+1$. Try to encode the XOR function in a Hopfield net with three neurons by storing the patterns $[-1,-1,-1]$, $[1,1,-1]$, $[-1,1,1]$, and $[1,-1,1]$ using Hebb's rule. Test whether the patterns are recognised or not. Discuss your findings.

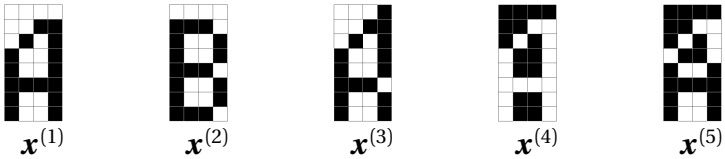


Figure 2.12: Each of the five patterns consists of 32 bits $x_i^{(\mu)}$. A black pixel i in pattern μ corresponds to $x_i^{(\mu)} = 1$, a white one to $x_i^{(\mu)} = -1$. Exercise 2.12.

3 Stochastic Hopfield nets

Two related problems became apparent in the previous Chapter. First, the Hopfield dynamics may get stuck in spurious minima. In fact, if there is a local minimum downhill from a given initial state, between this state and the correct attractor, then the dynamics arrests in the local minimum, so that the algorithm fails to converge to the correct attractor. Second, the energy function usually is a strongly varying function over a high-dimensional state space. Therefore it is difficult to predict the long-time dynamics of the network. Which is the first local minimum encountered on the down-hill path that the network takes?

Both problems are solved by introducing an element of stochasticity into the dynamics. This is a trick that works for many neural-network algorithms. In general, however, it is very challenging to analyse the stochastic dynamics. For the Hopfield network, by contrast, much is known. The reason is that the stochastic Hopfield network is closely related to systems studied in statistical mechanics, so-called spin glasses. Like these systems – and like many other physical systems – the stochastic Hopfield network exhibits an *order-disorder transition*. This transition becomes sharp in the limit of large N. This has important consequences. It may be that the network produces satisfactory results for a given number of patterns with a certain number of bits. But if one tries to store just one more pattern, the network may fail to recognise anything. The goal of this Chapter is to explain why this occurs, and how it can be avoided.

3.1 Stochastic dynamics

The asynchronous update rule (2.6) is called *deterministic*, because a given set of states s_j determines the outcome of the update of neuron m. To introduce noise, one replaces the rule (2.6) by an asynchronous *stochastic* rule:

$$s_m' = \begin{cases} +1 & \text{with probability} \quad p(b_m), \\ -1 & \text{with probability} \quad 1-p(b_m), \end{cases} \tag{3.1a}$$

with local field $b_m = \sum_j w_{mj} s_j - \theta_m$, and where the probability $p(b)$ is given by:

$$p(b) = \frac{1}{1+\mathrm{e}^{-2\beta b}}\,. \tag{3.1b}$$

The function $p(b)$ is plotted in Figure 3.1. The parameter β is the noise parameter. When β is large the noise level is small. In particular one obtains the deterministic dynamics (2.6) as β tends to infinity. In this limit, the function $p(b)$ approaches zero

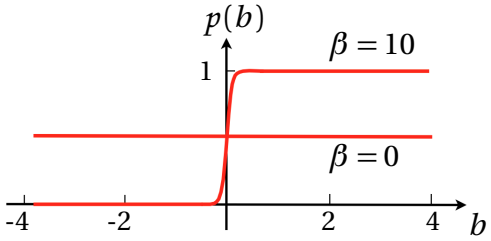


Figure 3.1: Probability function (3.1b) used in the definition of the stochastic rule (3.1), plotted for $\beta = 10$ and $\beta = 0$.

if b is negative, and it tends to unity if b is positive. So for $\beta \to \infty$, the stochastic update rule (3.1b) is the same as the deterministic rule (2.6). In the opposite limit, when $\beta = 0$, the function $p(b)$ simply equals $\frac{1}{2}$. In this case s_i is updated to -1 or $+1$ randomly, with equal probability. The dynamics does not depend upon the stored patterns (contained in the local field b_i).

The idea is to run the network for a small but finite noise level, that is at large values of β. Then the dynamics is very similar to the deterministic Hopfield dynamics analysed in the previous Chapter. But the noise allows the system to sometimes also go uphill, making it possible to escape spurious minima. Since the dynamics is stochastic, it is necessary to rephrase the convergence criterion (2.3). This is discussed next.

3.2 Order parameters

If we feed one of the stored patterns, $\boldsymbol{x}^{(1)}$ for example, then we want the stochastic dynamics to stay in the vicinity of $\boldsymbol{x}^{(1)}$. This can only work if the noise is weak enough, and even then it is not guaranteed. At time step t, bit i is correct if $s_i(t)x_i^{(1)} = 1$. All bits are correct when $\sum_{i=1}^{N} s_i(t)x_i^{(1)} = N$, otherwise the sum takes a value smaller than N. One measures success by averaging $\frac{1}{N}\sum_{i=1}^{N} s_i(t)x_i^{(1)}$ over the asynchronous stochastic dynamics of the network from $t = 0$ to $t = T$, for given bits $x_i^{(\mu)}$:

$$m_\mu(T) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} s_i(t)x_i^{(\mu)}\right). \tag{3.2a}$$

Since we decided to feed pattern $\boldsymbol{x}^{(1)}$ to the network, we have $m_1(t=0) = 1$ initially, and we want that $m_1(t)$ remains close to unity, so that the network recognises the pattern $\boldsymbol{x}^{(1)}$. In practice, the quantity $\frac{1}{N}\sum_{i=1}^{N} s_i(t)x_i^{(1)}$ settles into a *steady state*, where it fluctuates around a mean value with a definite distribution that becomes independent of the iteration number t. If the network works well, the finite-time

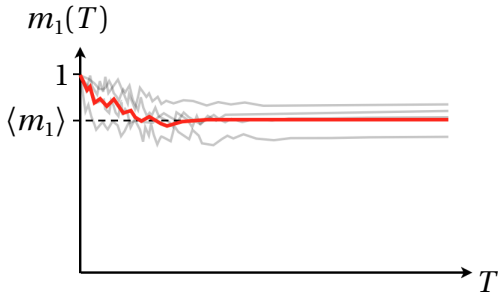


Figure 3.2: Illustrates how the finite-time average $m_1(T)$ depends upon the total iteration time T. The light grey lines show different realisations of $m_1(T)$ for different realisations of random patterns stored in the network, at a large but finite value of N. The thick red line is the average of $m_1(T)$ over the different realisations of random patterns.

average $m_1(T)$ converges to a value of order unity after a *transient* (Figure 3.2),

$$m_1 \equiv \lim_{T\to\infty} m_1(T). \tag{3.2b}$$

This limit, m_1, is called the *order parameter*. Since there is noise, the order parameter m_1 is usually smaller than unity.

Figure 3.2 also illustrates a subtlety. For finite values of N, the order parameter m_1 depends upon the stored patterns. Different realisations $\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(p)}$ of random patterns yield different values of m_1. In the limit of $N\to\infty$ this problem does not occur, the order parameter m_1 is independent of the stored patterns. We say that the system is *self averaging* in the limit $N\to\infty$. When N is finite, however, this is not the case. To obtain a definite value for the order parameter, one usually averages m_1 over different realisations of random patterns stored in the network (thick red line in Figure 3.2). The dashed line in Figure 3.2 shows $\langle m_1 \rangle$.

The other limits, $m_\mu = \lim_{T\to\infty} m_\mu(T)$ for $\mu > 1$, are expected to be small. This is certainly true for random patterns with many independent bits. Since the bits of the patterns $\boldsymbol{x}^{(2)}$ to $\boldsymbol{x}^{(p)}$ are independent from those of $\boldsymbol{x}^{(1)}$, the individual terms in the sum over i in Equation (3.2) cancel approximately upon summation if $s_i(t) \approx x_i^{(1)}$. In summary, if we feed pattern $\boldsymbol{x}^{(1)}$ and if the network works well, we expect that

$$m_\mu \approx \begin{cases} 1 & \text{if} \quad \mu = 1, \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

in the limit of large N. Whether this is the case or not depends on the values of p, N, and β. In the next Sections we determine how m_1 depends on these parameters.

3.3 Mean-field theory

The order parameter is defined as an average over the stochastic dynamics of the network in its steady state (Figure 3.2). It is a challenging task to compute this average because all neurons interact with each other in a nonlinear fashion. Consider neuron number i. The fate of s_i is determined by its local field b_i, through Equation (3.1). The difficulty is that the local field in turn depends on the states s_j of all other neurons in the network:

$$b_i(t) = \sum_{j=1}^{N} w_{ij} s_j(t) \tag{3.4}$$

(setting the thresholds to zero, consistent with Hebb's rule). When N is large, we may assume that $b_i(t)$ remains essentially constant in the steady state, independent of t, because fluctuations of $s_j(t)$ average out when summing over j:

$$b_i(t) = \langle b_i \rangle + \text{small fluctuations}. \tag{3.5}$$

The average local field $\langle b_i \rangle$ is called the *mean field*, and theories that neglect the small fluctuations in Equation (3.5) are called *mean-field theories*. Let us ignore the fluctuations in Equation (3.5) and write

$$b_i(t) \approx \langle b_i \rangle = \sum_{j=1}^{N} w_{ij} \langle s_j \rangle = \frac{1}{N} \sum_{\mu} \sum_{j \neq i} x_i^{(\mu)} x_j^{(\mu)} \langle s_j \rangle \tag{3.6}$$

for given patterns $\boldsymbol{x}^{(\mu)}$. We calculate the average $\langle s_j \rangle$ on the r.h.s. of Equation (3.6) self-consistently, using the update rule (3.1):

$$\begin{aligned} \langle s_i \rangle &= \text{Prob}(s_i = +1) - \text{Prob}(s_i = -1) = p(\langle b_i \rangle) - [1 - p(\langle b_i \rangle)] \\ &= \frac{e^{\beta \langle b_i \rangle}}{e^{\beta \langle b_i \rangle} + e^{-\beta \langle b_i \rangle}} - \frac{e^{-\beta \langle b_i \rangle}}{e^{\beta \langle b_i \rangle} + e^{-\beta \langle b_i \rangle}} = \tanh(\beta \langle b_i \rangle). \end{aligned} \tag{3.7}$$

In summary, one finds a set of N non-linear self-consistent equations for $\langle s_i \rangle$,

$$\langle s_i \rangle = \tanh(\beta \langle b_i \rangle) \quad \text{with} \quad \langle b_i \rangle = \frac{1}{N} \sum_{\mu} \sum_{j \neq i} x_i^{(\mu)} x_j^{(\mu)} \langle s_j \rangle, \tag{3.8}$$

for given patterns $\boldsymbol{x}^{(\mu)}$. The mean-field equations (3.8) were obtained neglecting fluctuations in the local field. An equivalent but slightly different description of the mean-field approximation is this: suppose we average s_i over the dynamics (3.1) at fixed $s_j \neq s_i$, and then we average all s_i over the dynamics. This gives $\langle s_i \rangle = \langle \tanh(\beta b_i) \rangle$. Comparing with Equation (3.8), we see that the mean-field approximation corresponds to approximating $\langle \tanh(\beta b_i) \rangle \approx \tanh(\beta \langle b_i \rangle)$.

Our aim is now to solve these equations to determine the time averages $\langle s_i \rangle$, and then the order parameters

$$m_\mu = \frac{1}{N} \sum_{j=1}^{N} \langle s_j \rangle x_j^{(\mu)}. \tag{3.9}$$

If we initially feed pattern $\boldsymbol{x}^{(1)}$ we want that $m_1 \approx 1$ while $m_\mu \approx 0$ for $\mu \neq 1$. To determine under which circumstances this works, we express the mean field in terms of the order parameters m_μ:

$$\langle b_i \rangle = \frac{1}{N} \sum_{\mu=1}^{p} \sum_{j \neq i} x_i^{(\mu)} x_j^{(\mu)} \langle s_j \rangle \approx \sum_{\mu} x_i^{(\mu)} m_\mu \,. \tag{3.10}$$

The last equality is only approximate because the j-sum in the definition of m_μ contains the term $j = i$. Whether or not to include this term makes only a small difference to m_μ, in the limit of large N.

Let us calculate m_1 assuming that $m_\mu \approx 0$ for $\mu \neq 1$, so that the corresponding terms in Equation (3.10) are negligible. It might happen that the small terms add up to make a difference, but let us assume here that p is small enough so that this does not happen. To ensure this, we must require that

$$\alpha = \frac{p}{N} \tag{3.11}$$

is small enough, at most $(\log N)/N$ for large N [21]. Keeping only the first term in Equation (3.10) yields together with Equation (3.8):

$$\langle s_i \rangle = \tanh(\beta \langle b_i \rangle) \approx \tanh(\beta m_1 x_i^{(1)}). \tag{3.12}$$

Applying the definition (3.9) of the order parameter one finds

$$m_1 = \frac{1}{N} \sum_{i=1}^{N} \tanh\left(\beta m_1 x_i^{(1)}\right) x_i^{(1)}. \tag{3.13}$$

Using that $\tanh(z) = -\tanh(-z)$ as well as the fact that the bits $x_i^{(\mu)}$ can only assume the values ± 1, one obtains:

$$m_1 = \tanh(\beta m_1). \tag{3.14}$$

This is a self-consistent equation for m_1. For $\beta \to 0$ there it has the solution $m_1 = 0$, but this is not the desired one because $m_1 = 0$ means that $\boldsymbol{x}^{(1)}$ is not recognised. For $\beta \to \infty$, by contrast, there are three solutions, $m_1 = 0, \pm 1$. Figure 3.3 shows results

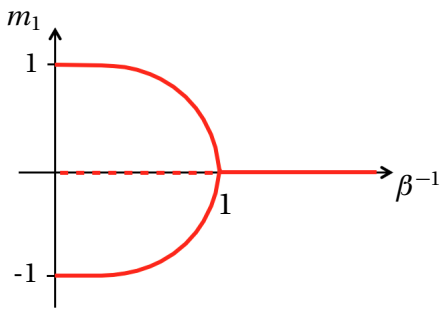


Figure 3.3: Solutions of the mean-field equation (3.14). The critical noise level is $\beta_c = 1$. The dashed line corresponds to an unstable solution.

of the numerical evaluation of Equation (3.14) for intermediate values of β. Below a critical noise level there are three solutions, namely for β larger than

$$\beta_c = 1\,. \tag{3.15}$$

For $\beta > \beta_c$, the solution $m_1 = 0$ is *unstable* (this can be shown by computing the derivatives of the *free energy* of the Hopfield network [1]). Even if we were to start with an initial condition that corresponds to $m_1 = 0$, the network would not stay there. The other two solutions are *stable*: when the network is initialised close to $\boldsymbol{x}^{(1)}$, then it converges to $m_1 = O(1)$.

The symmetry of the problem dictates that there must also be a solution with $m_1 = -m$ at small noise levels. This solution corresponds to the inverted pattern $-\boldsymbol{x}^{(1)}$ (Section 2.6). If we start in the vicinity of $\boldsymbol{x}^{(1)}$, then the network is unlikely to converge to $-\boldsymbol{x}^{(1)}$, provided that N is large enough. The probability of $\boldsymbol{x}^{(1)} \to -\boldsymbol{x}^{(1)}$ vanishes very rapidly as N increases and as the noise level decreases. If this transition were to happen in a simulation, the network would then stay near $-\boldsymbol{x}^{(1)}$ for a very long time. Consider the limit where T tends to ∞ at a finite but possibly large value of N. Then the network would (at a very small rate) jump back and forth between $\boldsymbol{x}^{(1)}$ and $-\boldsymbol{x}^{(1)}$, so that the order parameter would average to zero. This shows that the limits of large N and large T do not commute

$$\lim_{T\to\infty}\lim_{N\to\infty} m_1(T) \neq \lim_{N\to\infty}\lim_{T\to\infty} m_1(T)\,. \tag{3.16}$$

In practice the interesting limit is the left one, that of a large network run for a time T much longer than the initial transient, but not infinite. This is precisely where the mean-field theory applies. It correponds to taking the limit $N \to \infty$ first, at finite but large T. This describes simulations where the transition $\boldsymbol{x}^{(1)} \to -\boldsymbol{x}^{(1)}$ does not occur.

In summary, Equation (3.14) predicts that the order parameter converges to a definite value, m_1, independent of the stored patterns when N is large enough.

Figure 3.2 shows that the order parameter converges for large but finite values of N. However, the limiting value does depend on the stored patterns, as mentioned above. The system is not self averaging. When N is finite we should therefore average the result over different realisations of the stored patterns.

The value of $\langle m_1 \rangle$ determines the average number of correctly retrieved bits in the steady state:

$$\langle N_{\text{correct}} \rangle = \Big\langle \frac{1}{2} \sum_{i=1}^{N} \Big(1 + \langle s_i \rangle x_i^{(1)} \Big) \Big\rangle, \tag{3.17}$$

because $\frac{1}{2}(1 + s_i x_i^{(1)}) = 1$ if $x_i^{(1)}$ is retrieved correctly, and equal to zero otherwise. The outer average is over different realisations of random patterns (the inner average is over the network dynamics). It follows from Equation (3.17) that

$$\langle N_{\text{correct}} \rangle = \frac{N}{2}(1 + \langle m_1 \rangle). \tag{3.18}$$

Since $m_1 \to 1$ as $\beta \to \infty$ we see all bits are correctly retrieved in this limit,

$$\langle N_{\text{correct}} \rangle \to N\,. \tag{3.19}$$

This is expected since the stored patterns $\boldsymbol{x}^{(\mu)}$ are recognised for small enough values of α in the deterministic limit, because the cross-talk term is negligible. But it is important to know that the stochastic dynamics *slows down* as the noise level tends to zero. The lower the noise level, the longer the network remains stuck in local minima, so that it takes longer time to reach the steady state, and to sample the steady-state statistics of H.

In the opposite limit $\beta \to 0$ we have $m_1 \to 0$, so that

$$\langle N_{\text{correct}} \rangle \to \tfrac{1}{2} N\,. \tag{3.20}$$

In this noise-dominated limit the stochastic network ceases to function. If one were to assign N bits entirely randomly, then half of them would be correct, on average, so that $\langle N_{\text{correct}} \rangle = \frac{1}{2} N$.

We define the error probability in the steady state as

$$P_{\text{error}}^{t=\infty} = \frac{N - \langle N_{\text{correct}} \rangle}{N}\,. \tag{3.21}$$

From Equation (3.18) we find

$$P_{\text{error}}^{t=\infty} = \tfrac{1}{2}(1 - \langle m_1 \rangle). \tag{3.22}$$

In the deterministic limit the steady-state error probability approaches zero as m_1 tends to one. Let us compare this result with the one-step error probability $P_{\text{error}}^{t=1}$ derived in Chapter 2 in the deterministic limit. We should take the limit $\alpha = p/N \to 0$ in Equation (2.39) because the result (3.22) was derived assuming that α is very small. In this limit we find that the one-step and the steady-state error probabilities agree (they are both equal to zero). Above the critical noise level, for $\beta < \beta_c = 1$, the order parameter vanishes. In this case $P_{\text{error}}^{t=\infty}$ equals $\frac{1}{2}$. So when the noise is too large the network fails.

It is important to note that noise can also help, because mixed states have lower critical noise levels than the stored patterns $x_i^{(\mu)}$. This can be seen as follows [1, 22]. To derive the above mean-field result we assumed that $m_\mu = m\delta_{\mu 1}$. Mixed states correspond to solutions where an odd number of components of $\boldsymbol{m}$ is non-zero, for example:

$$\boldsymbol{m} = \begin{bmatrix} m \\ m \\ m \\ 0 \\ \vdots \end{bmatrix}. \tag{3.23}$$

Neglecting the cross-talk term, the mean-field equation reads

$$\langle s_i \rangle = \tanh\Big(\beta \sum_{\mu=1}^{p} m_\mu x_i^{(\mu)}\Big). \tag{3.24}$$

In the limit of $\beta \to \infty$, the $\langle s_i \rangle$ converge to the mixed states (2.53) when $\boldsymbol{m}$ is given by Equation (3.23). Using the definition of m_μ and averaging over the bits of the random patterns one finds:

$$m_\mu = \Big\langle x_i^{(\mu)} \tanh\Big(\beta \sum_{\nu=1}^{p} m_\nu x_i^{(\nu)}\Big)\Big\rangle. \tag{3.25}$$

The numerical solution of Equation (3.25) shows that there is a non-zero solution for $\beta > \beta_c = 1$. Yet this solution is unstable close to the critical noise level, more precisely for $1 < \beta < 2.17$ [22]. In other words, the mixed states have a lower critical noise level, $\beta^{-1} = 1/2.17$.

3.4 Storage capacity*

The preceding analysis replaced the sum (3.10) by its first term, $x_i^{(1)} m_1$. This corresponds to neglecting the cross-talk term. We expect that this can only work if p/N

is small enough. The influence of the cross-talk term was studied in Section 2.2, where the storage capacity

$$\alpha = \frac{p}{N}$$

was defined. When we computed $P_{\text{error}}^{t=1}$ in Section 2.2, only the first *initial* update step was considered, because it was too difficult to analyse the long-time limit of the deterministic dynamics. It is expected that the error probability increases as t increases, at least when α is large enough so that the cross-talk term matters.

The stochastic dynamics is simpler to analyse in the long-time limit because it approaches a steady state. The remainder of this Section describes the mean-field analysis of the steady state for larger values of α. We store p patterns in the network using Hebb's rule (2.26) and feed pattern $\boldsymbol{x}^{(1)}$ to the network. The aim is to determine the order parameter m_1 and the corresponding error probability in the steady state for $p \sim N$, so that α remains finite as $N \to \infty$. In this case we can no longer approximate the sum in Equation (3.10) just by its first term, because the other terms for $\mu > 1$ may sum up to a contribution that is of the same order as m_1. Instead we must evaluate all m_μ to compute the mean field $\langle b_i \rangle$.

The relevant calculation is summarised in Chapter 4 of Geszti [23]. It is also outlined in Section 2.5 of Hertz, Krogh and Palmer [1]. The remainder of this Section follows this outline quite closely. One starts by rewriting the mean-field equations (3.8) in terms of the order parameters m_μ. Using

$$\langle s_i \rangle = \tanh(\beta \sum_\mu x_i^{(\mu)} m_\mu) \tag{3.26}$$

we find

$$m_\nu = \frac{1}{N} \sum_i x_i^{(\nu)} \langle s_i \rangle = \frac{1}{N} \sum_i x_i^{(\nu)} \tanh\left(\beta \sum_\mu x_i^{(\mu)} m_\mu \right). \tag{3.27}$$

This coupled set of p non-linear equations is equivalent to the mean-field equation (3.8).

Now feed pattern $\boldsymbol{x}^{(1)}$ to the network. The strategy of solving Equation (3.27) is to assume that the network stays close to the pattern $\boldsymbol{x}^{(1)}$ in the steady state, so that m_1 remains of order unity. Since we cannot simply approximate the sum over μ on the r.h.s. of Equation (3.27) by its first term, we need to estimate the other order parameters m_μ for $\mu \neq 1$ as well. The trick is to estimate these order parameters assuming random patterns, so that the m_μ, $\mu = 2, \ldots, p$, become random numbers that fluctuate around zero with variance $\langle m_\mu^2 \rangle$ (this average is over random patterns). We use Equation (3.27) to compute the variance approximately. In the μ-sum on the r.h.s of Equation (3.27) we must treat the term $\mu = \nu$ separately (because the index ν appears also on the l.h.s. of this equation). Also the term $\mu = 1$ must be treated

separately, as before, because $\mu = 1$ is the index of the pattern that was fed to the network.

As a consequence, the calculations of m_1 and m_ν for $\nu \neq 1$ proceed slightly differently. We start with the second case and write

$$\begin{aligned} m_\nu &= \frac{1}{N}\sum_i x_i^{(\nu)} \tanh\Big(\beta\, x_i^{(1)} m_1 + \beta\, x_i^{(\nu)} m_\nu + \beta \sum_{\substack{\mu\neq 1\\ \mu\neq\nu}} x_i^{(\mu)} m_\mu\Big) \\ &= \frac{1}{N}\sum_i x_i^{(\nu)} x_i^{(1)} \tanh\Big(\underbrace{\beta\, m_1}_{①} + \underbrace{\beta\, x_i^{(1)} x_i^{(\nu)} m_\nu}_{②} + \underbrace{\beta \sum_{\substack{\mu\neq 1\\ \mu\neq\nu}} x_i^{(\mu)} x_i^{(1)} m_\mu}_{③}\Big). \end{aligned} \tag{3.28}$$

Now consider the three terms in the argument of tanh(…). The term ① is of order unity, it is independent of N. The term ③ may be of the same order, because the sum over μ contains $\sim pN$ terms. The term ②, by contrast, is small for large values of N. Therefore it is a good approximation to Taylor-expand the argument of tanh(…):

$$\tanh\Big(① + ② + ③\Big) \approx \tanh\Big(① + ③\Big) + ② \frac{\mathrm{d}}{\mathrm{d}x} \tanh\Big|_{①+③} + \dots . \tag{3.29}$$

Using $\frac{\mathrm{d}}{\mathrm{d}x}\tanh(x) = 1 - \tanh^2(x)$ one gets

$$\begin{aligned} m_\nu &= \frac{1}{N}\sum x_i^{(\nu)} x_i^{(1)} \tanh\Big(\underbrace{\beta\, m_1}_{①} + \underbrace{\beta \sum_{\substack{\mu\neq 1\\ \mu\neq\nu}} x_i^{(\mu)} x_i^{(1)} m_\mu}_{③}\Big) \\ &+ \frac{1}{N}\sum_i x_i^{(\nu)} x_i^{(1)} \underbrace{\beta\, x_i^{(1)} x_i^{(\nu)} m_\nu}_{②} \Big[1 - \tanh^2\Big(\beta\, m_1 + \beta \sum_{\substack{\mu\neq 1\\ \mu\neq\nu}} x_i^{(\mu)} x_i^{(1)} m_\mu\Big)\Big]. \end{aligned} \tag{3.30}$$

Using the fact that $x^{(\mu)} = \pm 1$ and thus $[x_i^{(\mu)}]^2 = 1$, this expression simplifies:

$$\begin{aligned} m_\nu &= \frac{1}{N}\sum_i x_i^{(\nu)} x_i^{(1)} \tanh\Big(\beta\, m_1 + \beta \sum_{\substack{\mu\neq 1\\ \mu\neq\nu}} x_i^{(\mu)} x_i^{(1)} m_\mu\Big) + \\ &+ \beta\, m_\nu \frac{1}{N}\sum_i \Big[1 - \tanh^2\Big(\beta\, m_1 + \beta \sum_{\substack{\mu\neq 1\\ \mu\neq\nu}} x_i^{(\mu)} x_i^{(1)} m_\mu\Big)\Big]. \end{aligned} \tag{3.31}$$

The next steps are similar to the analysis of the cross-talk term in Section 2.2. We assume that the patterns are random, that their bits $x_i^{(\mu)} = \pm 1$ are independently randomly distributed. Since the sums in Equation (3.31) contain many terms, we can estimate the sums using the central-limit theorem. The variable

$$z \equiv \sum_{\substack{\mu \neq 1 \\ \mu \neq \nu}} x_i^{(\mu)} x_i^{(1)} m_\mu \tag{3.32}$$

is then approximately Gaussian distributed, with mean zero. The variance of z is given by an average over a double sum. Since bits $x_i^{(\mu)}$ and $x_i^{(\mu')}$ are independent when $\mu \neq \mu'$, only the diagonal in this double sum contributes:

$$\sigma_z^2 = \sum_{\substack{\mu \neq 1 \\ \mu \neq \nu}} \langle m_\mu^2 \rangle \approx p \langle m_\mu^2 \rangle \quad \text{for any} \quad \mu \neq 1, \nu. \tag{3.33}$$

Here we assumed that p is large so that $p-2 \approx p$. Now return to Equation (3.31). The sum $\frac{1}{N}\sum_i$ in the second line can be approximated as an average over the Gaussian distributed variable z:

$$\beta m_\nu \int_{-\infty}^{\infty} \mathrm{d}z \frac{1}{\sqrt{2\pi}\sigma_z} \mathrm{e}^{-\frac{z^2}{2\sigma_z^2}} \left[1 - \tanh^2\left(\beta m_1 + \beta z\right)\right]. \tag{3.34}$$

We write the expression (3.34) as

$$\beta m_\nu \left[1 - \int_{-\infty}^{\infty} \mathrm{d}z \frac{1}{\sqrt{2\pi}\sigma_z} \mathrm{e}^{-\frac{z^2}{2\sigma_z^2}} \tanh^2\left(\beta m_1 + \beta z\right)\right] \equiv \beta m_\nu (1-q), \tag{3.35}$$

using the following definition of the parameter q:

$$q = \int_{-\infty}^{\infty} \mathrm{d}z \frac{1}{\sqrt{2\pi}\sigma_z} \mathrm{e}^{-\frac{z^2}{2\sigma_z^2}} \tanh^2\left(\beta m_1 + \beta z\right). \tag{3.36}$$

Returning to Equation (3.31) we see that it takes the form

$$m_\nu = \frac{1}{N} \sum_i x_i^{(\nu)} x_i^{(1)} \tanh\left(\beta m_1 + \beta \sum_{\substack{\mu \neq 1 \\ \mu \neq \nu}} x_i^{(\mu)} x_i^{(1)} m_\mu\right) + (1-q)\beta m_\nu. \tag{3.37}$$

Solving for m_ν we find:

$$m_\nu = \frac{\frac{1}{N}\sum_i x_i^{(\nu)} x_i^{(1)} \tanh\left(\beta m_1 + \beta \sum_{\mu \neq 1, \mu \neq \nu} x_i^{(\mu)} x_i^{(1)} m_\mu\right)}{1 - \beta(1-q)}, \tag{3.38}$$

for $\nu \neq 1$. This expression allows us to compute the variance σ_z, defined by Equation (3.33). Equation (3.38) shows that the average $\langle m_\nu^2 \rangle$ contains a double sum over the bit index, i. Since the bits are independent, only the diagonal terms contribute, so that

$$\langle m_\nu^2 \rangle \approx \frac{\frac{1}{N^2}\sum_i \tanh^2\left(\beta m_1 + \beta \sum_{\substack{\mu\neq 1 \\ \mu\neq\nu}} x_i^{(\mu)} x_i^{(1)} m_\mu\right)}{[1-\beta(1-q)]^2}, \tag{3.39}$$

independent of ν. The numerator is just q/N, from Equation (3.36). So the variance evaluates to

$$\sigma_z^2 = \frac{\alpha q}{[1-\beta(1-q)]^2}. \tag{3.40}$$

Up to now it was assumed that $\nu \neq 1$. One can derive an Equation for m_1 by repeating almost the same steps as above, but now for $\nu = 1$. The result is:

$$m_1 = \int \frac{\mathrm{d}z}{\sqrt{2\pi\sigma_z^2}} e^{-\frac{z^2}{2\sigma_z^2}} \tanh(\beta m_1 + \beta z). \tag{3.41}$$

This is a self-consistent equation for m_1. In summary there are three coupled equations, for m_1, q, and σ_z. Equations (3.35), (3.40), and (3.41). They must be solved together to determine how m_1 depends on β and α.

To compare with the results described in Section 2.2 we must take the deterministic limit, $\beta \to \infty$. In this limit q approaches unity, yet $\beta(1-q)$ remains finite. Setting $q = 1$ in Equation (3.40) but retaining $\beta(1-q)$ we find [1]:

$$\sigma_z^2 = \frac{\alpha}{[1-\beta(1-q)]^2}. \tag{3.42a}$$

The deterministic limits of Equations (3.35) and (3.41) become:

$$\beta(1-q) = \sqrt{\frac{2}{\pi\sigma_z^2}} \mathrm{e}^{-\frac{m_1^2}{2\sigma_z^2}}, \tag{3.42b}$$

$$m_1 = \mathrm{erf}\left(\frac{m_1}{\sqrt{2\sigma_z^2}}\right). \tag{3.42c}$$

Recall the definition (3.21) of the steady-state error probability. Inserting Equation (3.42c) for m_1 into this expression we find in the deterministic limit:

$$P_{\text{error}}^{t=\infty} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{m_1}{\sqrt{2\sigma_z^2}}\right)\right]. \tag{3.43}$$

Compare this with Equation (2.39) for the one-step error probability in the deterministic limit. That equation was derived for only one step in the dynamics of the

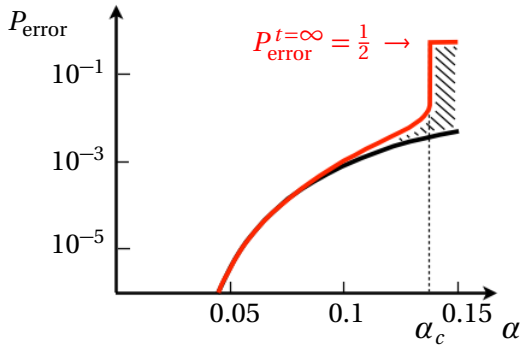


Figure 3.4: Error probability as a function of the storage capacity α in the deterministic limit. The one-step error probability $P_{\text{error}}^{t=1}$ [Equation (2.39)] is shown as a black line, the steady-state error probability $P_{\text{error}}^{t=\infty}$ [Equation (3.21)] is shown as a red line. In the hashed region error avalanches increase the error probability. After Figure 1 in Ref. [24].

network, while Equation (3.43) describes the long-time limit. Yet it turns out that Equation (3.43) reduces to (2.39) in the limit of $\alpha \to 0$. To see this one solves the set of Equations (3.42) by introducing the variable $y = m_1/\sqrt{2\sigma_z^2}$ [1]. One obtains the following one-dimensional equation for y [24]:

$$y(\sqrt{2\alpha} + (2/\sqrt{\pi})\, \mathrm{e}^{-y^2}) = \mathrm{erf}(y)\,. \tag{3.44}$$

The physical solutions are those satisfying $0 \le \mathrm{erf}(y) \le 1$, because the order parameter is restricted to this range (transitions to $-m_1$ do not occur in the limit $N \to \infty$). Figure 3.4 shows the steady-state error probability obtained from Equations (3.43) and (3.44). Also shown is the one-step error probability

$$P_{\text{error}}^{t=1} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{1}{\sqrt{2\alpha}}\right)\right]$$

derived in Section 2.2. You see that $P_{\text{error}}^{t=\infty}$ approaches $P_{\text{error}}^{t=1}$ for small α. This means that the error probability does not increase significantly as one iterates the network, at least for small α. In this case errors in earlier iterations have little effect on the probability that later errors occur. But the situation is different at larger values of α. In that case $P_{\text{error}}^{t=1}$ significantly underestimates the steady-state error probability. In the hashed region, errors in the dynamics increase the probability of errors in subsequent steps, giving rise to *error avalanches*.

Figure 3.4 illustrates that there is a critical value α_c where the steady-state error probability tends to $\frac{1}{2}$. Solution of the mean-field Equations gives

$$\alpha_c \approx 0.1379\,. \tag{3.45}$$

When $\alpha > \alpha_c$ the steady-state error probability equals $\frac{1}{2}$, in this region the network produces just noise. When α is small, the error probability is small, here the network

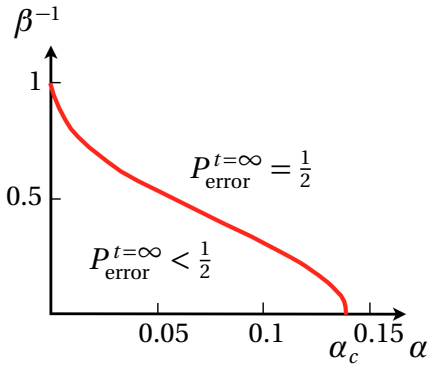


Figure 3.5: Phase diagram of the Hopfield network in the limit of large N (schematic). The region with $P_{\text{error}}^{t=\infty} < \frac{1}{2}$ is the ordered phase, the region with $P_{\text{error}}^{t=\infty} = \frac{1}{2}$ is the disordered phase. After Figure 2 in Ref. [24].

works well. Figure 3.4 shows that the steady-state error probability changes very abruptly near α_c. Assume you store 137 patterns with 1000 bits in a Hopfield network. Figure 3.4 demonstrates that the network can reliably retrieve the patterns. However, if you try to store one or two more patterns, the network fails to produce any output meaningfully related to the stored patterns. This rapid change is an example of a *phase transition*. In many physical systems one observes similar transitions between ordered and disordered phases.

What happens at higher noise levels? The numerical solution of Equations (3.35), (3.41), and (3.40) shows that the critical storage capacity α_c decreases as the noise level increases (smaller values of β). This is shown schematically in Figure 3.5. Below the red line the error probability is smaller than $\frac{1}{2}$ so that the network operates reliably (although less so as one approaches the phase-transition boundary). Outside this region the the error probability equals $\frac{1}{2}$. In this region the network fails. In the limit of small α the critical noise level is $\beta_c = 1$. In this limit the network is described by the theory explained in Section 3.3, Equation (3.14).

Often these two different phases of the Hopfield network are characterised in terms of the order parameter m_1. We see that $m_1 > 0$ in the hashed region, while $m_1 = 0$ outside.

3.5 Beyond mean-field theory*

The theory summarised in this Chapter rests on a mean-field approximation for the local field, Equation (3.6). The main result is the phase diagram shown in Figure 3.5. It is important to note that it was derived in the limit $N \to \infty$. For smaller values of N one expects the transition to be less sharp, so that m_1 is non-zero for values of α larger than α_c.

But even for large values of N the question remains how accurate the mean-field theory really is. To answer this question, one must take into account fluctuations. The corresponding calculation is more difficult than the ones outlined earlier in this Chapter, and it requires several steps. One starts from the steady-state distribution of $\boldsymbol{s}$ for fixed patterns $\boldsymbol{x}^{(\mu)}$. In Chapter 4 we will see that the steady-state distribution for the McCulloch-Pitts dynamics is the *Boltzmann distribution*

$$P_{\mathrm{B}}(\boldsymbol{s}) = Z^{-1}\mathrm{e}^{-\beta H(\boldsymbol{s})} \tag{3.46}$$

(the proof in Chapter 4 assumes that the diagonal weights are set to zero). The normalisation factor Z is called the *partition function*

$$Z = \sum_{\boldsymbol{s}} \mathrm{e}^{-\beta H(\boldsymbol{s})}. \tag{3.47}$$

One can compute the order parameter by adding a threshold term to the energy function (2.45)

$$H = -\frac{1}{2}\sum_{ij} w_{ij} s_i s_j + \sum_{\mu} \lambda_\mu \sum_i x_i^{(\mu)} s_i \,. \tag{3.48}$$

Then the order parameter m_μ is obtained by taking a derivative w.r.t λ_μ:

$$m_\mu = \Big\langle \frac{1}{N}\sum_i x_i^{(\mu)} \langle n_i \rangle \Big\rangle = -\frac{1}{N\beta}\frac{\partial}{\partial \lambda_\mu}\Big\langle \log Z \Big\rangle. \tag{3.49}$$

The outer average is over different realisations of random patterns. Since the logarithm of Z is difficult to average, one resorts to the *replica* trick. The idea is to represent the average of the logarithm as

$$\langle \log Z \rangle = \lim_{n\to 0} \tfrac{1}{n}(\langle Z^n \rangle - 1), \tag{3.50}$$

The function Z^n looks like the partition function of n copies of the system, thus the name *replica* trick. It is still debated when the replica trick works and when not [25]. Nevertheless, the most accurate theoretical result for the critical storage capacity is obtained in this way [26]

$$\alpha_{\mathrm{c}} = 0.138187\,. \tag{3.51}$$

The mean-field result (3.45) is different from (3.51), but it is very close. Most of the time, mean-field theories do not give such good results. Usually they yield at most a qualitative understanding of phase transitions, but not quantitatively accurate results as here. In the Hopfield model the mean-field theory works so well because the connections are global: every neuron is connected with every other neuron. This helps to average out the fluctuations in Equation (3.6).

The most precise Monte-Carlo simulations (Section 4.2) for finite values of N [27] yield upon extrapolation to $N = \infty$

$$\alpha_c = 0.143 \pm 0.002\,. \tag{3.52}$$

This is close to, yet significantly different from the best theoretical estimate, Equation (3.51), and also different from the mean-field result (3.45).

3.6 Correlated and non-random patterns

In the above Sections we assumed random patterns with independently identically distributed bits. This allowed us to calculate the storage capacity of the Hopfield network using the central-limit theorem. The hope is that the result describes what happens for typical, non-random patterns, or for random patterns with correlated bits.

Correlations affect the distribution of the cross-talk term, and thus the storage capacity of the Hopfield net. It has been argued that the storage capacity increases when the patterns are more strongly correlated, while others have claimed that the capacity decreases in this limit (see Ref. [28] for a discussion).

When we must deal with a definite set of patterns (no randomness to average over), the situation seems to be even more challenging. Yet there is a way of modifying Hebb's rule to deal with this problem, at least when the patterns are linearly independent (this requires $p \leq N$). The recipe is explained by Hertz, Krogh, and Palmer [1]. One simply incorporates the overlaps

$$Q_{\mu\nu} = \frac{1}{N} \boldsymbol{x}^{(\mu)} \cdot \boldsymbol{x}^{(\nu)}. \tag{3.53}$$

into Hebb's rule. To this end one defines the $p \times p$ *overlap matrix* $\mathbb{Q}$ with elements $Q_{\mu\nu}$ and writes:

$$w_{ij} = \frac{1}{N} \sum_{\mu\nu} x_i^{(\mu)} \left(\mathbb{Q}^{-1}\right)_{\mu\nu} x_j^{(\nu)}. \tag{3.54}$$

For orthogonal patterns ($Q_{\mu\nu} = \delta_{\mu\nu}$) this modified Hebb's rule is identical to Equation (2.25). For non-orthogonal patterns, the rule (3.54) ensures that all patterns are recognised. Equation (3.54) requires that the matrix $\mathbb{Q}$ is invertible: its columns must be linearly independent (and this implies that the rows are linearly independent too). This limits the number of patterns one can store with the rule (3.54), because $p > N$ implies linear dependence.

For linearly independent patterns one can find the weights w_{ij} iteratively, by successive improvement from an arbitrary starting point. We can say that the network learns the task through a sequence of weight changes. This is the idea used to solve classification tasks with perceptrons (Part II).

3.7 Summary

In this Chapter we analysed the dynamics of Hopfield nets. We asked under which circumstances the network dynamics can reliably retrieve stored patterns. For random patterns we explained how the the performance of the Hopfield net depends on its parameters: the number of stored patterns, the number of bits per pattern, and the noise level.

Hopfield networks share many properties with the networks discussed later on in these lectures. The most important point is perhaps that introducing noise in the dynamics allows to study the convergence and performance of the network: in the presence of noise there is a well-defined steady state that can be analysed. Without noise, in the deterministic limit, the network dynamics arrests in local minima of the energy function, and may not reach the stored patterns. Naturally the noise must be small enough for the network to function reliably. Apart from the noise level there is a second significant parameter, the storage capacity α, equal to the ratio of the number of patterns to the number of bits per pattern. When α is small then the network is reliable. A mean-field analysis of the $N \to \infty$-limit shows that there is a phase transition in the parameter plane (*phase diagram*) of the Hopfield network, Figure 3.5.

The building blocks of Hopfield networks are McCulloch-Pitts neurons and Hebb's rule for the weights. These elements are fundamental to the networks discussed in the coming Chapters.

3.8 Further reading

The statistical mechanics of Hopfield networks is explained in the book by Hertz, Krogh, and Palmer [1]. Starting from the Boltzmann distribution, Chapter 10 in this book explains how to compute the order parameters, and how to evaluate the stability of the corresponding solutions. For more details on the replica trick, see Ref. [21].

3.9 Exercises

3.1 Mixed states. Write a computer program that implements the stochastic dynamics of a Hopfield model. Compute how the order parameter for mixed states that are superpositions of the bits of three stored patterns depends on the noise level for $0.5 \leq \beta \leq 2.5$. Compare your numerical results with the predictions in Section 3.3. Repeat the exercise for mixed states that consist of superpositions of the bits of

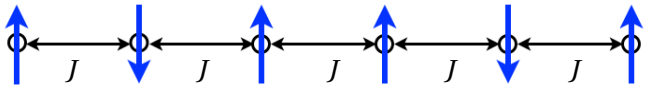


Figure 3.6: The Ising model is a model for ferromagnetism. It describes N spins that can either point up ($\uparrow$) or down ($\downarrow$), arranged on an integer lattice (here shown in one spatial dimension), interacting with their nearest neighbours with interaction strength J, and subject to an external magnetic field h. The state of spin i is described by the variable s_i, with $s_i = 1$ for $\uparrow$ and $s_i = -1$ for $\downarrow$.

five stored patterns. To this end, first derive the mean-field equation for the order parameter and solve this equation numerically. Second, perform your computer simulations and compare.

3.2 Order parameter. Derive the self-consistent Equation (3.41) for the order parameter m_1.

3.3 Deterministic limit. Derive the deterministic limit (3.42) of the three coupled Equations (3.35), (3.40), and (3.41) for m_1, q, and σ_z.

3.4 Phase diagram of the Hopfield network. Derive Equation (3.44) from Equation (3.42). Numerically solve (3.44) to find the critical storage capacity α_c in the deterministic limit. Quote your result with three-digit accuracy. To determine how the critical storate capacity depends on the noise level, numerically solve the three coupled Equations (3.41), (3.36), and (3.40). Compare your result with the schematic Figure 3.5.

3.5 Non-orthogonal patterns. Show that the rule (3.54) ensures that all patterns are recognised, for any set of non-orthogonal patterns that gives rise to an invertible matrix $\mathbb{Q}$. Demonstrate this by showing that the cross-talk term evaluates to zero, assuming that $\mathbb{Q}^{-1}$ exists.

3.6 Ising model. The Ising model is a model for ferromagnetism, N spins $s_i = \pm 1$ are arranged on a d-dimensional integer lattice as shown in Figure 3.6. The energy function for the Ising model is $H = -J \sum_{i,j=\mathrm{nn}(i)} s_i s_j - h \sum_i s_i$. Here J is the ferromagnetic coupling between nearest-neighbour spins, h is an external magnetic field, and $\mathrm{nn}(i)$ denotes the nearest-neighbours of site i on the lattice. In equilibrium at temperature T, the states are distributed according to the Boltzmann distribution with $\beta = 1/(k_B T)$. Derive a mean-field approximation for the magnetisation of the system, $m = \langle \frac{1}{N} \sum_i s_i \rangle$, assuming that N is large enough that the contribution of the boundary spins can be neglected. Derive an expression for the critical temperature below which mean-field theory predicts that the system becomes ferromagnetic,

$m \neq 0$. Discuss how the critical temperature depends on the dimension d. Note: mean-field theory fails for the one-dimensional Ising model. Its predictions become more accurate as d increases.

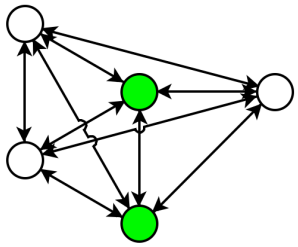


Figure 4.1: Boltzmann machine with five neurons. Two of them (green) are hidden. All weights are symmetric, the diagonal weights are set to zero. The states of the hidden neurons are denoted by $h_i = \pm 1$, and those of the visible neurons by $v_i = \pm 1$.

4 The Boltzmann distribution

Hopfield networks were introduced as models that can recognise patterns. In Section 2.5 we saw that the deterministic dynamics (2.6) of Hopfield nets admits the Lyapunov function

$$H = -\frac{1}{2}\sum_{ij} w_{ij} s_i s_j + \sum_i \theta_i s_i \,, \tag{4.1}$$

if the weights w_{ij} are symmetric, and provided that the diagonal weights are not negative.[1] This ensures that the stochastic McCulloch-Pitts dynamics (3.1) converges to a steady state where the state vector $\boldsymbol{s}$ follows the Boltzmann distribution

$$P_{\mathrm{B}}(\boldsymbol{s}) = Z^{-1} \mathrm{e}^{-\beta H(\boldsymbol{s})} \quad \text{with normalisation} \quad Z = \sum_{\boldsymbol{s}} \mathrm{e}^{-\beta H(\boldsymbol{s})}, \tag{4.2}$$

independent of time t. These properties allow to solve combinatorial optimisation tasks, by defining a Lyapunov function that has its global minimum at the solution of the optimisation problem. The network dynamics finds approximate solutions, in particular if one iteratively decreases the noise level by increasing β (*simulated annealing*).

The stochastic dynamics (3.1) is closely related to that of *Markov-chain Monte-Carlo* and *Metropolis* algorithms [29], which allow to sample from a distribution that is too expensive to compute. This is not a contradiction in terms, as we shall see in this Chapter. These algorithms are so widely used, that the corresponding paper [29] is considered *the most significant publication in the history of computational physics* [30].

Boltzmann machines [31–35] are stochastic Hopfield networks with *hidden neurons*, neurons that are neither used for inputs nor for outputs. Figure 4.1 shows

[1] In this Chapter we set the diagonal weights to zero.

the layout of a Boltzmann machine with hidden and visible neurons. Boltzmann machines can be *trained* to learn the properties of a distribution $P_{\text{data}}(\boldsymbol{x})$ of binary input patterns $\boldsymbol{x}$. The idea is to iteratively change the weights in Equation (4.1) until the Boltzmann distribution represents the input distribution. This idea, to iterate the weights until the network learns the input distribution P_{data}, is used in a slightly different form in *supervised learning* (Part II). Boltzmann machines are closely related to Hopfield nets. Both models learn to represent two-point correlations $\langle x_i^{(\mu)} x_j^{(\mu)} \rangle$ of pattern bits. When important information about the inputs is encoded in higher-order correlations, one uses hidden neurons to represent these correlations. Generally Boltzmann machines are hard to train, in particular if they have many hidden neurons. *Restricted Boltzmann machines* are Hopfield networks with hidden neurons, but with fewer connections: only connections between *visible* and hidden heurons are allowed. These neural nets can be fairly efficiently trained and can solve a number of different tasks. Apart from learning a distribution of input patterns, they can be trained to recognise incomplete input patterns, to classify inputs, and for sensory-motor control.

4.1 Convergence of the noisy dynamics

We begin by showing that the stochastic dynamics (3.1) has a steady state where $\boldsymbol{s}$ is distributed according to the Boltzmann distribution (4.2). To this end we introduce an alternative yet equivalent formulation of the network dynamics. It consists of two steps. First, choose a neuron randomly, number m say. Second, update s_m to $s'_m \neq s_m$ with probability

$$\text{Prob}(s_m \to s'_m) = \frac{1}{1+e^{\beta \Delta H_m}}, \tag{4.3a}$$

where

$$\Delta H_m = H(\ldots, s'_m, \ldots) - H(\ldots, s_m, \ldots) = -b_m(s'_m - s_m) \tag{4.3b}$$

with local field $b_m = \sum_j w_{mj} s_j - \theta_m$. In order to see why ΔH_m evaluates to $-b_m(s'_m - s_m)$, consider

$$H(\ldots, s_m = 1, \ldots) - H(\ldots, s_m = -1, \ldots) = -\sum_k (w_{mk} + w_{km}) s_m + 2\theta_m = -2b_m, \tag{4.4a}$$

and

$$H(\ldots, s_m = -1, \ldots) - H(\ldots, s_m = 1, \ldots) = \sum_k (w_{mk} + w_{km}) s_k - 2\theta_m = 2b_m. \tag{4.4b}$$

The left equality sign in Equations (4.4) assumes that the diagonal weights vanish [compare with Equation (2.48)], and the right one assumes that the weights are symmetric. Under these conditions we conclude that $\Delta H_m = -b_m(s'_m - s_m)$.

To explore the relation between Equations (4.3) and (3.1), we break up the prescription (4.3) up into different cases. The state of neuron m changes with probability

$$\text{if } s_m = -1 \qquad \text{obtain } s'_m = 1 \text{ with prob.} \qquad \frac{1}{1+e^{-2\beta b_m}} = p(b_m), \tag{4.5a}$$

$$\text{if } s_m = 1 \qquad \text{obtain } s'_m = 1 \text{ with prob.} \qquad \frac{1}{1+e^{2\beta b_m}} = 1 - p(b_m). \tag{4.5b}$$

In the second row we used that $1-p(b) = 1 - \frac{1}{1+\mathrm{e}^{-2\beta b}} = \frac{1+\mathrm{e}^{-2\beta b}-1}{1+\mathrm{e}^{-2\beta b}} = \frac{1}{1+\mathrm{e}^{2\beta b}}$. The state remains unchanged with probability:

$$\text{if } s_m = -1 \qquad \text{obtain } s_m = -1 \text{ with prob.} \qquad \frac{1}{1+e^{-\beta b_m}} = 1 - p(b_m), \tag{4.5c}$$

$$\text{if } s_m = 1 \qquad \text{obtain } s_m = 1 \text{ with prob.} \qquad \frac{1}{1+e^{\beta b_m}} = p(b_m). \tag{4.5d}$$

Comparing with Equation (3.1) we conclude that the two schemes (3.1) and (4.3) are equivalent if $w_{ij} = w_{ji}$ and $w_{ii} = 0$. Note that Equation (4.3) is more general than the stochastic Hopfield dynamics. Equation (4.3) does not require the energy function to be of the form (4.20), and in particular it is neither needed that the weights are symmetric, nor that the diagonal weights are non-negative. Equations (3.1) and (4.3) are not equivalent if these conditions are not satisfied (Exercise 4.3).

Equation (4.3) makes it clear that a little bit of noise achieves what we aimed for, namely to prevent the McCulloch-Pitts dynamics to arrest in local minima: moves with $\Delta H_m > 0$ are less likely when β is large. In the deterministic limit $\beta \to \infty$ the energy can only decrease, or remain constant. At large but finite values of β, the energy may sometimes increase, allowing the dynamics to escape local minima.

The transition probability (4.3) defines a *Markov chain* of states

$$\boldsymbol{s}_{t=0} \longrightarrow \boldsymbol{s}_{t=1} \longrightarrow \boldsymbol{s}_{t=2} \longrightarrow \ldots \tag{4.6}$$

As before, the index t counts the iteration steps. A Markov chain is a *memoryless* random sequence of states defined by *transition probabilities* $p_{l\to k}$ from state $\boldsymbol{s}^{(l)}$ to $\boldsymbol{s}^{(k)}$ [36]. The transition probability $p_{l\to k}$ connects arbitrary states, allowing for local moves (as in the previous Section where only one component of $\boldsymbol{s}$ was changed) or global moves – where many neurons are allowed to change their states in a single step. The rule (4.3) consists of two parts. First, a new state $\boldsymbol{s}^{(k)}$ is suggested with

probability $p^{(\mathrm{s})}_{l\to k}$. In (4.3), the dynamics is local because only a single neuron is picked randomly. Therefore $p^{(\mathrm{s})}$ is a constant,

$$p^{(\mathrm{s})} = N^{-1}\,, \tag{4.7}$$

where N is the number of neurons in the network. Second, the new state $\boldsymbol{s}^{(k)}$ is accepted with *acceptance probability*

$$p^{(\mathrm{a})}_{l\to k} = \frac{1}{1+\mathrm{e}^{\beta\Delta H}}\,. \tag{4.8}$$

where $\Delta H = H(\boldsymbol{s}^{(k)}) - H(\boldsymbol{s}^{(l)})$. As result, the transition probability is given by a product of two factors, $p_{l\to k} = p^{(\mathrm{s})}_{l\to k} p^{(\mathrm{a})}_{l\to k}$. These steps are repeated many times, creating the sequence of states (4.6).

Now we show that the Markov chain defined by Equations (4.7) and (4.8) has the Boltzmann distribution (4.2) as a steady-state distribution. This is ensured by the *detailed-balance* condition

$$P_{\mathrm{B}}(\boldsymbol{s}^{(l)}) p_{l\to k} = P_{\mathrm{B}}(\boldsymbol{s}^{(k)}) p_{k\to l}\,. \tag{4.9}$$

There are Markov chains that do not satisfy detailed balance but still have a steady state (Exercise 4.6).

Usually detailed balance implies not only that the Markov chain has $P_{\mathrm{B}}(\boldsymbol{s})$ as a steady-state distribution, but also that the distribution of states generated by the sequence (4.6) converges to $P(\boldsymbol{s})$ (see Ref. [36] for details). To prove that condition (4.9) holds, we use that $p^{(\mathrm{s})}$ is symmetric (because it is independent of l or k), and we use Equations (4.11) and (4.8):

$$\frac{p^{(\mathrm{s})}\mathrm{e}^{-\beta H(\boldsymbol{s}^{(l)})}}{1+\mathrm{e}^{\beta[H(\boldsymbol{s}^{(k)})-H(\boldsymbol{s}^{(l)})]}} = \frac{p^{(\mathrm{s})}}{\mathrm{e}^{\beta H(\boldsymbol{s}^{(l)})}+\mathrm{e}^{\beta H(\boldsymbol{s}^{(k)})}} = \frac{p^{(\mathrm{s})}\mathrm{e}^{-\beta H(\boldsymbol{s}^{(k)})}}{1+\mathrm{e}^{\beta[H(\boldsymbol{s}^{(l)})-H(\boldsymbol{s}^{(k)})]}}\,. \tag{4.10}$$

This demonstrates that the Boltzmann distribution is a steady state of the Markov chain defined by (4.3). If the simulation converges to the steady state (as it usually does), then states visited by the Markov chain are distributed according to the Boltzmann distribution. This also means that the steady-state distribution for the Hopfield model is the Boltzmann distribution, as stated in Section 3.5.

It is important to stress that Equation (4.9) is a condition for the transition probability $p_{l\to k} = p^{(\mathrm{s})}_{l\to k} p^{(\mathrm{a})}_{l\to k}$, not just for the acceptance probability $p^{(\mathrm{a})}_{l\to k}$. For the local moves discussed above, $p^{(\mathrm{s})}$ is a constant, so that $p_{l\to k} \propto p^{(\mathrm{a})}_{l\to k}$. In this case it is sufficient to check the detailed-balance condition for the acceptance probability. In general, and in particular for global moves, it is necessary to include $p^{(\mathrm{s})}_{l\to k}$ in the detailed-balance check [37].

4.2 Monte-Carlo simulation

The Markov chain described in the previous Section is the basis for the *Markov-chain Monte-Carlo* algorithm, which in turn is closely related to the *Metropolis* algorithm. This method is widely used in Statistical Physics and in Mathematical Statistics. It is therefore important to understand the connections between the different formulations.

The Boltzmann distributon describes the probabilities of observing configurations of a large class of physical systems in their steady states [18]. The statistical mechanics of systems with energy function (also called *Hamiltonian*) H shows that their configurations are distributed according to the Boltzmann distribution in thermodynamic equilibrium at a given temperature (in this context $\beta^{-1} = k_{\mathrm{B}} T$ where k_{B} is the Boltzmann constant), and free from any other constraints. If we denote the configuration of a system by the vector $\boldsymbol{s}$, then the Boltzmann distribution takes the form

$$P_{\mathrm{B}}(\boldsymbol{s}) = Z^{-1} \mathrm{e}^{-\beta H(\boldsymbol{s})} \tag{4.11}$$

Here $Z = \sum_{\boldsymbol{s}} \mathrm{e}^{-\beta H(\boldsymbol{s})}$ is a normalisation factor, called *partition function*. For systems with a large number of interacting degrees of freedom, the function $P_{\mathrm{B}}(\boldsymbol{s})$ can be very expensive to compute, because then the sum over $\boldsymbol{s}$ in the normalisation factor Z contains many terms. Therefore, instead of computing the distribution directly one generates a Markov chain of states using a suitable transition probability, for instance (4.3).

In practice one often uses a slightly different form of the transition probabilities (*Metropolis algorithm*). Assuming that $p^{(\mathrm{s})}$ is constant one takes:

$$p_{l\to k} = p^{(\mathrm{s})} \begin{cases} \mathrm{e}^{-\beta \Delta H} & \text{when} \quad \Delta H > 0\,, \\ 1 & \text{when} \quad \Delta H \le 0\,, \end{cases} \tag{4.12}$$

with $\Delta H = H(\boldsymbol{s}^{(k)}) - H(\boldsymbol{s}^{(l)})$. Equation (4.12) has the advantage that the transition probabilities are higher than in (4.8) so that moves are more frequently accepted. That the Metropolis rates obey the detailed-balance condition (4.9) can be seen using Equations (4.11) and (4.12):

$$\begin{aligned} P_{\mathrm{B}}(\boldsymbol{s}^{(l)}) p_{l\to k} &= Z^{-1} p^{(\mathrm{s})} \mathrm{e}^{-\beta H(\boldsymbol{s}^{(l)})} \begin{cases} \mathrm{e}^{-\beta[H(\boldsymbol{s}^{(k)}) - H(\boldsymbol{s}^{(l)})]} & \text{if} \quad H(\boldsymbol{s}^{(k)}) > H(\boldsymbol{s}^{(l)}) \\ 1 & \text{otherwise} \end{cases} \\ &= Z^{-1} p^{(\mathrm{s})} \mathrm{e}^{-\beta \max\{H(\boldsymbol{s}^{(k)}), H(\boldsymbol{s}^{(l)})\}} \\ &= Z^{-1} p^{(\mathrm{s})} \mathrm{e}^{-\beta H(\boldsymbol{s}^{(k)})} \begin{cases} \mathrm{e}^{-\beta[H(\boldsymbol{s}^{(l)}) - H(\boldsymbol{s}^{(k)})]} & \text{if} \quad H(\boldsymbol{s}^{(l)}) > H(\boldsymbol{s}^{(k)}) \\ 1 & \text{otherwise} \end{cases} \\ &= P_{\mathrm{B}}(\boldsymbol{s}^{(k)}) p_{k\to l}\,. \end{aligned} \tag{4.13}$$

The fact that the algorithm produces states distributed according to Equation (4.11) allows to solve optimisation tasks using *simulated annealing* [38]. Slowly lowering the noise level during the simulation mimics the slow cooling of a physical system. It passes through a sequence of quasi-equilibrium Boltzmann distributions with lower and lower temperatures, until the system finds the global minimum $H_{\min}$ of the energy function at zero temperature, where $P_{\mathrm{B}}(\boldsymbol{s}) = 0$ when $H(\boldsymbol{s}) > H_{\min}$, but $P_{\mathrm{B}}(\boldsymbol{s}) > 0$ when $H(\boldsymbol{s}) = H_{\min}$. Note that the Monte-Carlo algorithm applies to energy functions of arbitrary form, not just to the Hopfield nets.

In summary, while the Boltzmann distribution may be difficult to evaluate (since Z involves a sum over many states), ΔH is cheap to compute for local moves. But note also that the states in the sequence (4.6) are correlated, in particular when the moves are local, because then subsequent configurations are similar. Generating many quite strongly correlated samples from a distribution is not a very efficient way of sampling this distribution. Sometimes it may therefore be more efficient to suggest global moves instead, to avoid that subsequent states in the Markov chain are similar. But it is not guaranteed that global moves lead to weaker correlations. For global moves, ΔH may be more likely to assume large positive values, so that fewer suggested moves are accepted. As a consequence the Markov chain may stay at certain states, increasing correlations in the sequence. Usually a compromise is most efficient, moves that are neither local nor global.

4.3 Simulated annealing*

In this Section we discuss how to solve combinatorial optimisation problems using simulated annealing. Such problems admit 2^k or $k!$ configurations - too many to list and check in a serial approach when k is large. The idea is to write down an energy function that is at most quadratic in the state variables, like Equation (2.45). Then one can use the Hopfield dynamics to minimize H. The problem is of course that

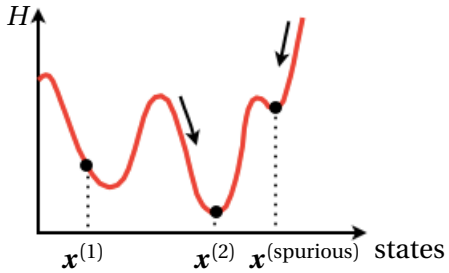


Figure 4.2: Deterministic gradient descent may not reach the desired minimum ($\boldsymbol{x}^{(2)}$ for example) because it arrests in a local minimum corresponding to spurious state $\boldsymbol{x}^{(\mathrm{spurious})}$.

city	coordinates
A	(0.1,0.15)
B	(0.4 0.2)
C	(0.5 0.7)
D	(0.2 0.1)
E	(0.1 0.8)
F	(0.8 0.9)
G	(0.9 0.3)

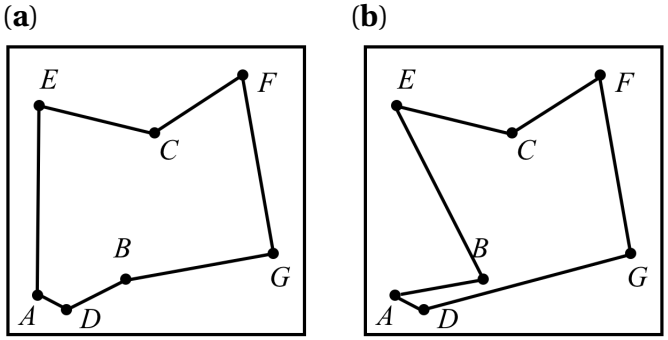


Figure 4.3: Traveling-salesman problem for $k = 7$. Given are the coordinates of k cities as points in the unit square. The problem is to find the shortest connected path that joins all cities, visits each city exactly once, and returns to the starting point. (**a**) optimal solution corresponding to the shortest path, (**b**) second solution with a longer path.

the deterministic network dynamics arrests in first local minimum encountered, usually not the desired optimum (Figure 4.2). Therefore it is important to introduce a little bit of noise. As discussed in Chapter 3, the noise helps the network to escape local minima.

A well-known combinatorial optimisation problem is the *travelling-salesman problem*. Given the coordinates of k cities, the goal is to determine the shortest journey visiting each city exactly once before returning to the starting point. The coordinates of seven cities A,...,F are given in Figure 4.3 (this Figure illustrates the problem for $k = 7$). The Figure shows two different solutions. We see that the path in panel (**a**) is the shorter one. Denoting the distance between city A and B by d_{AB} and so forth, the length of the path in panel (**a**) is

$$L = d_{AD} + d_{DB} + d_{BG} + d_{GF} + d_{FC} + d_{CE} + d_{EA}. \tag{4.14}$$

Paths are represented in terms of $k \times k$ matrices as follows. Each row of the matrix corresponds to a city, and the j-th element in this row has the entry 1 if the city is the j-th stop along the path. The other entries are zero. Path (**a**), for example, is represented by the matrix

$$\begin{array}{c} \text{(stop)} \\ \mathbb{M} = \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \\ \text{G} \end{array} \quad \text{(city)}. \tag{4.15}$$

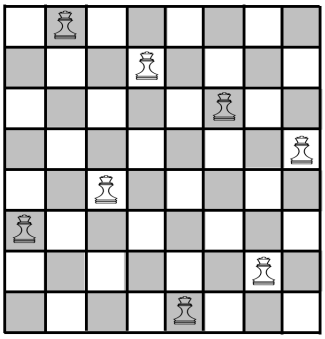


Figure 4.4: One solution of the k-queens problem for $k = 8$.

Since each city is visited only once, there can be only one 1 in each row. Since each visit corresponds to exactly one city, there can be only one 1 in each column. Any permutation of the elements that satisfies these constraints is an allowed path. There are $k!$ such permutations. They are $2k$-fold degenerate (there are k paths of the same length that differ by which city is visited first, and each path can be traveled clockwise or anti-clockwise). Therefore there are $k!/(2k)$ possible paths to consider in trying to determine the shortest one. This makes the problem hard. Note that *integer linear-programming* methods for solving the travelling-salesman problem usually use a different representation of the paths [39].

The *k-queens problem* is derived from the game of chess. The question is how to arrange k queens on a $k \times k$ chess board so that they cannot take each other out. This means that each row and column as well as each diagonal can have only one queen. The problem is illustrated in Figure 4.4. This Figure shows one solution for $k = 8$. The task is to find all solutions. Each arrangement of queens can be represented as a matrix $\mathbb{M}$, where one sets $M_{ij} = 1$ if there is a queen on site (i, j). All other elements are set to zero. To represent valid solutions, $\mathbb{M}$ must satisfy the following constraints [40]

$$\sum_{i=1}^{k}\sum_{j=1}^{k} M_{ij} = k\,, \tag{4.16}$$

$$\text{If} \quad M_{ij} = M_{pq} = 1 \quad \text{for} \quad (i,j) \neq (p,q) \quad \text{then}$$
$$i \neq p \quad \text{and} \quad j \neq q \quad \text{and} \quad i - j \neq p - q \quad \text{and} \quad i + j \neq p + q\,.$$

The *double-digest problem.* The Human Genome sequence was first assembled by piecing together overlapping DNA segments in the right order by making sure that overlapping segments share the same DNA sequence. To this end it is necessary to uniquely identify the DNA segments. The actual DNA sequence of a segment is a unique identifier. But it is sufficient and more efficient to identify a DNA segment

by a *fingerprint*, for example the sequence of *restriction sites*. These are short subsequences (four or six base pairs long) that are recognised by enzymes that cut (*digest*) the DNA strand precisely at these sites. A DNA segment is identified by the types and locations of restriction sites that it contains, the so-called *restriction map*.

When a DNA segment is cut by two different enzymes one can experimentally determine the lengths of the resulting fragments. Is it possible to determine how the cuts were ordered in the DNA sequence of the segment from the fragment lengths? This is the double-digest problem [41]. The order of cut sites is precisely the restriction map. In a double-digest experiment, a given DNA sequence is first digested by one enzyme (A say). Assume that this results in n fragments with lengths a_i ($i = 1, \ldots, n$). Second, the DNA sequence is digested by another enzyme, B. In this case m fragments are found, with lengths $b_1, b_2, \ldots, b_m$. Third, the DNA sequence is digested with both enzymes A and B, yielding l fragments with lengths $c_1, \ldots, c_l$. The question is now to determine all possible ordering of the a- and b-cuts that result in l fragments with lengths $c_1, c_2, \ldots, c_l$?

The first two problems introduced above are similar in that possible solutions can be represented in terms of $k \times k$ matrices $\mathbb{M}$ with 0/1 entries and certain constraints. It turns out that one can represent these problems in terms of a Hopfield network with $N = k^2$ neurons [42]. As an example, consider the travelling-salesman problem. We label the cities A to G by integers $m = 1, \ldots, k$, and denote their distances by d_{mn}. Then the path length to be minimised can be written as [42]

$$L = \tfrac{1}{2} \sum_{mnj} d_{mn} M_{mj} (M_{nj-1} + M_{nj+1}) \tag{4.17}$$

(Exercise 4.1). Here periodic boundary conditions in j were assumed ($j + 1 = 1$ for $j = k$). The column- and row-constraints upon the matrix $\mathbb{M}$

$$\sum_j M_{mj} = 1 \quad \text{row constraint}, \tag{4.18a}$$

$$\sum_m M_{mj} = 1 \quad \text{column constraint}, \tag{4.18b}$$

are incorporated using *Lagrange multipliers* λ_1 and λ_2 (both positive), so that the function to be minimised becomes

$$H = L + \tfrac{\lambda_1}{2} \sum_m \Big(1 - \sum_j M_{mj}\Big)^2 + \tfrac{\lambda_2}{2} \sum_j \Big(1 - \sum_m M_{mj}\Big)^2 . \tag{4.19}$$

When the constraints (4.18) are satisfied, their contributions to H vanish, otherwise they are positive. We conclude that H has a global minimum at the desired solution. If we use a stochastic method to minimise H, it is not guaranteed that the algorithm

finds the global minimum, either because the constraints are not exactly satisfied, or because the path found is not the shortest one. The magnitudes of the Lagrange multipliers λ_1 and λ_2 determine how strongly the constraints are enforced, during the search and in sub-optimal solutions.

The expression (4.19) is a quadratic function of M_{mj}. This suggests that we can write H as the energy function of a Hopfield net (Exercise 2.8):

$$H = -\frac{1}{2}\sum_{ijkl} w_{ijkl}\, s_{ij}\, s_{kl} + \sum_{ij} \theta_{ij}\, s_{ij} + \text{constant}\,. \tag{4.20}$$

The weights w_{ijkl} and thresholds θ_{ij} are determined by comparing Equations (4.17), (4.19), and (4.20). Note that the neurons carry two indices (not one as in Section 1.2). It turns out that the weights are symmetric, $w_{ijkl} = w_{klij}$ and that the diagonal weights are zero (Exercise 4.2). In general, since $s_{ij}^2 = 1$, we can always assume that the diagonal weights vanish, because they make only a constant contribution to H.

4.4 Boltzmann machines

As mentioned in the Introduction to this Chapter, Boltzmann machines are generalised Hopfield nets that can learn to approximate data distributions of binary input patterns. Boltzmann machines differ from Hopfield nets in two essential ways. First, instead of using Hebb's rule, the weights are adjusted until the Boltzmann machine approximates the data distribution precisely. The weights are iteratively refined to minimise the difference between the data distribution and the model (the Boltzmann distribution). Nevertheless, this procedure is closely related to Hebb's rule, as we shall see. Second, to represent higher-order correlations between bits of input patterns, Boltzmann machines employ hidden neurons. We begin by discussing how to train Boltzmann machines with only visible neurons, because it is simpler. Then we discuss why hidden neurons are necessary to learn the properties of general input distributions $P_{\text{data}}(\boldsymbol{x})$ of binary inputs $\boldsymbol{x}$. The goal of the training algorithm is to find weights so that the Boltzmann distribution

$$P_{\text{B}}(\boldsymbol{s} = \boldsymbol{x}) = Z^{-1} \exp\Big(\tfrac{1}{2}\sum_{i\neq j} w_{ij}\, x_i\, x_j\Big) \tag{4.21}$$

approximates the distribution $P_{\text{data}}(\boldsymbol{x})$ as precisely as possible. Here and in the remainder of this Chapter we set $\beta = 1$. The input patterns have N binary bits [Equation (2.1)] with values ± 1. The weight matrix $\mathbb{W}$ is symmetric, $w_{ij} = w_{ji}$, and its diagonal elements are set to zero, $w_{ii} = 0$. In this Section we also set the thresholds to zero.

The Boltzmann machine is trained by iteratively adjusting the weights w_{ij}, using a sequence of input patterns $\boldsymbol{x}^{(\mu)}$ ($\mu = 1,\ldots,p$) independently sampled from the data distribution $P_{\text{data}}(\boldsymbol{x})$. The goal is to change the weights in (4.21) to minimise the difference between the Boltzmann and the data distribution. This is achieved by maximising the *likelihood* $\mathscr{L} = \prod_{\mu=1}^{p} P_{\text{B}}(\boldsymbol{s} = \boldsymbol{x}^{(\mu)})$ that the Boltzmann machine produces the sequence $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$ of input patterns. Any pattern may appear more than once in the sequence, with frequency proportional to $P_{\text{data}}(\boldsymbol{x})$. Maximising $\mathscr{L}$ therefore corresponds to approximating the data distribution as accurately as possible.

Usually one maximises the logarithm of the likelihood, the *log-likelihood* function

$$\log \mathscr{L} = \log \prod_{\mu=1}^{p} P_{\text{B}}(\boldsymbol{s} = \boldsymbol{x}^{(\mu)}) = \sum_{\mu=1}^{p} \log P_{\text{B}}(\boldsymbol{s} = \boldsymbol{x}^{(\mu)}). \tag{4.22}$$

The logarithm is a monotonic function, so the log-likelihood has its maximum at the same weight-values as the likelihood. Taking the logarithm simplifies the analysis of the learning algorithm, essentially because a sum is easier to deal with than a product, and because $\log P_{\text{B}}(\boldsymbol{s} = \boldsymbol{s}^{(\mu)})$ is a quadratic function of $x_j^{(\mu)}$. Also, a learning algorithm based on the log-likelihood is usually more stable numerically, as explained below. A different reasoning behind maximising the log-likelihood starts from the *Kullback-Leibler divergence*

$$D_{\text{KL}} = \sum_{\mu=1}^{p} P_{\text{data}}(\boldsymbol{x}^{(\mu)}) \log[P_{\text{data}}(\boldsymbol{x}^{(\mu)}) / P_{\text{B}}(\boldsymbol{s} = \boldsymbol{x}^{(\mu)})]. \tag{4.23}$$

Terms in the sum with $P_{\text{data}}(\boldsymbol{x}^{(\mu)})$ are set to zero, and D_{KL} is defined to equal infinity when there are patterns for which $P_{\text{B}} = 0$ but $P_{\text{data}} \neq 0$. The Kullback-Leibler divergence is a measure of the difference between the two distributions: $D_{\text{KL}} \geq 0$ and it assumes its global minimum $D_{\text{KL}} = 0$ for $P_{\text{data}}(\boldsymbol{x}^{(\mu)}) = P_{\text{B}}(\boldsymbol{s} = \boldsymbol{x}^{(\mu)})$, see Excercise 4.8. We see from Equation (4.23) that minimising D_{KL} corresponds to maximising $\log \mathscr{L}$.

To find the global maximum of the log-likelihood we use *gradient ascent*: one repeatedly updates the weights by adding increments

$$w'_{mn} = w_{mn} + \delta w_{mn} \quad \text{with} \quad \delta w_{mn} = \eta \frac{\partial \log \mathscr{L}}{\partial w_{mn}}. \tag{4.24}$$

The small parameter $\eta > 0$ is the *learning rate*. The gradient points in the steepest uphill direction of $\mathscr{L}$. The idea is to take many uphill steps until one hopefully (but not necessarily) reaches the global maximum. Since the likelihood is a product of many possibly quite small factors, $\mathscr{L}$ can become very small. This can lead to

numerical instabilities. Maximising $\log \mathscr{L}$ instead of $\mathscr{L}$ can be more stable because it yields an additional factor $\mathscr{L}^{-1}$ in the gradient: $\partial \log \mathscr{L} / \partial w_{mn} = \mathscr{L}^{-1} \partial \mathscr{L} / \partial w_{mn}$.

To evaluate the gradient of $\mathscr{L}$ we start from Eq. (4.22)

$$\log \mathscr{L} = \sum_{\mu=1}^{p} \Big[-\log Z + \tfrac{1}{2} \sum_{i \neq j} w_{ij} x_i^{(\mu)} x_j^{(\mu)} \Big]. \tag{4.25}$$

Here we used that the diagonal weights vanish. The first step is to evaluate the derivative of

$$\log Z = \log \sum_{s_1=\pm 1, \ldots, s_N = \pm 1} \exp\Big(\tfrac{1}{2} \sum_{i \neq j} w_{ij} s_i s_j \Big). \tag{4.26}$$

To compute $\partial \log Z / \partial w_{mn}$ one uses the chain rule together with

$$\frac{\partial w_{ij}}{\partial w_{mn}} = \delta_{im} \delta_{jn} + \delta_{jm} \delta_{in} \,. \tag{4.27}$$

Here δ_{kl} is the Kronecker delta, $\delta_{kl} = 1$ if $k = l$ and zero otherwise. So $\delta_{im} \delta_{jn} = 1$ only if $i = m$ and $j = n$. Otherwise the product of Kronecker deltas equals zero. This is illustrated by the following story.

> *The linear function, x, and the constant function are going for a walk. When they suddenly see the derivative approaching, the constant function gets worried. "I'm not worried" says the function x confidently, "I'm not put to zero by the derivative." When the derivative comes closer, it says "Hi! I'm $\partial / \partial y$. How are you?"*

The moral is: since x and y are independent variables, $\partial x / \partial y = 0$. Equation (4.27) reflects the same principle: the weights w_{ij} and w_{mn} are independent variables unless their indices agree. That there are two terms in the r.h.s. of Equation (4.27) is a consequence of the fact that the weights are symmetric. Returning to the derivative of $\log Z$ with respect to w_{mn}, one finds using Equation (4.27):

$$\frac{\partial \log Z}{\partial w_{mn}} = \sum_{s_1=\pm 1, \ldots, s_N = \pm 1} s_m s_n P_{\mathrm{B}}(\boldsymbol{s}) \equiv \langle s_m s_n \rangle_{\mathrm{model}} \,, \tag{4.28}$$

where the last equality defines the two-point correlations of the model, $\langle s_m s_n \rangle_{\mathrm{model}}$, computed using the steady-state distribution (4.21) of the Boltzmann machine. Evaluating the derivative of the second term in Equation (4.25) gives:

$$\frac{\partial}{\partial w_{mn}} \tfrac{1}{2} \sum_{i \neq j} w_{ij} x_i^{(\mu)} x_j^{(\mu)} = x_m^{(\mu)} x_n^{(\mu)} \,. \tag{4.29}$$

In summary we find

$$\frac{\partial \log \mathscr{L}}{\partial w_{mn}} = \sum_{\mu=1}^{p} \left(x_m^{(\mu)} x_n^{(\mu)} - \langle s_m s_n \rangle_{\text{model}} \right) = p \left(\langle x_m x_n \rangle_{\text{data}} - \langle s_m s_n \rangle_{\text{model}} \right). \tag{4.30}$$

Here $\langle x_m x_n \rangle_{\text{data}} = p^{-1} \sum_{\mu=1}^{p} x_m^{(\mu)} x_n^{(\mu)}$ is the two-point correlation of the input data. Using (4.24), the *learning rule* becomes:

$$\delta w_{mn} = \eta \left(\langle x_m x_n \rangle_{\text{data}} - \langle s_m s_n \rangle_{\text{model}} \right), \tag{4.31}$$

where we have dropped a factor of p which only affects the numerical value of the learning rate η. The weight changes are determined by the two-point pattern correlations, just like Hebb's rule (2.25). The first term on the r.h.s. of Eq. (4.31) has precisely the same form as Equation (2.25), a sum over two-point correlations of the input patterns. The second average is over the steady-state distribution (4.21) of the Boltzmann machine. The learning rule takes the form of the difference between two two-point correlations because the task is to minimise the difference between two distributions. It is plausible that the learning rule may converge because the weight increments vanish when the model correlations equal the data correlations.

The average $\langle s_m s_n \rangle_{\text{model}}$ can be approximated by numerical simulation of the McCulloch-Pitts dynamics

$$v_i' = \begin{cases} 1 & \text{with probability} \quad p(b_i), \\ -1 & \text{with probability} \quad 1 - p(b_i), \end{cases} \tag{4.32}$$

with $b_i = \sum_j w_{ij} v_j$ and $p(b_i) = \frac{1}{1+\mathrm{e}^{-2b_i}}$. One must iterate Equation (4.32) until the system has reached its steady state, and long enough so that any initial transient becomes negligible.

The training algorithm can be summarised as follows. One initialises all weights and computes $\langle x_m x_n \rangle_{\text{data}}$ from the given sequence of input patterns. One estimates $\langle s_m s_n \rangle_{\text{model}}$ by numerical simulation of the dynamics of the Boltzmann machine, and updates the weights using (4.31). One iterates this step, either with a sequence of new inputs, or with the same inputs but in permuted sequence. In each iteration one must compute $\langle s_m s_n \rangle_{\text{model}}$ again, because the weights were updated. This procedure is quite slow however, because it usually takes long simulations to estimate $\langle x_m x_n \rangle_{\text{model}}$ accurately, in each iteration of the learning algorithm.

There is a more fundamental problem. Like Hebb's rule, the learning rule (4.31) relies entirely upon two-point correlations of the input bits. This means that the Boltzmann machine cannot learn higher-order correlations between inputs. However, two-point correlations may not be sufficient to represent the information

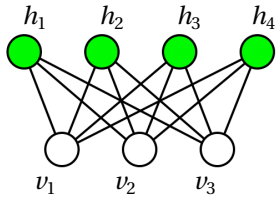


Figure 4.5: Restricted Boltzmann machine with three visible and four hidden neurons.

encoded in the input data. To illustrate this point, consider the Boolean XOR function (Exercise 2.15).It can be encoded in the four patterns $[-1,-1,-1]$, $[1,1,-1]$, $[-1,1,1]$, and $[1,-1,1]$. The first two components represent the input to the XOR function. The third component represents the output, which depends on both input variables as prescribed by the XOR function. Let us define an input distribution that reflects these three-point correlations by assigning $P_{\text{data}} = \frac{1}{4}$ to these four patterns, and setting $P_{\text{data}} = 0$ otherwise. A Boltzmann machine with three neurons cannot represent this input distribution, because there is no energy function of the form (4.1) that has four global minima at these patterns. Also Hopfield nets fail for the XOR function: the four states are not attractors of a Hopfield net with three neurons (Exercise 2.15).

So the three-point correlations encoded in the four patterns cannot be represented in terms of a Boltzmann machine in its simplest form. Also Hopfield nets fail for the XOR function: the four states are not attractors of a Hopfield net with three neurons. One could consider neural networks with third- or higher-order couplings [35],

$$H = -\tfrac{1}{2}\sum_{ij} w^{(2)}_{ij} s_i s_j - \tfrac{1}{6}\sum_{ijk} w^{(3)}_{ijk} s_i s_j s_k + \dots \tag{4.33}$$

(Exercise 2.7). But the number of weights proliferates as the order increases, rendering the training very slow.

An alternative is to use Boltzmann machines with *hidden* neurons, that are neither input nor output units (Figure 4.1). The idea is that the hidden neurons can learn to represent such correlations [35]. The learning rule for the Boltzmann machines with hidden neurons is very similar to Equation (4.31), but when the number of hidden neurons is large, the Boltzmann machine is very slow to train. It is more efficient to remove all weights between visible neurons, and between hidden neurons. This is described in the next Section.

4.5 Restricted Boltzmann machines*

Restricted Boltzman machines [43] consist of visible and hidden neurons arranged in an undirected bipartite graph (Figure 4.5): the only connections are between

neurons of different kinds, there are no connections between visible neurons, no connections between hidden neurons either. So the energy function for a restricted Boltzmann machine for N visible neurons v_j and M hidden neurons h_i takes the form

$$H = -\sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij} h_i v_j + \sum_{j=1}^{N} \theta_j^{(\mathrm{v})} v_j + \sum_{i=1}^{M} \theta_i^{(\mathrm{h})} h_i \,, \tag{4.34}$$

with thresholds $\theta_j^{(\mathrm{v})}$ and $\theta_i^{(\mathrm{h})}$. The McCulloch-Pitts dynamics reads

$$h_i' = \begin{cases} 1 & \text{with probability} \quad p(b_i^{(\mathrm{h})}) \\ -1 & \text{with probability} \quad 1-p(b_i^{(\mathrm{h})}) \end{cases} \qquad \text{with} \quad b_i^{(\mathrm{h})} = \sum_{j=1}^{N} w_{ij} v_j - \theta_i^{(\mathrm{h})}, \tag{4.35a}$$

and

$$v_j' = \begin{cases} 1 & \text{with probability} \quad p(b_j^{(\mathrm{v})}) \\ -1 & \text{with probability} \quad 1-p(b_j^{(\mathrm{v})}) \end{cases} \qquad \text{with} \quad b_j^{(\mathrm{v})} = \sum_{i=1}^{M} h_i w_{ij} - \theta_j^{(\mathrm{v})}. \tag{4.35b}$$

The diagonal weights are assumed to vanish, but the weight matrix is not required to be symmetric. Since most often $M \gg N$, it is usually not even a square matrix (Exercise 4.11).

The learning rule for the weights of the restricted Boltzmann machine is derived using gradient ascent on the log-likelihood

$$\log P(\boldsymbol{x}^{(\mu)}) = \log \sum_{h_1=\pm1,\ldots,h_M=\pm1} P_{\mathrm{B}}(\boldsymbol{v}=\boldsymbol{x}^{(\mu)}, \boldsymbol{h}) \tag{4.36}$$

for a single pattern $\boldsymbol{x}^{(\mu)}$. Proceeding as in the previous Section one finds:

$$\delta w_{mn}^{(\mu)} = \eta\big(\langle h_m x_n^{(\mu)}\rangle_{\mathrm{data}} - \langle h_m v_n\rangle_{\mathrm{model}}\big). \tag{4.37}$$

The first average

$$\langle h_m x_n^{(\mu)}\rangle_{\mathrm{data}} = \sum_{h_1=\pm1,\ldots,h_M=\pm1} h_m x_n^{(\mu)} \left[\prod_{i=1}^{M} P(h_i|\boldsymbol{v}=\boldsymbol{x}^{(\mu)})\right] \tag{4.38}$$

can be evaluated further, using the fact that there are no links between the hidden units. Making use of the update rule (4.35a) we find

$$\sum_{h_m=\pm1} h_m P(h_m|\boldsymbol{v}=\boldsymbol{x}^{(\mu)}) = p(b_m^{(\mathrm{h})}) - [1-p(b_m^{(\mathrm{h})})] = \tanh(b_m^{(\mathrm{h})}), \tag{4.39}$$

just like Equation (3.7). Using the normalisation condition $1 = \sum_{h_k=\pm 1} P(h_k|\boldsymbol{v} = \boldsymbol{x}^{(\mu)})$ yields:

$$\langle h_m x_n^{(\mu)} \rangle_{\text{data}} = \tanh(b_m^{(\text{h})})\, x_n^{(\mu)} = \tanh\Big(\sum_{j=1}^{N} w_{mj} x_j^{(\mu)} - \theta_m^{(\text{h})}\Big)\, x_n^{(\mu)}.$$

The second average on the r.h.s. of Equation (4.37) simplifies to

$$\langle h_m v_n \rangle_{\text{model}} = \Big\langle \tanh\Big(\sum_{j=1}^{N} w_{mj} v_j - \theta_m^{(\text{h})}\Big)\, v_n \Big\rangle_{\text{model}}. \tag{4.40}$$

The average is computed by Monte-Carlo sampling, using the McCulloch-Pitts dynamics (4.35) to generate the sequence

$$\boldsymbol{v}_{t=0} \to \boldsymbol{h}_{t=0} \to \boldsymbol{v}_{t=1} \to \boldsymbol{h}_{t=1} \to \boldsymbol{v}_{t=2} \to \cdots. \tag{4.41}$$

In the limit $t \to \infty$ the steady state of this sequence is distributed according to the model distribution, the Boltzmann distribution with energy function (4.34). In general only the asynchronous McCulloch-Pitts dynamics can be proven to converge (Sections 2.5 and 4.1). Here, however, the Markov chain can be generated more efficiently by updating all hidden neurons $\boldsymbol{h}_t$ at the same time, given $\boldsymbol{v}_t$, because the different h_i are independent since there are no links between them. In the same way the visible neurons $\boldsymbol{v}_t$ are updated in parallel. To speed up the computation further, one usually only iterates for a finite number of steps, up to $t = k$, and initialises the chain with $\boldsymbol{v}_{t=0} = \boldsymbol{x}^{(\mu)}$. After k steps one approximates

$$\Big\langle \tanh\Big(\sum_{j=1}^{N} w_{mj} v_j - \theta_m^{(\text{h})}\Big)\, v_n \Big\rangle_{\text{model}} \approx \tanh\Big(\sum_{j=1}^{N} w_{mj} v_{j,t=k} - \theta_m^{(\text{h})}\Big)\, v_{n,t=k}\,. \tag{4.42}$$

This algorithm is called *contrastive-divergence* or CD-k algorithm. Since the average over the model distribution is approximated [Equation (4.42)], CD-k does not precisely correspond to gradient ascent (Algorithm 2). In summary,

$$\delta w_{mn} = \eta\Big[\tanh\Big(\sum_j w_{mj} v_{j,t=0} - \theta_m^{(\text{h})}\Big) v_{n,t=0} - \tanh\Big(\sum_j w_{mj} v_{j,t=k} - \theta_m^{(\text{h})}\Big) v_{n,t=k}\Big]. \tag{4.43}$$

The analogous update rules for the weights read:

$$\delta\theta_n^{(\text{v})} = -\eta\big(v_{n,t=0} - v_{n,t=k}\big), \tag{4.44a}$$

$$\delta\theta_m^{(\text{h})} = -\eta\Big[\tanh\Big(\sum_j w_{mj} v_{j,t=0} - \theta_m^{(\text{h})}\Big) - \tanh\Big(\sum_j w_{mj} v_{j,t=k} - \theta_m^{(\text{h})}\Big)\Big]. \tag{4.44b}$$

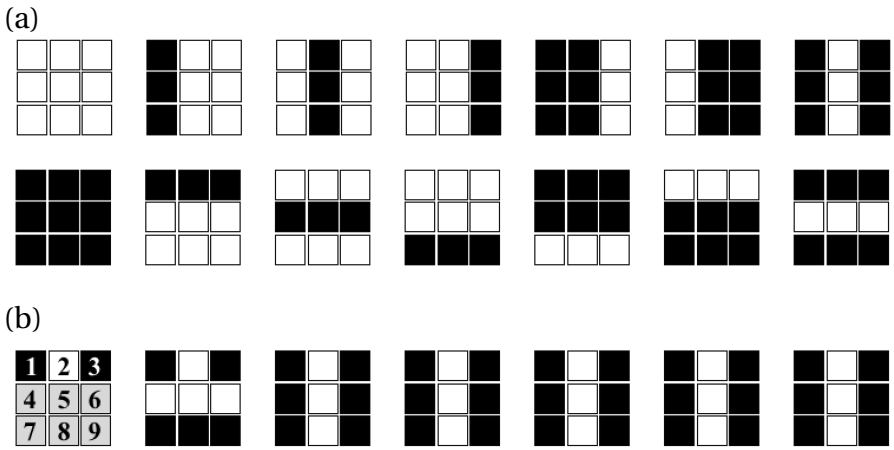


Figure 4.6: Pattern completion for bars-and-stripes data set [35]. (a) All patterns in the 3×3 bars-and-stripes data set, □ corresponds to -1, ■ to $+1$. (b) The three visible units $[v_1, v_2, v_3]$ corresponding to the first row are clamped to $[+1,-1,+1]$ and remain fixed to these values. The remaining units are initially set to 0 (gray bits), their states are allowed to change while sampling from the restricted Boltzmann machine using After a short transient of the McCulloch-Pitts dynamics, the pattern is correctly completed. Schematic, after Figure 7 in Ref. [14].

The derivation is left as an exercise (Exercise 4.12). Usually restricted Boltzmann machines have 0/1 neurons with state values 0 and 1 instead of -1 and 1. For 0/1 neurons the CD-k algorithm is slightly different (Exercise 4.13).

Figure 4.6 illustrates how a restricted Boltzmann machine can learn to complete patterns, using the bars-and-stripes ensemble as an example. The restricted Boltzmann machine is trained using the CD-k algorithm. Now consideer a partially obscured pattern: only the upper row is known, with $v_1 = +1$ (■), $v_2 = -1$ (□), and $v_3 = +1$ (■). The remaining bits $v_4, \ldots, v_9$ are obscured, their states are set to zero as shown in Figure 4.6(b). To complete the pattern, one samples from the Boltzmann distribution $P_{\mathrm{B}}(v_4, \ldots, v_9 | v_1 = +1, v_2 = -1, v_3 = +1)$ keeping $v_1 = +1, v_2 = -1, v_3 = +1$ fixed (*clamping* these neurons), and iterating the McCulloch-Pitts dynamics for the remaining ones. Panel (b) shows how the machine outputs the correct completed pattern after some McCulloch-Pits steps.

This requires hidden neurons, because three-point correlations are needed to discriminate between bar and stripe patterns. In general a restricted Boltzmann machine can approxiate a distribution P_{data} of binary input data better with more hidden neurons. How many are needed? The answer is not known in general, but $M \sim 2^N$ hidden neurons are sufficient, because each hidden neuron can encode one of the binary input patterns (*winning neuron*, Section 7.1). Figure 4.7 illustrates how well a restricted Boltzmann machine approximates the XOR distribution introduced

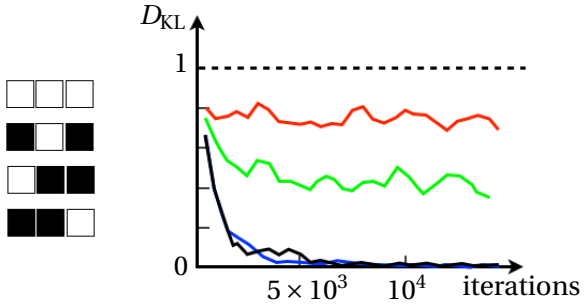


Figure 4.7: Restricted-Boltzmann-machine learning for the XOR problem (left), see Section 4.4. Right: Kullback-Leibler divergence D_{KL} during training as a function of the iteration number, for different numbers of hidden neurons: $M = 0$ (red), 2 (green), 4 (blue) and 8 (black). Schematic, based on simulations performed by Arvid Wenzel Wartenberg using the CD-k algorithm for $k = 100$, with learning rate $\eta = 0.1$, averaging over 500 realisations.

in Section 4.4. The Figure shows how the Kullback-Leibler divergence decreases during training, for different numbers of hidden neurons (Exercise 4.9). In this example there are $N = 3$ inputs. We see that $M = 2^3 = 8$ hidden neurons allow the restricted Boltzmann machine to approximate the data distribution very precisely, as expected. The machine works well also with four hidden neurons. Two hidden neurons are not sufficient, as the Figure shows.

Restricted Boltzmann machines are *generative models*, they can be used to sample from a distribution the machine has learned [14]. As illustrated above (Figure 4.6), they can complete missing information. Restricted Boltzmann machines can also learn to classify patterns, by learning a distribution of binary inputs together with their labels. To this end one splits the visible neurons into input neurons (inputs) and output neurons (labels). This is a supervised-learning task, the subject of Part II. Recently, restricted Boltzmann machines were used to represent and analyse ground-state wave functions of quantum many-body systems [44]. Restricted Boltzmann machines have also been used as models for *parsimonious control* [45]. The question can be phrased as: how complex does the brain of an insect have to be to allow it to learn to walk?

4.6 Summary

This Chapter dealt with the Boltzmann distribution. It was shown that the stochastic McCulloch-Pitts dynamics (3.1) has the Boltzmann distribution as a steady state. We also showed that the update rule (3.1) is a special case of the Markov-chain Monte-Carlo algorithm, for Hopfield nets with energy function (2.45). Since these

algorithms tend to decrease the energy function, they can be used to solve complex optimisation problems. In simulated annealing one gradually reduces the noise level as the simulation proceeds. This mimics the slow cooling of a physical system, an efficient way of bringing the system into its global optimum.

Boltzmann machines are generalisations of Hopfield nets that can learn distributions of binary data by iteratively changing the weights and thresholds until the corresponding Boltzmann distribution approximates the data distribution. The learning rule is derived using gradient ascent on a target function, in this case the log-likelihood. A related idea is used for training deep neural networks with stochastic gradient descent (Part II). To learn general input distributions of binary patterns requires hidden neurons, also this is a central topic of Part II. Since Boltzmann machines with many hidden neurons are hard to train, one removes connections that are not needed. Restricted Boltzmann machines have connections only beween visible and hidden neurons.

4.7 Further reading

Older but still good references for Monte-Carlo methods in Statistical Physics are the books *Monte Carlo methods in Statistical Physics* edited by Binder [46] and Sokal's lecture notes [47]. For a concise introduction to Boltzmann machines refer to the book by MacKay [35], or the one by Murphy [48]. Ref. [49] is a more mathematical review of restricted Boltzmann machines. How many hidden neurons should one allow for in a restricted Boltzmann machine? The general answer is not known, but upper bounds for a sufficient number of hidden neurons are derived and discussed in Ref. [50], together with upper bounds for the Kullback-Leibler divergence. *Deep-belief networks* consist of layers of restricted Boltzmann machines [2]. Contrastive-divergence training for such deep architectures (nets with many layers) is one of the first examples of deep-learning algorithms [51] (Chapter 7).

4.8 Exercises

4.1 Travelling-salesman problem. Derive Equation (4.17) for the path length in the travelling-salesman problem.

4.2 Weights and thresholds for travelling-salesman problem. Derive expressions for the weights and the thresholds for the energy function (4.20) for the travelling-salesman problem. See Ref. [39].

4.3 Asymmetric weights. Show that Equations (3.1) and (4.3) are not equivalent for the network shown in Figure 2.9. The reason is that the weights are not symmetric.

4.4 Simulated annealing with 0/1-neurons. Write down the equivalent of the stochastic dynamics (3.1) for 0/1-neurons with states $n_j = 0$ or 1. Write down the equivalent form (4.3) with energy function $H = -\frac{1}{2}\sum_{ij} w_{ij} n_i n_j + \sum_i \mu_i n_i$. Why are the two formulations *not* equivalent if the weights are asymmetric, or if some w_{ii} are positive?

4.5 Metropolis algorithm. Use the Metropolis algorithm to generate a Markov chain that samples the exponential distribution $P(x) = \exp(-x)$.

4.6 Markov chain. Figure 4.8 illustrates the transition probabilities $p_{l\to k}$ for a Markov chain on a state space with three states. Find the steady state of this Markov chain. Does this chain satisfy detailed balance?

4.7 Double-digest problem. Implement the Metropolis algorithm for the double-digest problem. Denote the ordered set of fragment lengths produced by digesting with enzyme A by $a = \{a_1, \ldots, a_n\}$, where $a_1 \geq a_2 \geq \ldots \geq a_n$. Similarly $b = \{b_1, \ldots, b_m\}$ ($b_1 \geq b_2 \geq \ldots \geq b_m$) for fragment lengths produced by digesting with enzyme B, and $c = \{c_1, \ldots, c_l\}$ ($c_1 \geq c_2 \geq \ldots \geq c_l$) for fragment lengths produced by digesting first with A and then with B. Given permutations σ and μ of the sets a and b correspond to a set of c-fragments we denote it by $\hat{c}(\sigma,\mu)$. Use the energy function $H(\sigma,\mu) = \sum_j c_j^{-1}[c_j - \hat{c}_j(\sigma,\mu)]^2$. Configuration space is the space of all permutation pairs (σ,μ). Local moves correspond to inversions of short subsequence of σ and/or μ. Check that the scheme of suggesting new states is symmetric. This is necessary for the algorithm to converge. The solutions of the double-digest problem are degenerate. Determine the degeneracy of the solutions for the fragment sets shown in Table 4.1.

4.8 Kullback-Leibler divergence. Show that the Kullback-Leibler divergence D_{KL}, Equation (4.23), is non-negative, and that it assumes its global minimum $D_{\mathrm{KL}} = 0$ at

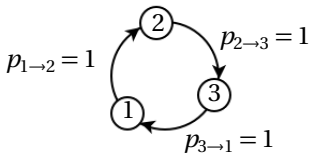


Figure 4.8: Markov chain in a three-dimensional state space. The transition probabilities are denoted by $p_{l\to k}$ (Exercise 4.6).

$L = 10000$

$a =$ [5976, 1543, 1319, 1120, 42]

$b =$ [4513, 2823, 2057, 607]

$c =$ [4513, 1543, 1319, 1120, 607, 514, 342, 42]

$L = 20000$

$a =$ [8479, 4868, 3696, 2646, 169, 142]

$b =$ [11968, 5026, 1081, 1050, 691, 184]

$c =$ [8479, 4167, 2646, 1081, 881, 859, 701, 691, 184, 169, 142]

$L = 40000$

$a =$ [9979, 9348, 8022, 4020, 2693, 1892, 1714, 1371, 510, 451]

$b =$ [9492, 8453, 7749, 7365, 2292, 2180, 1023, 959, 278, 124, 85]

$c =$ [7042, 5608, 5464, 4371, 3884, 3121, 1901, 1768, 1590, 959, 899, 707, 702, 510, 451, 412, 278, 124, 124, 85]

Table 4.1: Example configurations for the double-digest problem for three different chromosome lengths L. For each example, three ordered fragment sets are given, corresponding to the result of digestion with A, with B, and with both A and B.

$P_{\text{data}}(\boldsymbol{x}^{(\mu)}) = P_{\text{B}}(\boldsymbol{s} = \boldsymbol{x}^{(\mu)})$. Show that minimising the D_{KL} is equivalent to maximising the log-likelihood (4.22).

4.9 XOR function. Program a restricted Boltzmann machine to learn the XOR function, by approximating the following data distribution over three-bit binary patterns: $P_{\text{data}} = \frac{1}{4}$ for $[-1,-1,-1]$, $[1,1,-1]$, $[-1,1,1]$, and $[1,-1,1]$, and $P_{\text{data}} = 0$ otherwise. Plot the Kullback-Leibler divergence as a function of iteration number for different numbers of hidden neurons: 0, 2, 4, and 8.

4.10 Shifter ensemble. Explain why the shifter ensemble [31, 35] cannot be approximated by a Boltzmann machine without hidden neurons.

4.11 McCulloch-Pitts dynamics for restricted Boltzmann machine. Write down the deterministic analogue of the update rule (4.35) and show that the energy function of the restricted Boltzmann machine cannot increase under this rule. Note that it is not required that the weight matrix is symmetric, or that the diagonal elements are non-positive.

4.12 Thresholds in restricted Boltzmann machines. Derive the update rule (4.44) for the thresholds for a restricted Boltzmann machine.

4.13 Restricted Boltzmann machine with 0/1 neurons. Derive the contrastive divergence algorithm for training a restricted Boltzmann machine with 0/1 neurons.

Algorithm 2 contrastive divergence CD-k for ± 1 neurons

initialise weights and thresholds;
for $\nu = 1, \ldots, \nu_{\max}$ **do**
  sample p_0 patterns from the data distribution ($p_0 \leq p$);
  for $\mu = 1, \ldots, p_0$ **do**
    initialise $\boldsymbol{v}(0) \leftarrow \boldsymbol{x}^{(\mu)}$;
    update all hidden neurons: $\boldsymbol{b}^{(\mathrm{h})}(0) \leftarrow \mathbb{W}\boldsymbol{v}(0) - \boldsymbol{\theta}^{(\mathrm{h})}$;
    for $i = 1, \ldots, M$ **do**
      $h_i(0) \leftarrow +1$ with probability $p\big(b_i^{(\mathrm{h})}(0)\big)$ otherwise $h_i(0) \leftarrow -1$;
    end for
    for $t = 1, \ldots, k$ **do**
      update all visible neurons: $\boldsymbol{b}^{(\mathrm{v})}(t-1) \leftarrow \boldsymbol{h}(t-1) \cdot \mathbb{W} - \boldsymbol{\theta}^{(\mathrm{v})}$;
      for $j = 1, \ldots, N$ **do**
        $v_j(t) \leftarrow +1$ with probability $p\big(b_j^{(\mathrm{v})}(t-1)\big)$ otherwise $v_j(t) \leftarrow -1$;
      end for
      update all hidden neurons: $\boldsymbol{b}^{(\mathrm{h})}(t) \leftarrow \mathbb{W}\boldsymbol{v}(t) - \boldsymbol{\theta}^{(\mathrm{h})}$;
      for $i = 1, \ldots, M$ **do**
        $h_i(t) \leftarrow +1$ with probability $p\big(b_i^{(\mathrm{h})}(t)\big)$ otherwise $h_i(t) \leftarrow -1$;
      end for
    end for
    compute weight and theshold updates:
    $\delta w_{mn} \leftarrow \eta\Big[\tanh\big(b_m^{(\mathrm{h})}(0)\big)v_n(0) - \tanh\big(b_m^{(\mathrm{h})}(k)\big)v_n(k)\Big]$;
    $\delta\theta_n^{(\mathrm{v})} \leftarrow -\eta[v_n(0) - v_n(k)]$;
    $\delta\theta_m^{(\mathrm{h})} \leftarrow -\eta\big[\tanh\big(b_m^{(\mathrm{h})}(0)\big) - \tanh\big(b_m^{(\mathrm{h})}(k)\big)\big]$;
  end for
  for $\mu = 1, \ldots, p_0$ **do**
    update weights and thresholds;
  end for
end for

PART II
SUPERVISED LEARNING

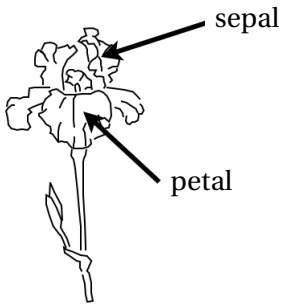
The Hopfield nets described in Part I recognise (or retrieve) patterns. Its neurons act as inputs and outputs. In the pattern-recognition problem, a distorted pattern is fed into the network, the recursive network dynamics is run until a steady state is reached. The aim is that the steady-state values of the neurons converge to those of the correct pattern *associated* with the distorted one.

A related type of problem that is very common are *classification* tasks. The *machine-learning repository* [52] at the University of California Irvine contains a large number of such problems. A well-known example is the *iris data set*. It lists attributes of 150 iris plants. The data set was described by the geneticist R. A. Fisher [53]. For each plant four attributes are given (Figure 5.1): its sepal length, sepal width, petal length, and petal width. Also, each plant is classified into one of three classes: *iris setosa*, *iris versicolor*, or *iris virginica*. The task is to program a neural network that determines the class of a plant from its attributes. To each input (attributes of an iris plant) the network should associate the correct output, the class of the plant. The correct output is referred to as the *target*.

In *supervised learning* one uses a *training* data set of correct input/output pairs. One feeds an input from the training data into the input terminals of the network and compares the states of the output neurons to the target values. The weights and thresholds are changed to minimise the differences between network outputs and targets for all input patterns in the training set. In this way the network learns to associate input patterns in the training set with the correct target values. A crucial question is whether the trained network can *generalise*: does it find the correct targets for input patterns that were not in the training set?

The networks used for supervised learning are called *perceptrons* [10, 11]. They consist of layers of McCulloch-Pitts neurons: an input layer, a number of layers of *hidden* neurons (Chapter 4), and an output layer. The layers are usually arranged from the left (input) to the right (output). All connections are one-way, from neurons in one layer to neurons in the layer immediately to the right. There are no connections between neurons in a given layer, or back to layers on the left. This arrangement ensures convergence of the training algorithm (*stochastic gradient descent*). During training with this algorithm the weights are updated iteratively. In each step, an input is applied and the weights of the network are updated to reduce the error in the output. In a sense each step corresponds to adding a little bit of Hebb's rule to the weights. This is repeated until the network classifies the training set correctly.

Stochastic gradient descent for multi-layer perceptrons has received much attention recently, after it was realised that networks with many hidden layers can be trained to reliably recognise and classify image data, for self-driving cars for instance but also for other applications (*deep learning*).



sepal length	sepal width	petal length	petal width	classification
6.3	2.5	5.0	1.9	virginica
5.1	3.5	1.4	0.2	setosa
5.5	2.6	4.4	1.2	versicolor
4.9	3.0	1.4	0.2	setosa
6.1	3.0	4.6	1.4	versicolor
6.5	3.0	5.2	2.0	virginica

Figure 5.1: Left: petals and sepals of the iris flower. Right: six entries of the iris data set [52]. All lengths in cm. The whole data set contains 150 entries.

5 Perceptrons

Perceptrons [10, 11] are layered feed-forward networks, illustrated in Figure 5.2. The leftmost layer consists of input terminals, drawn in black in Figure 5.2. To the right follows a number of layers of McCulloch-Pitts neurons. The right-most layer of neurons consists of output neurons. The intermediate layers are called hidden layers, because their states are not read out. All connections w_{ij} are one-way: every neuron (or input terminal) feeds only to neurons in the layer immediately to the right. There are no connections within layers, no back connections, no connections that skip a layer.

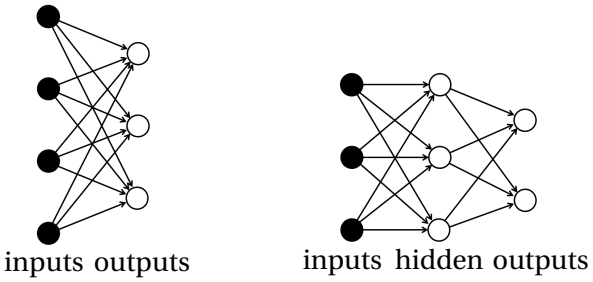


Figure 5.2: Feed-forward network without hidden layer (left), and with one hidden layer (right). The input terminals are coloured black.

There are N input terminals. As in Part I we denote the input patterns by

$$\boldsymbol{x}^{(\mu)} = \begin{bmatrix} x_1^{(\mu)} \\ x_2^{(\mu)} \\ \vdots \\ x_N^{(\mu)} \end{bmatrix} . \tag{5.1}$$

The index $\mu = 1, \ldots, p$ labels the different input patterns in the training set. The output neurons in the network on the left of Figure 5.2 perform the computation

$$O_i = g(B_i) \quad \text{with} \quad B_i = \sum_j W_{ij} x_j - \Theta_i \tag{5.2}$$

The index $i = 1, \ldots, M$ labels the output neurons. Each output neuron has a threshold, Θ_i. In the literature on *deep learning* the thresholds are sometimes referred to as *biases*, defined as $-\Theta_i$. In this book we use the convention (5.2), with a minus sign. The function g is an activation function (Section 1.2).

Now consider the network on the right of Figure 5.2. The states of the neurons in the hidden layer are denoted by V_j, with thresholds θ_j and weights w_{jk}. The net computes:

$$V_j = g\left(b_j\right) \quad \text{with} \quad b_j = \sum_k w_{jk} x_k - \theta_j , \tag{5.3a}$$

$$O_i = g(B_i) \quad \text{with} \quad B_i = \sum_j W_{ij} V_j - \Theta_i . \tag{5.3b}$$

A classification problem is given by a training set of input patterns $\boldsymbol{x}^{(\mu)}$ and corresponding *target values*

$$\boldsymbol{t}^{(\mu)} = \begin{bmatrix} t_1^{(\mu)} \\ t_2^{(\mu)} \\ \vdots \\ t_M^{(\mu)} \end{bmatrix} . \tag{5.4}$$

The perceptron is trained by choosing its weights and thresholds so that the network produces the desired output.

$$O_i^{(\mu)} = t_i^{(\mu)} \quad \text{for all} \quad i \quad \text{and} \quad \mu . \tag{5.5}$$

If we take $t_i^{(\mu)} = x_i^{(\mu)}$ for $i = 1, \ldots, N$ then the task is pattern recognition, as discussed in Part I. In the Hopfield networks described in Part I, the weights were assigned using Hebb's rule (2.26). Perceptrons, by contrast, are trained by iteratively updating

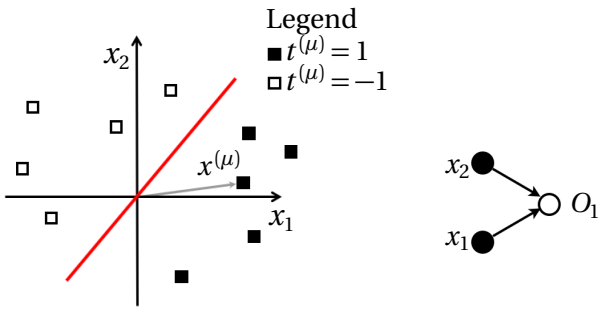


Figure 5.3: Left: classification problem with two-dimensional real-valued inputs and targets equal to ±1. The red line is the decision boundary (see text). Right: corresponding perceptron.

their weights and thresholds until Equation (5.5) is satisfied. This is achieved by repeatedly adding small multiples of Hebb's rule to the weights (Section 5.2). An equivalent approach is to define an energy function, a function of the weights of the network, that has a global minimum when Equation (5.5) is satisfied. The network is trained by taking small steps in weight space that reduce the energy function (gradient descent, Section 5.3).

5.1 A classification problem

To illustrate how perceptrons can solve classification problems, we consider a very simple example (Figure 5.3). There are ten patterns, each has two real-valued components:

$$\boldsymbol{x}^{(\mu)} = \begin{bmatrix} x_1^{(\mu)} \\ x_2^{(\mu)} \end{bmatrix}. \tag{5.6}$$

In Figure 5.3 the patterns are drawn as points in the x_1-x_2 plane, the *input plane*. There are two classes of patterns, with targets ±1

$$t^{(\mu)} = 1 \quad \text{for} \quad \blacksquare \quad \text{and} \quad t^{(\mu)} = -1 \quad \text{for} \quad \square. \tag{5.7}$$

The activation function consistent with the possible target values is the signum function, $g(b) = \text{sgn}(b)$. The perceptron has two input terminals connected to a single output neuron. Since there is only one neuron, we can arrange the weights into a weight vector

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}. \tag{5.8}$$

The network performs the computation

$$O = \text{sgn}(w_1 x_1 + w_2 x_2 - \theta) = \text{sgn}(\boldsymbol{w} \cdot \boldsymbol{x} - \theta). \tag{5.9}$$

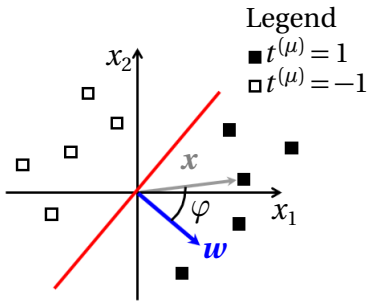


Figure 5.4: The perceptron classifies the patterns correctly for the weight vector $\boldsymbol{w}$ shown, orthogonal to the decision boundary.

Here $\boldsymbol{w}\cdot\boldsymbol{x}=w_1x_1+w_2x_2$ is the scalar product between the vectors $\boldsymbol{w}$ and $\boldsymbol{x}$.

This very simple example allows us to find a geometrical interpretation of the classification problem. We see in Figure 5.3 that the patterns fall into two clusters: □ to the right and ■ to the left. We can classify the patterns by drawing a line that separates the two clusters, so that everything on the right of the line has $t=1$, while the patterns on the left of the line have $t=-1$. This line is called the *decision boundary*. To find the geometrical significance of Equation (5.9), let us ignore the threshold for a moment, so that

$$O=\mathrm{sgn}(\boldsymbol{w}\cdot\boldsymbol{x}). \tag{5.10}$$

The classification problem takes the form

$$\mathrm{sgn}\left(\boldsymbol{w}\cdot\boldsymbol{x}^{(\mu)}\right)=t^{(\mu)}. \tag{5.11}$$

To evaluate the scalar product we write the vectors as

$$\boldsymbol{w}=|\boldsymbol{w}|\begin{pmatrix}\cos\beta\\ \sin\beta\end{pmatrix}\quad\text{and}\quad\boldsymbol{x}=|\boldsymbol{x}|\begin{pmatrix}\cos\alpha\\ \sin\alpha\end{pmatrix}. \tag{5.12}$$

Here $|\boldsymbol{w}|=\sqrt{w_1^2+w_2^2}$ denotes the norm of the vector $\boldsymbol{w}$, and α and β are the angles of the vectors with the x_1-axis. Then $\boldsymbol{w}\cdot x=|\boldsymbol{w}||\boldsymbol{x}|\cos(\alpha-\beta)=|\boldsymbol{w}||\boldsymbol{x}|\cos\varphi$, where φ is the angle between the two vectors. When φ is between $-\pi/2$ and $\pi/2$, the scalar product is positive, otherwise negative. As a consequence, the network classifies the patterns in Figure 5.3 correctly if the weight vector is orthogonal to the decision boundary drawn in Figure 5.4.

What is the role of the threshold θ? Equation (5.9) shows that the decision boundary is parameterised by $\boldsymbol{w}\cdot\boldsymbol{x}=\theta$, or

$$x_2=-(w_1/w_2)\,x_1+\theta/w_2. \tag{5.13}$$

Therefore the threshold determines the intersection of the decision boundary with the x_2-axis (equal to θ/w_2). This is illustrated in Figure 5.5.

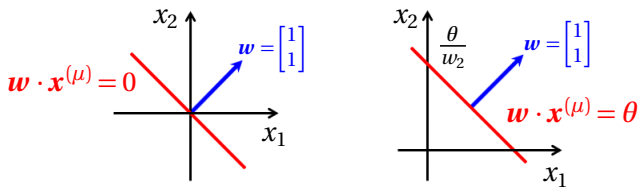


Figure 5.5: Decision boundaries without and with threshold.

The decision boundary – the straight line orthogonal to $\boldsymbol{w}$ – should divide inputs with positive and negative targets. If no such line can be found, then the problem cannot be solved with a single neuron. Conversely, if such a line exists, the problem can be solved, and it is called *linearly separable*). Otherwise the problem is not linearly separable. This can occur only when $p > N$. Examples of problems that are linearly separable and not linearly separable are shown in Figure 5.6.

Other examples are *Boolean functions*. A Boolean function takes N binary inputs and has one binary output. The Boolean AND function (two inputs) is illustrated in Figure 5.7. The value table of the function is shown on the left. The graphical representation is shown in the centre of the Figure (□ corresponds to $t = -1$ and ■ to $t = +1$). Also shown is the decision boundary, the weight vector $\boldsymbol{w}$, and the network layout with the corresponding values of the weights and the threshold. It is important to note that the decision boundary is not unique, neither are the weight and threshold values that solve the problem. The norm of the weight vector, in particular, is arbitrary. Neither is its direction uniquely specified.

Figure 5.8 shows that the Boolean XOR function is not linearly separable [54]. There are 16 different Boolean functions of two variables. Only two are not linearly separable, the XOR and the NOT XOR function.

Up to now we discussed only one output unit. If the classification problem requires several output units, each has its own weight vector $\boldsymbol{w}_i$ and threshold θ_i. We can group the weight vectors into a weight matrix as in Part I, so that the $\boldsymbol{w}_i$ are the rows of $\mathbb{W}$.

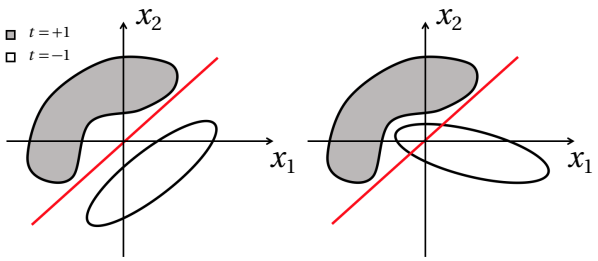
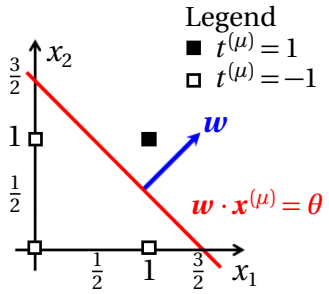


Figure 5.6: Linearly separable and non-separable data.

x_1	x_2	t
0	0	-1
0	1	-1
1	0	-1
1	1	+1



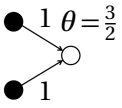


Figure 5.7: Boolean AND function: value table, geometrical representation, and network layout. The weight values are written next to the connections.

5.2 Iterative learning algorithm

In the previous Section we determined the weights and threshold for the Boolean AND function by inspection. Now we discuss an algorithm that allows a computer to find the weights iteratively. How this works is illustrated in Figure 5.9. In panel (**a**), the pattern $\boldsymbol{x}^{(8)}$ ($t^{(8)}=1$) is on the wrong side of the decision boundary. To turn the decision boundary anti-clockwise one *adds* a small multiple of the pattern vector $\boldsymbol{x}^{(8)}$ to the weight vector

$$\boldsymbol{w}'=\boldsymbol{w}+\delta\boldsymbol{w} \quad \text{with} \quad \delta\boldsymbol{w}=\eta\boldsymbol{x}^{(8)}. \tag{5.14}$$

The parameter $\eta>0$ is called the *learning rate*. It must be small, so that the decision boundary is not rotated too far. The result is shown in panel (**b**). Panel (**c**) shows another case, where pattern $\boldsymbol{x}^{(4)}$ ($t^{(4)}=-1$) is on the wrong side of the decision boundary. In order to turn the decision boundary in the right way, anti-clockwise, one *subtracts* a small multiple of $\boldsymbol{x}^{(4)}$:

$$\boldsymbol{w}'=\boldsymbol{w}+\delta\boldsymbol{w} \quad \text{with} \quad \delta\boldsymbol{w}=-\eta\boldsymbol{x}^{(4)}. \tag{5.15}$$

Note the minus sign. These two learning rules combine to

$$\boldsymbol{w}'=\boldsymbol{w}+\delta\boldsymbol{w}^{(\mu)} \quad \text{with} \quad \delta\boldsymbol{w}^{(\mu)}=\eta\, t^{(\mu)}\boldsymbol{x}^{(\mu)}. \tag{5.16}$$

x_1	x_2	t
0	0	-1
0	1	+1
1	0	+1
1	1	-1

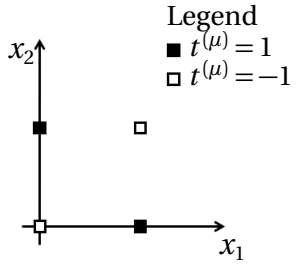


Figure 5.8: The Boolean XOR function is not linearly separable.

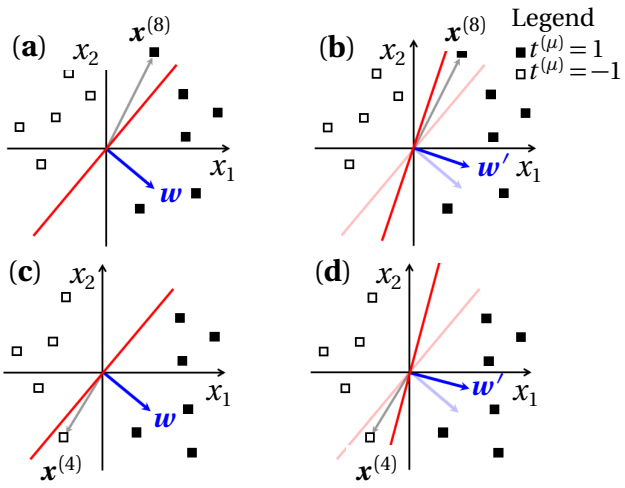


Figure 5.9: Illustration of the learning algorithm. In panel (**a**) the $t=1$ pattern $x^{(8)}$ is on the wrong side of the decision boundary. To correct the error the weight must be rotated anti-clockwise [panel (**b**)]. In panel (**c**) the $t=-1$ pattern $x^{(4)}$ is on the wrong side of the decision boundary. To correct the error the weight must be rotated anti-clockwise [panel (**d**)].

For more than one output unit the rule reads

$$w'_{ij} = w_{ij} + \delta w_{ij}^{(\mu)} \quad \text{with} \quad \delta w_{ij}^{(\mu)} = \eta t_i^{(\mu)} x_j^{(\mu)}. \tag{5.17}$$

This rule is reminiscent of Hebb's rule (2.9), except that here inputs and outputs are associated with distinct units. Therefore we have $t_i^{(\mu)} x_j^{(\mu)}$ instead of $x_i^{(\mu)} x_j^{(\mu)}$. One applies (5.17) iteratively for a sequence of randomly chosen patterns μ, until the problem is solved. This corresponds to adding a little bit of Hebb's rule in each iteration. To ensure that the algorithm stops when the problem is solved, one can use

$$\delta w_{ij}^{(\mu)} = \eta (t_i^{(\mu)} - O_i^{(\mu)}) x_j^{(\mu)}. \tag{5.18}$$

5.3 Gradient descent for linear units

In this Section the learning algorithm (5.18) is derived in a different way, by minimising an energy function using gradient descent. This requires differentiation, therefore we must choose a differentiable activation function. The simplest choice is $g(b) = b$. We set $\theta = 0$, so that the network computes:

$$O_i^{(\mu)} = \sum_k w_{ik} x_k^{(\mu)}. \tag{5.19}$$

A neuron with a linear activation function is called a *linear unit*. The outputs $O_i^{(\mu)}$ assume continuous values, but not necessarily the targets $t_i^{(\mu)}$. For linear units, the classification problem

$$O_i^{(\mu)} = t_i^{(\mu)} \quad \text{for} \quad i = 1, \ldots, N \quad \text{and} \quad \mu = 1, \ldots, p \tag{5.20}$$

has the formal solution

$$w_{ik} = \frac{1}{N} \sum_{\mu\nu} t_i^{(\mu)} \left(\mathbb{Q}^{-1} \right)_{\mu\nu} x_k^{(\nu)}, \tag{5.21}$$

as you can verify by inserting Equation (5.21) into (5.19). Here $\mathbb{Q}$ is the overlap matrix with elements

$$Q_{\mu\nu} = \tfrac{1}{N} \boldsymbol{x}^{(\mu)} \cdot \boldsymbol{x}^{(\nu)} \tag{5.22}$$

(Section 10.2). For the solution (5.21) to exist, the matrix $\mathbb{Q}$ must be invertible. As explained in Section 10.2, this requires that $p \leq N$, because otherwise the pattern vectors are *linearly dependent*, and thus also the columns (and rows) of $\mathbb{Q}$. If the matrix $\mathbb{Q}$ has linearly dependent columns or rows it cannot be inverted.

Let us assume that the patterns are linearly independent, so that the solution (5.21) exists. In this case we can find the solution iteratively. To this end one defines the energy function

$$H(\{w_{ij}\}) = \frac{1}{2} \sum_{i\mu} \left(t_i^{(\mu)} - O_i^{(\mu)} \right)^2 = \frac{1}{2} \sum_{i\mu} \left(t_i^{(\mu)} - \sum_j w_{ij} x_j^{(\mu)} \right)^2. \tag{5.23}$$

Here H is regarded as a function of the weights w_{ij}, unlike the energy function in Part I which is a function of the state-variables of the neurons. The energy function (5.23) is non-negative, and it vanishes for the optimal w_{ij} if the pattern vectors $\boldsymbol{x}^{(\mu)}$ are linearly independent. This solution of the classification problem corresponds to the global minimum of H. To find it one uses *gradient descent*

$$w'_{mn} = w_{mn} + \delta w_{mn} \quad \text{with} \quad \delta w_{mn} = -\eta \frac{\partial H}{\partial w_{mn}}. \tag{5.24}$$

with learning rate $\eta > 0$. The idea is the same as in Chapter 4: one takes many downhill steps in search of the global minimum. To evaluate the derivatives one uses the chain rule together with Equation (4.27), just as in Section 4.4. This yields

$$\delta w_{mn} = \eta \sum_{\mu} \left(t_m^{(\mu)} - O_m^{(\mu)} \right) x_n^{(\mu)}. \tag{5.25}$$

This learning rule is very similar to Equation (5.18). The difference is that Equation (5.25) contains a sum over all patterns (*batch training* or *batch mode*). An advantage

Figure 5.10: Left: 5 points in general position in the plane. Right: these points are not in general position because three points lie on a straight line.

of the rule (5.25) is that it is derived from an energy function. This allows to analyse the convergence of the algorithm.

Linear units [Equation (5.19)] are special. You cannot solve the Boolean AND problem (Figure 5.7) with a linear unit, even though the problem is linearly separable: since the pattern vectors $\boldsymbol{x}^{(\mu)}$ are linearly dependent, the solution (5.21) does not exist. Shifting the patterns or introducing a threshold does not change this fact. Linear separability does not imply linear independence (but the converse is true).

In Section 5.5 we discuss how to solve problems that are not linearly separable using a hidden layer of neurons with non-linear activation functions. Note that introducing hidden layers with linear units does not help, because the resulting input-output mapping is still linear if all neurons have linear activation functions, so that only problems with $p \leq N$ can be solved. This is the main reason for using hidden layers non-linear activation functions. There are four points to keep in mind. First, hidden layers are required, because a single neuron with a continuous non-linear activation function can only solve problems with linearly independent patterns (Exercise 5.11). Second, if the problem is linearly separable then we can use gradient descent to determine suitable weights (and thresholds). Third, for gradient descent we must require that the activation function $g(b)$ is differentiable, or at least piecewise differentiable. Fourth, we calculate the gradients using the chain rule, resulting in factors of derivatives $\frac{\mathrm{d}}{\mathrm{d}b}g(b)$. This is the origin of the *vanishing-gradient problem* (Chapter 7).

5.4 Classification capacity

Consider the binary classfication problem introduced in the beginning of this Chapter, in Section 5.1. In Chapter 3 we analysed the storage capacity of Hopfield nets. The analogous question for the classification problem is: how many patterns can a single neuron with activation function $g(b) = \mathrm{sgn}(b)$ classify? As in the case of Hopfield nets one can find a general answer for random classification problems.

Consider p points with coordinate vectors $\boldsymbol{x}^{(\mu)}$ in N-dimensional space. Assign

random targets as follows:

$$t^{(\mu)} = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}. \end{cases} \tag{5.26}$$

This random classification problem is *homogeneously* linearly separable if we can find an N-dimensional weight vector $\boldsymbol{w}$, so that $\boldsymbol{w} \cdot \boldsymbol{x} = 0$ is a valid decision boundary that goes through the origin:

$$\boldsymbol{w} \cdot \boldsymbol{u}^{(\mu)} > 0 \quad \text{if} \quad t^{(\mu)} = 1 \quad \text{and} \quad \boldsymbol{w} \cdot \boldsymbol{u}^{(\mu)} < 0 \quad \text{if} \quad t^{(\mu)} = -1 . \tag{5.27}$$

So *homogeneously linearly separable* problems are classification problems that are linearly separable by a hyperplane that contains the origin.

Now assume that the points (including the origin) are in *general position* (Figure 5.10). In this case *Cover's theorem* [55] gives an expression for the probability that the random classification problem of p patterns in dimension N is homogeneously linearly separable:

$$P(p,N) = \begin{cases} \left(\frac{1}{2}\right)^{p-1} \sum_{k=0}^{N-1} \binom{p-1}{k} & \text{for } p > N , \\ 1 & \text{otherwise}. \end{cases} \tag{5.28}$$

Here $\binom{l}{k} = \frac{l!}{(l-k)!k!}$ are the binomial coefficients, for $l \geq k \geq 0$. If one *defines* $\binom{l}{k} = 0$ for $l < k$ then Equation (5.28) can be written as $P(p,N) = (\frac{1}{2})^{p-1} \sum_{k=0}^{N-1} \binom{p-1}{k}$. Equation (5.28) is proven by recursion, starting from a set of $p-1$ points in general position. Assume that the number $C(p-1,N)$ of homogeneously linearly separable classification problems given these points is known. After adding one more point, one can compute the $C(p,N)$ in terms of $C(p-1,N)$. Recursion yields Equation (5.28). Figure 5.11 shows this probability as a function of $\alpha = p/N$ for different values of N. For $p \leq N$ any random classification problem is homogeneously linearly separable. In this case the pattern vectors are linearly independent, so that the problem can also be solved by a linear unit (Section 5.3). But a neuron with activation function $\mathrm{sgn}(b)$ can classify problems with more than N patterns. In the limit of $N \to \infty$, the function $P(\alpha N, N)$ approaches a step function $\theta_{\mathrm{H}}(2-\alpha)$ (Exercise 5.12). In this limit the maximal classification capacity is $\alpha_{\max} = 2$.

What is the expected classification capacity for finite values of N? To answer this question, consider a random sequence of patterns $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ and targets $t_1, t_2, \ldots$ and ask [55]: what is the distribution of the largest integer so that the problem $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ is separable in dimension N, but $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}$ is not? $P(n,N)$ is the probability that n patterns are linearly separable in N-dimensional input space. We can write $P(n+1,N) = q(n+1|n)P(n,N)$ where $q(n+1|n)$ is the conditional

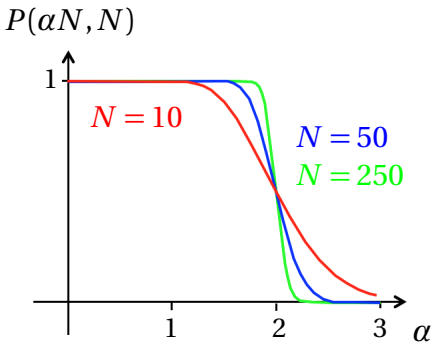


Figure 5.11: Probability (5.28) of separability as a function of $\alpha = p/N$ for three different values of the dimension N of input space. Note the pronounced threshold near $\alpha = 2$, for large values of m.

probability that $n+1$ patterns are linearly separable if the n patterns were. Then the probability that $n+1$ patterns are not separable (but n patterns are) reads $(1-q)P(n,N) = P(n,N) - P(n+1,N)$. We can interpret the right-hand side of this Equation as a distribution p_n of the random variable n, the maximal number of separable patterns in dimension N:

$$p_n = P(n,N) - P(n+1,N) = \left(\frac{1}{2}\right)^n \binom{n-1}{N-1} \quad \text{for} \quad n = 0,1,2,\ldots .$$

It follows that the expected maximal number of separable patterns is

$$\langle n \rangle = \sum_{n=0}^{\infty} n\, p_n = 2N\,. \tag{5.29}$$

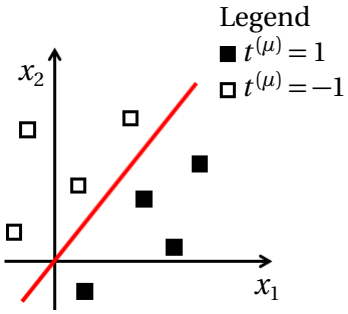
So the expected classification capacity is twice the input dimension:

$$\langle \alpha_{\max} \rangle = 2\,. \tag{5.30}$$

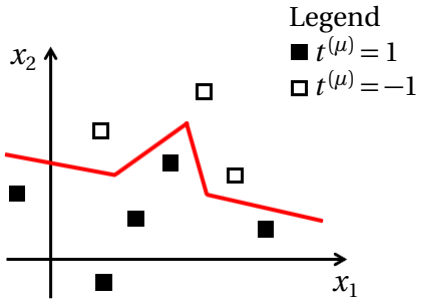
This quantifies the notion that it is easier to separate patterns in higher embedding dimensions.

5.5 Multi-layer perceptrons

In Sections 5.1 and 5.2 we discussed how to solve linearly separable problems [Figure 5.12(**a**)]. The aim of this Section is to show that non-separable problems like the one in Figure 5.12(**b**) can be solved by a perceptron with one hidden layer. A network that does the trick for the classification problem in Figure 5.12(**b**) is depicted in Figure 5.13. Here the hidden neurons are 0/1 units, but the output neuron gives ±1,



(a) Linearly separable problem.



(b) Problem not linearly separable.

Figure 5.12: Problems that are not linearly separable can be solved by a piecewise linear decision boundary.

as in the previous Section. The network computes with the following rules:

$$\begin{aligned} V_j^{(\mu)} &= \theta_{\mathrm{H}}\left(b_j^{(\mu)}\right) \quad \text{with} \quad b_j^{(\mu)} = \sum_k w_{jk} x_k^{(\mu)} - \theta_j \,, \\ O_1^{(\mu)} &= \mathrm{sgn}\left(B_1^{(\mu)}\right) \quad \text{with} \quad B_1^{(\mu)} = \sum_j W_{1j} V_j^{(\mu)} - \Theta_1 \,. \end{aligned} \tag{5.31}$$

Here $\theta_{\mathrm{H}}(b)$ is the Heaviside function (Figure 2.10). Each of the three neurons in the hidden layer has its own decision boundary. The idea is to choose weights and thresholds in such a way that the three decision boundaries partition the input plane into distinct regions, so that each region contains either only $t=0$ patterns or $t=1$ patterns. We shall see that the values of the hidden neurons encode the different regions. Finally, the output neuron associates the correct target value with each region.

How this construction works is shown in Figure 5.14. The left part of the Figure shows the three decision boundaries. The indices of the corresponding hidden neurons are drawn in blue. Also shown are the weight vectors. The regions are

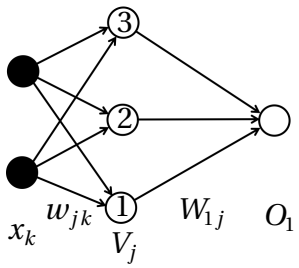
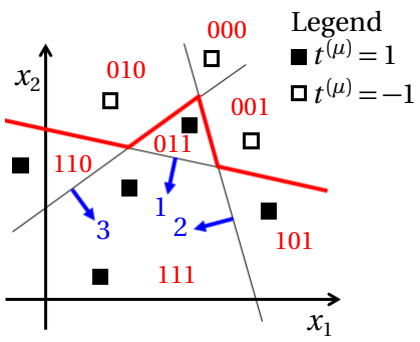


Figure 5.13: Hidden-layer perceptron to solve the problem shown in Figure 5.12 (**b**). The three hidden neurons are 0/1 neurons, the output neuron produces ±1.



V_1	V_2	V_3	target
0	0	0	-1
1	0	0	-
0	1	0	-1
0	0	1	-1
1	1	0	+1
1	0	1	+1
0	1	1	+1
1	1	1	+1

Figure 5.14: Left: decision boundaries and regions. Right: encoding of the regions and corresponding targets. The region 100 does not exist.

encoded with a three-digit binary code. The value of the j-th digit is the value of the j-th hidden neuron: $V_j = 1$ if the pattern is on the weight-vector side of the decision boundary, and $V_j = 0$ on the other side. The Table shows the targets associated with each region, together with the code of the region.

A graphical representation of the output problem is shown in Figure 5.15. The problem is linearly separable. The following function computes the correct output for each region:

$$O_1^{(\mu)} = \mathrm{sgn}\left(V_1^{(\mu)} + V_2^{(\mu)} + V_3^{(\mu)} - \tfrac{3}{2}\right). \tag{5.32}$$

This completes the construction of a solution. It is not unique. In summary, one can solve non-linearly separable problems by adding a hidden layer. The neurons in the hidden layer define segments of a piecewise linear decision boundary. More neurons are needed if the decision boundary is very wiggly.

Figure 5.16 shows another example, how to solve the Boolean XOR problem with a perceptron that has two 0/1 neurons in a hidden layer, with thresholds $\frac{1}{2}$ and $\frac{3}{2}$, and all weights equal to unity. The output neuron has weights $+1$ and -1 and

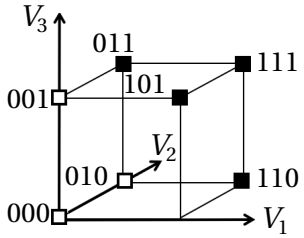
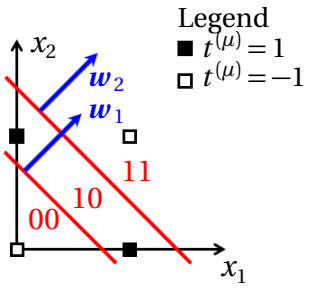
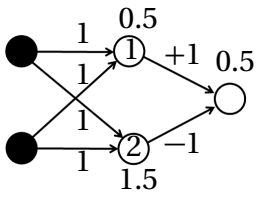


Figure 5.15: Graphical representation of the output problem for the classification problem shown in Figure 5.14.





V_1	V_2	t
0	0	-1
1	0	+1
0	1	+1
1	1	-1

Figure 5.16: Boolean XOR function: geometrical representation, network layout, and value table for the output neuron. The two hidden neurons as 0/1 neurons, the output produces ±1.

threshold $\frac{1}{2}$:

$$O_1 = \text{sgn}(V_1 - V_2 - \tfrac{1}{2}). \tag{5.33}$$

Minsky and Papert [54] proved in 1969 that all Boolean functions can be represented by multilayer perceptrons, but that at least one hidden neuron must be connected to *all* input terminals. This means that not all neurons in the network are *locally* connected (have only a few incoming weights). Since fully connected networks are much harder to train, Minsky and Papert offered a somewhat pessimistic view of learning with perceptrons, resulting in a controversy [56]. Now, almost 50 years later, the perspective has changed. Convolutional networks (Chapter 7) have only local connections to the inputs and can be trained to recognise objects in images with high accuracy.

In summary, perceptrons are trained on a training set $[\boldsymbol{x}^{(\mu)}, \boldsymbol{t}^{(\mu)}]$ with $\mu = 1, \ldots, p$ by moving the decision boundaries into the correct positions. This is achieved by repeatedly applying Hebb's rule to adjust all weights. This corresponds to using gradient-descent learning on the energy function (5.23). We have not discussed how to update the *thresholds* yet, but it is clear that they too can be updated with gradient-descent learning.

Once all decision boundaries are in the right place we must ask: what happens if we apply the trained network to a new dataset? Does it classify the new inputs correctly? In other words, can the network *generalise*? An example is shown in Figure 5.17. Panel (**a**) shows the result of training the network on a training set. The decision boundary separates $t = -1$ patterns from $t = 1$ patterns, so that the network classifies all patterns in the training set correctly. In panel (**b**) the trained network is applied to patterns in a *validation set*. We see that most patterns are correctly classified, save for one error. This means that the energy function (5.23) is not exactly zero for the validation set. Nevertheless, the network does quite a good job. Usually it is not a good idea to try to precisely classify all patterns near the decision boundary, because real-world data sets are subject to noise. It is a futile

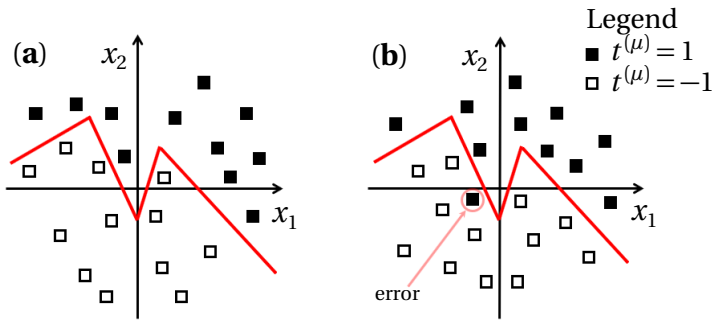


Figure 5.17: (**a**) Result of training the network on a training set. (**b**) Validation by feeding the patterns of a validation set.

effort to try to learn and predict noise.

5.6 Summary

Perceptrons are layered feed-forward networks that can learn to classify data in a training set $(\boldsymbol{x}^{(\mu)}, \boldsymbol{t}^{(\mu)})$. For each input pattern $\boldsymbol{x}^{(\mu)}$ the network finds the correct targets $\boldsymbol{t}^{(\mu)}$. We discussed the learning algorithm for a simple example: real-valued patterns with just two components, and one binary target. This allowed us to represent the classification problem graphically, and to see how linearly separable classification problems can be solved by a simple perceptron. There are three different ways of understanding how the perceptron learns. First, geometrically, to learn means to move the decision boundaries into the right places. Second, this can be achieved by repeatedly adding a little bit of Hebb's rule. Third, this algorithm corresponds to gradient descent on the energy function (5.23). Cover's theorem quantifies the capacity of a simple perceptron to separate patterns with binary targets. Finally we discussed how to solve non-linearly separable classification problems with perceptrons with a hidden layer.

5.7 Further reading

A short account of the history of perceptron research is the review by Kanal [56]. He discusses the work of Rosenblatt [10, 11], McCulloch and Pitts [9], as well as the early controversy around the book by Minsky and Papert [54]. For a proof of Cover's theorem see Ref. [57].

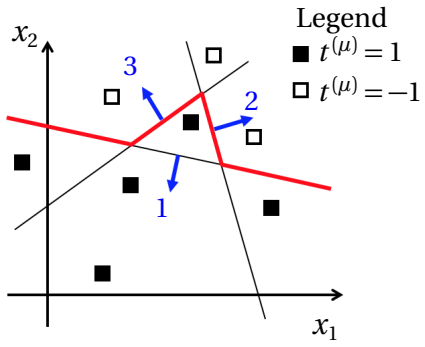


Figure 5.18: Alternative solution of the classification problem shown in Figure 5.14. Exercise 5.4.

5.8 Exercises

5.1 Boolean AND problem. Show that the Boolean AND problem (Figure 5.7) cannot be solved by the rule (5.21).

5.2 Boolean functions. How many Boolean functions with three-dimensional inputs are there? How many of them are linearly separable?

5.3 Output problem for binary classification. The binary classification problem shown in Figure 5.14 can be solved with a network with one hidden layer and one output neuron. Figure 5.15 shows the problem that the output neuron has to solve. Show that such output problems are linear separable if the decision boundaries corresponding to the hidden units allow to partition the input plane into distinct regions that contain either only $t = 1$ or only $t = -1$ patterns.

5.4 Piecewise linear decision boundary. Find an alternative solution for the classification problem shown in Figure 5.14, where the weight vectors are chosen as depicted in Figure 5.18.

5.5 Boolean functions. Any N-dimensional Boolean function can be represented using a perceptron with one hidden layer consisting of 2^N neurons. For $N = 3$ the problem is shown in Figure 5.19. The input bits $x_k^{(\mu)}$ for $k = 1, 2, 3$ are either $+1$ or -1. The output $O^{(\mu)}$ of the network is $+1$ if there is an odd number of positive bits in $\boldsymbol{x}^{(\mu)}$, and -1 if the number of positive bits are even. In one solution, the state $V_j^{(\mu)}$ of neuron $j = 1, \ldots, 2^N$ in the hidden layer is given by:

$$V_j^{(\mu)} = \begin{cases} 1 & \text{if} \quad -\theta_j + \sum_k w_{jk} x_k^{(\mu)} > 0, \\ 0 & \text{if} \quad -\theta_j + \sum_k w_{jk} x_k^{(\mu)} \leq 0, \end{cases} \tag{5.34}$$

where the weights and thresholds are given by $w_{jk} = x_k^{(j)}$ and $\theta_j = 2$. The network

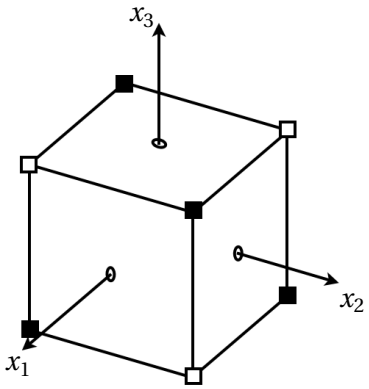


Figure 5.19: Three-dimensional parity problem, with targets $t^{(\mu)} = 1$ (■), $t^{(\mu)} = -1$ (□). Exercise 5.5.

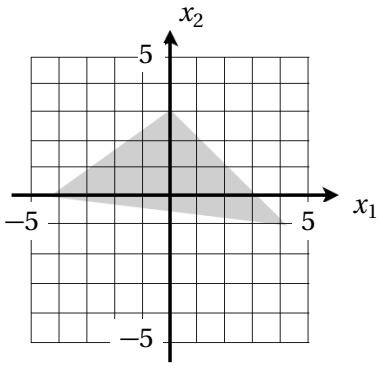


Figure 5.20: Classification problem. Exercise 5.6.

output is computed as $O^{(\mu)} = \sum_j W_j V_j^{(\mu)}$. Determine the weights W_j.

5.6 Linearly inseparable problem. A classification problem is given in Figure 5.20. Inputs $\boldsymbol{x}^{(\mu)}$ inside the gray triangle have targets $t^{(\mu)} = 1$, inputs outside the triangle have targets is $t^{(\mu)} = 0$. The problem can be solved by a perceptron with one hidden layer with three neurons $V_j^{(\mu)} = \theta_{\mathrm{H}}\big(-\theta_j + \sum_{k=1}^{2} w_{jk} x_k^{(\mu)}\big)$, for $j = 1,2,3$. The network output is computed as $O^{(\mu)} = \theta_{\mathrm{H}}(-\Theta + \sum_{j=1}^{3} W_j V_j^{(\mu)})$. Find weights w_{jk}, W_j and thresholds θ_j, Θ that solve the classification problem.

5.7 Perceptron with one hidden layer. A perceptron has one input layer, one layer of hidden neurons, and one output unit. It receives two-dimensional input patterns $\boldsymbol{x}^{(\mu)} = [x_1^{(\mu)}, x_2^{(\mu)}]^{\mathsf{T}}$. They are mapped to four hidden neurons $V_i^{(\mu)}$ as

$$V_j^{(\mu)} = \begin{cases} 0 & \text{if} \quad -\theta_j + \sum_k w_{jk} x_k^{(\mu)} \le 0\,, \\ 1 & \text{if} \quad -\theta_j + \sum_k w_{jk} x_k^{(\mu)} > 0\,. \end{cases} \tag{5.35}$$

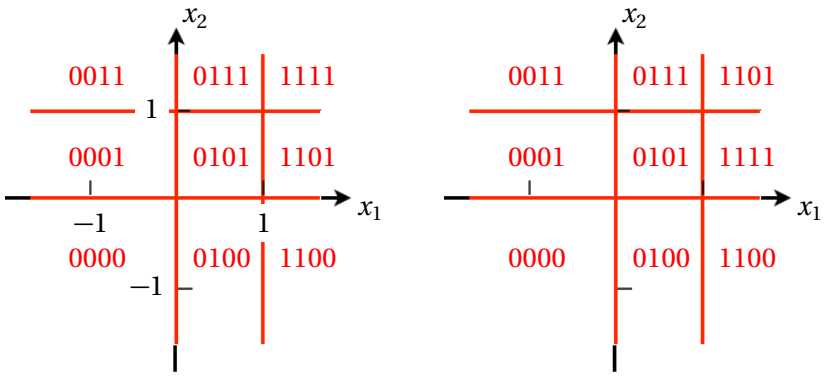


Figure 5.21: Left: input plane with decision boundaries of hidden neurons V_j (red lines). The boundaries partition input space into nine regions labeld by the binary code $V_1 V_2 V_3 V_4$. Right: same, but for a different labeling. Exercise 5.7.

The network output is computed as

$$O^{(\mu)} = \begin{cases} 0 & \text{if} \quad -\Theta + \sum_j W_j V_j^{(\mu)} \leq 0\,, \\ 1 & \text{if} \quad -\Theta + \sum_j W_j V_j^{(\mu)} > 0\,, \end{cases} \tag{5.36}$$

with $W_1 = W_3 = W_4 = 1$, $W_2 = -1$, and $\Theta = \frac{1}{2}$. Figure 5.21(left) shows how input space is mapped to the the hidden neurons. Draw the decision boundary of the network. Give values for w_{ij} and θ_i that yield the pattern in Figure 5.21(left). Show that one cannot map the input space to the space of hidden neurons as in Figure 5.21(right).

5.8 Multilayer perceptron. A classification problem is shown in Figure 5.22. It can be solved by a multilayer perceptron with two inputs, three hidden neurons $V_j^{(\mu)} = \theta_H\left(\sum_{i=1}^{2} w_{jk} x_k^{(\mu)} - \theta_j\right)$, and one output $O^{(\mu)} = \theta_H\left(\sum_{j=1}^{3} W_j V_j^{(\mu)} - \Theta\right)$. A possible solution is illustrated in Fig. 5.22. Compute weights w_{jk} and thresholds θ_j of the hidden neurons that determine the three decision boundaries (red lines). Draw a representation of the problem in the space with axes V_1, V_2, and V_3. Find output weights W_j and threshold Θ that solve the problem.

5.9 Expected maximal number of separable patterns. Show that the sum in Equation (5.29) sums to $2m$.

5.10 Cover's theorem. Prove that $P(3,2) = \frac{3}{4}$ by complete enumeration of all cases. Some cases (not all) are shown in Figure 5.23.

5.11 Non-linear activation function. Consider a single neuron with continuous, non-linear, and monotonically increasing activation function $g(b)$, and with N

x_1	x_2	$t^{(\mu)}$
0.1	0.95	0
0.2	0.85	0
0.2	0.9	0
0.3	0.75	1
0.4	0.65	1
0.4	0.75	1
0.6	0.45	0
0.8	0.25	0
0.1	0.65	1
0.2	0.75	1
0.7	0.2	1

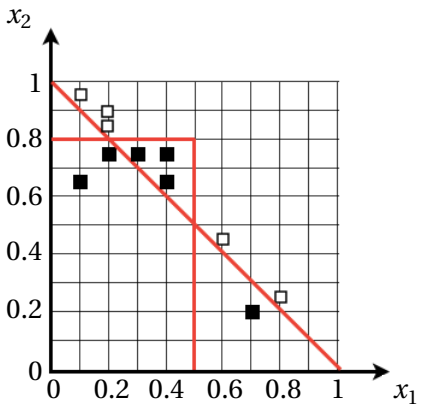


Figure 5.22: Inputs and targets for a classification problem. The targets are either $t = 0$ (□) or $t = 1$ (■). The three decision boundaries (red lines) illustrate a solution to the problem by a multilayer perceptron. Exercise 5.8.

inputs $x_1, \ldots, x_N$. Show that this neuron cannot solve binary classication problems $[\boldsymbol{x}^{(\mu)}, t^{(\mu)}]$ $(\mu = 1, \ldots, p)$ if $p > N$.

5.12 Random classification problem. The probability $P(p, N)$ that a random binary classification problem with p patterns in input dimension N is homogeneously linearly separable is given in Equation (5.28). Show that [1]

$$P(p,N) \sim \tfrac{1}{2}\Big\{1 + \mathrm{erf}\Big[\sqrt{\tfrac{\alpha N}{2}}\Big(\tfrac{2}{\alpha} - 1\Big)\Big]\Big\} \tag{5.37}$$

in the limit of $N \to \infty$ at fixed $\alpha = p/N$.

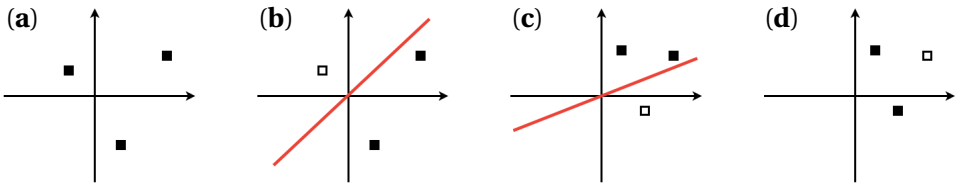


Figure 5.23: Cover's theorem for $p = 3$ and $m = 2$. Examples for problems that are homogeneously linearly separable, (**b**) and (**c**), and for problems that are not homogeneously linearly separable (**a**) and (**d**). Exercise 5.10.

6 Stochastic gradient descent

In Chapter 5 we discussed how a hidden layer helps to classify problems that are not linearly separable. We explained how the decision boundary in Figure 5.14 is represented in terms of the weights and thresholds of the hidden neurons, and introduced a training algorithm based on gradient descent. In this Section, the training algorithm is discussed in more detail.

Figure 6.1 shows the layout of the network to be trained. There are p input patterns $\boldsymbol{x}^{(\mu)}$ with N components each, as before. The output of the network has M components:

$$\boldsymbol{O}^{(\mu)} = \begin{bmatrix} O_1^{(\mu)} \\ O_2^{(\mu)} \\ \vdots \\ O_M^{(\mu)} \end{bmatrix}, \tag{6.1}$$

to be matched to the targets $\boldsymbol{t}^{(\mu)}$. The activation functions must be differentiable (or at least piecewise differentiable), but apart from that there is no need to specify them further at this point. The network shown in Figure 6.1 performs the computation

$$V_j^{(\mu)} = g\left(b_j^{(\mu)}\right) \quad \text{with} \quad b_j^{(\mu)} = \sum_k w_{jk} x_k^{(\mu)} - \theta_j\,, \tag{6.2a}$$

$$O_i^{(\mu)} = g\left(B_i^{(\mu)}\right) \quad \text{with} \quad B_i^{(\mu)} = \sum_j W_{ij} V_j^{(\mu)} - \Theta_i\,. \tag{6.2b}$$

In other words, the outputs are computed in terms of nested activation functions:

$$O_i^{(\mu)} = g\Bigg(\sum_j W_{ij}\, \underbrace{g\Bigg(\sum_k w_{jk} x_k^{(\mu)} - \theta_j\Bigg)}_{V_j^{(\mu)}} - \Theta_i\Bigg). \tag{6.3}$$

This is a consequence of the network layout of the perceptron: all incoming connections to a given neuron are from the layer immediately to the left, all outgoing connections to the layer immediately to the right. The more hidden layers a network has, the deeper is the nesting of the activation functions.

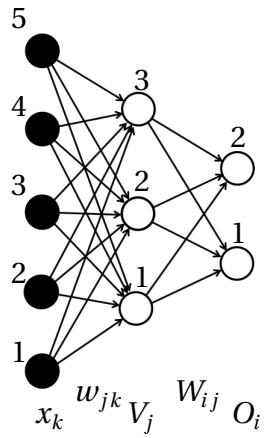


Figure 6.1: Neural network with one hidden layer. Illustrates the notation used in Section 6.1.

6.1 Chain rule and error backpropagation

The network is trained by gradient-descent learning on the energy function (5.23), in the same way as in Section 5.3:

$$H = \frac{1}{2}\sum_{\mu i}\left(t_i^{(\mu)} - O_i^{(\mu)}\right)^2 . \tag{6.4}$$

The weights are updated as follows:

$$\delta W_{mn} = -\eta \frac{\partial H}{\partial W_{mn}} \quad \text{and} \quad \delta w_{mn} = -\eta \frac{\partial H}{\partial w_{mn}} . \tag{6.5}$$

As in Section 5.3, the small parameter $\eta > 0$ is the learning rate. The derivatives of the energy function are evaluated with the *chain rule*. For the weights connecting to the output layer we apply the chain rule once

$$\frac{\partial H}{\partial W_{mn}} = -\sum_{\mu i}\left(t_i^{(\mu)} - O_i^{(\mu)}\right)\frac{\partial O_i^{(\mu)}}{\partial W_{mn}} , \tag{6.6a}$$

and then once more:

$$\frac{\partial O_i^{(\mu)}}{\partial W_{mn}} = g'(B_i^{(\mu)})\delta_{im} V_n^{(\mu)} . \tag{6.6b}$$

Here $g'(B) = \mathrm{d}g/\mathrm{d}B$ is the derivative of the activation function with respect to the local field B. An important point is that the states V_j of the neurons in the hidden layer do not depend on W_{mn}, because these neurons do not have incoming connections with these weights, a consequence of the *feed-forward layout* of the network. In summary we obtain for the increments of the weights connecting to the output layer:

$$\delta W_{mn} = -\eta \frac{\partial H}{\partial W_{mn}} = \eta \sum_{\mu=1}^{p} \left(t_m^{(\mu)} - O_m^{(\mu)}\right) g'\left(B_m^{(\mu)}\right) \equiv \eta \sum_{\mu=1}^{p} \Delta_m^{(\mu)} V_n^{(\mu)}. \tag{6.7a}$$

The quantity

$$\Delta_m^{(\mu)} = (t_m^{(\mu)} - O_m^{(\mu)}) g'(B_m^{(\mu)}) \tag{6.7b}$$

is a weighted output *error*: it vanishes when $O_m^{(\mu)} = t_m^{(\mu)}$. The weights connecting to the hidden layer are updated in a similar fashion, by applying the chain rule four times:

$$\frac{\partial H}{\partial w_{mn}} = -\sum_{\mu i} \left(t_i^{(\mu)} - O_i^{(\mu)}\right) \frac{\partial O_i^{(\mu)}}{\partial w_{mn}}, \tag{6.8a}$$

$$\frac{\partial O_i^{(\mu)}}{\partial w_{mn}} = \sum_l \frac{\partial O_i^{(\mu)}}{\partial V_l^{(\mu)}} \frac{\partial V_l^{(\mu)}}{\partial w_{mn}}, \tag{6.8b}$$

$$\frac{\partial O_i^{(\mu)}}{\partial V_l^{(\mu)}} = g'(B_i^{(\mu)}) W_{il}, \tag{6.8c}$$

$$\frac{\partial V_l^{(\mu)}}{\partial w_{mn}} = g'(b_l^{(\mu)}) \delta_{lm} x_n^{(\mu)}. \tag{6.8d}$$

Here we used Equation (4.27) which takes the form

$$\frac{\partial w_{ij}}{\partial w_{mn}} = \delta_{im} \delta_{jn} \tag{6.9}$$

for asymmetric weights. As before, δ_{im} is the Kronecker delta: δ_{im}=1 if $i = m$ and zero otherwise. Using the definition of the output error, $\Delta_i^{(\mu)}$, one obtains:

$$\delta w_{mn} = \eta \sum_{\mu} \sum_{i} \Delta_i^{(\mu)} W_{im} g'\left(b_m^{(\mu)}\right) x_n^{(\mu)} \equiv \eta \sum_{\mu} \delta_m^{(\mu)} x_n^{(\mu)}, \tag{6.10}$$

with

$$\delta_m^{(\mu)} = \sum_i \Delta_i^{(\mu)} W_{im} g'\left(b_m^{(\mu)}\right). \tag{6.11}$$

The quantities $\delta_m^{(\mu)}$ are errors associated with the hidden layer, they vanish when the output errors $\Delta_i^{(\mu)}$ are zero. Equation (6.11) shows that the errors are determined recursively. The neuron states are also updated recursively, Equation (6.2), but there is an important difference between Equations (6.11) and (6.2). The feed-forward structure of the layered network implies that the neurons are updated from left to right. Equation (6.11), by contrast, says that the errors are updated from the right to the left, from the output layer to the hidden layer. The term *backpropagation* refers to this difference: the neurons are updated forward, the errors are updated backward.

In terms of the errors $\Delta_m^{(\mu)}$ and $\delta_m^{(\mu)}$, the weight increments have the same form for both layers:

$$\delta W_{mn} = \eta \sum_{\mu=1}^{p} \Delta_m^{(\mu)} V_n^{(\mu)} \quad \text{and} \quad \delta w_{mn} = \eta \sum_{\mu=1}^{p} \delta_m^{(\mu)} x_n^{(\mu)} . \tag{6.12}$$

The rule (6.12) is also called δ-rule. It is local in the following sense: the increments of the weights feeding into a certain layer are determined by the errors associated with that layer, and by the states of the neurons in the layer immediately to the left. The thresholds are updated in a similar way:

$$\delta\Theta_m = -\eta \frac{\partial H}{\partial \Theta_m} = \eta \sum_{\mu=1}^{p} \left(t_m^{(\mu)} - O_m^{(\mu)}\right)\left[-g'\left(B_m^{(\mu)}\right)\right] = -\eta \sum_{\mu=1}^{p} \Delta_m^{(\mu)} , \tag{6.13a}$$

$$\delta\theta_m = -\eta \frac{\partial H}{\partial \theta_m} = \eta \sum_{\mu=1}^{p} \sum_i \Delta_i^{(\mu)} W_{im} \left[-g'\left(b_m^{(\mu)}\right)\right] = -\eta \sum_{\mu=1}^{p} \delta_m^{(\mu)} . \tag{6.13b}$$

The general form for the threshold increments is analogous to Equation (6.12)

$$\delta\Theta_m = -\eta \sum_{\mu=1}^{p} \Delta_m^{(\mu)} \quad \text{and} \quad \delta\theta_m = -\eta \sum_{\mu=1}^{p} \delta_m^{(\mu)} , \tag{6.14}$$

but without the state variables of the neurons (or the inputs), as expected. A way to remember the difference between Equations (6.12) and (6.14) is to note that the formula for the threshold increments looks like the one for the weight increments if one sets the values of the neurons to -1.

The backpropagation rules (6.12) and (6.14) contain sums over patterns. This corresponds to feeding all patterns at the same time to compute the increments of weights and thresholds (*batch* training). Alternatively one may choose a single pattern, update the weights by backpropagation, and then continue to iterate these

training steps many times (*sequential* training). One iteration corresponds to feeding a single pattern, p iterations are called one *epoch* (in batch training, one iteration corresponds to one epoch). If one chooses the patterns randomly, then sequential training results in *stochastic gradient descent*:

$$\delta W_{mn} = \eta \Delta_m^{(\mu)} V_n^{(\mu)} \quad \text{and} \quad \delta w_{mn} = \eta \delta_m^{(\mu)} x_n^{(\mu)}, \tag{6.15a}$$

$$\delta \Theta_m = -\eta \Delta_m^{(\mu)} \quad \text{and} \quad \delta \theta_m = -\eta \delta_m^{(\mu)}. \tag{6.15b}$$

Since the sum over pattern is absent, the steps do not necessarily decrease the energy function. Their directions fluctuate, but the average weight increment (averaged over all patterns) points downhill. The result is a *stochastic path* through weight and threshold space, less prone to getting stuck in local minima (Chapter 3).

6.2 Stochastic gradient-descent algorithm

The stochastic-gradient descent formulae derived in the previous Section apply to networks with any number of layers (Algorithm 3). This Section describes the details of the stochastic-gradient algorithm for deep networks with many hidden layers. To this end we need some notation, illustrated in 6.2. We label the layers by the index ℓ. The layer of input terminals has label $\ell = 0$, while the $\ell = L$ denotes the layer of output neurons. The state variables for the neurons in layer ℓ are $V_j^{(\ell)}$, the weights connecting into these neurons from the left are $w_{jk}^{(\ell)}$, the errors associated with layer l are denoted by $\delta_k^{(\ell)}$. In this notation Equations (6.2) read:

$$V_j^{(\ell)} = g\Big(\sum_k w_{jk}^{(\ell)} V_k^{(\ell-1)} - \theta_j^{(\ell)}\Big), \tag{6.16}$$

where $b_j^{(\ell)} = \sum_k w_{jk}^{(\ell)} V_k^{(\ell-1)} - \theta_j^{(\ell)}$ is the local field for $V_j^{(\ell)}$. It involves the matrix-vector product between the weight matrix $\mathbb{W}^{(\ell)}$ and the vector $\boldsymbol{V}^{(\ell-1)}$.

Weights and thresholds are updated using Equation (6.15a):

$$\delta w_{mn}^{(\ell)} = \eta \delta_m^{(\ell)} V_n^{(\ell-1)} \quad \text{and} \quad \delta \theta_m^{(\ell)} = -\eta \delta_m^{(\ell)}, \tag{6.17}$$

with errors

$$\delta_j^{(\ell-1)} = \sum_i \big(t_i - V_i^{(L)}\big) \sum_j \frac{\partial V_i^{(L)}}{\partial V_j^{(\ell-1)}} g'(b_j^{(\ell-1)}). \tag{6.18}$$

Evaluating the gradient $\partial V_i^{(L)} / \partial V_j^{(\ell-1)}$ using the chain rule, one arrives at the recursion (6.11):

$$\delta_j^{(\ell-1)} = \sum_i \delta_i^{(\ell)} w_{ij}^{(\ell)} g'(b_j^{(\ell-1)}) \tag{6.19}$$

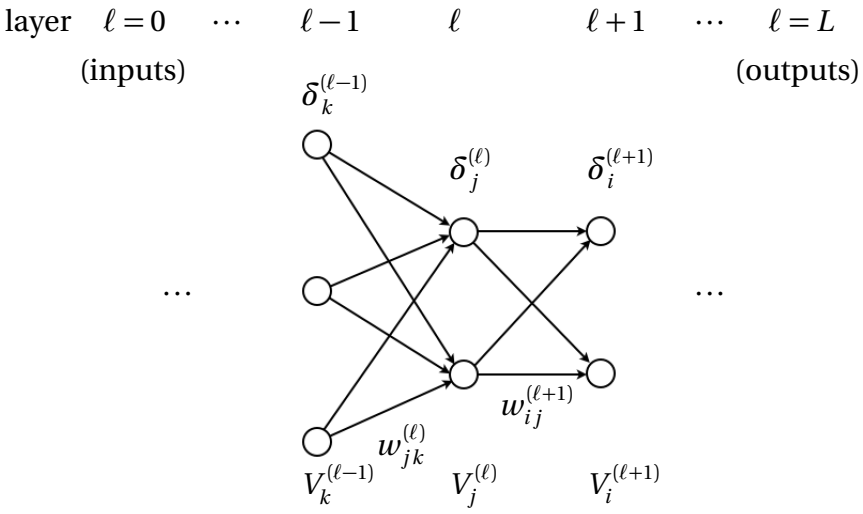


Figure 6.2: Illustrates the notation used in Algorithm 3.

Algorithm 3 stochastic gradient descent

initialise weights $w_{mn}^{(\ell)}$ to random numbers, thresholds to zero, $\theta_m^{(\ell)}=0$;
for $\nu=1,\ldots,\nu_{\max}$ **do**
  choose a value of μ and apply pattern $\boldsymbol{x}^{(\mu)}$ to input layer, $\boldsymbol{V}^{(0)}\leftarrow\boldsymbol{x}^{(\mu)}$;
  for $\ell=1,\ldots,L$ **do**
    propagate forward: $V_j^{(\ell)}\leftarrow g\Big(\sum_k w_{jk}^{(\ell)}V_k^{(\ell-1)}-\theta_j^{(\ell)}\Big)$;
  end for
  compute errors for output layer: $\delta_i^{(L)}\leftarrow g'(b_i^{(L)})(t_i-V_i^{(L)})$;
  for $\ell=L,\ldots,2$ **do**
    propagate backward: $\delta_j^{(\ell-1)}\leftarrow\sum_i\delta_i^{(\ell)}w_{ij}^{(\ell)}g'(b_j^{(\ell-1)})$;
  end for
  for $\ell=1,\ldots,L$ **do**
    update: $w_{mn}^{(\ell)}\leftarrow w_{mn}^{(\ell)}+\eta\delta_m^{(\ell)}V_n^{(\ell-1)}$ and $\theta_m^{(\ell)}\leftarrow\theta_m^{(\ell)}-\eta\delta_m^{(\ell)}$;
  end for
end for

with initial condition $\delta_i^{(L)} = (t_i - V_i^{(L)}) g'(b_i^{(L)})$. The result of this recursion is a vector with components $\delta_j^{(\ell-1)}$, is obtained by component-wise multiplication of $[\mathbb{W}^{(\ell)\mathsf{T}} \boldsymbol{\delta}^{(\ell)}]_j$ with $g'(b_j^{(\ell-1)})$. Component-wise multiplication of vectors is somtimes called Schur or Hadamard product [58], denoted by $\boldsymbol{a} \odot \boldsymbol{b} = [a_1 b_1, \ldots, a_N b_N]^\mathsf{T}$. It does not have a geometric meaning like the scalar or the cross product of vectors, and therefore there is little point in using it. It is more important to note that the vector $\boldsymbol{\delta}^{(\ell)}$ is multiplied by the transpose of the weight matrix, $\mathbb{W}^{(\ell)\mathsf{T}}$, rather than by the weight matrix itself.

The stochastic-gradient algorithm is summarised in Algorithm 3. One feeds an input $\boldsymbol{x}^{(\nu)}$, updates the weights using (6.17), and iterates these steps until the energy function (5.23) is deemed sufficiently small (this corresponds to the sum over ν in Algorithm 3). The resulting weights and thresholds are not unique. In Figure 5.16 all weights for the Boolean XOR function are equal to ± 1. But the training algorithm (6.7a), (6.10), and (6.13) corresponds to repeatedly adding weight increments. This may cause the weights to grow.

In practice, the stochastic gradient-descent dynamics may be too noisy. It is often better to average over a small number of randomly chosen patterns. Such a set is called *mini batch*, of size m_B say. In *stochastic gradient descent with mini batches* one replaces Equations (6.12) and (6.14) by

$$\delta W_{mn} = \eta \sum_{\mu=1}^{m_B} \Delta_m^{(\mu)} V_n^{(\mu)} \quad \text{and} \quad \delta\Theta_m = -\eta \sum_{\mu=1}^{m_B} \Delta_m^{(\mu)}, \tag{6.20}$$
$$\delta w_{mn} = \eta \sum_{\mu=1}^{m_B} \delta_m^{(\mu)} x_n^{(\mu)} \quad \text{and} \quad \delta\theta_m = -\eta \sum_{\mu=1}^{m_B} \delta_m^{(\mu)}.$$

Sometimes the mini-batch rule is quoted with prefactors of m_B^{-1} before the sums. This does not make any fundamental difference, the factors m_B^{-1} can just be absorbed in the learning rate. But when you compare learning rates for different implementations, it is important to check whether or not there are factors of m_B^{-1} in front of the sums in Equation (6.20). How does one assign inputs to mini batches? This is discussed in Section 6.3: at the beginning of each epoch, one should randomly *shuffle* the sequence of the input patterns in the training set. Then the first mini batch contains patterns $\mu = 1, \ldots, m_B$, and so forth.

Common choices for the activation functions $g(b)$ are the *sigmoid* function or tanh:

$$g(b) = \frac{1}{1+e^{-b}} \equiv \sigma(b), \tag{6.21a}$$
$$g(b) = \tanh(b). \tag{6.21b}$$

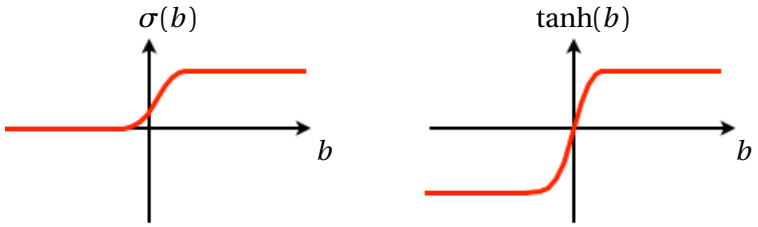


Figure 6.3: Saturation of the activation functions (6.21): the derivative $g'(b)$ tends to zero for large values of $|b|$.

In both cases, the derivatives $g'(b)$ can be expressed in terms of the function itself:

$$\frac{\mathrm{d}}{\mathrm{d}b}\sigma(b)=\sigma(b)[1-\sigma(b)], \quad \frac{\mathrm{d}}{\mathrm{d}b}\tanh(b)=\left[1-\tanh^2(b)\right]. \tag{6.22}$$

As illustrated in Figure 6.3, the activation functions (6.21) saturate at large values of $|b|$, so that the derivative $g'(b)$ tends to zero. Since the backpropagation rule (6.18) contains factors of $g'(b)$, this implies that the algorithm slows down. For the same reason, the initial weights and thresholds should be chosen so that the local fields b are not too large in modulus, to avoid that $g'(b)$ becomes too small. A standard procedure is to take all weights to be initially randomly distributed, for example Gaussian with zero mean, and with a suitable variance. The performance of networks with many hidden layers (*deep* networks) can be sensitive to the initialisation of the weights (Section 7.6). The initial values of the thresholds are not so critical, they are often learned more rapidly than the weights, at least initially. The thresholds are initialised to zero.

6.3 Preprocessing the input data

It can be useful to preprocess the input data, although any preprocessing may remove information from the data. Nevertheless, it is usually advisable to shift the data so the mean of each component over all p patterns vanishes:

$$\langle x_k\rangle=\frac{1}{p}\sum_{\mu=1}^{p}x_k^{(\mu)}=0. \tag{6.23}$$

There are several reasons for this. First, large mean values can cause steep cliffs in the energy function that are difficult to navigate with gradient descent. Different input-data variances in different directions have a similar effect. Therefore one *scales* the inputs so that the input-data distribution has the same variance in all directions (Figure 6.4), equal to unity for instance:

$$\sigma_k^2=\frac{1}{p}\sum_{\mu=1}^{p}\left(x_k^{(\mu)}-\langle x_k\rangle\right)^2=1. \tag{6.24}$$

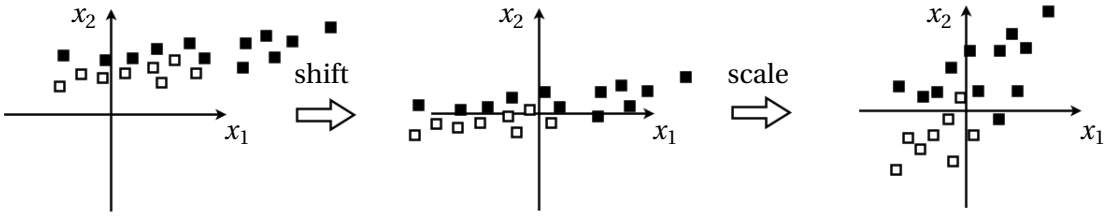


Figure 6.4: Shift and scale the input data to achieve zero mean and unit variance.

Second, to avoid saturation of the neurons connected to the inputs, their local fields must not be too large (Section 6.2). If one initialises the weights to Gaussian random numbers with mean zero and unit variance, large activations are quite likely if the distribution of input patterns has a large mean or a large variance. Third, enforcing zero input mean by shifting the input data avoids that the weights of the neurons in the first hidden layer must decrease or increase together [59]. Equation (6.20) shows that the components of $\delta \boldsymbol{w}_m \propto \delta_m \boldsymbol{x}$ into hidden neuron m are likely to have the same signs if the input data has a large mean. This means that the weight increments have the same signs. This makes it difficult for the network to learn to differentiate. In summary, one shifts and scales the input-data distribution so that it has mean zero and unit variance, as illustrated in Figure 6.4. The same transformation (using the mean values and scaling factors determined for the training set) should be applied to any new data set that the network is supposed to classify after it has been trained on the training set.

Figure 6.5 shows a distribution of inputs that falls into two distinct clusters. The difference between the clusters is sometimes called *covariate shift*, here *covariate* is just another term for input. Imagine feeding first just inputs from one of the clusters to the network. It will learn local properties of the decision boundary, instead of its global features. Such global properties are efficiently learned if the network is more frequently confronted with unfamiliar data. For sequential training (stochastic gradient descent) this is not a problem, because the sequence of input patterns

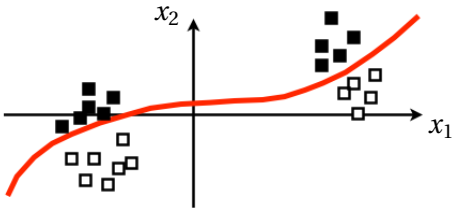


Figure 6.5: When the input data falls into clusters as shown in this Figure, one should randomly pick data from either cluster, to avoid that patterns become too familiar. The decision boundary is shown in red.

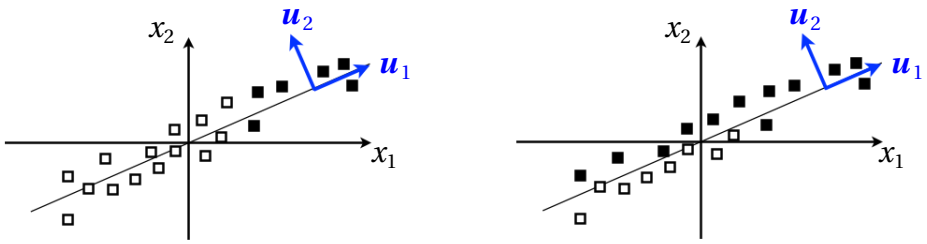


Figure 6.6: Principal-component analysis (schematic). The data set on the left can be classified keeping only the principal component $\boldsymbol{u}_1$ of the data. This is not true for the data set on the right.

presented to the network is random. However, if one trains with mini batches, the mini batches should contain randomly chosen patterns in order to avoid covariate shifts. To this end one randomly *shuffles* the sequence of the input patterns in the training set, at the beginning of each epoch.

It is also recommended [59] to observe the output errors during training. If the errors are similar for a number of subsequent learning steps, the corresponding inputs appear familiar to the network. Larger errors correspond to unfamiliar inputs, and Ref. [59] suggests to feed such inputs more often.

6.3.1 Dimensionality reduction

Often the input data is very high dimensional, requiring many input terminals. This usually means that one should use many neurons in the hidden layers. This can be problematic because it increases the risk of overfitting the input data. To avoid this as far as possible, one can reduce the dimensionality of the input data by *principal-component analysis*. This method allows to project high-dimensional data to a lower dimensional subspace (Figure 6.6).

The data in Figure 6.6(left) falls approximately onto a straight line, the principal direction $\boldsymbol{u}_1$. The coordinate orthogonal to the principal direction is not useful in classifying the data. Consequently this coordinate can be disregarded, reducing the dimensionality of the data set. But note that the data set shown on the right of Figure 6.6 is much harder to classify if we use only the principal component alone. This illustrates a potential problem: we may lose important information by projecting the data on its principal component.

The idea of principal component analysis is to rotate the basis in input space so that the variance of the data along the first axis of the new coordinate system is maximised. The data variance along a direction $\boldsymbol{v}$ reads

$$\sigma_v^2 = \langle(\boldsymbol{x}\cdot\boldsymbol{v})^2\rangle - \langle\boldsymbol{x}\cdot\boldsymbol{v}\rangle^2 = \boldsymbol{v}\cdot\mathbb{C}\boldsymbol{v}\,. \tag{6.25}$$

Here

$$\mathbb{C} = \langle \delta \boldsymbol{x}\, \delta \boldsymbol{x}^{\mathsf{T}} \rangle \quad \text{with} \quad \delta \boldsymbol{x} = \boldsymbol{x} - \langle \boldsymbol{x} \rangle \tag{6.26}$$

is the data *covariance matrix*. The variance σ_v^2 is maximal when $\boldsymbol{v}$ points in the direction of the leading eigenvector of the covariance matrix $\mathbb{C}$. This can be seen as follows. The covariance matrix is symmetric, therefore its eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ form an orthonormal basis of input space. This allows us to express the matrix $\mathbb{C}$ as

$$\mathbb{C} = \sum_{\alpha=1}^{N} \lambda_\alpha \boldsymbol{u}_\alpha \boldsymbol{u}_\alpha^{\mathsf{T}}. \tag{6.27}$$

The eigenvalues λ_α are non-negative. This follows from Equation (6.26) and the eigenvalue equation $\mathbb{C}\boldsymbol{u}_\alpha = \lambda_\alpha \boldsymbol{u}_\alpha$. We arrange the eigenvalues by magnitude, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$. Using Equation (6.27) we can write for the variance

$$\sigma_v^2 = \sum_{\alpha=1}^{N} \lambda_\alpha v_\alpha^2 \tag{6.28}$$

with $v_\alpha = \boldsymbol{v} \cdot \boldsymbol{u}_\alpha$. We want to show that σ_v^2 is maximal for $\boldsymbol{v} = \pm \boldsymbol{u}_1$ subject to the constraint that $\sum_\alpha v_\alpha^2 = 1$. To ensure that the constraint is satisfied one introduces a Lagrange multiplier λ as in Chapter 4. The function to maximise reads

$$\mathscr{L} = \sum_\alpha \lambda_\alpha v_\alpha^2 - \lambda\big(1 - \sum_\alpha v_\alpha^2\big). \tag{6.29}$$

Variation of $\mathscr{L}$ yields that $v_\beta(\lambda_\beta + \lambda) = 0$. This means that all components v_β must vanish, except one which must equal unity. The maximum of $\mathscr{L}$ is obtained by $\lambda = -\lambda_1$, where λ_1 is the maximal eigenvalue of $\mathbb{C}$ with eigenvector $\boldsymbol{u}_1$. This shows that the variance σ_v^2 is maximised by the principal direction.

In more than two dimensions there is commonly more than one direction along which the data varies significantly. These k principal directions correspond to the k eigenvectors of $\mathbb{C}$ with the largest eigenvalues. This can be shown recursively. One projects the data to the subspace orthogonal to $\boldsymbol{u}_1$ by applying the projection matrix $\mathbb{P}_1 = \mathbb{1} - \boldsymbol{u}_1 \boldsymbol{u}_1^{\mathsf{T}}$. Then one repeats the procedure outlined above, and finds that the data varies maximally along $\boldsymbol{u}_2$. Iterating one obtains the k principal directions $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$. Often there is a gap between the k largest eigenvalues and the small ones (all close to zero). Then one can safely project the data onto the subspace spanned by the k principal directions. If there is no gap then it is less clear what to do.

6.4 Cross validation

The goal of supervised learning is to generalise from a training set to new data. Only general properties of the training set can be generalised, not specific ones that are

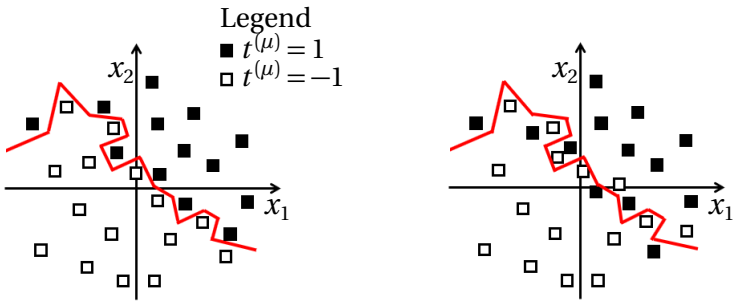


Figure 6.7: Overfitting. Left: accurate representation of the decision boundary in the training set, for a network with 15 neurons in the hidden layer. Right: this new data set differs from the first one just by a little bit of noise. The points in the vicinity of the decision boundary are not correctly classified.

particular to the training set and that could be very different in new data. A network with more neurons may classify the training data better, because it accurately represents all specific features of the data. But those specific properties could look quite different in new data (Figure 6.7). As a consequence, we must look for a compromise: between accurate classification of the training set and the ability of the network to generalise. The problem illustrated in Figure 6.7 is also referred to as *overfitting*: the network fits too fine details (for instance noise in the training set) that have no general meaning. The tendency to overfit is larger for networks with more neurons.

One way of avoiding overfitting is to use *cross validation* and *early stopping*. One splits the training data into two sets: a *training set* and a *validation set*. The idea is that these sets share the general features to be learnt. But although training and validation data are drawn from the same distribution, they differ in details that are not of interest. The network is trained on the training set. During training one monitors not only the energy function for the training set, but also the energy function evaluated on the validation data. As long as the network learns general features of the input distribution, both training and validation energies decrease. But when the network starts to learn specific features of the training set, then the validation energy saturates, or may start to increase. At this point the training should be stopped. This scheme is illustrated in Figure 6.8.

Often the possible values of the output neurons are continuous while the targets assume only discrete values. Then it is important to also monitor the *classification error* of the validation set. The definition of the classification error depends on the type of the classification problem. For one single output unit with targets $t = 0/1$ the classification error is defined as

$$C = \frac{1}{p} \sum_{\mu=1}^{p} \left| t^{(\mu)} - \theta_{\mathrm{H}}(O^{(\mu)} - \tfrac{1}{2}) \right| . \tag{6.30a}$$

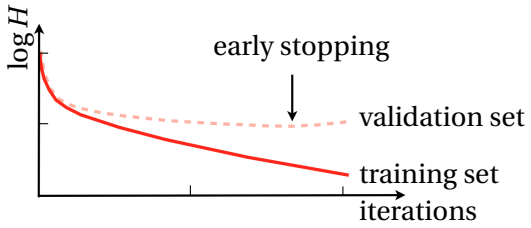


Figure 6.8: Progress of training and validation errors. The plot is schematic, and the data is smoothed. Based on simulations performed by Oleksandr Balabanov. Shown is the natural logarithm of the energy functions for the training set (solid line) and the validation set (dashed line) as a function of the number of training iterations. The training is stopped when the validation energy begins to increase.

If, by contrast, the targets take the values $t = \pm 1$, then the classification error reads:

$$C = \frac{1}{2p} \sum_{\mu=1}^{p} \left| t^{(\mu)} - \mathrm{sgn}(O^{(\mu)}) \right| . \tag{6.30b}$$

Now consider a classification problem where inputs must be classified into M mutually exclusive classes. An example is the MNIST data set of hand-written digits (Section 8.3) where $M = 10$. Another example is given in Table 6.1, with $M = 3$. In both examples one of the targets $t_i^{(\mu)} = 1$ while the others equal zero, for a given input $\boldsymbol{x}^{(\mu)}$. As a consequence, $\sum_i^M t_i^{(\mu)} = 1$. Assume that the network has sigmoid outputs, $O_i^{(\mu)} = \sigma(b_i^{(\mu)})$. To classify input $\boldsymbol{x}^{(\mu)}$ from the network outputs $O_i^{(\mu)}$ we compute for the given value of μ:

$$y_i^{(\mu)} = \begin{cases} 1 & \text{if } O_i^{(\mu)} \text{ is the largest of all outputs } i = 1, \ldots, M, \\ 0 & \text{otherwise.} \end{cases} \tag{6.31a}$$

In this case the classification error can be computed from

$$C = \frac{1}{2p} \sum_{\mu=1}^{p} \sum_{i=1}^{M} \left| t_i^{(\mu)} - y_i^{(\mu)} \right| . \tag{6.31b}$$

In all cases, the *classification accuracy* is defined as $(1 - C)\,100\%$, it is usually quoted in percent.

While the classification error is designed to show the fraction of inputs that are classified wrongly, it contains less information than the energy function (which is in fact a mean-squared error of the outputs). This is illustrated by the two problems in Table 6.1. Both problems have the same classification error, but the energy function is much lower for the second problem, reflecting the better quality of its solution. Yet another measure of classification success is the *cross-entropy error*. It is discussed in Chapter 7.

output			targets				correct?
0.4	0.4	0.55	0	0	1	setosa	yes
0.4	0.55	0.4	0	1	0	versicolor	yes
0.1	0.2	0.8	1	0	0	virginica	no

output			targets				correct?
0.1	0.2	0.8	0	0	1	setosa	yes
0.1	0.8	0.2	0	1	0	versicolor	yes
0.4	0.4	0.55	1	0	0	virginica	no

Table 6.1: Illustrates the difference between energy function and classification error. Each table shows network outputs for three different inputs from the iris data set, as well as the correct classfications.

6.5 Adaptation of the learning rate

It is tempting to choose larger learning rates, because they enable the network to escape more efficiently from shallow minima. But clearly this causes problems when the energy function varies rapidly. As a result the training may fail because the training dynamics starts to oscillate. This can be avoided by changing the learning rule

$$\delta w_{mn}^{(t)} = -\eta \left.\frac{\partial H}{\partial w_{mn}}\right|_{\{w_{ij}\}=\{w_{ij}^{(t)}\}} + \alpha \delta w_{mn}^{(t-1)}. \tag{6.32}$$

Here $t = 1,2,\ldots,T$ labels the iteration number, and You see that the increment at step t depends not only on the instantaneous gradient, but also on the weight change $\delta w_{mn}^{(t-1)}$ of the previous iteration. We say that the dynamics becomes *inertial*, the weights gain *momentum*. The parameter $\alpha \geq 0$ is called momentum constant. It determines how strong the inertial effect is. Obviously $\alpha = 0$ corresponds to the usual backpropagation rule. When α is positive, then how does inertia change the learning rule? Iterating Equation (6.32) yields

$$\delta w_{mn}^{(T)} = -\eta \sum_{t=0}^{T} \alpha^{T-t} \frac{\partial H}{\partial w_{mn}^{(t)}}. \tag{6.33}$$

Here and in the following we use the short-hand notation

$$\frac{\partial H}{\partial w_{mn}^{(t)}} \equiv \left.\frac{\partial H}{\partial w_{mn}}\right|_{\{w_{ij}\}=\{w_{ij}^{(t)}\}}.$$

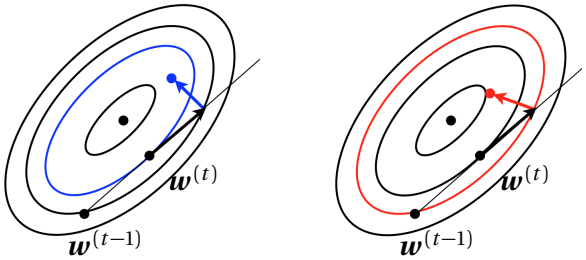


Figure 6.9: Left: Momentum method (6.32). The blue arrow represents the increment $-\eta(\partial H/\partial w_{mn})|_{\{w_{ij}^{(t)}\}}$. Right: Nesterov's accelerated gradient method (6.36). The red arrow represents $-\eta(\partial H/\partial w_{mn})|_{\{w_{ij}^{(t)}+\alpha_{t-1}\delta w_{ij}^{(t-1)}\}}$. The location of $\boldsymbol{w}^{(t+1)}$ (red point) is closer to the minimum (black point) than in the Figure on the left.

Equation (6.33) shows that $\delta w_{mn}^{(T)}$ is a weighted average of the gradients encountered during training. Now assume that the training is stuck in a shallow minimum. Then the gradient $\partial H/\partial w_{mn}^{(t)}$ remains roughly constant through many time steps. To illustrate what happens, let us assume that $\partial H/\partial w_{mn}^{(t)} = \partial H/\partial w_{mn}^{(0)}$ for $t = 1,\ldots,T$. In this case we can write

$$\delta w_{mn}^{(T)} \approx -\eta \frac{\partial H}{\partial w_{mn}^{(0)}} \sum_{t=0}^{T} \alpha^{T-t} = -\eta \frac{\alpha^{T+1}-1}{\alpha-1} \frac{\partial H}{\partial w_{mn}^{(0)}}. \tag{6.34}$$

In this situation, convergence is accelerated when α is close to unity. We also see that it is necessary that $\alpha < 1$ for the sum in Equation (6.34) to converge. The other limit to consider is that the gradient changes rapidly from iteration to iteration. How is the learning rule modified in this case? As an example, let us assume that the gradient remains of the same magnitude, but that its sign oscillates, $\partial H/\partial w_{mn}^{(t)} = (-1)^t \partial H/\partial w_{mn}^{(0)}$ for $t = 1,\ldots,T$. Inserting this into Equation (6.33), we obtain:

$$\delta w_{mn}^{(T)} \approx -\eta \frac{\partial H}{\partial w_{mn}^{(0)}} \sum_{t=0}^{T} (-1)^t \alpha^{T-t} = -\eta \frac{\alpha^{T+1}+(-1)^T}{\alpha+1} \frac{\partial H}{\partial w_{mn}^{(0)}}, \tag{6.35}$$

so that the increments are much smaller compared with those in Equation (6.34). This shows that introducing inertia can substantially accelerate convergence without sacrificing accuracy. The disadvantage is, of course, that there is yet another parameter to choose, namely the momentum constant α.

Nesterov's *accelerated gradient* method [60] is another way of implementing momentum. The algorithm was developed for smooth optimisation problems, but it is often used in stochastic gradient descent when training deep neural networks.

The algorithm can be summarised as follows [61]:

$$\delta w_{mn}^{(t)} = -\eta \frac{\partial H}{\partial w_{mn}}\bigg|_{\{w_{ij}^{(t)}+\alpha_{t-1}\delta w_{ij}^{(t-1)}\}} + \alpha_{t-1}\delta w_{mn}^{(t-1)}. \tag{6.36}$$

A suitable sequence of coefficients α_t is defined by recursion [61]. The coefficients α_t approach unity from below as t increases.

Nesterov's accelerated-gradient method is more efficient than the simple momentum method, because the accelerated-gradient method evaluates the gradient at an extrapolated point, not at the initial point. Figure 6.9 illustrates a situation where Nesterov's method converges more rapidly. Since Nesterov's method often works better than the simple momentum scheme and because it is not much more difficult to implement, it is used quite frequently. There are other ways of adapting the learning rate during training, see Section 4.10 in Haykin's book [2]. Finally, the learning rate need not be the same for all neurons. If the weights of neurons in different layers change at very different speeds (Section 7.2), it may be advantageous to define a layer-dependent learning rate η_ℓ that is larger for neurons with smaller gradients.

6.6 Summary

Backpropagation is an efficient algorithm for stochastic gradient-descent on the energy function in weight space, because it refers only to quantities that are local to the weight to be updated. Networks with many hidden neurons have many free parameters (their weights and thresholds). This increases the risk of overfitting, which reduces the power of the network to generalise. The tendency of networks to overfit can be reduced by cross-validation (Section 6.4). Deep networks with many hidden layers are particularly prone to overfitting (Chapter 7).

6.7 Further reading

The backpropagation algorithm is explained in Section 6.1. of Hertz, Krogh and Palmer [1], and in Chapter 4 of Haykin's book [2]. The paper [59] by LeCun *et al.* predates deep learning, but it is still a very nice collection of recipes for making backpropagation more efficient. Historical note: one of the first papers on error backpropagation is the one by Rumelhart *et al.* [62]. Have a look! The paper gives an excellent explanation and summary of the backpropagation algorithm. The authors also describe results of different numerical experiments, one of them introduces convolutional nets (Chapter 8) to learn to tell the difference between the letters T and C (Figure 6.10).

Figure 6.10: Patterns detected by the convolutional net of Ref. [62]. After Fig. 13 in Ref. [62].

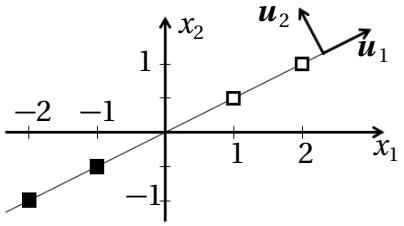


Figure 6.11: The principal direction of this data set is $\boldsymbol{u}_1$.

6.8 Exercises

6.1 Covariance matrix. Show that the eigenvalues of the data covariance matrix $\mathbb{C}$ defined in Equation (6.26) are real and non-negative.

6.2 Principal-component analysis. Compute the data covariance matrix $\mathbb{C}$ for the example shown in Figure 6.11 and determine the principal direction. Determine the principal direction for the data shown in Figure 6.12.

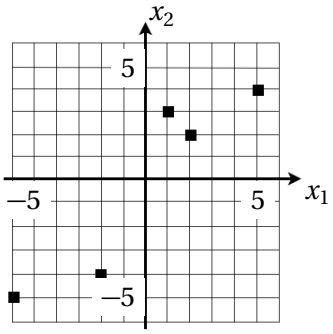


Figure 6.12: Calculate the principal component of this data set. Exercise 10.4.

6.3 Nesterov's accelerated-gradient method. The version (6.36) of Nesterov's algorithm is slightly different from the original formulation [60]. This point is discussed in [61]. Show that both versions are equivalent.

6.4 Skipping layers. Show how the backpropagation algorithm can be generalised for feed-forward networks that allow for connections from the two nearest layers to the left, not only from the nearest layer to the left (Section 7.4).

6.5 Momentum. Section 6.5 describes how to speed up gradient descent by introducing momentum. To explain how this works it was assumed that the gradient is constant throughout, for all time steps $t = 0, \ldots, T$. A more realistic assumption is that the gradient approaches a constant from a certain time step $t_{\min} > 0$ onwards, not from $t = 0$. Rephrase the arguments in Section 6.5 assuming that the gradient approaches a constant and then remains constant for $t \geq t_{\min}$.

6.6 Backpropagation. Explain how to train a multi-layer perceptron by backpropagation. Draw a flow-chart of the algorithm. In your discussion, refer to and explain the following terms: *forward propagation, backward propagation, hidden layer, energy function, gradient descent, local energy minima, batch training, training set, validation set, classification error,* and *overfitting.* Your answer must not be longer than one A4 page.

6.7 Stochastic gradient descent. To train a multi-layer perceptron using stochastic gradient descent one needs update formulae for the weights and thresholds in the network. Derive these update formulae for *sequential training* using backpropagation for the network shown in Fig. 6.13. The weights for the first and second hidden layer, and for the output layer are denoted by $w_{jk}^{(1)}$, $w_{mj}^{(2)}$, and W_{1m}. The corresponding thresholds are denoted by $\theta_j^{(1)}$, $\theta_m^{(2)}$, and Θ_1, and the activation function by $g(\cdots)$. The target value for input pattern $\boldsymbol{x}^{(\mu)}$ is $t_1^{(\mu)}$, and the pattern index μ ranges from 1 to p. The energy function is $H = \frac{1}{2}\sum_{\mu=1}^{p}(t_1^{(\mu)} - O_1^{(\mu)})^2$.

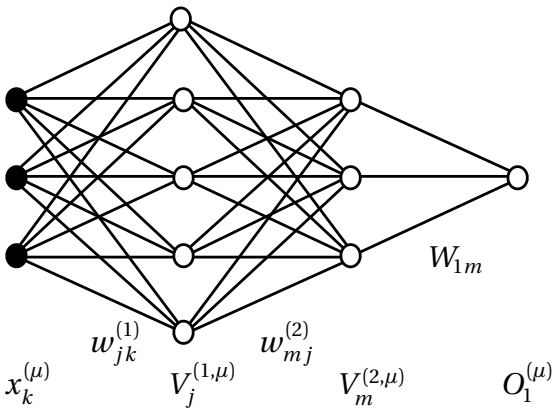


Figure 6.13: Multi-layer perceptron with three input terminals, two hidden layers, and one output unit. Exercise 6.7.

6.8 Multi-layer perceptron. A perceptron has hidden layers $\ell = 1, \ldots, L-1$ and

output layer $l = L$. Neuron $V_j^{(\ell,\mu)}$ in layer ℓ computes $V_j^{(\ell,\mu)} = g(b_j^{(\ell,\mu)})$ with $b_j^{(\ell,\mu)} = -\theta_j^{(\ell)} + \sum_k w_{jk}^{(\ell)} V_k^{(\ell-1,\mu)}$, where $\boldsymbol{V}^{(0,\mu)} = \boldsymbol{x}^{(\mu)}$, $w_{jk}^{(\ell)}$ are weights, $\theta_j^{(\ell)}$ are thresholds, $g(b)$ is the activation function, and $V_i^{(L)} = O_i^{(\mu)} = g(b_i^{(L,\mu)})$. Draw this network. Indicate where the elements $x_k^{(\mu)}$, $b_j^{(\ell,\mu)}$, $V_j^{(\ell,\mu)}$, $O_i^{(\mu)}$, $w_{jk}^{(\ell)}$ and $\theta_j^{(\ell)}$ for $\ell = 1,\ldots,L$ belong. Determine how the derivatives $\partial V_i^{(\ell,\mu)}/\partial w_{mn}^{(p)}$ depend upon the derivatives $\partial V_j^{(\ell-1,\mu)}/\partial w_{mn}^{(p)}$ for $p < \ell$. Evaluate the derivative $\partial V_j^{(\ell,\mu)}/\partial w_{mn}^{(p)}$ for $p = \ell$. Using gradient descent on the energy function $H = \frac{1}{2}\sum_{i\mu}\left(t_i^{(\mu)} - O_i^{(\mu)}\right)^2$, find the update rule for the weight $w_{mn}^{(L-2)}$ with learning rate η.

Figure 7.1: Images of iris flowers. From left to right: iris setosa (copyright T. Monto), iris versicolor (copyright R. A. Nonemacher), and iris virginica (copyright A. Westermoreland). All images are copyrighted under the creative commons license.

7 Deep learning

7.1 How many hidden layers?

In Chapter 5 we saw why it is sometimes necessary to have a hidden layer: this makes it possible to solve problems that are not linearly separable. Under which circumstances is one hidden layer sufficient? Are there problems that require more than one hidden layer? Even if not necessary, may additional hidden layers improve the performance of the network? The second question is more difficult to answer than the first, so we start with the first question. To understand how many hidden layers are necessary it is useful to view the classification problem as an *approximation problem* [63]. Consider the classification problem $[\boldsymbol{x}^{(\mu)}, t^{(\mu)}]$ for $\mu = 1, \ldots, p$. This problem defines a *target function* $t(\boldsymbol{x})$. Training a network to solve this task corresponds to approximating the target function $t(\boldsymbol{x})$ by the output function $O(\boldsymbol{x})$ of the network, from N-dimensional input space to one-dimensional output space.

How many hidden layers are necessary or sufficient to approximate a given set of functions to a certain accuracy, by choosing weights and thresholds? The answer

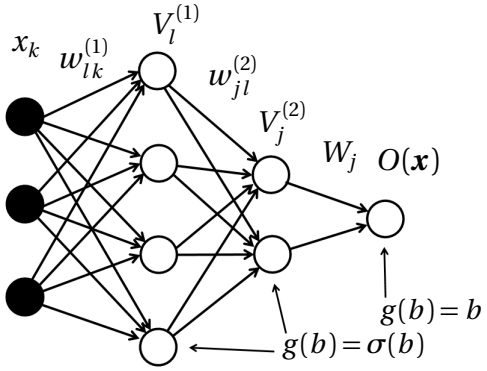


Figure 7.2: Multi-layer perceptron for function approximation.

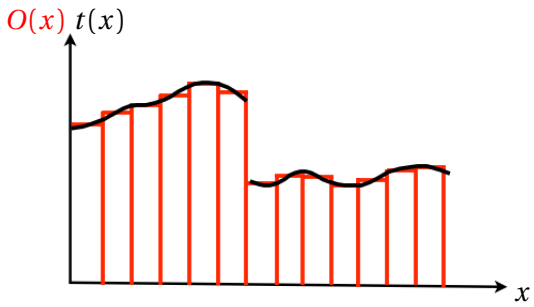


Figure 7.3: The neural-net output $O(x)$ approximates the target function $t(x)$.

depends on the nature of the set of functions. Are they real-valued or do they assume only discrete values? If the functions are real-valued, are they continuous or not?

We start with real-valued inputs and output [1, 64]. Consider the network drawn in Figure 7.2. The neurons in the hidden layers have sigmoid activation functions $\sigma(b)=(1+e^{-b})^{-1}$. The output is continuous, with activation function $g(b)=b$. With two hidden layers the task is to approximate the function $t(\boldsymbol{x})$ by

$$O(\boldsymbol{x})=\sum_j W_j g\Big(\sum_l w^{(2)}_{jl} g\Big(\sum_k w^{(1)}_{lk} x_k-\theta^{(1)}_l\Big)-\theta^{(2)}_j\Big)-\Theta\,. \tag{7.1}$$

In the simplest case the inputs are one-dimensional (Figure 7.3). The training set consists of pairs $[x^{(\mu)}, t^{(\mu)}]$. The task is then to approximate the corresponding target function $t(x)$ by the network output $O(x)$:

$$O(x)\approx t(x)\,. \tag{7.2}$$

We approximate the real-valued function $t(x)$ by linear combinations of the basis functions $\mathscr{B}(x)$ shown in Figure 7.4. Any reasonable real-valued function $t(x)$ can be approximated by a sums of such basis functions, each suitably shifted and scaled. Furthermore, these basis functions can be expressed as scaled differences of activation functions

$$\mathscr{B}(x)=W[\sigma(w^{(1)}x-\theta^{(1)}_1)-\sigma(w^{(1)}x-\theta^{(1)}_2)]\,. \tag{7.3}$$

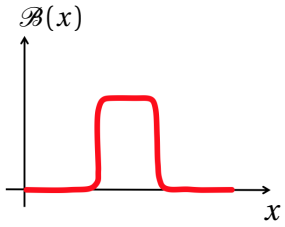


Figure 7.4: Basis function used to approximate a one-dimensional target function $t(x)$.

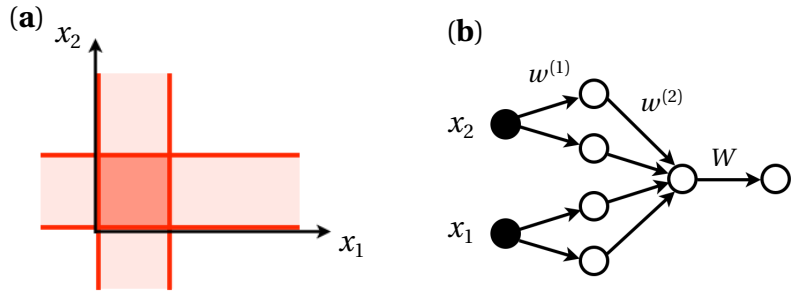


Figure 7.5: Two-dimensional basis functions. (**a**) To make a localised basis function with two inputs, one needs two hidden layers of neurons with sigmoid activation functions. One layer determines the lightly shaded cross in terms of a linear combination of four sigmoid outputs. The second layer localises the final output to the darker square [Equation (7.4)]. (**b**) Network layout for one basis function, after Fig. 8 in Ref. [64].

Comparing with Equation (7.1) shows that one hidden layer is sufficient to construct the function $O(x)$ in this way. Now consider two-dimensional inputs. In this case, a suitable basis function is (Figure 7.5):

$$\mathscr{B}(\boldsymbol{x}) = W\sigma\left\{w^{(2)}[\sigma(w^{(1)}x_1) - \sigma(w^{(1)}x_1 - \theta_2^{(1)}) + \sigma(w^{(1)}x_2) - \sigma(w^{(1)}x_2 - \theta_4^{(1)})] - \theta^{(2)}\right\}. \quad (7.4)$$

So for two input dimensions, two hidden layers are sufficient. For each basis function we require four neurons in the first hidden layer and one neuron in the second hidden layer. The construction is the same for more than two inputs, with $2N$ neurons in the first and one neuron in second layer, for each basis function. In conclusion, two hidden layers are sufficient to approximate a real-valued input-output function.

Yet it is not always necessary to use two layers for real-valued functions. For continuous functions, one hidden layer is sufficient. This is ensured by the *universal approximation theorem* [2]. This theorem says any continuous function can be approximated to arbitrary accuracy by a network with a single hidden layer, for sufficiently many neurons in the hidden layer.

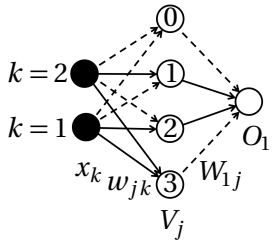
In Chapter 5 we considered discrete Boolean functions. It turns out that any Boolean function with N-dimensional inputs can be represented by a network with one hidden layer, using 2^N neurons in the hidden layer. An example for such a network is discussed in Ref. [1]:

$$\begin{array}{ll} x_k \in \{+1,-1\} & k = 1,\ldots,N \quad \text{inputs} \\ V_j & j = 0,\ldots,2^N-1 \quad \text{hidden neurons} \\ g(b) = \tanh(b) & \text{activation function of hidden neurons} \\ g(b) = \operatorname{sgn}(b) & \text{activation function of output unit} \end{array} \quad (7.5)$$

(**a**)

x_1	x_2	t
-1	-1	-1
+1	-1	+1
-1	+1	+1
+1	+1	-1

(**b**)



(**c**)

j	digit 1	digit 2
0	0	0
1	0	1
2	1	0
3	1	1

Figure 7.6: Boolean XOR function. (**a**) value table, (**b**) network layout. For the weights feeding into the hidden layer, dashed lines correspond to $w_{jk} = -\delta$, solid lines to $w_{jk} = \delta$. For the weights feeding into the output neuron, dashed lines correspond to $W_{1j} = -\gamma$, and solid lines to $W_{1j} = \gamma$ (**c**) construction principle for the weights of the hidden layer.

A difference compared with the Boolean networks in Section 5.5 is that here the inputs take the values ± 1. The reason is that this simplifies the proof, which is by construction [1]. For each hidden neuron one assigns the weights as follows

$$w_{jk} = \begin{cases} \delta & \text{if the } k^{\text{th}}\text{digit of binary representation of } j \text{ is } 1\,, \\ -\delta & \text{otherwise,} \end{cases} \tag{7.6}$$

with $\delta > 1$ (see below). The thresholds θ_j of all hidden neurons are the same, equal to $N(\delta - 1)$. The idea is that each input pattern turns on exactly one neuron in the hidden layer (called the *winning neuron*). This requires that δ is large enough, as we shall see. The weights feeding into the output neuron are assigned as follows. If the output for the pattern represented by neuron V_j is $+1$, let $W_{1j} = +1$, otherwise $W_{1j} = -1$. The threshold is $\Theta = \sum_j W_{1j}$.

To show how this construction works, consider the Boolean XOR function as an example. First, for each pattern only the corresponding winning neuron gives a positive signal. For pattern $\boldsymbol{x}^{(1)} = [-1,-1]^{\mathsf{T}}$, for example, this is the first neuron in the hidden layer ($j = 0$). To see this, compute the local fields for this input pattern:

$$\begin{aligned} b_0^{(1)} &= 2\delta - 2(\delta - 1) = 2\,, \\ b_1^{(1)} &= -2(\delta - 1) = 2 - 2\delta\,, \\ b_2^{(1)} &= -2(\delta - 1) = 2 - 2\delta\,, \\ b_3^{(1)} &= -2\delta - 2(\delta - 1) = 2 - 4\delta\,. \end{aligned} \tag{7.7}$$

If we choose $\delta > 1$ then the output of the first hidden neuron gives a positive output ($V_0 > 0$), the other neurons produce negative outputs, $V_j < 0$ for $j = 1,2,3$. Now

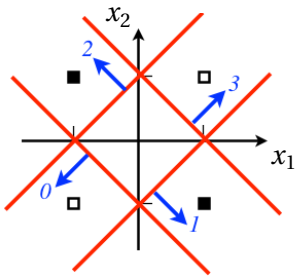


Figure 7.7: Shows how the XOR network depicted in Figure 7.6 partitions the input plane. Target values are encoded as in Figure 5.8: □ corresponds to $t=-1$ and ■ to $t=+1$).

consider $\boldsymbol{x}^{(3)}=[-1,+1]^{\mathsf{T}}$. In this case

$$\begin{aligned} b_0^{(3)} &= -2(\delta-1)=2-2\delta\,, \\ b_1^{(3)} &= -2\delta-2(\delta-1)=2-4\delta\,, \\ b_2^{(3)} &= 2\delta-2(\delta-1)=2\,, \\ b_3^{(3)} &= -2(\delta-1)=2-2\delta\,. \end{aligned} \tag{7.8}$$

Now the third hidden neuron gives a positive output, while the others yield negative outputs. It works in the same way for the other two patterns, $\boldsymbol{x}^{(2)}$ and $\boldsymbol{x}^{(4)}$. In summary, there is a unique winning neuron for each pattern.[1] How do the decision boundaries corresponding to V_j for $j=0,\ldots,3$ partition the input plane? This is shown in Figure 7.7.

According to the scheme outlined above, the output neuron computes

$$O_1=\mathrm{sgn}(-V_1+V_2+V_3-V_4) \tag{7.9}$$

with $\Theta=\sum_j W_{1j}=0$. For $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(4)}$ we find the correct result $O_1=-1$. The same is true for $\boldsymbol{x}^{(2)}$ and $\boldsymbol{x}^{(3)}$, we obtain $O_1=1$. In summary, this example illustrates how an N-dimensional Boolean function is represented by a network with one hidden layer, with 2^N neurons. The problem is of course that this network is expensive to train for large N because the number of hidden neurons is very large.

There are more efficient layouts if one uses more than one hidden layer. As an example, consider the *parity function* for N binary inputs equal to 0 or 1. The function measures the parity of the input sequence. It gives 1 if there is an odd number of ones in the input, otherwise 0. A construction similar to the above yields a network layout with 2^N neurons in the hidden layer. If one instead wires together

[1] That pattern $\mu=k$ gives the winning neuron $j=k-1$ is of no importance, it is just a consequence of how the patterns are ordered in the value table in 7.6

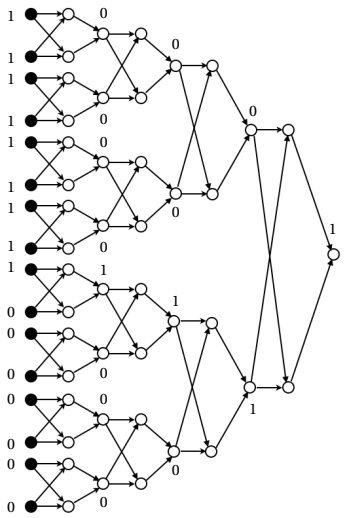


Figure 7.8: Solution of the parity problem for N-dimensional inputs. The network is built from XOR units (Figure 5.16). Each XOR unit has a hidden layer with two neurons. Above only the states of the inputs and outputs of the XOR units are shown, not those of the hidden neurons. In total, the whole network has $O(N)$ neurons.

the XOR networks shown in Figure 5.16, one can solve the parity problem with $O(N)$ neurons, as Figure 7.8 demonstrates. When N is a power of two then this network has $3(N-1)$ neurons. To see this, set the number of inputs to $N = 2^k$. Figure 7.8 shows that the number $\mathcal{N}_k$ of neurons satisfies the recursion $\mathcal{N}_{k+1} = 2\mathcal{N}_k + 3$ with $\mathcal{N}_1 = 3$. The solution of this recursion is $\mathcal{N}_k = 3(2^k - 1)$.

This example also illustrates a second reason why it may be useful to have more than one hidden layer. To design a network for a certain task it is often convenient to build the network from building blocks. One wires them together, often in a hierarchical fashion. In Figure 7.8 there is only one building block, the XOR network from Figure 5.16.

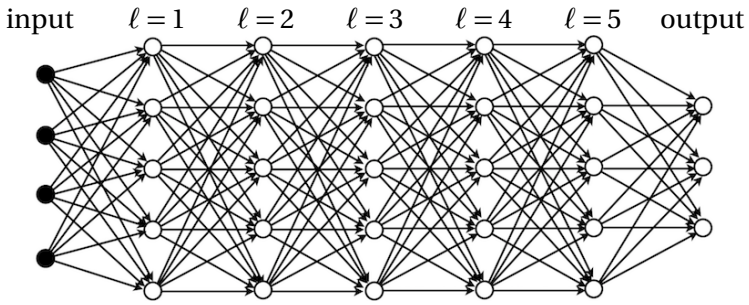


Figure 7.9: Fully connected deep network with five hidden layers. How deep is deep? Usually one says: deep networks have two or more hidden layers.

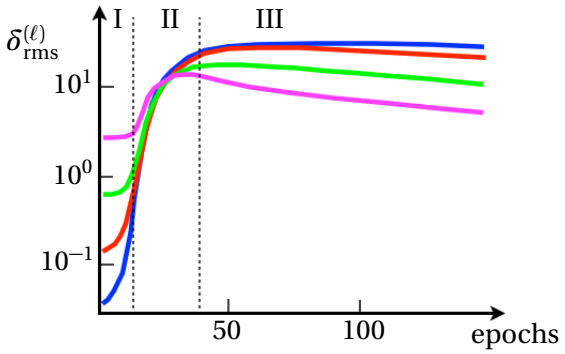


Figure 7.10: Vanishing-gradient problem for a network with four fully connected hidden layers. The Figure illustrates schematically how the r.m.s. error $\delta^{(\ell)}_{\rm rms}$ in layer ℓ (see text) depends on the number of training epochs, for $\ell = 1$ (blue), $\ell = 2$ (red), $\ell = 3$ (green), and $\ell = 4$ (magenta). During phase I the vanishing-gradient problem is severe, during phase II the network starts to learn, phase III is the convergence phase where the errors decline. Schematic, based on simulations performed by Ludvig Storm.

Another example are convolutional networks for image analysis (Chapter 8). Here the fundamental building blocks are *feature maps*, they recognise different geometrical features in the image, such as edges or corners.

7.2 Vanishing gradients

In Chapter 6 it was pointed out that learning slows down when the factors $g'(b)$ become small. When the network has many hidden layers, this problem is more severe. One finds that the neurons in hidden layers close to the input layer (small values of ℓ in Figure 7.9) change only by small amounts, the smaller the more hidden layers the network has.

Figure 7.10 shows that the r.m.s. errors averaged over different realisations of random initial weights, $\big(\langle N_\ell^{-1}\sum_{j=1}^{N_\ell}[\delta_j^{(\ell)}]^2\rangle\big)^{1/2}$, tend to be very small for the first 20 training epochs. In this regime the gradient (and thus the speed of training) vanishes exponentially as $\ell \to 1$. This *slowing down* is the result of the diminished effect of the neurons in layer ℓ upon the output, when ℓ is small. This is the *vanishing-gradient problem*. To explain this phenomenon, consider the very simple case shown in Figure 7.11: a deep network with only one neuron per layer. The output $V^{(L)}$ is given by nested activation functions

$$V^{(L)} = g\Big(w^{(L)} g\Big(w^{(L-1)} \cdots g\big(w^{(2)} g(w^{(1)} x - \theta^{(1)}) - \theta^{(2)}\big) \ldots - \theta^{(L-1)}\Big) - \theta^{(L)}\Big). \qquad (7.10)$$

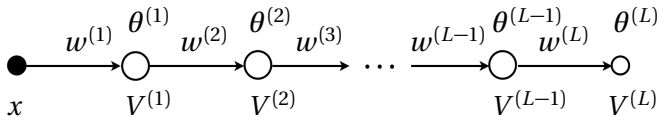


Figure 7.11: 'Network' illustrating the vanishing-gradient problem, with neurons $V^{(\ell)}$, weights $w^{(\ell)}$, and thresholds $\theta^{(\ell)}$.

The effects of the neurons in Figure 7.11 are computed using the chain rule:

$$\begin{aligned}
\frac{\partial V^{(L)}}{\partial V^{(L-1)}} &= g'(b^{(L)})w^{(L)} \\
\frac{\partial V^{(L)}}{\partial V^{(L-2)}} &= \frac{\partial V^{(L)}}{\partial V^{(L-1)}}\frac{\partial V^{(L-1)}}{\partial V^{(L-2)}} = g'(b^{(L)})w^{(L)}g'(b^{(L-1)})w^{(L-1)} \\
&\vdots
\end{aligned} \tag{7.11}$$

where $b^{(k)} = w^{(k)}V^{(k-1)} - \theta^{(k)}$ is the local field for neuron k. This yields the following expression for $\partial V^{(L)}/\partial V^{(\ell)}$:

$$\frac{\partial V^{(L)}}{\partial V^{(\ell)}} = \prod_{k=L}^{\ell+1}[g'(b^{(k)})w^{(k)}]. \tag{7.12}$$

Using Equation (6.18), $\delta^{(\ell)} = [t - V^{(L)}]\partial V^{(L)}/\partial V^{(\ell)}g'(b^{(\ell)})$, one finds:

$$\delta^{(\ell)} = [t - V^{(L)}(x)]g'(b^{(L)})\prod_{k=L}^{\ell+1}[w^{(k)}g'(b^{(k-1)})]. \tag{7.13}$$

One can also obtain this expression by applying the recursion from Algorithm 3, $\delta^{(\ell)} = \delta^{(\ell+1)}w^{(\ell+1)}g'(b^{(\ell)})$.

Now consider the early stages of training. If one initialises the weights as described in Chapter 6 to Gaussian random variables with mean zero and variance σ_w^2, and the thresholds to zero, then the factors $w^{(k)}g'(b^{(k-1)})$ are usually smaller than unity (for the activation functions (6.21), the maximum of $g'(b)$ is $\frac{1}{2}$ and 1, respectively). The product of these factors vanishes quickly as ℓ decreases. So the slowing down is a consequence of multiplying many small numbers to get something really small (*vanishing-gradient problem*). This is phase I in Figure 7.10.

What happens at later times? One might argue that the weights may grow during training, as a function of ℓ. If that happened, the problem might become worse still, because $g'(b)$ tends to zero exponentially as $|b|$ grows. This indicates that the first layers may continue to learn slowly. Figure 7.10 shows that the effect persists

for about 20 epochs. But then even the first layers begin to learn faster (phase II). This does not contradict the above discussion, because it assumed random weights. As the network learns, the weights are no longer independent random numbers. But there is to date no mathematical theory describing how this transition occurs. Much later in training, the errors decay during the convergence phase (phase III in Figure 7.10).

In summary, Equation (7.13) demonstrates that different layers of the network learn at different speeds initially, when the weights are still random.

There is a second, equivalent, point of view: the learning is slow in a layer far from the output because the output is not very sensitive to the state of these neurons. To measure the effect of a given neuron on the output, we calculate how the output of the network changes when changing the *state* of a neuron in a particular layer. For the example shown in Figure 7.11, the output $V^{(L)}$ is given by nested activation functions

$$V^{(L)}=g\Big(w^{(L)}g\Big(w^{(L-1)}\cdots g\big(w^{(2)}g(w^{(1)}x-\theta^{(1)})-\theta^{(2)}\big)\ldots-\theta^{(L-1)}\Big)-\theta^{(L)}\Big). \tag{7.14}$$

The effects of the neurons in Figure 7.11 are computed using the chain rule:

$$\begin{aligned}\frac{\partial V^{(L)}}{\partial V^{(L-1)}} &= g'(b^{(L)})w^{(L)}\\ \frac{\partial V^{(L)}}{\partial V^{(L-2)}} &= \frac{\partial V^{(L)}}{\partial V^{(L-1)}}\frac{\partial V^{(L-1)}}{\partial V^{(L-2)}}=g'(b^{(L)})w^{(L)}g'(b^{(L-1)})w^{(L-1)}\\ &\vdots\end{aligned} \tag{7.15}$$

where $b^{(k)}=w^{(k)}V^{(k-1)}-\theta^{(k)}$ is the local field for neuron k. This yields the following expression for $\partial V^{(L)}/\partial V^{(\ell)}$:

$$\frac{\partial V^{(L)}}{\partial V^{(\ell)}}=\prod_{k=L}^{\ell+1}[g'(b^{(k)})w^{(k)}]. \tag{7.16}$$

Again it is a product of $g'(b)$ that determines how the effect of $V^{(\ell)}$ on the output decreases as ℓ decreases. Equations (7.16) and (7.13) are equivalent. This follows from Equation (6.18), $\delta^{(\ell)}=[t-V^{(L)}]\partial V^{(L)}/\partial V^{(\ell)}g'(b^{(\ell)})$.

More generally, note that the product in Equation (7.16) is unlikely to remain of order unity when L is large. To see this, assume that the weights are independently distributed random numbers. Taking the logarithm and using the *central-limit theorem* shows that the distribution of the product is log normal. This means that the learning speed can be substantially different in different layers. This is also referred to the problem of *unstable gradients*. The example shown in Figure 7.11

illustrates the origin of this problem: it is due to the fact that multiplying many small numbers together produces a result that is very small. Multiplying many numbers that are larger than unity, by contrast, yields a large result.

In networks like the one shown in Figure 7.9 the principle is the same, but instead of multiplying numbers one multiplies matrices. The product (7.16) of random numbers becomes of product of random matrices. Assume that all layers $\ell = 1, \ldots, L$ have N neurons. Using the chain rule we find:

$$\frac{\partial V_i^{(L)}}{\partial V_j^{(\ell)}} = \sum_{mn\cdots p} \frac{\partial V_i^{(L)}}{\partial V_m^{(L-1)}} \frac{\partial V_m^{(L-1)}}{\partial V_n^{(L-2)}} \cdots \frac{\partial V_p^{(\ell+1)}}{\partial V_j^{(\ell)}} . \tag{7.17}$$

Using the update rule

$$V_m^{(k)} = g\Big(\sum_j w_{ij}^{(k)} V_j^{(k-1)} - \theta_i^{(k)} \Big) \tag{7.18}$$

we can evaluate each factor:

$$\frac{\partial V_m^{(k)}}{\partial V_n^{(k-1)}} = g'(b_m^{(k)}) w_{mn}^{(k)} . \tag{7.19}$$

In summary, this yields the following expression for the matrix $\mathbb{J}_{\ell,L}$ with elements $J_{ij}^{(\ell,L)} = \partial V_i^{(L)} / \partial V_j^{(\ell)}$:

$$\mathbb{J}_{\ell,L} = \mathbb{D}^{(L)} \mathbb{W}^{(L)} \mathbb{D}^{(L-1)} \mathbb{W}^{(L-1)} \cdots \mathbb{D}^{(\ell+1)} \mathbb{W}^{(\ell+1)} , \tag{7.20}$$

where $\mathbb{W}^{(k)}$ is the matrix of weights feeding into layer k, and

$$\mathbb{D}^{(k)} = \begin{bmatrix} g'(b_1^{(k)}) & & \\ & \ddots & \\ & & g'(b_N^{(k)}) \end{bmatrix} . \tag{7.21}$$

This expression is analogous to Equation (7.16). The eigenvalues of the matrix $\mathbb{J}_{0,k}$ describe how small changes $\delta \boldsymbol{V}^{(0)}$ to the inputs $\boldsymbol{V}^{(0)}$ (or small differences between the inputs) grow as they propagate through the layers. If the maximal eigenvalue is larger than unity, then $|\delta \boldsymbol{V}^{(0)}|$ grows exponentially as a function of layer index k. This is quantified by the maximal *Lyapunov exponent* [65]

$$\lambda = \lim_{k\to\infty} \frac{1}{2k} \Big\langle \log \mathrm{tr} \big(\mathbb{J}_{0,k}^{\mathsf{T}} \mathbb{J}_{0,k} \big) \Big\rangle \tag{7.22}$$

where the average is over realisations of weights and thresholds. The matrix $\mathbb{J}_{0,k}^{\mathsf{T}} \mathbb{J}_{0,k}$ is called the *right Cauchy-Green* matrix, and tr denotes the *trace* of this matrix, the

sum of its diagonal elements. The right Cauchy-Green matrix is symmetric, and it is positive definite. The eigenvectors of $\mathbb{J}_{0,k}^{\mathsf{T}}\mathbb{J}_{0,k}$ are called *forward Lyapunov vectors*. They describe how small corrections to the inputs rotate, shrink, or stretch as they propagate through the network.

If we multiply the matrix $\mathbb{J}_{k,L}$ from the left with the transpose of the vector $\boldsymbol{\delta}^{(L)}$ of output errors, we see how the errors change as they propagate backwards from layer k to the leftmost hidden layer, how this vector rotates, shrinks, or stretches.

There are a number of different tricks that help to suppress vanishing gradients, to some extent at least. First, it is usually argued that it helps to use an activation function that does not saturate at large b, such as the ReLU function introduced in Section 7.3. But the results of Ref. [66] show that the effect is perhaps not as strong as originally thought. Second, batch normalisation (Section 7.7.5) may help against the unstable gradient problem. Third, introducing connections that skip layers (*residual network*) can also reduce the unstable-gradient problem. This is discussed in Section 7.4.

7.3 Rectified linear units

Glorot *et al.* [67] suggested to the piecewise activation function, the ReLu function $\max\{0, b\}$ (Chapter 1). Note that the derivative of the ReLU function is discontinuous at $b = 0$. A common convention is to set the derivative to zero at $b = 0$. What is the point of using ReLU neurons? When training a deep network with ReLU functions, many of the hidden neurons produce output zero. This means that the network of active neurons (non-zero output) is *sparsely* connected. It is thought that sparse networks have desirable properties, and sparse representations of a classification problem are more likely to be linearly separable (as shown in Section 5.4). Figure 7.12 illustrates that for a given input pattern only a certain fraction of hidden neurons is active. For these neurons the computation is linear, yet different input patterns give different sets of active neurons. The product in Equation (7.20) acquires a particularly simple structure: the matrices $\mathbb{D}^{(k)}$ are diagonal with 0/1 entries. But while the weight matrices are independent, the $\mathbb{D}^{(k)}$-matrices are correlated: which elements vanish depends on the states of the neurons in the corresponding layer, which in turn depend on the weights to the right of $\mathbb{D}^{(k)}$ in the matrix product. A hidden layer with only one or very few active neurons might act as a *bottleneck* preventing efficient backpropagation of output errors which could in principle slow down training. For the examples given in Ref. [67] this does not occur.

The ReLU function is unbounded for large positive local fields. Therefore, the vanishing-gradient problem (Section 7.2) is thought to be less severe in networks made of rectified linear units, but see Ref. [66]. Since the ReLU function does not

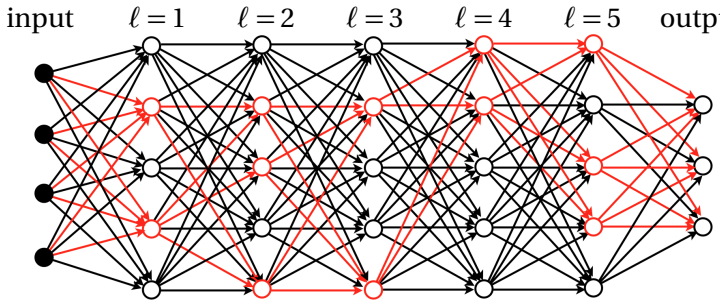


Figure 7.12: Sparse network of active neurons with ReLU activation functions. The red paths correspond to *active* neurons with positive local fields.

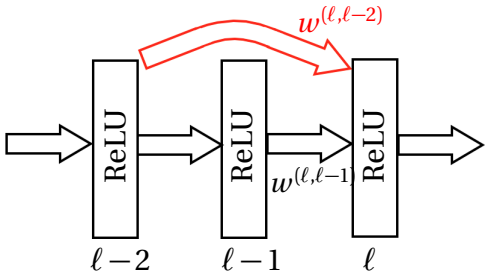


Figure 7.13: Schematic illustration of a network with skipping connections.

saturate, the weights tend to increase. Glorot *et al.* [67] suggested to use L_1-weight decay (Section 7.7.1) to make sure that the weights do not grow.

Finally, using ReLU functions instead of sigmoid functions speeds up the training, because the ReLU function has piecewise constant derivatives. Such function calls are faster to evaluate than sigmoid functions, for example.

7.4 Residual networks

One way of reducing the vanishing-gradient problem is to introduce connections that skip layers [68] (Exercise 6.4). Deep neural nets with skipping layers are called *residual nets*. Empirical evidence shows that residual nets are easier to train than standard multilayer perceptrons. This Section explains how to train residual nets. The layout is illustrated schematically in Figure 7.13. Black arrows stand for usual feed-forward connections. The notation differs somewhat from that of Algorithm 3. Here the weights from layer $\ell-1$ to ℓ are denoted by $w_{jk}^{(\ell,\ell-1)}$, and those from layer $\ell-2$ to ℓ by $w_{ij}^{(\ell,\ell-2)}$ (red arrow in Figure 7.13). Note that the superscripts are ordered in the same way as the subscripts: the *right* index refers to the layer on the *left*. Neuron j in layer ℓ computes

$$V_j^{(\ell)} = g\Big(\sum_k w_{jk}^{(\ell,\ell-1)} V_k^{(\ell-1)} - \theta_j^{(\ell)} + \sum_n w_{jn}^{(\ell,\ell-2)} V_n^{(\ell-2)}\Big). \tag{7.23}$$

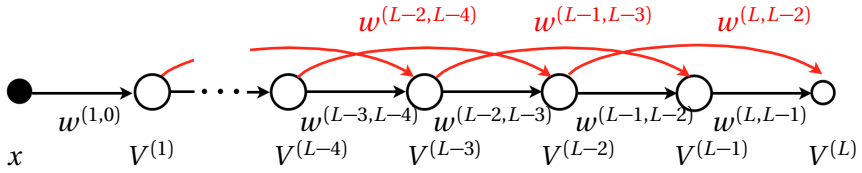


Figure 7.14: 'Network' with connections that skip layers.

The weights of connections that skip layers are trained in the usual fashion, by stochastic gradient descent. To illustrate the structure of the resulting formulae consider a 'network' with just one neuron per layer (Figure 7.14). We calculate the weight updates using Equations (**??**) and (6.18). The recursion (6.19) applies only to standard feed-forward nets without skipping layers. In order to determine how to update the weights for the network shown in Figure 7.14, we need to evaluate the gradients $\partial V^{(L)}/\partial V^{(\ell)}$.

To begin with consider the update rule for $w^{(L,L-1)}$. Using Equations (**??**) and (6.18) one finds:

$$\delta w^{(L,L-1)} = \eta \delta^{(L)} V^{(L-1)} \quad \text{with} \quad \delta^{(L)} = (t - V^{(L)}) g'(b^{(L)}), \tag{7.24}$$

as in Algorithm 3. In the same way one finds

$$\delta w^{(L,L-2)} = \eta \delta^{(L)} V^{(L-2)} \quad \text{with} \quad \delta^{(L)} = (t - V^{(L)}) g'(b^{(L)}), \tag{7.25}$$

Now consider the update rule for $w^{(L-1,L-2)}$. Using $\partial V^{(L)}/\partial V^{(L-1)} = w^{(L,L-1)}$ gives

$$\delta w^{(L-1,L-2)} = \eta \delta^{(L-1)} V^{(L-2)} \quad \text{with} \quad \delta^{(L-1)} = \delta^{(L)} w^{(L,L-1)} g'(b^{(L-1)}). \tag{7.26}$$

But the update for $w^{(L-2,L-3)}$ is different because the short cuts come into play. The extra connection from layer $L-2$ to L gives rise to an extra term in the gradient

$$\frac{\partial V^{(L)}}{\partial V^{(L-2)}} = \Big(\frac{\partial V^{(L)}}{\partial V^{(L-1)}} \frac{\partial V^{(L-1)}}{\partial V^{(L-2)}} + \frac{\partial V^{(L)}}{\partial V^{(L-2)}} \Big). \tag{7.27}$$

Evaluating the partial derivatives yields:

$$\begin{aligned} \delta w^{(L-2,L-3)} = \eta \delta^{(L-2)} V^{(L-3)} \quad \text{with} \quad \delta^{(L-2)} &= \delta^{(L-1)} w^{(L-1,L-2)} g'(b^{(L-2}) \\ &+ \delta^{(L)} w^{(L,L-2)} g'(b^{(L-2)}), \end{aligned} \tag{7.28}$$

and so forth. In general, the error-backpropagation rule reads

$$\delta^{(\ell-1)} = \delta^{(\ell)} w^{(\ell,\ell-1)} g'(b^{(\ell-1)}) + \delta^{(\ell+1)} w^{(\ell+1,\ell-1)} g'(b^{(\ell-1)}) \tag{7.29}$$

for $\ell = L, L-1, \ldots$. The first term is the same as in step 9 of Algorithm 3. The second term is due to the skipping connections. These connections reduce the vanishing-gradient problem. To see this, note that we can write the error $\delta^{(\ell)}$ as

$$\delta^{(\ell)} = \delta^{(L)} \sum_{\ell_1, \ell_2, \ldots, \ell_n} w^{(L,\ell_n)} g'(b^{(\ell_n)}) \cdots w^{(\ell_2,\ell_1)} g'(b^{(\ell_1)}) w^{(\ell_1,\ell)} g'(b^{(\ell)}) \tag{7.30}$$

where the sum is over all paths $L > \ell_n > \ell_{n-1} > \cdots > \ell_1 > \ell$ back through the network. The smallest gradients are dominated by the product corresponding to the path with the smallest number of steps (factors), resulting in a smaller probability to get small gradients. Introducing connections that skip more than one layer tends to increase the small gradients, as Equation (7.30) shows. Recently it has been suggested to randomise the layout by randomly short-circuiting the network. Equation (7.30) remains valid for this case too. Finally, the network described in Ref. [68] used unit weights for the skipping connections,

$$V_j^{(\ell)} = g\Big(\sum_k w_{jk}^{(\ell,\ell-1)} V_k^{(\ell-1)} - \theta_j^{(\ell)} + V_j^{(\ell-2)} \Big), \tag{7.31}$$

so that the hidden layer $V_k^{(\ell-1)}$ learns the difference between the input $V_j^{(\ell-2)}$ and the output $V_j^{(\ell)}$. Therefore such networks are called *residual networks*.

7.5 Outputs and energy functions

Up to now we discussed networks that have the same activation functions for all neurons in all layers, either sigmoid or tanh activation functions (Equation 6.21), or ReLU functions (Section 7.3). These networks are trained by stochastic gradient descent on the quadratic energy function (5.23). It has been shown that it may be advantageous to employ a different energy function, and to use slightly different activation functions for the neurons in the output layer, so-called *softmax* outputs, defined as

$$O_i = \frac{\mathrm{e}^{\alpha b_i^{(L)}}}{\sum_{k=1}^{M} \mathrm{e}^{\alpha b_k^{(L)}}}. \tag{7.32}$$

Here $b_i^{(L)} = \sum_j w_{ij}^{(L)} V_j^{(L-1)} - \theta_i^{(L)}$ are the local fields in the output layer. Usually the constant α is taken to be unity. In the limit $\alpha \to \infty$, you see that $O_i = \delta_{i i_0}$ where i_0 is the index of the winning output neuron, the one with the largest value $b_i^{(L)}$ (Chapter 10). Usually one takes $\alpha = 1$, then Equation (7.32) is a *soft* version of this maximum criterion, thus the name *softmax*. Three important properties of softmax outputs

are, first, that $0 \leq O_i \leq 1$. Second, the values of the outputs sum to one

$$\sum_{i=1}^{M} O_i = 1 \,. \tag{7.33}$$

This means that the outputs of softmax units can be interpreted as probabilities. Third, when $b_i^{(L)}$ increases then O_i increases but the values O_k of the other output neurons $k \neq i$ decrease.

Softmax units can simplify interpreting the network output for classification problems where the inputs must be assigned to one of M classes. In this problem, the output $O_i^{(\mu)}$ of softmax unit i represents the probability that the input $\boldsymbol{x}^{(\mu)}$ is in class i (in terms of the targets: $t_i^{(\mu)} = 1$ while $t_k^{(\mu)} = 0$ for $k \neq i$). Softmax units are often used in conjunction with a different energy function. It is defined in terms of *negative log likelihoods*

$$H = -\sum_{i\mu} t_i^{(\mu)} \log O_i^{(\mu)} \,. \tag{7.34}$$

Here and in the following log stands for the natural logarithm. The function (7.34) is minimal when $O_i^{(\mu)} = t_i^{(\mu)}$. Since the function (7.34) is different from the energy function used in Chapter 6, the details of the backpropagation algorithm are slightly different. To find the correct formula for backpropagation, we need to evaluate

$$\frac{\partial H}{\partial w_{mn}} = -\sum_{i\mu} \frac{t_i^{(\mu)}}{O_i^{(\mu)}} \frac{\partial O_i^{(\mu)}}{\partial w_{mn}} \,. \tag{7.35}$$

Here I did not write out the labels L that denote the output layer, and in the following equations I also drop the index μ that refers to the input pattern. Using the identities

$$\frac{\partial O_i}{\partial b_l} = O_i(\delta_{il} - O_l) \quad \text{and} \quad \frac{\partial b_l}{\partial w_{mn}} = \delta_{lm} V_n \,, \tag{7.36}$$

one obtains

$$\frac{\partial O_i}{\partial w_{mn}} = \sum_l \frac{\partial O_i}{\partial b_l} \frac{\partial b_l}{\partial w_{mn}} = O_i(\delta_{im} - O_m) V_n \,. \tag{7.37}$$

So

$$\delta w_{mn} = -\eta \frac{\partial H}{\partial w_{mn}} = \eta \sum_{i\mu} t_i^{(\mu)} (\delta_{im} - O_m^{(\mu)}) V_n^{(\mu)} = \eta \sum_{\mu} (t_m^{(\mu)} - O_m^{(\mu)}) V_n^{(\mu)} \,, \tag{7.38}$$

since $\sum_{i=1}^{M} t_i^{(\mu)} = 1$ for the type of classification problem where each input belongs to precisely one class. The corresponding expression for the threshold updates reads

$$\delta \theta_m = -\eta \frac{\partial H}{\partial \theta_m} = -\eta \sum_{\mu} (t_m^{(\mu)} - O_m^{(\mu)}) \,. \tag{7.39}$$

Equations (7.38) and (7.39) highlight a further advantage of softmax output neurons (apart from the fact that they allow the output to be interpreted in terms of probabilities). The weight and threshold increments for the output layer derived in Section 6 [Equations (6.7a) and (6.13a)] contain factors of derivatives $g'(B_m^{(\mu)})$. As noted earlier, these derivatives tend to zero when the activation function saturates, slowing down the learning. But here the rate at which the neuron learns is simply proportional to the error, $(t_m^{(\mu)} - O_m^{(\mu)})$, no small factors reduce this rate.

Softmax units are normally only used in the output layer. First, the derivation shows that the learning speedup mentioned above is coupled to the use of the log likelihood function (7.34). Second, one usually tries to avoid dependence between the neurons in a given hidden layer, but Equation (7.32) shows that the output of neuron i depends on all local fields in the hidden layer. A better alternative is usually the ReLU activation function discussed in Section 7.3.

There is an alternative way of choosing the energy function that is very similar to the above, but works with sigmoid activation functions and 0/1 targets:

$$H = -\sum_{i\mu} t_i^{(\mu)} \log O_i^{(\mu)} + (1 - t_i^{(\mu)}) \log(1 - O_i^{(\mu)}), \tag{7.40}$$

with $O_i = \sigma(b_i)$ where σ is the sigmoid function (6.21a). The function (7.40) is called *cross-entropy* function. To compute the weight increments, we apply the chain rule:

$$\frac{\partial H}{\partial w_{mn}} = \sum_{i\mu} \left(\frac{t_i^{(\mu)}}{O_i^{(\mu)}} - \frac{1 - t_i^{(\mu)}}{1 - O_i^{(\mu)}} \right) \frac{\partial O_l}{\partial w_{mn}} = \sum_{i\mu} \frac{t_i^{(\mu)} - O_i^{(\mu)}}{O_i^{(\mu)}(1 - O_i^{(\mu)})} \frac{\partial O_l}{\partial w_{mn}}. \tag{7.41}$$

Using Equation (6.22) we obtain

$$\delta w_{mn} = \eta \sum_{\mu} (t_m^{(\mu)} - O_m^{(\mu)}) V_n^{(\mu)}, \tag{7.42}$$

identical to Equation (7.38). The threshold increments are also updated in the same way, Equation (7.39). Yet the interpretation of the outputs is slightly different, since the values of the softmax units in the output layers sum to unity, while those with sigmoid activation functions do not. In either case you can use the definition (6.31) for the classification error.

7.6 Weight initialisation

The results of Section 7.2 point to the importance of initialising the weights in the right way, to avoid that the learning slows down. This is significant because it is often found that the initial transient learning phase poses a substantial bottleneck

to learning [69]. For this initial transient, correct weight initialisation can give a substantial improvement. Moreover, when training deep networks with sigmoid activation functions in the hidden layers, it was observed that the values of the output neurons remain very close to zero for many training iterations (Figure 2 in Ref. [70]), slowing down the training substantially. It is argued that this is a consequence of the way the weights are initialised, in combination with the particular shape of the sigmoid activation function.

So, how should the weights be initialised? The standard choice is to initialise the weights to independent Gaussian random numbers with mean zero and unit variance and the thresholds to zero (Section 6.1). However, this scheme may fail for networks with hidden layers with many neurons. Consider a neuron i in the first hidden layer with N incoming connections. Its local field

$$b_i = \sum_{j=1}^{N} w_{ij} x_j \tag{7.43}$$

is a sum of many independently identically distributed random numbers. Assume that the input patterns have independent random bits, equal to 0 or 1 with probability $\frac{1}{2}$. From the central-limit theorem we find that the local field is Gaussian distributed in the limit of large N, with mean zero and variance

$$\sigma_b^2 = \sigma_w^2 N/2\,. \tag{7.44}$$

This means that the local field is typically quite large, of order $\sqrt{N}$, and this implies that the neurons of the first hidden layer saturate – slowing down the learning. This conclusion rests on our particular assumption concerning the input patterns, but it is in general much better to initialise the weights uniformly or Gaussian with mean zero and with variance

$$\sigma_w^2 \propto N^{-1}\,, \tag{7.45}$$

to cancel the factor of N in Equation (7.44). The thresholds can be initialised to zero, as described in Section 6.1. The normalisation (7.45) ensures that the weights are not too large initially, but it does not circumvent the vanishing-gradient problem discussed in Section 7.2. There the problem was illustrated for $N = 1$, so unit variance for the initial weight distribution corresponds to Equation (7.45).

7.7 Regularisation*

Deeper networks have more neurons, so the problem of overfitting (Figure 6.7) tends to be more severe for deeper networks. Therefore *regularisation* schemes that

limit the tendency to overfit are important for deeper networks. In training deep networks, a number of other regularisation schemes have proved useful: *weight decay*, *pruning*, *drop out*, and expanding the training set. This Section summarises the most important aspects of these methods.

7.7.1 Weight decay

Recall Figure 5.16 which shows a solution of the classification problem defined by the Boolean XOR function. All weights are of unit modulus, and also the thresholds are of order unity. If one uses the backpropagation algorithm to find a solution to this problem, one may find that the weights continue to grow during training. This can be problematic, if it means that the local fields become too large, so that the algorithm reaches the plateau of the activation function. Then training slows down, as explained in Section 6.1.

One solution to this problem is to reduce the weights by some factor during training, either at each iteration or in regular intervals, $w_{ij} \to (1-\varepsilon)w_{ij}$ for $0 < \varepsilon < 1$, or

$$\delta w_{mn} = -\varepsilon w_{mn} \quad \text{for} \quad 0 < \varepsilon < 1 \,. \tag{7.46}$$

This is achieved by adding a term to the energy function

$$H = \underbrace{\frac{1}{2}\sum_{i\mu}\left(t_i^{(\mu)} - O_i^{(\mu)}\right)^2}_{\equiv H_0} + \frac{\gamma}{2}\sum_{ij} w_{ij}^2 \,. \tag{7.47}$$

Gradient descent on H gives:

$$\delta w_{mn} = -\eta \frac{\partial H_0}{\partial w_{mn}} - \varepsilon w_{mn} \tag{7.48}$$

with $\varepsilon = \eta\gamma$. One can add a corresponding term for the thresholds, but this is usually not necessary. The scheme summarised here is sometimes called L_2*-regularisation*. An alternative scheme is L_1*-regularisation*. It amounts to

$$H = \frac{1}{2}\sum_{i\mu}\left(t_i^{(\mu)} - O_i^{(\mu)}\right)^2 + \frac{\gamma}{2}\sum_{ij} |w_{ij}| \,. \tag{7.49}$$

This gives the update rule

$$\delta w_{mn} = -\eta \frac{\partial H_0}{\partial w_{mn}} - \varepsilon \,\mathrm{sgn}(w_{mn}) \,. \tag{7.50}$$

The discontinuity of the update rule at $w_{mn} = 0$ is cured by defining $\mathrm{sgn}(0) = 0$. Comparing Equations (7.48) and (7.50) we see that L_1-regularisation reduces small

weights much more than L_2-regularisation. We expect therefore that the L_1-scheme puts more weights to zero, compared with the L_2-scheme.

An alternative to these two methods is *max-norm regularisation* [71], where the weights are constrained to remain smaller than a given constant: $|w_{ij}| \leq c$. If a $|w_{ij}|$ exceeds the positive constant c, then w_{ij} is rescaled so that $|w_{ij}| = c$.

These weight-decay schemes are referred to as *regularisation* schemes because they tend to help against overfitting. How does this work? Weight decay adds a constraint to the problem of minimising the energy function. The result is a compromise, depending upon the value γ, between a small value of H and small weight values. The idea is that a network with smaller weights is more robust to the effect of noise. When the weights are small, then small changes in some of the patterns do not give a substantially different training result. When the network has large weights, by contrast, it may happen that small changes in the input give significant differences in the training result that are difficult to *generalise* (Figure 6.7).

7.7.2 Pruning

The term *pruning* refers to removing unnecessary weights or neurons from the network, to improve its efficiency. The simplest approach is *weight elimination* by *weight decay* [72]. Weights that tend to remain very close to zero during training are removed by setting them to zero and not updating them anymore. Neurons that have zero weights for all incoming connections are effectively removed (*pruned*). Pruning is a regularisation method: by removing unnecessary weights, one reduces the risk of overfitting. As opposed to drop out, where hidden neurons are only temporarily ignored, pruning refers to permanently removing hidden neurons. The idea is to train a large network, and then to prune a large fraction of neurons to obtain a much smaller network. It is usually found that such pruned nets generalise much better than small nets that were trained without pruning. Up to 90% of the hidden neurons can be removed.

An efficient pruning algorithm is based on the idea to remove weights in such a way that the effect upon the energy function is as small as possible [73]. The idea is to find the optimal weight, to remove it, and to change the other weights in such a way that the energy function increases as little as possible. The algorithm works as follows. Assume that the net was trained, so that it reached a (local) minimum of the energy function H. One expands the energy function around this minimum: The expansion of H reads:

$$H = H_{\min} + \tfrac{1}{2}\delta\boldsymbol{w} \cdot \mathbb{M}\delta\boldsymbol{w} \quad + \quad \text{higher orders in} \quad \delta\boldsymbol{w}\,. \tag{7.51}$$

The term linear in $\delta\boldsymbol{w}$ vanishes because we expand around a local minimum. The

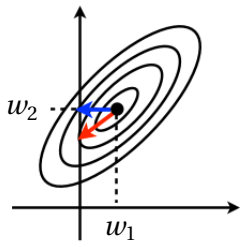


Figure 7.15: Pruning algorithm (schematic). The minimum of H is located at (w_1, w_2). The contours of the quadratic approximation to H are represented as solid black lines. The weight change $\delta \boldsymbol{w} = [-w_1, 0]^\mathsf{T}$ (blue) leads to a smaller increase in H than $\delta \boldsymbol{w} = [0, -w_2]^\mathsf{T}$. The red arrow represents the optimal $\delta \boldsymbol{w}_q^*$ which leads to an even smaller increase in H.

matrix $\mathbb{M}$ is the *Hessian*, the matrix of second derivatives of the energy function.

For the next step it is convenient to adopt the following notation [73]. One groups all weights in the network into a long weight vector $\boldsymbol{w}$ (as opposed to grouping them into a weight matrix $\mathbb{W}$ as we did in Chapter 2). A particular component w_q is extracted from the vector $\boldsymbol{w}$ as follows:

$$w_q = \hat{\mathbf{e}}_q \cdot \boldsymbol{w} \quad \text{where} \quad \hat{\mathbf{e}}_q = \begin{bmatrix} \vdots \\ 1 \\ \vdots \end{bmatrix} \leftarrow q\,. \tag{7.52}$$

Here $\hat{\mathbf{e}}_q$ is the Cartesian unit vector in the direction q, with components $[\hat{\mathbf{e}}_q]_j = \delta_{qj}$. In this notation, the elements of $\mathbb{M}$ are $M_{pq} = \partial^2 H / \partial w_p \partial w_q$.

Now, eliminating the weight w_q amounts to setting

$$\delta w_q = -w_q\,. \tag{7.53}$$

To minimise the damage to the network one wants to eliminate the weight that has least effect upon H, changing the other weights at the same time so that H increases as little as possible (Figure 7.15). This is achieved by minimising

$$\min_q \min_{\delta \boldsymbol{w}} \{\tfrac{1}{2} \delta \boldsymbol{w} \cdot \mathbb{M} \delta \boldsymbol{w}\} \quad \text{subject to the constraint} \quad \hat{\mathbf{e}}_q \cdot \delta \boldsymbol{w} + w_q = 0\,. \tag{7.54}$$

The constant term $H_{\min}$ was dropped because it does not matter. Now we first minimise H w.r.t. $\delta \boldsymbol{w}$, for a given value of q. The linear constraint is incorporated using a *Lagrange multiplier* as in Section 6.3.1 and in Chapter 4, to form the *Lagrangian*

$$\mathscr{L} = \tfrac{1}{2} \delta \boldsymbol{w} \cdot \mathbb{M} \delta \boldsymbol{w} + \lambda (\hat{\mathbf{e}}_q \cdot \delta \boldsymbol{w} + w_q)\,. \tag{7.55}$$

A necessary condition for a minimum $(\delta \boldsymbol{w}, \lambda)$ satisfying the constraint is

$$\frac{\partial \mathscr{L}}{\partial \delta \boldsymbol{w}} = \mathbb{M}\delta \boldsymbol{w} + \lambda \hat{\mathbf{e}}_q = 0 \quad \text{and} \quad \frac{\partial \mathscr{L}}{\partial \lambda} = \hat{\mathbf{e}}_q \cdot \delta \boldsymbol{w} + w_q = 0\,. \tag{7.56}$$

We denote the solution of these Equations by $\delta \boldsymbol{w}^*$ and λ^*. It is obtained by solving the linear system

$$\begin{bmatrix} \mathbb{M} & \hat{\mathbf{e}}_q \\ \hat{\mathbf{e}}_q^\mathsf{T} & 0 \end{bmatrix} \begin{bmatrix} \delta \boldsymbol{w}^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} 0 \\ -w_q \end{bmatrix}. \tag{7.57}$$

If $\mathbb{M}$ is invertible, then the top rows of Eq. (7.57) give

$$\delta \boldsymbol{w}^* = -\mathbb{M}^{-1}\hat{\mathbf{e}}_q \lambda^*\,. \tag{7.58}$$

Inserting this result into $\hat{\mathbf{e}}_q^\mathsf{T}\delta \boldsymbol{w}^* + w_q = 0$ we find

$$\delta \boldsymbol{w}^* = -\mathbb{M}^{-1}\hat{\mathbf{e}}_q\, w_q (\hat{\mathbf{e}}_q^\mathsf{T}\mathbb{M}^{-1}\hat{\mathbf{e}}_q)^{-1} \text{and} \quad \lambda^* = w_q (\hat{\mathbf{e}}_q^\mathsf{T}\mathbb{M}^{-1}\hat{\mathbf{e}}_q)^{-1}\,. \tag{7.59}$$

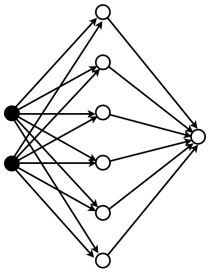
You see that $\hat{\mathbf{e}}_q \cdot \delta \boldsymbol{w}^* = -w_q$, so that the weight w_q is eliminated. The other weights are also changed (red arrow in Figure 7.15). The final step is to find the optimal q by minimising

$$\mathscr{L}(\delta \boldsymbol{w}^*, \lambda^*; q) = \frac{1}{2} w_q^2 (\hat{\mathbf{e}}_q^\mathsf{T}\mathbb{M}^{-1}\hat{\mathbf{e}}_q)^{-1}\,. \tag{7.60}$$

The Hessian of the energy function is expensive to evaluate, and so is the inverse of this matrix. Usually one resorts to an approximate expression for $\mathbb{M}^{-1}$ [73]. One possibility is to set the off-diagonal elements of $\mathbb{M}$ to zero [74]. But in this case the other weights are not adjusted, because $\hat{\mathbf{e}}_{q'} \cdot \delta \boldsymbol{w}_q^* = 0$ for $q' \neq q$, if $\mathbb{M}$ is diagonal. In this case it is necessary to retrain the network after weight elimination.

The algorithm is summarised in Algorithm 4. It succeeds better than elimination by weight decay in removing the unnecessary weights in the network [73]. Weight decay eliminates the smallest weights. One obtains weight elimination of the smallest weights by substituting $\mathbb{M} = \mathbb{I}$ in the algorithm described above [Equation (7.60)]. But small weights are often needed to achieve a small training error. So this is usually not a good approximation.

Pruning is applied with success to deep networks, as illustrated by the following example: Frankle & Carbin [75] used the Boolean XOR function to analyse the effectiveness of pruning. Figure 5.16 shows that the XOR function can be represented by a hidden layer with two neurons. Suitable weights and thresholds are given in this Figure. Frankle & Carbin [75] point out that backpropagation takes a long time to find a valid solution, for random initial weights. They observe that a network with many more neurons in the hidden layer usually learns better. Figure 7.16 lists the fraction of successful trainings for networks with different numbers of neurons in the



n	10	8	6	4	2
Training success					
without pruning	98.5	96.8	92.5	78.3	49.1
pruned network	-	-	-	97.9	83.3

Figure 7.16: Boolean XOR problem. The network has one hidden layer with n ReLU neurons. The output neuron has a sigmoid activation function. The network is trained with stochastic gradient descent for 10 000 iterations. The initial weights were Gaussian random numbers with mean zero, standard deviation 0.1, and max-norm regularisation $|w_{ij}| < 2$. The thresholds were initially zero. Training success was measured in an ensemble of 1000 independent training realisations. Data from Ref. [75].

hidden layer. With two hidden neurons, only 49.1% of the networks learned the task in 10 000 training steps of stochastic gradient descent. Networks with more neurons in the hidden layer ensure better training success. The Figure also shows the training success of pruned networks, that were initially trained with $n = 10$ neurons. Then networks were pruned iteratively during training, removing the neurons with the largest average magnitude. After training, the weights and threshold were reset to their initial values, the values before training began. One can draw three conclusions from this data (from Ref. [75]). First, iterative pruning during training singles out neurons in the hidden layer that had initial weights and thresholds resulting in the correct decision boundaries. Second, the pruned network with two hidden neurons has much better training success than the network that was trained with only two hidden neurons. Third, despite pruning more than 50% of the hidden neurons, the network with $n = 4$ hidden neurons performs almost as well as the one with $n = 10$

Algorithm 4 pruning least important weight

1: train the network to reach $H_{\min}$;
2: compute $\mathbb{M}^{-1}$ approximately;
3: determine q^* as the value of q for which $\mathscr{L}(\delta\boldsymbol{w}^*, \lambda^*; q)$ is minimal;
4: **if** $\mathscr{L}(\delta\boldsymbol{w}^*, \lambda^*; q^*) \ll H_{\min}$ **then**
5: update all weights using $\delta\boldsymbol{w} = -w_{q^*}\mathbb{M}^{-1}\hat{\mathbf{e}}_{q^*}(\hat{\mathbf{e}}_{q^*}^{\mathsf{T}}\mathbb{M}^{-1}\hat{\mathbf{e}}_{q^*})^{-1}$;
6: goto 2;
7: **else**
8: end;
9: **end if**

hidden neurons. When training deep networks it is common to start with many neurons in the hidden layers, and to prune up to 90% of them. This results in small trained networks that can efficiently and reliably classify.

7.7.3 Drop out

In this regularisation scheme some neurons are ignored during training [71]. Usually this regularisation technique is applied to hidden neurons. The procedure is illustrated in Figure 7.17. In each step of the training algorithm (for each mini batch, or for each individual pattern) one ignores at random a fraction q of neurons from each hidden layer, and updates the weights in the remaining, diluted network in the usual fashion. The weights coming into the dropped neurons are not updated, and as a consequence neither are their outputs. For the next step in the training algorithm, the removed neurons at put back, and another set of hidden neurons is removed. Once the training is completed, all hidden neurons are activated, but their outputs are multiplied by q.

Srivastava *et al.* [71] motivate this method by remarking that the performance of machine-learning algorithms is usually improved by combining the results of several learning attempts, for instance by separately training several networks with different layouts, and then to average over their outputs. However, for deep networks this is computationally very expensive. Drop out is an attempt to achieve the same goal more efficiently. The idea is that dropout corresponds to effectively training a large number of different networks. If there are k hidden neurons, then there are 2^k different combinations of neurons that are turned on or off. The hope is that the network learns more robust features of the input data in this way, and that this reduces overfitting. In practice the method is usually applied together with max-norm regularisation (Section 7.7.1).

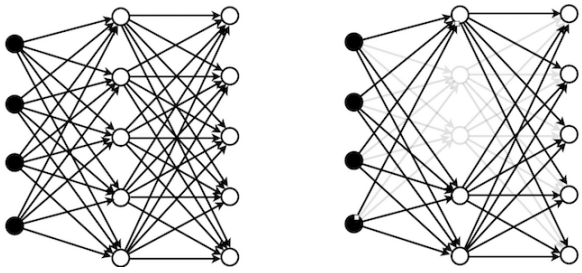


Figure 7.17: Illustrates regularisation by drop out.

7.7.4 Expanding the training set

If one trains a network with a fixed number of hidden neurons on larger training sets, one observes that the network generalises with higher accuracy (better classification success). The reason is that overfitting is reduced when the training set is larger. Thus, a way of avoiding overfitting is to *expand* or *augment* the training set. It is sometimes argued that the recent success of deep neural networks in image recognition and object recognition is in large part due to larger training sets. One example is ImageNet, a database of more than 10^7 hand-classified images, into more than 20 000 categories [76]. Naturally it is expensive to improve training sets in this way. Instead, one can expand a training set *artificially*. For digit recognition (Figure 2.1), for example, one can expand the training set by randomly shifting, rotating, and shearing the digits.

7.7.5 Batch normalisation

Batch normalisation [77] can significantly speed up the training of deep networks with backpropagation. The idea is to shift and normalise the input data for each hidden layer, not only for the input patterns (Section 6.3). This is done separately for each mini batch (Section 6.2), and for each component of the inputs $V_j^{(\mu)}$, $j = 1, \ldots$ (Algorithm 5). One calculates the average and variance over each mini batch

$$\overline{V}_j = \frac{1}{m_B} \sum_{\mu=1}^{m_B} V_j^{(\mu)} \quad \text{and} \quad \sigma_B^2 = \frac{1}{m_B} \sum_{\mu=1}^{m_B} (V_j^{(\mu)} - \overline{V}_j)^2 , \tag{7.61}$$

subtracts the mean from the $V_j^{(\mu)}$, and divides by $\sqrt{\sigma_B^2 + \epsilon}$. The parameter $\epsilon > 0$ is added to the denominator to avoid division by zero. There are two additional parameters in Algorithm 5, γ_j and β_j. They are are learnt by backpropagation, just like the weights and thresholds. In general the new parameters are allowed to differ from layer to layer, $\gamma_j^{(\ell)}$ and $\beta_j^{(\ell)}$.

Batch normalisation was originally motivated by arguing that it reduces possible covariate shifts faced by hidden neurons in layer ℓ: as the parameters of the neurons in the preceding layer $\ell - 1$ change, their outputs shift thus forcing the neurons in layer ℓ to adapt. However in Ref. [78] it was argued that batch normalisation does not reduce the internal covariate shift, but that it speeds up the training by effectively smoothing the energy landscape.

Batch normalisation helps to combat the *vanishing-gradient problem* because it prevents local fields of hidden neurons to grow. This makes it possible to use sigmoid functions in deep networks, because the distribution of inputs remains normalised. It is sometimes argued that batch normalisation has a regularising

effect, and it has been suggested [77] that batch normalisation can replace drop out (Section 7.7).

It is also argued that batch normalisation may help the network to generalise better, in particular if each mini batch contains randomly picked inputs. Then batch normalisation corresponds to randomly transforming the inputs to each hidden neuron (by the randomly changing means and variances). This may help to make the learning more robust. There is no theory that proves either of these claims, but it is an empirical fact that batch normalisation often speeds up the training.

Algorithm 5 batch normalisation

1: **for** $j = 1, \ldots$ **do**
2: calculate mean $\overline{V}_j \leftarrow \frac{1}{m_B} \sum_{\mu=1}^{m_B} V_j^{(\mu)}$
3: calculate variance $\sigma_B^2 \leftarrow \frac{1}{m_B} \sum_{\mu=1}^{m_B} (V_j^{(\mu)} - \overline{V}_j)^2$
4: normalise $\hat{V}_j^{(\mu)} \leftarrow (V_j^{(\mu)} - \overline{V}_j)/\sqrt{\sigma_B^2 + \epsilon}$
5: calculate outputs as: $g(\gamma_j \hat{V}_j^{(\mu)} + \beta_j)$
6: **end for**

7.8 Summary

Neural nets with many layers of hidden neurons are called deep networks. Error backpropagation in deep networks suffers from the vanishing-gradient problem. It can be reduced by using ReLU units, by initialising the weights in certain ways, and by networks with connections that skip layers. Yet vanishing or exploding gradients remain a fundamental difficulty, slowing learning down in the initial phase of training. Brute force (computer power) helps to alleviate the problem. As a consequence, convolutional neural networks have become immensely successful in object recognition, outperforming other algorithms significantly.

Since deep networks contain many free parameters, deep networks tend to overfit the training data. There are different ways to regularise the problem, for example weight decay, drop out, pruning, and data-set augmentation, Section 7.7).

7.9 Further reading

Deep networks suffer from *catastrophic forgetting*: when you train a network on a new input distribution that is quite different from the one the network was originally trained on, then the network tends to forget what it learned initially. If you want to read more, good starting point is Ref. [79].

The stochastic-gradient descent algorithm (with or without minibatches) samples the input-data distribution uniformly randomly. As mentioned in Section 6.3, it may be advantageous to sample those inputs more frequently that initially cause larger output errors. More generally, the algorithm may use other criteria to choose certain input data more often, with the goal to speed up learning. It may even suggest how to augment a given training set most efficiently, by asking to specifically label certain types of input data (*active learning*) [80].

Another question concerns the structure of the energy landscape for multilayer perceptrons. It seems that local minima are perhaps less important for deep networks, because their energy functions tend to have more saddle points than minima [81].

7.10 Exercises

7.1 Pruning. Show that the expression (7.59) for the weight increment $\delta \boldsymbol{w}^*$ minimises the Lagrangian (7.55) subject to the constraint (7.53).

7.2 Decision boundaries for XOR problem. Figure 7.6 shows the layout of a network that solves the Boolean XOR problem. Draw the decision boundaries for the four hidden neurons in the input plane, and label the boundaries and the regions as in Figure 5.14.

7.3 Vanishing-gradient problem. Train the network shown in Figure 7.9 on the iris data set, available from the Machine learning repository of the University of California Irvine. Measure the effects upon of the neurons in the different layers, by calculating the derivative of the energy function H w.r.t. the thresholds of the neurons in question.

7.4 Residual network. Derive Equation (7.30) for the error $\delta^{(\ell)}$ in layer ℓ of the residual 'network' shown in Figure 7.14.

7.5 Cross-entropy function. The cross-entropy function (7.40) is an energy function for sigmoid output neurons. Write down a cross-entropy function for tanh-output units, and show that it has a global minimum at $\boldsymbol{O}^{(\mu)} = \boldsymbol{t}^{(\mu)}$, where the cross-

entropy function takes the value zero.

7.6 Softmax outputs. Consider a perceptron with L layers and softmax output units. For pattern μ, the state of the output neuron i is given by

$$O_i^{(\mu)} = \frac{e^{\,b_i^{(L,\mu)}}}{\sum_j e^{\,b_j^{(L,\mu)}}}, \tag{7.62}$$

where $b_j^{(L,\mu)}$ denotes the *local field* $b_j^{(L,\mu)} = -\theta_j^{(L)} + \sum_k w_{jk}^{(L)} V_k^{(L-1,\mu)}$. Here $\theta_j^{(L)}$ and $w_{jk}^{(L)}$ are thresholds and weights, and $V_k^{(L-1,\mu)}$ is the state of the k^{th} neuron in layer $L-1$, evaluated for pattern μ. Compute the derivative of output $O_i^{(\mu)}$ with respect to the local field $b_m^{(L,\mu)}$ of output neuron m. The network is trained by gradient descent on the negative log-likelihood function,

$$H = -\sum_{i\mu} t_i^{(\mu)} \log\bigl(O_i^{(\mu)}\bigr). \tag{7.63}$$

The targets $t_i^{(\mu)}$ satisfy the constraint $\sum_i t_i^{(\mu)} = 1$, for all patterns μ. When updating, the increment of a weight $w_{mn}^{(\ell)}$ in layer ℓ is given by

$$\delta w_{mn}^{(\ell)} = -\eta \frac{\partial H}{\partial w_{mn}^{(\ell)}}, \tag{7.64}$$

where η denotes the learning rate. Derive the learning rule $\delta w_{mn}^{(L)}$ for weight $w_{mn}^{(L)}$ in layer L.

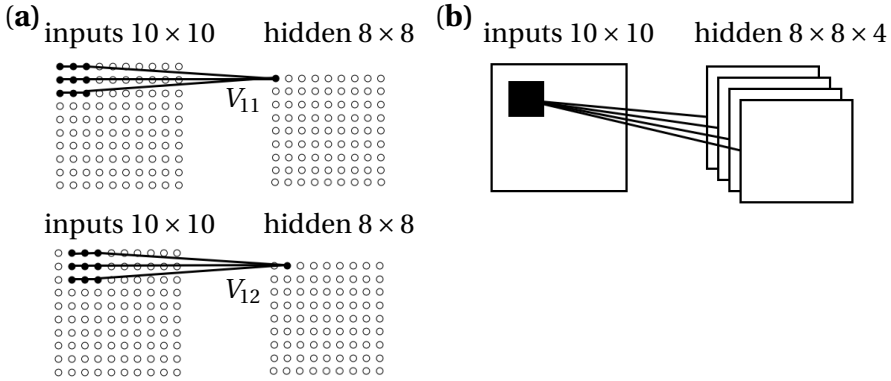


Figure 8.1: (**a**) Feature map, kernel, and receptive field (schematic). A feature map (the 8×8 array of hidden neurons) is obtained by translating a kernel (filter), here with a 3×3 receptive field, over the input image, here a 10×10 array of pixels. (**b**) A convolution layer consists of a number of feature maps, each corresponding to a given kernel that detects a certain feature in parts of the input image.

8 Convolutional networks

Convolutional networks have been around since the 1980's. They became widely used after Krizhevsky *et al.* [82] won the ImageNet challenge (Section 8.5) with a convolutional net. One reason for the recent success of convolutional networks is that they have fewer neurons. This has two advantages. Firstly, such networks are obviously cheaper to train. Secondly, as pointed out above, reducing the number of neurons regularises the network, it reduces the risk of overfitting.

Convolutional neural networks are designed for object recognition and pattern detection. They take images as inputs (Figure 7.1), not just a list of attributes (Figure 5.1). Convolutional networks have important properties in common with networks of neurons in the visual cortex of the Human brain [4]. First, there is a spatial array of input terminals. For image analysis this is the two-dimensional array of bits shown in Figure 8.1(**a**). Second, neurons are designed to detect local features of the image (such as edges or corners for instance). The maps learned by such neurons, from inputs to output, are referred to as *feature maps*. Since these features occur in different parts of the image, one uses the same *kernel* (or *filter*) for different parts of the image, always with the same weights and thresholds for different parts of the image. Since these kernels are local, and since they act in a translational-invariant way, the number of neurons from the two-dimensional input array is greatly reduced, compared with fully connected networks. Feature maps are obtained by convolution of the kernel with the input image. Therefore, layers consisting of a number of feature maps corresponding to different kernels are also referred to as *convolution layers*,

Figure 8.1(**b**).

Convolutional networks can have hierarchies of convolution layers. The idea is that the additional layers can learn more abstract features. Apart from feature maps, convolutional networks contain other types of layers. *Pooling layers* connect directly to the convolution layer(s), their task is to simplify the output of the convolution layers. Connected to the pooling layers, convolutional networks may also contain several fully connected layers.

8.1 Convolution layers

Figure 8.1(**a**) illustrates how a feature map is obtained by convolution of the input image with a kernel which reads a 3×3 part of the input image. In analogy with the terminology used in neuroscience, this area is called the *local receptive field* of the kernel. The outputs of the kernel from different parts of the input image make up the feature map, here an 8×8 array of hidden neurons: neuron V_{11} connects to the 3×3 area in the upper left-hand corner of the input image. Neuron V_{12} connects to a shifted area, as illustrated in Figure Figure 8.1(**a**), and so forth. Since the input has 10×10 pixels, the dimension of the feature map is 8×8 in this example. The important point is that the neurons V_{11} and V_{12}, and all other neurons in this convolution layer, share their weights and the threshold. In the example shown in Figure 8.1(**a**) there are thus only 9 independent weights, and one threshold. Since the different hidden neurons share weights and thresholds, their computation rule takes the form of a discrete *convolution*:

$$V_{ij} = g\Big(\sum_{p=1}^{3}\sum_{q=1}^{3} w_{pq}\, x_{p+i-1,q+j-1} - \theta\Big). \tag{8.1}$$

In Figure 8.1(**a**) the local receptive field is shifted by one pixel at a time. Sometimes it is useful to use a different *stride* (s_1, s_2), to shift the receptive field by s_1 pixels horizontally and by s_2 pixels vertically. Also, the local receptive regions need not have size 3×3. If we assume that their size is $Q \times P$, and that $s_1 = s_2 = s$, the rule (8.1) takes the form

$$V_{ij} = g\Big(\sum_{p=1}^{P}\sum_{q=1}^{Q} w_{pq}\, x_{p+s(i-1),q+s(j-1)} - \theta\Big). \tag{8.2}$$

Figure 8.1(**a**) depicts a two-dimensional input array. For colour images there are usually three colour *channels*, in this case the input array is three-dimensional, and the input bits are labeled by three indices: two for position and the last one for colour, x_{pqr}. Usually one connects several feature maps (corresponding to different

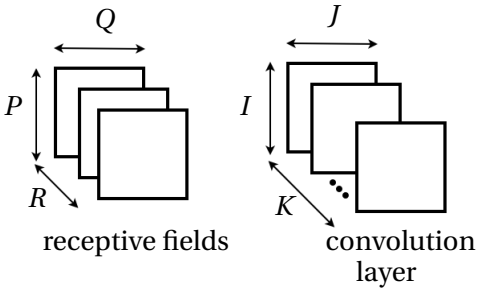


Figure 8.2: Illustration for Equation (8.3). Each feature map as a receptive field of dimension $P \times Q \times R$. There are K feature maps, each of dimension $I \times J$.

kernels) to the input layer, as shown in Figure 8.1(**b**). The different kernels detect different features of the input image, one detects edges for example, and another one detects corners, and so forth. To account for these extra dimensions, one groups weights (and thresholds) into higher-dimensional arrays (*tensors*). The convolution takes the form:

$$V_{ijk} = g\Big(\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} w_{pqrk}\, x_{p+s(i-1),q+s(j-1),r} - \theta_k\Big) \tag{8.3}$$

(see Figure 8.2). All neurons in a given convolution layer have the same threshold. The software package *TensorFlow* [83] is designed to efficiently perform tensor operations as in Equation (8.3).

If one couples several convolution layers together, the number of neurons in these layers decreases rapidly as one moves to the right. In this case one can *pad* the image (and the convolution layers) by adding rows and columns of bits set to zero. In Figure 8.1(**a**), for example, one obtains a convolution layer of the same dimension as the original image by adding adding one column each on the left-hand and right-hand sides of the image, as well as two rows, one at the bottom and one at the top. In general, the numbers of rows and columns need not be equal, so the amount of padding is specified by four numbers, (p_1, p_2, p_3, p_4).

Convolution layers are trained with backpropagation. Consider the simplest case, Equation (8.1). As usual, we use the chain rule to evaluate the gradients:

$$\frac{\partial V_{ij}}{\partial w_{mn}} = g'(b_{ij})\frac{\partial b_{ij}}{\partial w_{mn}} \tag{8.4}$$

with local field $b_{ij} = \sum_{pq} w_{pq}\, x_{p+i-1,q+j-1} - \theta$. The derivative of b_{ij} is evaluated by applying rule (6.9):

$$\frac{\partial b_{ij}}{\partial w_{mn}} = \sum_{pq} \delta_{mp}\delta_{nq}\, x_{p+i-1,q+j-1} \tag{8.5}$$

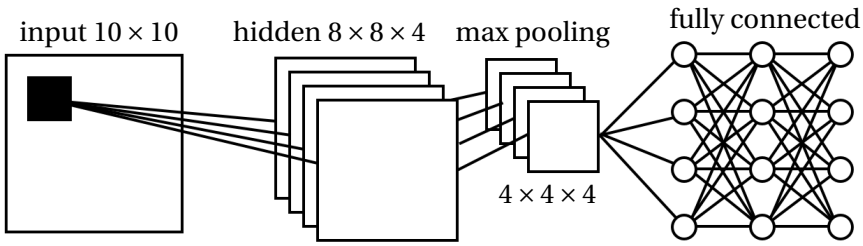


Figure 8.3: Layout of a convolutional neural network for object recognition and image classification. The inputs are stored in a 10×10 array. They feed into a convolution layer with four different feature maps, with 3×3 kernels, stride $(1,1)$, and zero padding. Each convolution layer feeds into its own max-pooling layer, with stride $(2,2)$ and zero padding. Between these and the output layer are a couple of fully connected hidden layers.

In this way one can train several stacked convolution layers too. It is important to keep track of the summation boundaries. To that end it helps to pad out the image and the convolution layers, so that the upper bounds remain the same in different layers.

Details aside, the fundamental principle of feature maps is that the map is applied in the same form to different parts of the image (*translational invariance*). In this way the learning of parameters is shared between pixels, each weight in a given feature map is trained on different parts of the image. This effectively increases the training set for the feature map and combats overfitting.

8.2 Pooling layers

Pooling layers process the output of convolution layers. A neuron in a pooling layer takes the outputs of several neighbouring feature maps and summarises their outputs into a single number. *Max-pooling units*, for example, summarise the outputs of nearby feature maps (in a 2×2 square for instance) by taking the maximum over the feature-map outputs. Instead, one may compute the root-mean square of the map values (L_2-*pooling*). There are no weights or thresholds associated with the pooling layers, they compute the output from the inputs using a pre-defined prescription. Other ways of pooling are discussed in Ref. [4]. Just as for convolution layers, we need to specify stride and padding for pooling layers.

Usually several feature maps are connected to the input. Pooling is performed separately on each of them. The network layout looks like the one shown schematically in Figure 8.3. In this Figure, the pooling layers feed into a number of fully connected hidden layers that connect to the output neurons. There are as many output neurons as there are classes to be recognised. This layout is qualitatively

Figure 8.4: Examples of digits from the MNIST data set of handwritten digits [84]. The images were produced using MATLAB. But note that by default MATLAB displays the digits white one black background. Copyright for the data set: Y. LeCun and C. Cortes.

similar to the layout used by Krizhesvky *et al.* [82] in the ImageNet challenge (see Section 8.5 below).

8.3 Learning to read handwritten digits

Figure 8.4 shows patterns from the MNIST data set of handwritten digits [84]. The data set derives from a data set compiled by the National Institute of Standards and Technology (NIST), of digits handwritten by high-school students and employees of the United States Census Bureau. The data contains a data set of 60 000 images of digits with 28×28 pixels, and a *test set* of 10 000 digits. The images are grayscale with 8-bit resolution, so each pixel contains a value ranging from 0 to 255. The images in the database were preprocessed. The procedure is described on the MNIST home page. Each original binary image from the National Institute of Standards and Technology was represented as a 20×20 gray-scale image, preserving the aspect ratio of the digit. The resulting image was placed in a 28×28 image so that the centre-of-mass of the image coincided with its geometrical centre. These steps can make a crucial difference (Section 8.4).

The goal of this Section is to show how the principles described in Chapters 6, 7, and in this Chapter allow to learn the MNIST data with low classification error, as outlined in Ref. [5].

One divides the data set into a *training set* with 50 000 digits and a *validation set* with 10 000 digits. The latter is used for cross-validation. The test data is used for measuring the classification error after training. For this purpose one must use a data set that was not involved in the training. As described in Section 6.3, the inputs are preprocessed further by by subtracting the mean image averaged over the whole training set from each input image [Equation (6.23)].

To find appropriate parameter values and network layouts is one of the main difficulties when training a neural network, and it usually requires a fair deal of experimenting. There are recipes for finding certain parameters [85], but the general

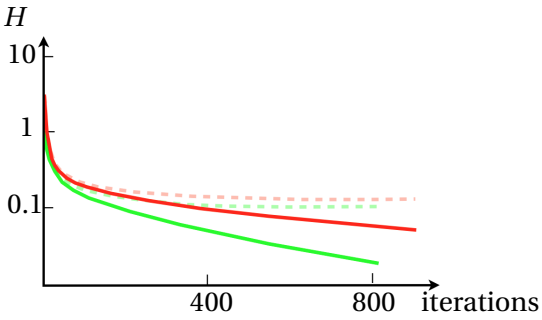


Figure 8.5: Energy functions for the MNIST training set (solid lines) and for the validation set (dashed lines) for a fully connected hidden layer with 30 neurons (red lines) and for a similar algorithm, but with 100 neurons in the hidden layer, green lines. The data was smoothed and the plot is schematic. The x-axis shows iterations. One iteration corresponds to feeding one minibatch of patterns. One epoch consists of $50000/8192 \approx 6$ iterations. Schematic, based on simulations performed by Oleksandr Balabanov.

approach is still trial and error [5]. Consider first a network with one hidden layer with ReLU activation functions (Section 7.3), and a softmax output layer (Section 7.5) with ten outputs O_i and energy function (7.34), so that output O_i is the probability that the pattern fed to the network falls into category i. The networks are trained with stochastic gradient descent with momentum, Equation (6.32). The learning rate is set to $\eta = 0.001$, and the momentum constant to $\alpha = 0.9$. The mini-batch size [Equation (6.20)] equals 8912. Cross validation and early stopping is implemented as follows: during training, the algorithm keeps track of the smallest validation error observed so far. Training stops when the validation error was larger than the minimum for a specified number of times, equal to 5 in this case.

Figure 8.5 shows how the training and the validation energies decrease during training, for nets with 30 and 100 hidden units. One epoch corresponds to applying p patterns or $p/m_{\mathrm{B}} = 50000/8192$ iterations (Section 6.1). The energies are a little lower for the network with 100 hidden neurons. But one observes overfitting in

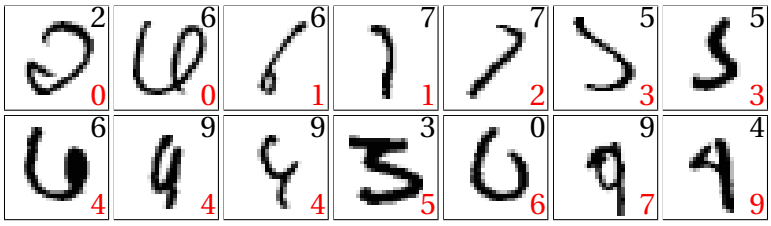


Figure 8.6: Some hand-written digits from the MNIST test set, misclassified by a convolutional net that achieved an overall classification accuracy of 98%. Correct classification (top right), misclassification (bottom right). Data from Oleksandr Balabanov, reproduced with permission

both cases: after many training steps the validation energy is much higher than the training energy. Early stopping caused the training of the larger network to abort after 135 epochs, this corresponds to 824 iterations. The resulting classification accuracy is about 97.2% for the net with 100 hidden neurons.

It is difficult to increase the increase the classification accuracy by adding more hidden layers, most likely because the network overfits the data (Section 6.4). This problem becomes more acute as you add more hidden neurons. The tendency of the network to overfit is reduced by regularisation (Section 7.7). For the network with one hidden layer with 100 ReLU units, L_2-regularisation improves the classification accuracy to almost 98%.

Deep convolutional networks can be optimised to yield higher classification accuracies than those quoted above. A convolutional net with one convolution layer with 20 feature maps, a max-pooling layer, and a fully connected hidden layer with 100 ReLU units, similar to the network shown in Figure 8.3, gives classification accuracy only slightly above 98%, after training for 60 epochs. Adding a second convolution layer and batch normalisation (Section 7.7.5) gives a classification accuracy is 98.99% after 30 epochs (this layout is similar to te one from MathWorks. The accuracy can be improved further by tuning parameters and network layout, and by using ensembles of convolutional neural networks [84]. The best classification accuracy found in this way is 99.77% [86]. Several of the MNIST digits are difficult to classify for Humans too (Figure 8.6), so we conclude that convolutional nets really work very well. Yet the above examples show also that it takes much experimenting to find the right parameters and network layout as well as long training times to reach the best classification accuracies. It could be argued that one reaches a stage of *diminishing returns* as the classification error falls below a few percent.

8.4 Coping with deformations of the input distribution

How well does a MNIST-trained convolutional net classify your own hand-written digits? Suppose you create your own data set by drawing the digits with GoodNotes or a similar program. Preprocess the digits in the same way as the MNIST data. Figure 8.7 shows digits obtained in this way.

Using a MNIST-trained convolutional net on these digits yields a classification accuracy of about 90%, substantially lower than the classification errors quoted in the previous Section. What is going on? Compare Figures 8.4 and 8.7. The digits in Figure 8.7 have a more slender stroke. It was suggested in Ref. [87] the line thickness of hand-written text can make a difference for algorithms that read hand-written text [88]. There are different methods for normalising the line thickness of hand-written text. Applying the method proposed in Ref. [88] to our digits results in Figure

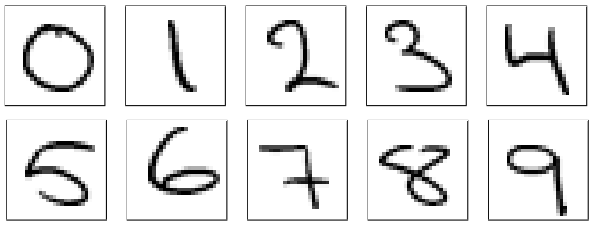


Figure 8.7: Non-MNIST hand-written digits, preprocessed like the MNIST digits. Data from Oleksandr Balabanov, reproduced with permission.

8.8. The algorithm of Ref. [88] has a free parameter, T, that specifies the resulting line thickness. In Figure 8.8 it was taken to be $T = 10$, close to the line thickness of the MNIST digits, we measured the latter to $T \approx 9.7$ using the method described in Ref. [88]. If we run a MNIST-trained convolutional net on a data set of 60 digits with normalised line thickness, it fails on only two digits. This corresponds to a classification accuracy of roughly 97%, not so bad – but not as good as the best results in Section 8.3. Note that we can only make a rough comparison. In order to to obtain a better estimate of the classification accuracy we need to test many more than 60 digits. A question is of course whether there are perhaps other differences between our own hand-written digits and those in the MNIST data. It would also be of interest to try digits that were drawn using *Paint*, or a similar program. How does do MNIST-trained convolutional nets perform on computer-drawn digits?

At any rate, the results of this Section show that the way the input data are processed can make a big difference. This raises a point of fundamental importance. We have seen that convolutional nets can be trained to represent a distribution of input patterns with very high accuracy. But if you test the network on a data set that has a slightly different distribution, perhaps because it was preprocessed differently, the network may not work as well.

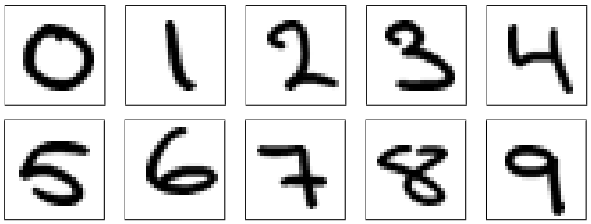


Figure 8.8: Same digits as in Figure 8.7. The difference is that the thickness of the stroke was normalised (see text). Data from Oleksandr Balabanov, reproduced with permission.

Figure 8.9: Object recognition using a deep convolutional network. Shown is a frame from a movie recorded on a telephone. The network was trained on the Pascal VOC data set [89] using YOLO [90]. Details on how to obtain the weights and how to install the software are given on the YOLO website.

8.5 Deep learning for object recognition

Deep learning has become so popular in the last few years because deep convolutional networks are good at recognising objects in images. Figure 8.9 shows a frame from a movie taken from a car with my mobile telephone. A deep convolutional network trained on the Pascal VOC training set [89] recognises objects in the movie by putting bounding boxes around the objects and classifying them. The Pascal VOC data set is a training set for object-class recognition in images. It contains circa 20 000 images, each annotated with one of 20 classes. The people behind this data set ran image classification challenges from 2005 to 2012. A more recent challenge is the *ImageNet large-scale visual recognition challenge* (ILSVRC) [91], a competition for image classification and object recognition using the ImageNet database [76]. The challenge is based on a subset of ImageNet. The training set contains more than 10^6 images manually classified into one of 1000 classes. There are approximately 1000 images for each class. The validation set contains 50 000 images.

The ILSVRC challenge consists of several tasks. One task is *image classification*, to list the object classes found in the image. A common measure for accuracy is the so-called *top-5 error* for this classification task. The algorithm lists the five object classes it identified with highest probabilities. The result is considered correct if the annotated class is among these five. The error equals the fraction of incorrectly classified images. Why does one not simply judge whether the most probable class is the correct one? The reason is that the images in the ImageNet database are annotated by a single-class identifier. Often this is not unique. The image in Figure

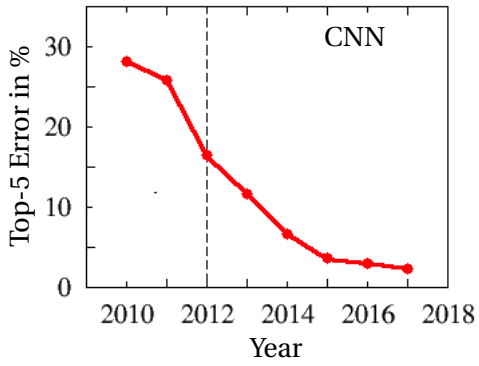


Figure 8.10: Smallest classification error for the ImageNet challenge [91]. The data up to 2014 comes from Ref. [91]. The data for 2015 comes from Ref. [68], for 2016 from Ref. [92], and for 2017 from Ref. [93]. From 2012 onwards the smallest error was achieved by convolutional neural networks (CNN). After Fig. 1.12 in Goodfellow *et al.* [4].

8.4, for example, shows not only a car but also trees, yet the image is annotated with the class label *car*. This is ambiguous. The ambiguity is significantly smaller if one considers the top five classes the algorithm gives, and checks whether the annotated class is among them.

The tasks in the ILSVRC challenge are significantly more difficult than the digit recognition described in Section 8.3, and also more difficult than the VOC challenges. One reason is that the ImageNet classes are organised into a deep hierarchy of subclasses. This results in highly specific sub classes that can be very difficult to distinguish. The algorithm must be very sensitive to small differences between similar sub classes. We say that the algorithm must have high *inter-class variability* [94]. Different images in the same sub class, on the other hand, may look quite different. The algorithm should nevertheless recognise them as similar, belonging to the same class, the algorithm should have small *intra-class variability* [94].

Since 2012, algorithms based on deep convolutional networks won the ILSVRC challenge. Figure 8.10 shows that the error has significantly decreased until 2017, the last year of the challenge in the form described above. We saw in previous Sections that deep networks are difficult to train. So how can these algorithms work so well? It is generally argued that the recent success of deep convolutional networks is mainly due to three factors.

First, there are now much larger and better annotated training sets available. ImageNet is an example. Excellent training data is now recognised as one of the most important factors, and companies developing software for self-driving cars and systems that help to avoid accidents recognise that good training sets is one of the most important factors, and difficult to achieve: to obtain reliable training data one must *manually* collect and annotate the data (Figure 8.11). This is costly,

Figure 8.11: Reproduced from xkcd.com/1897 under the creative commons attribution-noncommercial 2.5 license.

but at the same time it is important to have as large data sets as possible, to reduce overfitting. In addition one must aim for a large variability in the collected data.

Second, the hardware is much better today. Deep networks are nowadays implemented on single or multiple GPUs. There are also dedicated chips, such as the tensor processing unit [95].

Third, improved regularisation techniques (Section 7.7) and weight sharing in convolution layers help to fight overfitting, and ReLU units (Section 7.3) render the networks less susceptible to the vanishing-gradient problem (Section 7.2).

The winning algorithm for 2012 was based on a network with five convolution layers and three fully connected layers, using drop out, ReLU units, and data-set augmentation [82]. The algorithm was implemented on GPU processors. The 2013 ILSVRC challenge was also won by a convolutional network [96], with 22 layers. Nevertheless, the network has substantially fewer free parameters (weights and thresholds) than the 2012 network: 4×10^6 instead of 60×10^6. In 2015, the winning algorithm [68] had 152 layers. One significant new element in the layout were connections that skip layers (*residual networks*, Section 7.4). The 2016 [97] and 2017 [93] winning algorithms used ensembles of convolutional networks.

8.6 Summary

Convolutional nets can be trained to recognise objects in images with high accuracy. It is sometimes stated that convolutional networks are now *better than Humans*, in that they recognise objects with lower classification errors than Humans [98]. This statement is problematic for several reasons. To start with, the article refers to the 2015 ILSVRC competition, and the company mentioned in the Guardian article was later caught out cheating. At any rate, this and similar statements refer to an experiment showing that the Human classification error in recognising objects in

the ImageNet database is about 5.1% [99], worse than the most recent convolutional neural-network algorithms (Figure 8.10).

We have seen that training deep networks suffers from overfitting[**?**]. In this regard convolutional nets have an advantage over fully connected networks because they have fewer weights, and the weights of a given feature map are trained on different parts of the input images, effectively increasing the training set.

It is clear that these algorithms learn in quite a different way from Humans. They can detect local features, but since these convolutional networks rely on translational invariance, they do not easily understand global features, and can mistake a leopard-patterned sofa for a leopard [100]. It may help to include more sofas in the training data set, but the essential difficulty remains: translational invariance imposes constraints on what convolutional networks can learn [100]. More fundamentally one may argue that Humans learn differently, by abstraction instead of going through vast training sets. Just try it out for yourself, this website [101] allows you to learn like a convolutional network. Nevertheless, the examples described in this Chapter illustrate the tremendous success of deep convolutional networks.

The examples described in Section 8.3 show that convolutional nets are sensitive to differences in how the input data are preprocessed. You may run into problems if you train a network on given training and validation sets, but apply it to a test set that was preprocessed in a different way – so that the test set corresponds to a different input distribution. Convolutional nets excel at learning the properties of a given input distribution, but they may have difficulties in recognising patterns sampled from a slightly different distribution, even if the two distributions appear very similar to the Human eye. Note also that this problem cannot be solved by cross-validation, because training and validation sets are drawn from the same input distribution, but here we are concerned with what happens when the network is applied to a input distribution different from the one that was trained on. Here is another example illustrating this point: the authors of Ref. [102] trained a convolutional network on perturbed grayscale images from the ImageNet data base, adding a little bit of noise independently to each pixel (*white noise*) before training. This network failed to recognise images that were weakly perturbed in a different way, by setting a small number of pixels to white or black. When we look at the images we have no difficulties seeing through the noise.

Refs. [103, 104] illustrate intriguing failures of convolutional networks. Szegedy *et al.* [103] demonstrate that the way convolutional nets partition input space can lead to surprising results. The authors took an image that the network classifies correctly with high confidence, and it perturbed slightly. The difference between the original and perturbed images (*adversarial images*) is undetectable to the Human eye, yet the network misclassifies the perturbed image with high confidence [103]. This indicates that decision boundaries are always close in input space, not intuitive but

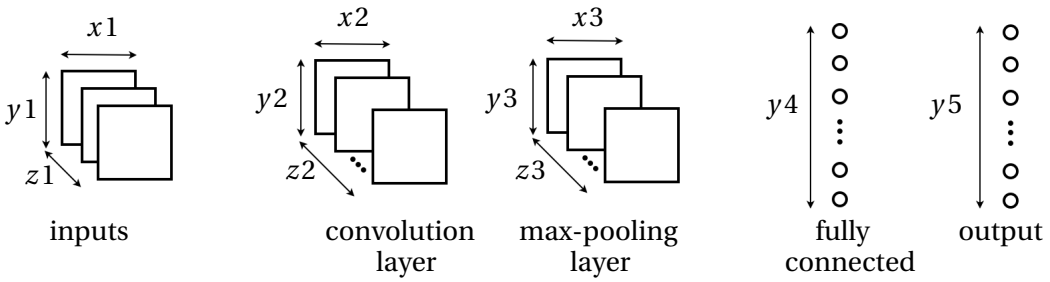


Figure 8.12: Layout of convolutional net for Exercise 8.1.

possible in high dimensions. Figure 1 in Ref. [104] shows images that are completely unrecognisable to the Human eye. Yet a convolutional network classifies these images with high confidence. This illustrates that there is no telling what a network may do if the input is far away from the training distribution. Unfortunately the network can sometimes be highly confident yet wrong.

To conclude, convolutional networks are very good at recognising objects in images. But we should not imagine that they *understand* what they see in the same way as Humans. The theory of deep learning has somewhat lagged behind the performance in practice. But some progress has been made in recent years, and there are many interesting open questions.

8.7 Further reading

What do the hidden layers in a convolutional layer actually compute? Feature maps that are directly coupled to the inputs detect local features, such as edges or corners. Yet it is unclear precisely how hidden convolutional layers help the network to learn. Therefore it is interesting to visualise the activity of deep layers by asking: which input patterns maximise the outputs of the neurons in a certain layer [105]?

8.8 Exercises

8.1 Number of parameters of a convolutional net. A convolutional net has the following layout (Figure 8.12): an input layer of size $21 \times 21 \times 3$, a convolutional layer with ReLU activations with 16 kernels with local receptive fields of size 2×2, stride $(1,1)$, and padding $(0,0,0,0)$, a max-pooling layer with local receptive field of size 2×2, stride $= (2,2)$, padding $= (0,0,0,0)$, a fully connected layer with 20 neurons with sigmoid activations, and a fully connected output layer with 10 neurons. In one or two sentences, explain the function of each of the layers. Enter the values of the parameters $x1, y1, z1, x2, \ldots, y5$ into Figure 8.12 and determine the number of

trainable parameters (weights and thresholds) for the connections into each layer of the network.

8.2 Parity function. The parity function outputs 1 if and only if the input sequence of N binary numbers has an odd number of ones, and zero otherwise. The parity function for $N = 2$ is the XOR function. The XOR function can be represented using a multilayer perceptron with two inputs, a fully connected hidden layer with two hidden neurons, and one output unit. Determine suitable weights w_{jk} and thresholds θ_j for the two hidden units ($j = 1, 2$), as well as the weights W_j and the threshold Θ for the output unit. Draw the corresponding decision boundaries in the input plane. Also draw the problem in the V_1-V_2 plane, with the decision boundary corresponding to the output neuron. Describe how to combine several such XOR units to represent the parity function for $N > 2$. Explain how the total number of neurons in the network grows with the input dimension N, for $N = 2^k$.

8.3 Convolutional neural net. Figure 8.3 shows a schematic layout of a convolutional net. Explain how a *convolution layer* works. In your discussion, refer to the terms *convolution*, *colour channel*, *receptive field*, *feature map*, *stride*, and explain the meaning of the parameters in the computation rule

$$V_{ij} = g\Big(\sum_{p=1}^{P}\sum_{q=1}^{Q} w_{pq}\, x_{p+s(i-1),q+s(j-1)} - \theta\Big). \tag{8.6}$$

Explain how a *pooling layer* works, and why it is useful.

8.4 Feature map. The two patterns shown in Figure 8.13(a) are processed by a very simple convolutional network that has one convolution layer with one single 3×3 kernel with ReLU units, zero threshold, weights as given in Figure 8.13(b), and stride (1,1). The resulting feature map is fed into a 3×3 max-pooling layer with stride (1,1). Finally there is a fully connected classification layer with two output units with Heaviside activation functions.
For both patterns determine the resulting feature map and the output of the max-pooling layer. Determine weights and thresholds of the classification layer that allow to classify the two patterns into different classes.

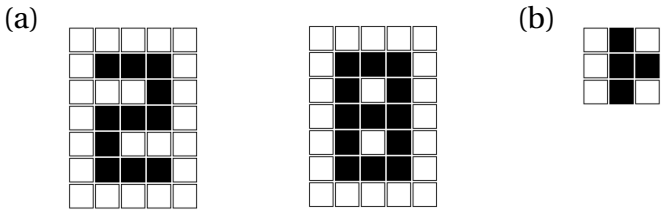


Figure 8.13: (a) Input patterns with 0/1 bits (□ corresponds to $x_i = 0$ and ■ to $x_i = 1$). (b) 3×3 kernel of a feature map. ReLU units, zero threshold, weights either 0 or 1 (□ corresponds to $w = 0$ and ■ to $w = 1$). (Exercise 8.4).

9 Supervised recurrent networks

The layout of the perceptrons analysed in the previous Chapters is special. All connections are one way, and only to the layer immediately to the right, so that the update rule for the i-th neuron in layer ℓ becomes

$$V_i^{(\ell)} = g\Big(\sum_j w_{ij}^{(\ell)} V_j^{(\ell-1)} - \theta_i^{(\ell)}\Big). \tag{9.1}$$

The backpropagation algorithm relies on this *feed-forward* layout. It means that the derivatives $\partial V_j^{(\ell-1)}/\partial w_{mn}^{(\ell)}$ vanish. This ensures that the outputs are nested functions of the inputs, which in turn implies the simple iterative structure of the backpropagation algorithm on page 99.

In some cases it is necessary or convenient to use networks that do not have this simple layout. The Hopfield networks discussed in part I are examples where all connections are symmetric. More general networks may have a feed-forward layout with *feedbacks*, as shown in Figure 9.1. Such networks are called *recurrent networks*. There are many different ways in which the feedbacks can act: from the output layer to hidden neurons for example (Figure 9.1), or there could be connections between the neurons in a given layer. Neurons 3 and 4 in Figure 9.1 are output units, they are associated with targets just as in Chapters 5 to 7. The layout of recurrent networks is very general, but because of the feedback links we must consider how such networks can be trained.

Unlike multi-layer perceptrons, which represent an input-to-output mapping in terms of nested activation functions, recurrent networks are *dynamical networks*, where the iteration index t replaces the layer index ℓ:

$$V_i(t) = g\Big(\sum_j w_{ij}^{(vv)} V_j(t-1) + \sum_k w_{ik}^{(vx)} x_k - \theta_i^{(v)}\Big) \quad \text{for} \quad t = 1, 2, \ldots. \tag{9.2}$$

See Figure 9.1 for a definition of the different weights. Furthermore, the parameters $\theta_i^{(v)}$ are thresholds. Equation (9.2) is analogous to the deterministic McCulloch-Pitts dynamics of Hopfield nets and Boltzmann machines [c.f. Equation (1.5)]. As in the case of Hopfield nets (Exercise 2.10), one may also consider a continuous network dynamics:

$$\tau \frac{\mathrm{d}V_i}{\mathrm{d}t} = -V_i + g\Big(\sum_j w_{ij}^{(vv)} V_j(t) + \sum_k w_{ik}^{(vx)} x_k - \theta_i^{(v)}\Big), \tag{9.3}$$

with time constant τ. We shall see in a moment why it is convenient to assume that the dynamics is continuous in t.

Recurrent networks can learn in different ways. One possibility is to use a training set of pairs $(\boldsymbol{x}^{(\mu)}, \boldsymbol{y}^{(\mu)})$ with $\mu = 1, \ldots, p$. To avoid confusion with the iteration index t,

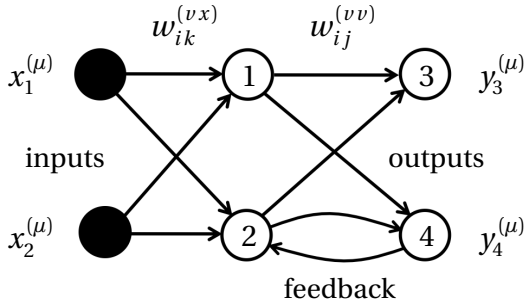


Figure 9.1: Network with a feedback connection. Neurons 1 and 2 are hidden neurons. The weights from the input x_k to the neurons V_i are denoted by $w_{ik}^{(vx)}$, the weight from neuron V_j to neuron V_i is $w_{ij}^{(vv)}$. Neurons 3 and 4 are output neurons, with prescribed target values y_i. To avoid confusion with the iteration index t, the targets are denoted by y in this Chapter.

the targets are denoted by y in this Chapter. One feeds a pattern from this set and runs the dynamics (9.2) or (9.3) for the given $\boldsymbol{x}^{(\mu)}$ until it reaches a steady state $\boldsymbol{V}^*$ (if this does not happen, the training fails). Then one updates the weights by gradient descent using the energy function

$$H = \frac{1}{2}\sum_k E_k^2 \quad \text{where } E_k = \begin{cases} y_k - V_k & \text{if } V_k \text{ is an output unit,} \\ 0 & \text{otherwise,} \end{cases} \tag{9.4}$$

evaluated at $\boldsymbol{V} = \boldsymbol{V}^*$, that is $H^* = \frac{1}{2}\sum_k (E_k^*)^2$ with $E_k^* = y_k - V_k^*$. Instead of defining the energy function in terms of the mean-squared output errors, one could also use the negative log-likelihood function (7.40). Then one feeds another pattern $\boldsymbol{x}^{(\{mu)}$, finds the steady state $\boldsymbol{V}^*$, and updates the weights. These steps are repeated until the steady-state outputs yield the correct targets for all input patterns. This is reminiscent of the algorithms discussed in Chapters 5 to 7, and we shall see that the backpropagation algorithm can be modified (*recurrent backpropagation*) to make the recurrent net learn in an analogous fashion.

Another possibility is that inputs and targets change as functions of time t while the network dynamics runs. In this way the network can solve *temporal association tasks* where it learns to output certain targets in response to the sequence $\boldsymbol{x}(t)$ of input patterns, and targets $\boldsymbol{y}(t)$. In this way recurrent networks can translate written text or recognise speech. Such networks can be trained by unfolding their dynamics in time as explained in Section 9.2 (*backpropagation in time*), although this algorithm suffers from the vanishing-gradient problem discussed in Chapter 7.

9.1 Recurrent backpropagation*

Recall Figure 9.1. We want to train a network with N real-valued units V_i with sigmoid activation functions, and weights w_{ij} from V_j to V_i. Several of the units may be connected to inputs $x_k^{(\mu)}$. Other units are output units with associated target values $y_i^{(\mu)}$. We take the dynamics to be continuous in time, Equation (9.3), and assume that $\boldsymbol{V}(t)$ runs into a steady state

$$\boldsymbol{V}(t) \to \boldsymbol{V}^* \quad \text{so that} \quad \frac{\mathrm{d}V_i^*}{\mathrm{d}t} = 0\,. \tag{9.5}$$

From Equation (9.3) we deduce

$$V_i^* = g\Big(\sum_j w_{ij}^{(\nu\nu)} V_j^* + \sum_k w_{ik}^{(\nu x)} x_k - \theta_i^{(\nu)}\Big). \tag{9.6}$$

We assume that $\boldsymbol{V}^*$ is a *linearly stable* steady state of the dynamics (9.3), so that small perturbations $\delta\boldsymbol{V}$ away from $\boldsymbol{V}^*$ decay with time. The synchronous discrete dynamics (9.2) can exhibit undesirable stable periodic solutions [106], as mentioned in Section 1.3. This is a reason for using the continuous dynamics (9.3), yet convergence to the steady state is not guaranteed in this case either. Equation (9.6) is a non-linear self-consistent condition for the components of $\boldsymbol{V}^*$, in general difficult to solve. However, if the fixed point $\boldsymbol{V}^*$ is stable, we can use the dynamics (9.3) to automatically pick out the steady-state solution $\boldsymbol{V}^*$. This solution depends on the pattern $\boldsymbol{x}^{(\mu)}$, but in Equations (9.5) and (9.6) and also in the remainder of this Section the superscript (μ) is left out.

The goal is to find weights so that the outputs give the correct target values in the steady state, those associated with $\boldsymbol{x}^{(\mu)}$. To this end we use gradient descent on the energy function (9.4). Consider first how to update the weights $w_{ij}^{(\nu\nu)}$:

$$\delta w_{mn}^{(\nu\nu)} = -\eta \frac{\partial H}{\partial w_{mn}^{(\nu\nu)}} = \eta \sum_k E_k^* \frac{\partial V_k^*}{\partial w_{mn}^{(\nu\nu)}}\,. \tag{9.7}$$

To calculate the gradients of $\boldsymbol{V}^*$ we use Equation (9.6):

$$\frac{\partial V_i^*}{\partial w_{mn}^{(\nu\nu)}} = g'(b_i^*)\frac{\partial b_i^*}{\partial w_{mn}^{(\nu\nu)}} = g'(b_i^*)\Big(\delta_{im} V_n^* + \sum_j w_{ij}^{(\nu\nu)} \frac{\partial V_j^*}{\partial w_{mn}^{(\nu\nu)}}\Big), \tag{9.8}$$

where $b_i^* = \sum_j w_{ij}^{(\nu\nu)} V_j^* + \sum_k w_{ik}^{(\nu x)} x_k - \theta_i^{(\nu)}$. Equation (9.8) is a self-consistent equation for the gradient, as opposed to the explicit equations we found in Chapter 6. The reason for the difference is that the recurrent network has feedbacks.

Equation (9.8) is linear in the gradients and can therefore be solved by matrix inversion, at least formally. In terms of the matrix $\mathbb{L}$ with elements

$$L_{ij} = \delta_{ij} - g'(b_i^*) w_{ij}^{(vv)}, \tag{9.9}$$

Equation (9.8) can be written as

$$\sum_j L_{ij} \frac{\partial V_j^*}{\partial w_{mn}^{(vv)}} = \delta_{im} g'(b_i^*) V_n^*. \tag{9.10}$$

If $\mathbb{L}$ is invertible, one applies $\sum_i \left(\mathbb{L}^{-1}\right)_{ki}$ to both sides. Using the fact that $\sum_i \left(\mathbb{L}^{-1}\right)_{ki} L_{ij} = \delta_{kj}$ one finds:

$$\frac{\partial V_k^*}{\partial w_{mn}^{(vv)}} = \left(\mathbb{L}^{-1}\right)_{km} g'(b_m^*) V_n^*. \tag{9.11}$$

Inserting this result into (9.7) one obtains:

$$\delta w_{mn}^{(vv)} = \eta \sum_k E_k^* \left(\mathbb{L}^{-1}\right)_{km} g'(b_m^*) V_n^*. \tag{9.12}$$

This learning rule can be written in the form of the backpropagation rule (6.12) by introducing the error

$$\Delta_m^* = g'(b_m^*) \sum_k E_k^* \left(\mathbb{L}^{-1}\right)_{km}. \tag{9.13}$$

Then the learning rule (9.12) takes the form

$$\delta w_{mn}^{(vv)} = \eta \Delta_m^* V_n^*. \tag{9.14}$$

A problem is that the matrix $\mathbb{L}$ must be invertible for the solution (9.11) to exist. Also, matrix inversion is an expensive operation.

As described in Chapter 5, one can try find the inverse iteratively. The trick is now to write down a dynamical equation for Δ_i that has a steady state at the solution of Equation (9.13):

$$\tau \tfrac{\mathrm{d}}{\mathrm{d}t} \Delta_j = -\Delta_j + g'(b_j^*) E_j^* + \sum_i \Delta_i w_{ij}^{(vv)} g'(b_j^*). \tag{9.15}$$

It is left as an exercise for the reader to verify that the dynamics (9.15) has a steady state satisfying Equation (9.13). Equation (9.15) is written in a form to stress that (9.15) and (9.3) exhibit the same *duality* as Algorithm 3, between forward propagation of states of neurons (step 5) and backpropagation of errors (step 9). The sum in Equation (9.15) has the same form as the recursion for the errors in Algorithm 3 (step 9), except that there are no layer indices ℓ here.

Equation (9.15) admits the steady state (9.13). But does $\Delta_i(t)$ converge to Δ_i^*? For convergence it is necessary that the fixed point is linearly stable. Whether or not this is the case is determined by *linear stability analysis* [107]. One asks: does a small deviation from the fixed point increase or decrease under Equation (9.15)? To answer this question one writes

$$\boldsymbol{V}(t) = \boldsymbol{V}^* + \delta\boldsymbol{V}(t) \quad \text{and} \quad \boldsymbol{\Delta}(t) = \boldsymbol{\Delta}^* + \delta\boldsymbol{\Delta}(t), \tag{9.16}$$

inserts this *ansatz* into (9.3) and (9.15), and linearises:

$$\tau \tfrac{\mathrm{d}}{\mathrm{d}t}\delta V_i = -\delta V_i + g'(b_i^*)\sum_j w_{ij}^{(\nu\nu)}\delta V_j \approx -\sum_j L_{ij}\delta V_j\,, \tag{9.17a}$$

$$\tau \tfrac{\mathrm{d}}{\mathrm{d}t}\delta\Delta_j = -\delta\Delta_j + \sum_i \delta\Delta_i\, w_{ij}^{(\nu\nu)} g'(b_j^*) \approx -\sum_i \delta\Delta_i g'(b_i^*)L_{ij}/g'(b_j^*). \tag{9.17b}$$

Equation (9.17a) shows: whether or not the norm of $\delta\boldsymbol{V}(t)$ grows is determined by the eigenvalues of the matrix $\mathbb{L}$. We say that $\boldsymbol{V}^*$ is a linearly stable fixed point of Equation (9.3) if all eigenvalues of $\mathbb{L}$ have negative real parts. In this case $|\delta\boldsymbol{V}(t)| \to 0$. If at least one eigenvalue has a positive real part then δV_i grows. In this case we say that $\boldsymbol{V}^*$ is linearly unstable. Since the matrix with elements $g'(b_i^*)L_{ij}/g'(b_j^*)$ has the same eigenvalues as $\mathbb{L}$, $\boldsymbol{\Delta}^*$ is a stable fixed point of (9.15) if $\boldsymbol{V}^*$ is a stable fixed point of (9.3). This was assumed in the beginning, Equation (9.5). If it does not hold, the algorithm does not converge. Finally, the update rule for the weights $w_{mn}^{(\nu x)}$ is derived in an analogous fashion. The result is:

$$\delta w_{mn}^{(\nu x)} = \eta\Delta_m^* x_n\,. \tag{9.18}$$

In summary, recurrent backpropagation (Algorithm 6) is analogous to backpropagation (Algorithm 3) for layered feed-forward networks, save for two differences. First, the non-linear network dynamics is no longer a simple input-to-output mapping with nested activation functions, but a non-linear dynamics that may (or may not) converge to a steady state. Second, the feedbacks give rise to linear self-consistent equations for the steady-state gradients $\partial V_j^*/\partial w_{mn}$, which can be viewed as steady-state conditions for a dual dynamics of the errors. Convergence of the network and error dynamics is not guaranteed. For a layered feed-forward network, recurrent backpropagation simplifies to Algorithm 3 (Exercise 9.1).

Algorithm 6 recurrent backpropagation

initialise all weights;
for $t = 1, \dots, T$ **do**
  choose a value of μ and apply $\boldsymbol{x}^{(\mu)}$ to the inputs;
  find V_n^* by relaxing $\tau \frac{\mathrm{d}V_n}{\mathrm{d}t} = -V_n + g\big(\sum_j w_{nj}^{(vv)} V_j + \sum_k w_{nk}^{(vx)} x_k - \theta_n^{(v)}\big)$;
  compute $E_k^* = y_k - V_k^*$ for all output units;
  find Δ_m^* by relaxing $\tau \frac{\mathrm{d}\Delta_m}{\mathrm{d}t} = -\Delta_m + \sum_j \Delta_j w_{jm} g'(b_m^*) + g'(b_m^*) E_m^*$;
  update all weights: $w_{mn}^{(vv)} \leftarrow w_{mn} + \delta w_{mn}^{(vv)}$ with $\delta w_{mn}^{(vv)} = \eta \Delta_m^* V_n^*$ and $w_{mn}^{(vx)} \leftarrow w_{mn}^{(vx)} + \delta w_{mn}^{(vx)}$ with $\delta w_{mn}^{(vx)} = \eta \Delta_m^* x_n$;
end for

9.2 Backpropagation through time

Recurrent networks can be used to learn sequential inputs, as in speech recognition and machine translation. The training set is a time sequence of inputs and targets $[\boldsymbol{x}(t), \boldsymbol{y}(t)]$. The network is trained on the sequence and learns to predict the targets. In this context the layout is changed a little bit compared with the one described in the previous Section. There are two main differences. Firstly, the inputs and targets depend on t and one uses a discrete-time update rule. Secondly, separate output units $O_i(t)$ are added to the layout. The update rule takes the form

$$V_i(t) = g\Big(\sum_j w_{ij}^{(vv)} V_j(t-1) + \sum_k w_{ik}^{(vx)} x_k(t) - \theta_i^{(v)}\Big), \tag{9.19a}$$

$$O_i(t) = g\Big(\sum_j w_{ij}^{(ov)} V_j(t) - \theta_i^{(o)}\Big). \tag{9.19b}$$

The activation function of the outputs O_i can be different from that of the hidden neurons V_j. Often the softmax function is used for the outputs [108, 109].

To train recurrent networks with time-dependent inputs and targets and with the dynamics (9.19) one uses *backpropagation through time*. The idea is to unfold the network in time to get rid of the feedbacks, at the expense of as many copies of the original neurons as there are time steps.

This is illustrated in Figure 9.2 for a recurrent network with one hidden neuron, one input, and one output. The unfolded network has T inputs and outputs. It can be trained in the usual way with *stochastic gradient descent*. The errors are calculated using backpropagation as in Algorithm 3, but here the error is propagated back in time, not from layer to layer. The energy function is the squared error summed over all time steps

$$H = \frac{1}{2} \sum_{t=1}^{T} E_t^2 \quad \text{with} \quad E_t = y_t - O_t \,. \tag{9.20}$$

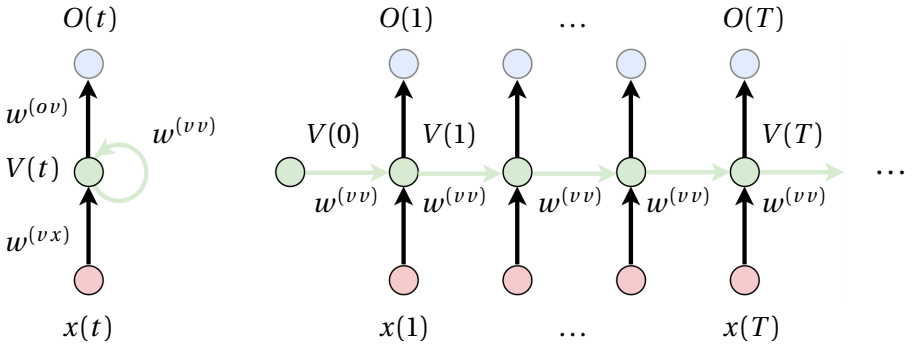


Figure 9.2: Left: recurrent network with one hidden neuron (green) and one output neuron (blue). The input terminal is drawn red. Right: same network but unfolded in time. The weights $w^{(vv)}$ remain unchanged as drawn, also the weights $w^{(vx)}$ and $w^{(ov)}$ remain unchanged (not drawn).

Since there is only one hidden neuron in our example, and since the inputs and outputs are also one-dimensional, it is simpler to write the time argument as a subscript, O_t instead of $O(t)$ for example. Note also that one could use the negative log-likelihood function (7.34) instead of Equation (9.20).

Consider first how to update the weight $w^{(vv)}$. The gradient-descent rule (5.24) gives a result that is of the same form as Equation (9.7):

$$\delta w^{(vv)} = \eta \sum_{t=1}^{T} E_t \frac{\partial O_t}{\partial w^{(vv)}} = \eta \sum_{t=1}^{T} \Delta_t \, w^{(ov)} \frac{\partial V_t}{\partial w^{(vv)}} \,. \tag{9.21}$$

Here $\Delta_t = E_t g'(B_t)$ is an output error, $B_t = w^{(ov)} V_t - \theta^{(o)}$ is the local field of the output neuron at time t [Equation (9.19)], and $\partial V_t / \partial w^{(vv)}$ is evaluated with the chain rule, as usual. Equation (9.19a) yields the recursion

$$\frac{\partial V_t}{\partial w^{(vv)}} = g'(b_t) \Big(V_{t-1} + w^{(vv)} \frac{\partial V_{t-1}}{\partial w^{(vv)}} \Big) \tag{9.22}$$

for $t \geq 1$. Since $\partial V_0 / \partial w^{(vv)} = 0$, Equation (9.22) implies:

$$\begin{aligned}
\frac{\partial V_1}{\partial w^{(vv)}} &= g'(b_1) V_0 \,, \\
\frac{\partial V_2}{\partial w^{(vv)}} &= g'(b_2) V_1 + g'(b_2) w^{(vv)} g'(b_1) V_0 \,, \\
\frac{\partial V_3}{\partial w^{(vv)}} &= g'(b_3) V_2 + g'(b_3) w^{(vv)} g'(b_2) V_1 + g'(b_3) w^{(vv)} g'(b_2) w^{(vv)} g'(b_1) V_0 \\
&\vdots \\
\frac{\partial V_{T-1}}{\partial w^{(vv)}} &= g'(b_{T-1}) V_{T-2} + g'(b_{T-1}) w^{(vv)} g'(b_{T-2}) V_{T-3} + \ldots \\
\frac{\partial V_T}{\partial w^{(vv)}} &= g'(b_T) V_{T-1} + g'(b_T) w^{(vv)} g'(b_{T-1}) V_{T-2} + \ldots
\end{aligned}$$

Equation (9.21) says that we must sum over t. Regrouping the terms in this sum yields:

$$\begin{aligned}
&\Delta_1 \frac{\partial V_1}{\partial w^{(vv)}} + \Delta_2 \frac{\partial V_2}{\partial w^{(vv)}} + \Delta_3 \frac{\partial V_3}{\partial w^{(vv)}} + \ldots \\
&= [\Delta_1 g'(b_1) + \Delta_2 g'(b_2) w^{(vv)} g'(b_1) + \Delta_3 g'(b_3) w^{(vv)} g'(b_2) w^{(vv)} g'(b_1) + \ldots] V_0 \\
&+ [\Delta_2 g'(b_2) + \Delta_3 g'(b_3) w^{(vv)} g'(b_2) + \Delta_4 g'(b_4) w^{(vv)} g'(b_3) w^{(vv)} g'(b_2) + \ldots] V_1 \\
&+ [\Delta_3 g'(b_3) + \Delta_4 g'(b_4) w^{(vv)} g'(b_3) + \Delta_5 g'(b_5) w^{(vv)} g'(b_4) w^{(vv)} g'(b_3) + \ldots] V_2 \\
&\vdots \\
&+ [\Delta_{T-1} g'(b_{T-1}) + \Delta_T g'(b_T) w^{(vv)} g'(b_{T-1})] V_{T-2} \\
&+ [\Delta_T g'(b_T)] V_{T-1} \,.
\end{aligned}$$

To write the learning rule in the usual form, we define *errors* δ_t recursively:

$$\delta_t = \begin{cases} \Delta_T w^{(ov)} g'(b_T) & \text{for } t = T, \\ \Delta_t w^{(ov)} g'(b_t) + \delta_{t+1} w^{(vv)} g'(b_t) & \text{for } 0 < t < T. \end{cases} \tag{9.23}$$

Then the learning rule takes the form

$$\delta w^{(vv)} = \eta \sum_{t=1}^{T} \delta_t V_{t-1} \,, \tag{9.24}$$

just like Equation (6.11), or like the recursion in step 9 of Algorithm 3. The factor $w^{(vv)} g'(b_{t-1})$ in the recursion (9.23) gives rise to a product of many such factors in δ_t when T is large, exactly as described in Section 7.2 for multilayer perceptrons. This

means that the training of recurrent nets suffers from *unstable gradients*, as backpropagation of multilayer perceptrons does (Section 7.2). If the factors $|w^{(vv)}g'(b_p)|$ are smaller than unity then the errors δ_t become very small when t becomes small (*vanishing-gradient problem*). This means that the early states of the hidden neuron no longer contribute to the learning, causing the network to forget what it has learned about early inputs. When $|w^{(vv)}g'(b_p)| > 1$, on the other hand, exploding gradients make learning impossible. In summary, the *unstable gradients* in recurrent neural networks occurs much in the same way as in multilayer perceptrons (Section 7.2). The resulting difficulties for training recurrent neural networks are discussed in more detail in Ref. [110].

A slight variation of the above algorithm (*truncated backpropagation through time*) suffers less from the exploding-gradient problem. The idea is that the exploding gradients are tamed by truncating the memory. This is achieved by limiting the error propagation backwards in time, errors are computed back to $T-\tau$ and not further, where τ is the truncation time [2]. Naturally this implies that long-time correlations cannot be learnt.

Finally, the update formulae for the weights $w^{(vx)}$ are obtained in a similar fashion. Equation (9.19a) yields the recursion

$$\frac{\partial V_t}{\partial w^{(vx)}} = g'(b_t)\Big(x_t + w^{(vv)}\frac{\partial V_{t-1}}{\partial w^{(vx)}}\Big). \tag{9.25}$$

This looks just like Equation (9.22), except that V_{t-1} is replaced by x_t. As a consequence we have

$$\delta w^{(vx)} = \eta \sum_{t=1}^{T} \delta_t x_t\,. \tag{9.26}$$

The update formula for $w^{(ov)}$ is simpler to derive. From Equation (9.19b) we find by differentiation w.r.t. $w^{(ov)}$:

$$\delta w^{(ov)} = \eta \sum_{t=1}^{T} E_t g'(B_t) V_t\,. \tag{9.27}$$

How are the thresholds updated? Going through the above derivation we see that we must replace V_{t-1} and x_t in Equations (9.24) and (9.26) by -1. It works in the same way for the output threshold.

In order to keep the formulae simple, I only described the algorithm for a single hidden and a single output neuron, so that I could leave out the indices referring to different hidden neurons and/or different output components. You can add those indices yourself, the structure of the Equations remains exactly the same, save for a

number of extra sums over those indices:

$$\delta w_{mn}^{(vv)} = \eta \sum_{t=1}^{T} \delta_m^{(t)} V_n^{(t-1)} \tag{9.28}$$

$$\delta_j^{(t)} = \begin{cases} \sum_i \Delta_i^{(t)} w_{ij}^{(ov)} g'(b_j^{(t)}) & \text{for } t = T, \\ \sum_i \Delta_i^{(t)} w_{ij}^{(ov)} g'(b_j^{(t)}) + \sum_i \delta_i^{(t+1)} w_{ij}^{(vv)} g'(b_j^{(t)}) & \text{for } 0 < t < T. \end{cases}$$

The second term in the recursion for $\delta_j^{(t)}$ is analogous to the recursion in step 9 of Algorithm 3. The time index t here plays the role of the layer index ℓ in Algorithm 3. A difference is that the weights in Equation (9.28) are the same for all time steps.

In summary you see that backpropagation through time for recurrent networks is similar to backpropagation for multilayer perceptrons. After the recurrent network is unfolded to get rid of the feedback connections it can be trained by backpropagation. The time index t takes the role of the layer index ℓ. Backpropagation through time is the standard approach for training recurrent nets, despite the fact that it suffers from the vanishing-gradient problem. The next Section describes how improvements to the layout make it possible to efficiently train recurrent networks.

9.3 Vanishing gradients

Hochreiter and Schmidhuber [111] suggested to replace the hidden neurons of the recurrent network with computation units that are specially designed to reduce the vanishing-gradient problem. The method is referred to as *long short-term memory* (LSTM). The basic ingredient is the same as in *residual networks* (Section 7.4): short cuts reduce the vanishing-gradient problem. For our purposes we can think of LSTMs as units that replace the hidden neurons. For a detailed description of LSTMs see Ref. [112].

Gated recurrent units [113] serve the same purpose as LSTMs, and they function in a similar way. It has been argued that LSTMs outperform gated recurrent units for certain tasks, but since they are simpler than LSTMs, the remainder of this Section focuses on gated recurrent units. As illustrated in Figure 9.3, these units replace the hidden units of a recurrent neural net with update rule (9.19a). The update rule for

a gated recurrent unit reads:

$$z_m = \sigma\Big(\sum_k w_{mk}^{(zx)} x_k(t) + \sum_j w_{mj}^{(zv)} V_j(t-1)\Big), \tag{9.29a}$$

$$r_n = \sigma\Big(\sum_k w_{nk}^{(rx)} x_k(t) + \sum_j w_{nj}^{(rv)} V_j(t-1)\Big), \tag{9.29b}$$

$$h_i = g\Big(\sum_k w_{ik}^{(hx)} x_k(t) + \sum_j w_{ij}^{(hv)} r_j V_j(t-1)\Big), \tag{9.29c}$$

$$V_i(t) = (1-z_i)h_i + z_i V_i(t-1). \tag{9.29d}$$

The first two Equations are referred to as *gates* because they regulate how the values of the hidden state variables V_i are passed through the unit. Here $\sigma(b)$ is the sigmoid function, Equation (6.21a). If all z_m are equal to zero and all $r_n = 1$, Equation (9.29) coincides with the standard update rule (9.19a), save for the thresholds which were left out in Equation (9.29). As explained above, the resulting recurrent net suffers from the vanishing-gradient problem. This means that states in the past history, $\boldsymbol{V}(0), \boldsymbol{V}(1), \ldots$, have little effect upon the present state $\boldsymbol{V}(t)$ for $t \gg 1$. The recurrent net forgets early inputs, so that it cannot learn from them. If by contrast all $z_m = 1$ then the input is passed right through the unit. Since $\partial V_i(t)/\partial V_j(t-1) = \delta_{ij}$, the gradients do not decrease as explores the history. However, since $\boldsymbol{V}(t-1) = \boldsymbol{V}(t)$, the recurrent net reproduces previous states. This is analogous to skipping layers in a residual net, although comparison with Equation (7.31) reveals some differences in detail.

For the recurrent net to learn in a meaningful way from past inputs, the weights in Equation (9.29) (and the thresholds) are adjusted so that the gated recurrent unit operates between these two extreme limits. This is achieved by including the weights and thresholds of the gated recurrent unit in the gradient-descent optimisation of

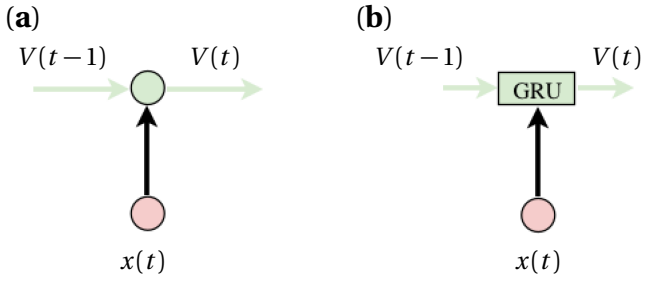


Figure 9.3: Gated recurrent unit. (**a**) The symbol refers to the standard recursion (9.19a) for the hidden variable, as in the right panel of Figure 9.2. (**b**) To combat the vanishing-gradient problem, the standard unit is replaced by a gated recurrent unit (9.29).

the energy function (9.20). The weight updates are calculated in the same as before. Using Equation (9.19b) one has:

$$\delta w_{mn}^{(ab)} = \eta \sum_{t=1}^{T} \sum_{i} \Delta_i(t) w_{ij}^{(ov)} \frac{\partial V_j(t)}{\partial w_{mn}^{(ab)}}, \tag{9.30}$$

where $w^{(ab)}$ stands for $w^{(zx)}$, $w^{(zv)}$, $w^{(rx)}$,.... The derivatives $\partial V_j(t)/\partial w_{mn}^{(ab)}$ are evaluated using the chain rule and Equations (9.29).

It is instructive to inspect the values of z_i and r_i when the recurrent net operates after training. Suppose that a unit assumes small values of z_m and r_n. This means that the update of the state variable $V_i(t)$ is determined entirely by the instantaneous inputs $x_k(t)$. Since the unit does not refer to the past history of the hidden-state variables, it truncates the dynamical memory. In the opposite limit, when $z_i \approx r_i \approx 1$, the unit can contribute to building up long-term dynamical memory. These arguments suggest that a unit with just one gate may achieve the same goal [114]:

$$z_m = \sigma\Big(\sum_k w_{mk}^{(zx)} x_k(t) + \sum_j w_{mj}^{(zv)} V_j(t-1)\Big), \tag{9.31a}$$

$$h_i = g\Big(\sum_k w_{ik}^{(hx)} x_k(t) + \sum_j w_{ij}^{(hv)} z_j V_j(t-1)\Big), \tag{9.31b}$$

$$V_i(t) = (1-z_i)h_i + z_i V_i(t-1). \tag{9.31c}$$

This unit is easier to train because it has fewer parameters than the standard gated recurrent unit (9.29). Yet, the additional parameters may help to represent and exploit correlations on different time scales. LSTMs have even more parameters. How this tradeoff between ease of training and accurate representation of time correlations works out may well depend on the problem at hand. In the following Section we describe recurrent nets with LSTM units, following Refs. [108, 109].

9.4 Recurrent networks for machine translation*

Recurrent networks are used for machine translation [109]. How does this work? The networks are trained using backpropagation through time. The vanishing-gradient problem is dealt with by using LSTMs (Section 9.3).

How are the network inputs and outputs coded? For machine translation one respresents all words in a given dictionary in terms of a code. The simplest code is a binary code where 100... represents the first word in the dictionary, 010... the second word, and so forth. Each input is a vector with as many components as there are words in the dictionary. A sentence corresponds to a sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$. Each sentence ends with an end-of-sentence tag, <EOS>. Softmax outputs give

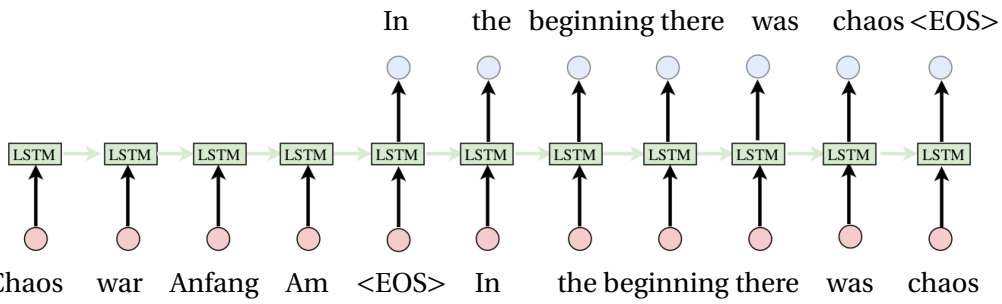


Figure 9.4: Schematic illustration of unfolded recurrent network for machine translation, after Refs. [108, 109]. The green rectangular boxes represent the hidden states in the form of long short-term memory (LSTM) units, see Section 9.3. Otherwise the network layout is like the one shown in Figure 9.2. Sutskever *et al.* [108] found that the network translates much better if the sentence is read in reverse order, from the end. The tag <EOS> denotes the end-of-sentence tag. Here it denotes the beginning of the sentence.

the probability $p(\boldsymbol{O}_1,\ldots,\boldsymbol{O}_{T'}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)$ of an output sequence conditional on the input sequence. The translated sentence is the one with the highest probability (it also contains the end-of-sentence tag <EOS>). So both inputs and outputs are represented by high-dimensional vectors $\boldsymbol{x}_t$ and $\boldsymbol{O}_t$. Other encoding schemes are described in Ref. [109].

What is the role of the hidden states, represented in terms of an LSTM? The network encodes the input sequence $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T$ in these states. Upon encountering the <EOS> tag in the input sequence, the network outputs the first word of the translated sentence using the information about the input sequence stored in $\boldsymbol{V}_T$ as shown in Figure 9.4. The first output is fed into the next input, and the network continues to translate until it produces an <EOS> tag for the output sequence. In short, the network calculates the probabilities

$$p(\boldsymbol{O}_1,\ldots,\boldsymbol{O}_{T'}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)=\prod_{t=1}^{T'} p(\boldsymbol{O}_t|\boldsymbol{O}_1,\ldots,\boldsymbol{O}_{t-1};\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T), \tag{9.32}$$

where $p(O_t|O_1,\ldots,O_{t-1};\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)$ is the probability of the next word in the output sequence give the inputs and the output sequence up to O_{t-1} [115].

There is a large number of recent papers on machine translation with recurrent neural nets. Most studies are based on the training algorithm described in Section 9.2, backpropagation through time. The different approaches mainly differ in their network layouts. Google's machine translation system uses a deep network with layers of LSTMs [115]. Different hidden states are unfolded forward as well as backwards in time, as illustrated in Figure 9.5. In this Figure the hidden states are

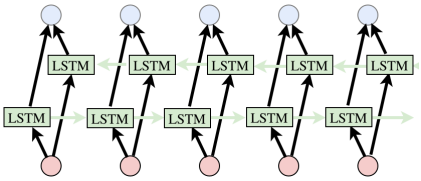


Figure 9.5: Schematic illustration of a bidirectional recurrent network . The net consists of two hidden states that are unfolded in different ways. The hidden states are represented by LSTMs.

represented by LSTMs. In the simplest case the hidden states are just encoded in hidden neurons, as in Figure 9.2 and Equation (9.19). If we represent the hidden states by neurons, as in Section 9.2, then the corresponding bidirectional network has the dynamics

$$\begin{aligned} V_i(t) &= g\Big(\sum_j w_{ij}^{(vv)} V_j(t-1) + \sum_k w_{ik}^{(vx)} x_k(t) - \theta_i^{(v)}\Big), \\ U_i(t) &= g\Big(\sum_j w_{ij}^{(uu)} U_j(t+1) + \sum_k w_{ik}^{(ux)} x_k(t) - \theta_i^{(u)}\Big), \\ O_i(t) &= g\Big(\sum_j w_{ij}^{(ov)} V_j(t) + \sum_j w_{ij}^{(ou)} U_j(t) - \theta_i^{(o)}\Big). \end{aligned} \tag{9.33}$$

It is natural to use bidirectional nets for machine translation because correlations go either way in a sentence, forward and backwards. In German, for example, the finite verb form is usually at the end of the sentence.

Different schemes for scoring the accuracy of a translation are described by Lipton *et al.* [109]. One difficulty is that there are often several different valid translations of a given sentence, and the score must compare the machine translation with all of them. More recent papers on machine translation usually use the so-called BLEU score to evaluate the translation accuracy. The acronym stands for *bilingual evaluation understudy*. The scheme was proposed by Papieni *et al.* [116]. It is argued to evaluate the accuracy of a translation not too differently from Humans.

9.5 Reservoir computing*

An alternative to backpropagation through time for recurrent nets is *reservoir computing* [117]. This method has been used with success to predict chaotic dynamics [118, 119], and rare transitions in stochastic bi-stable systems [120]. Consider input data in the form of a time series $\boldsymbol{x}(1),\ldots\boldsymbol{x}(T)$ of N-dimensional vectors $\boldsymbol{x}(t)$, and a corresponding series of M-dimensional targets $\boldsymbol{y}(t)$. The goal is to train the recurrent net so that its ouputs $\boldsymbol{O}(t)$ approximate the targets as precisely as possible,

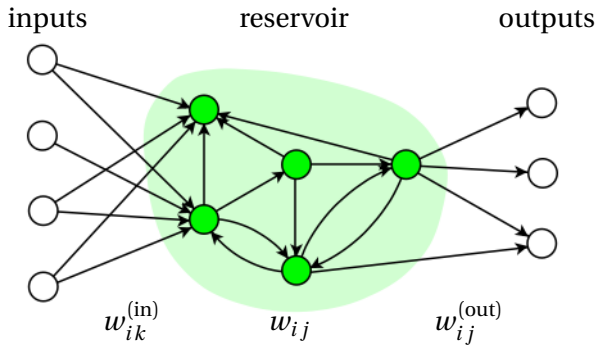


Figure 9.6: Reservoir computing (schematic). Not all connections are drawn. There can be connections from all inputs to all neurons in the reservoir (green), and from all reservoir neurons to all output neurons.

by minimising the energy function $H = \frac{1}{2}\sum_{t=1}^{T}[E_i(t)]^2$, where $E_i(t) = y_i(t) - O_i(t)$ is the output error. For time-series prediction, the targets are $\boldsymbol{y}(t) = \boldsymbol{x}(t)$. After training the recurrent net, one continues to iterate the network dynamics with inputs $\boldsymbol{y}(\tau), \ldots, \boldsymbol{y}(T+\tau)$ to predict the time series for $t = T+\tau$ with $\tau = 1, 2, \ldots$.

Figure 9.6 shows the layout for this task, N input terminals are connected with weights $w_{jk}^{(\mathrm{in})}$ to a reservoir of hidden neurons with state variables r_j. The reservoir is linked to M linear output units O_i with weights $w_{ij}^{(\mathrm{out})}$. The reservoir is a recurrent net with weights w_{ij}. The update rule is similar to Equation (9.19):

$$r_i(t) = g\Big(\sum_j w_{ij} r_j(t-1) + \sum_k w_{ik}^{(\mathrm{in})} x_k(t)\Big), \tag{9.34}$$

$$y_i(t) = \sum_j w_{ij}^{(\mathrm{out})} r_j(t). \tag{9.35}$$

The main difference to the algorithms described in the previous Sections of this Chapter is that the input weights $w_{jk}^{(\mathrm{in})}$ and the reservoir weights $w_{jk}^{(\mathrm{in})}$ are randomly initialised and kept constant. Only the ouput weights $w_{jk}^{(\mathrm{out})}$ are trained by gradient descent. The idea is that the dynamics of a sufficiently large reservoir finds non-linear, high-dimensional representations of the input data [117], not unlike sparse representations of binary classification problems embedded in a high-dimensional space (Section 5.4) that become linearly separable in this way. In addition, and this is a difference to the problem described in Section 5.4, the reservoir serves as a dynamical memory.

This requires that the reservoir states represent the input sequence, that similar input sequences yield similar activations of the reservoir, provided one iterates it long enough. However, for random weights the recurrent reservoir dynamics can

be chaotic [107], so that the state of the reservoir bears no relation to the input sequence. To avoid this, one requires that the resevoir dynamics is linearly stable, that the maximal eigenvalue of its linearisation is smaller than unity in modulus.[1] Linearising the reservoir dynamics (9.34) gives

$$\delta \boldsymbol{r}(t) = \mathbb{D}[\mathbb{W}\delta r(t-1) + \mathbb{W}^{(\mathrm{in})}\delta \boldsymbol{x}(t)] \tag{9.36}$$

where $\mathbb{D}$ is a diagonal matrix with entries $D_{ii} = \mathrm{d}g(b_i(t))/\mathrm{d}b$, just like in Section 7.2. If one uses tanh-activation functions and ensures that the local fields $b_i(t)$ remain small, then the diagonal elements (eigenvalues) of $\mathbb{D}$ are close to unity. In this case the stability condition for the reservoir dynamics is determined by the weight matrix $\mathbb{W}$. Whether or not $\delta \boldsymbol{r}$ grows is determined by $\mathbb{W}^{\mathsf{T}}\mathbb{W}$ (Section 7.2). Denoting the eigenvalues of this symmetric matrix by $\Lambda_\alpha = \exp(2\lambda_\alpha)$, the stability condition reads eigenvalue of $\mathbb{W}$ must be smaller than unity in modulus,

$$\lambda_{\max}(\mathbb{W}) < 0\,. \tag{9.37}$$

But note that this is not a necessary condition, because the derivative of $\tanh(b)$ is smaller than unity.

For inputs with long time correlations, the reservoir must not decay too quickly, so that it can represent the dynamical correlations in the input sequence. This means that one should try to adjust the modulus of the maximal eigenvalue $\lambda_{\max}$ of $\mathbb{W}$ to be as close to unity as possible. In summary, achieving a stable high-dimensional representation and reliable dynamic memory requires a delicate balance, in the same way as one must strike a balance between growing and vanishing gradients in backpropagation of multilayer perceptrons (Section 7.2). In practice one draws the elements of $\mathbb{W}$ independently from a uniform distribution and rescales the matrix by $\mathbb{W}' = \mathbb{W}/\exp(\lambda_{\max})$.

There are many different recipes for how to set up a reservoir. Some take the weights to be uniformly distribution, some Gaussian, or assume that $w_{ij} = \pm 1$ with equal probability. Usually one takes the reservoir to be sparse, with only a small fraction of weights non-zero. The weight matrix $\mathbb{W}^{(\mathrm{in})}$ is usually taken to be a full matrix, and its elements are drawn from the same distribution as those of the reservoir. Lukosevicius [121] gives a practical overview over the different recipes for setting up reservoir computers. In order to represent complex spatio-temporal patterns it is necessary to use multiple reservoirs [118]. Tanaka *et al.* [122] describe different physical implementations of reservoir computers, based on electronic RC-circuits, optical cavities or resonators, spin-torque oscillators, or mechanical devices.

[1] For time-continuous dynamics (Section 9.1), linear stability is ensured when all eigenvalues have negative real parts, for discrete dynamics they must all be in modulus smaller than unity [107].

9.6 Summary

It is sometimes said that recurrent networks learn *dynamical systems* while multilayer perceptrons learn *input-output maps*. This notion refers to backpropagation in time. I would emphasise, by contrast, that both networks are trained in similar ways, by backpropagation. Neither is it given that the tasks must differ: recurrent networks are also used to learn time-independent data. It is true though that tools from *dynamical-systems theory* have been used with success to analyse the dynamics of recurrent networks [110, 123].

Recurrent neural networks are trained by stochastic gradient descent after unfolding the network in time to get rid of feedback connections. This algorithm suffers from the vanishing-gradient problem. To overcome this difficulty, the hidden states in the recurrent network are usually represented by LSTMs. Recent layouts for machine translation use deep bidirectional networks with layers of LSTMs.

9.7 Further reading

The training of recurrent networks is discussed in Chapter 15 of Ref. [2]. Recurrent backpropagation is described by Hertz, Krogh and Palmer [1], for a slightly different network layout. The standard references for backpropagation through time are Refs. [124, 125]. How LSTMs combat the vanishing-gradient problem is explained in Ref. [112]. For a recent review of recurrent neural nets, see Ref. [109]. This webpage [126] gives a very enthusiastic overview about what recurrent nets can do. A more pessimistic view is expressed in this blog. For a review of reservoir computing, see Ref. [117].

9.8 Exercises

9.1 Recurrent backpropagation. Derive Eq. (9.18) for the weight updates $\delta w_{mn}^{(vx)}$ in recurrent backpropagation. Show how the recurrent-backpropagation algorithm simplifies to Algorithm 3 for layered feed-forward networks when there are no feedbacks,

9.2 Learning rules for backpropagation through time. Derive the learning rules (9.26) and (9.27) from Equation (9.19).

9.3 Recurrent network. Figure 9.7 shows a simple recurrent network with one hidden neuron $V(t)$, one input $x(t)$ and one output $O(t)$. The network learns a time series of input-output pairs $[x(t), y(t)]$ for $t = 1, 2, 3, \ldots, T$. Here t is a discrete time

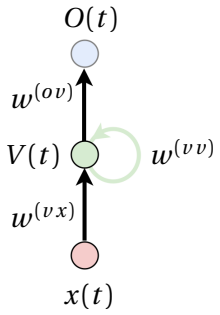


Figure 9.7: Recurrent network with one input unit $x(t)$ (red), one hidden neuron $V(t)$ (green) and one output neuron $O(t)$ (blue). Exercise 9.3.

index and $y(t)$ is the target value at time t (the targets are denoted by y). The hidden unit is initialised to $V(0)$ at $t = 0$. This network can be trained by backpropgation by *unfolding it in time*. Draw the unfolded network, label the connections using the labels shown in Figure 9.7, and discuss the layout (max half an A4 page). Write down the dynamical rules for this network, the rules that determine $V(t)$ in terms of $V(t-1)$ and $x(t)$, and $O(t)$ in terms of $V(t)$. Assume that both $V(t)$ and $O(t)$ have the same activation function $g(b)$. Derive the update rule for $w^{(ov)}$ for gradient descent on the energy function $H = \frac{1}{2}\sum_{t=1}^{T} E(t)^2$ with $E(t) = y(t) - O(t)$. Denote the learning rate by η. *Hint:* the update rule for $w^{(ov)}$ is much simpler to derive than those for $w^{(vx)}$ and $w^{(vv)}$. Explain how recurrent networks are used for machine translation. Draw the layout, describe how the inputs are encoded. How is the *unstable-gradient problem* overcome? (Max one A4 page).

9.4 Backpropagation through time. A recurrent network with two hidden neurons is shown in Figure 9.8. Write down the dynamical rules for this network. Assume that all neurons have the same activation function $g(b)$. Draw the unfolded network. Derive the update rules for the weights.

9.5 Dual dynamics for recurrent backpropagation. Show that the dynamics (9.15) admits a steady state satisfying (9.13).

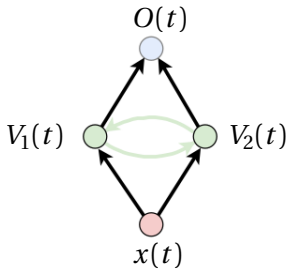


Figure 9.8: . Recurrent network used in Exercise 9.4.

PART III
LEARNING WITHOUT LABELS

Figure 9.9: Supervised learning finds decision boundaries for labeled data, like in the binary classification problem shown on the left. Unsupervised learning can find clusters in the input data (right).

Chapters 5 to 9 describe supervised learning of labeled data. The neural net is trained to reproduce the correct labels (targets) for each input pattern. Naturally this does not work when the data is unlabeled. The analysis of unlabeled data requires different methods. There are many questions where machine learning can be applied with success to large data sets of high-dimensional unlabeled data. The machine can for instance mark patterns that are typical for the given distribution, or detect outliers. Other tasks are to detect similariy, to find clusters in the data (Figure 9.9), and to determine non-linear, low-dimensional representations of high-dimensional data. More recently unsupervised learning algorithms have been used to create synthetic data, patterns that resemble those in a certain data set. One possible application is data-set augmentation for supervised learning.

Learning without labels is also called *unsupervised learning*. In this case there there are no targets that tell the network whether it has learnt correctly or not, there is no obvious function to fit, or dynamics to learn. Instead the network organises the input data in relevant ways, and this requires *redundancy* in the input data. It is sometimes said that unsupervised learning corresponds to learning without a teacher. This could be taken to mean that the network itself discovers suitable ways of organising the input data. This not accurate, because unsupervised nets operate with a pre-determined learning rule (instead of adapting the weights to minimise the difference between actual and target outputs).

Part III this book is organised as follows. Chapter 10 describes unsupervised-learning algorithms, starting with unsupervised Hebbian learning to detect familiarity and similarity of input patterns (Sections 10.1 and 10.2). Related algorithms can be used to find low-dimensional non-linear projections of high-dimensional input data (self-organised maps, Section 10.3). In Section 10.4, these algorithms are compared and contrasted with a standard unsupervised clustering algorithm, K-means clustering. Section 10.5 introduces radial-basis function nets, they learn using a hybrid algorithm with supervised and unsupervised learning. Sections and describe how to use layered feedforward nets for unsupervised learning. Chapter 11 deals with learning tasks that lie in between supervised and unsupervised learning, problems where the machine receives partial feedback on its performance in the form of a penalty or a reward. The machine learns to reproduce (reinforce) outputs

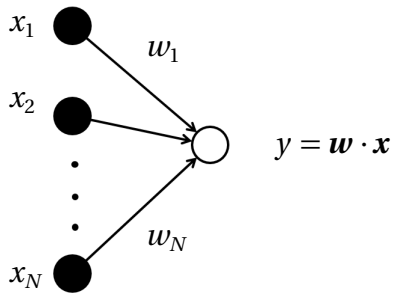


Figure 10.1: Network for unsupervised Hebbian learning, with a single linear output unit that has weight vector $\boldsymbol{w}$. The network output is denoted by y in this Chapter.

that tend to give positive rewards (*reinforcement learning*).

10 Unsupervised learning

10.1 Oja's rule

A simple example for an unsupervised-learning algorithm uses a single McCulloch-Pitts neuron with linear activation function (Figure 10.1). The neuron computes[1] $y = \boldsymbol{w} \cdot \boldsymbol{x}$ with weight vector $\boldsymbol{w} = [w_1, \ldots, w_N]^{\mathsf{T}}$. Now consider a distribution $P_{\text{data}}(\boldsymbol{x})$ of input patterns $\boldsymbol{x} = [x_1, \ldots, x_N]^{\mathsf{T}}$ with continuous-valued components x_i. Patterns are drawn from this distribution at random and fed one after another to the net. The idea is to adjust the weights $\boldsymbol{w}$ iteratively, in such a way that the output of the network becomes the larger the more frequently the input pattern occurs in $P_{\text{data}}(\boldsymbol{x})$. This is achieved by Hebb's rule:

$$\boldsymbol{w}' = \boldsymbol{w} + \delta\boldsymbol{w} \quad \text{with} \quad \delta\boldsymbol{w} = \eta\, y\, \boldsymbol{x}\,, \tag{10.1}$$

where $y = \boldsymbol{w} \cdot \boldsymbol{x}$ is the output. The rule (10.1) is also called *Hebbian unsupervised learning*, because it is reminiscent of Hebb's rule (Chapter 2). As usual, $0 < \eta \ll 1$ is the learning rate.

What can this rule learn about the input distribution $P_{\text{data}}(\boldsymbol{x})$? Since we keep adding multiples of the pattern vectors $\boldsymbol{x}$ to the weights, the magnitude of the output $|y|$ becomes the larger the more often the input pattern occurs in the distribution $P_{\text{data}}(\boldsymbol{x})$. So the most familiar pattern produces the largest output. In this way the net can detect how *familiar* certain input patterns are.

[1] In this Chapter we follow a common convention [1] and denote the output of unsupervised-learning algorithms by y.

Algorithm 7 Oja's rule

initialise weights randomly;
for $t = 1, \ldots, T$ **do**
  draw an input pattern $\boldsymbol{x}$ from $P_{\text{data}}(\boldsymbol{x})$ and apply it to the network;
  update all weights using $\delta \boldsymbol{w} = \eta\, y(\boldsymbol{x} - y\,\boldsymbol{w})$;
end for

A problem is that the weight vector continues to grow as we keep on adding increments. This means that the simple Hebbian learning rule (10.1) does not converge to a steady state. To obtain definite learning outcomes we require the network to approach a steady state. This is achieved by adding a weight-decay term with coefficient proportional to y^2 to Equation (10.1), ensuring that the weights remain normalised. The update rule with weight decay reads:

$$\delta \boldsymbol{w} = \eta\, y(\boldsymbol{x} - y\,\boldsymbol{w}). \tag{10.2}$$

Making use of $y = \boldsymbol{w} \cdot \boldsymbol{x} = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x} = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{w}$ Equation (10.2) can be rewritten in the following form:

$$\delta \boldsymbol{w} = \eta \left\{ \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \boldsymbol{w} - [\boldsymbol{w} \cdot (\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}})\boldsymbol{w}]\boldsymbol{w} \right\}. \tag{10.3}$$

This learning rule is called *Oja's rule* [127]. To see why Equation (10.3) ensures that $\boldsymbol{w}$ remains normalised, consider an analogy: a vector $\boldsymbol{q}$ that obeys the differential equation

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{q} = \mathbb{A}(t) \boldsymbol{q}. \tag{10.4}$$

For a general matrix $\mathbb{A}(t)$, the norm $|\boldsymbol{q}|$ may increase or decrease. We can ensure that $\boldsymbol{q}$ remains normalised by adding a term to Equation (10.4):

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{w} = \mathbb{A}(t) \boldsymbol{w} - [\boldsymbol{w} \cdot \mathbb{A}(t) \boldsymbol{w}]\boldsymbol{w}\,. \tag{10.5}$$

The vector $\boldsymbol{w}$ turns in the same way as $\boldsymbol{q}$, and if we set $|\boldsymbol{w}| = 1$ initially, then $\boldsymbol{w}$ remains normalised, $\boldsymbol{w} = \boldsymbol{q}/|\boldsymbol{q}|$, see Exercise 10.1. Equation (10.5) describes the dynamics of the normalised orientation vector of a small rod in turbulence [128], where $\mathbb{A}(t)$ is the matrix of fluid-velocity gradients.

Returning to Equation (10.2), we conclude that $\boldsymbol{w}$ remains normalised when the learning rate is small enough. Oja's algorithm is summarised in Algorithm 7. One draws a pattern $\boldsymbol{x}$ from the distribution $P_{\text{data}}(\boldsymbol{x})$ of input patterns, applies it to the network, and updates the weights as prescribed in Equation (10.2). This is repeated many times. In the following we denote the average over T input patterns as $\langle \cdots \rangle = \frac{1}{T} \sum_{t=1}^{T} \cdots$.

While the rule (10.2) does not have a steady state, Oja's rule does. For zero-mean input data, its steady state $\boldsymbol{w}^*$ corresponds to the principal component of the input

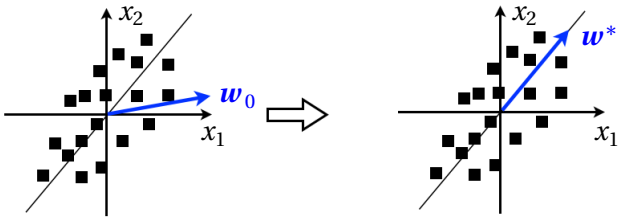


Figure 10.2: Oja's rule finds the principal component of zero-mean data (schematic). The initial weight vector is $\boldsymbol{w}_0$, the steady-state weight vector is $\boldsymbol{w}^*$.

data, as illustrated in Figure 10.2. This can be seen by analysing the steady-state condition

$$0 = \langle \delta \boldsymbol{w} \rangle_{\boldsymbol{w}^*} . \tag{10.6}$$

Here $\langle \cdots \rangle_{\boldsymbol{w}^*}$ is an average over t at fixed $\boldsymbol{w}^*$, the presumed steady state. Equation (10.6) says that the weight increments $\delta \boldsymbol{w}$ must average to zero in the steady state, to ensure that the weights neither grow nor decrease in the steady state. Equation (10.6) is a condition upon $\boldsymbol{w}^*$, using the learning rule (10.2) it can be written as:

$$0 = \mathbb{C}' \boldsymbol{w}^* - (\boldsymbol{w}^* \cdot \mathbb{C}' \boldsymbol{w}^*) \boldsymbol{w}^* \quad \text{with} \quad \mathbb{C}' = \langle \boldsymbol{x} \boldsymbol{x}^{\mathsf{T}} \rangle . \tag{10.7}$$

Equation (10.7) shows that $\boldsymbol{w}^*$ must be an eigenvector of the matrix[2] $\mathbb{C}'$, normalised to unity, $|\boldsymbol{w}^*| = 1$. But which one?

We denote the eigenvectors and eigenvalues of $\mathbb{C}'$ by $\boldsymbol{u}_\alpha$ and λ_α, and investigate the stability of $\boldsymbol{w}^* = \boldsymbol{u}_\alpha$ for different values of α by linear stability analysis, just as in Section 9.1. To this end, consider a small perturbation $\boldsymbol{\varepsilon}_t$ away from $\boldsymbol{w}^* = \boldsymbol{u}_\alpha$:

$$\boldsymbol{w}_t = \boldsymbol{u}_\alpha + \boldsymbol{\varepsilon}_t . \tag{10.8}$$

A difference to the analysis in Section 9.1 is that the dynamics is discrete in time. The perturbation at the next time step, $\boldsymbol{\varepsilon}_{t+1}$, is defined by $\boldsymbol{w}_{t+1} = \boldsymbol{u}_\alpha + \boldsymbol{\varepsilon}_{t+1}$. A second difference is that the weight update depends on the randomly chosen input pattern. In order to determine the linear stability one should iterate and then linearise the random dynamics (10.3), to see whether $\boldsymbol{\varepsilon}_t$ grows or not. However, in the limit of small learning rate it is sufficient to average over $\boldsymbol{x}$ before iterating (Exercise 10.4). To linear order in $\boldsymbol{\varepsilon}_t$ one finds:

$$\boldsymbol{\varepsilon}_{t+1} \approx \boldsymbol{\varepsilon}_t + \eta \Big[\mathbb{C}' \boldsymbol{\varepsilon}_t - 2(\boldsymbol{\varepsilon}_t \cdot \mathbb{C}' \boldsymbol{u}_\alpha) \boldsymbol{u}_\alpha - (\boldsymbol{u}_\alpha \cdot \mathbb{C} \boldsymbol{u}_\alpha) \boldsymbol{\varepsilon}_t \Big] = \mathbb{M}^{(\alpha)} \boldsymbol{\varepsilon}_t , \tag{10.9}$$

where the last equality sign defines the matrix $\mathbb{M}^{(\alpha)}$. The steady state $\boldsymbol{w}^* = \boldsymbol{u}_\alpha$ is linearly stable if all eigenvalues of $\mathbb{M}^{(\alpha)}$ have real parts smaller than unity in modulus. To determine the eigenvalues of $\mathbb{M}^{(\alpha)}$, we use the fact that $\mathbb{M}^{(\alpha)}$ has the same

[2] For zero-mean input data, $\mathbb{C}'$ equals the data-covariance matrix, Equation (6.26).

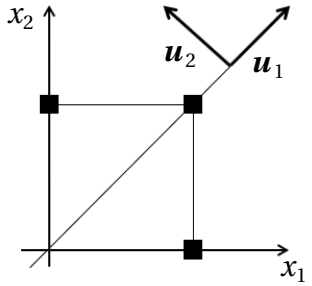


Figure 10.3: Input data with non-zero mean. Algorithm 7 converges to $\boldsymbol{u}_1$, but the principal direction is $\boldsymbol{u}_2$.

eigenvectors as $\mathbb{C}'$. Since $\mathbb{C}'$ is symmetric, these eigenvectors form an orthonormal basis, $\boldsymbol{u}_\alpha \cdot \boldsymbol{u}_\beta = \delta_{\alpha\beta}$. As a consequence, the eigenvalues of $\mathbb{M}^{(\alpha)}$ are simply given by

$$\Lambda_\beta^{(\alpha)} = \boldsymbol{u}_\beta \cdot \mathbb{M}^{(\alpha)} \boldsymbol{u}_\beta = 1 + \eta[(\lambda_\beta - \lambda_\alpha) - 2\lambda_\alpha \delta_{\alpha\beta}]. \tag{10.10}$$

Since $\mathbb{C}'$ is a positive-semidefinite matrix (its eigenvalues λ_α cannot be negative), Equation (10.10) shows that there are eigenvalues with $|\Lambda_\beta^{(\alpha)}| > 1$ unless $\boldsymbol{w}^*$ is the leading eigenvector of $\mathbb{C}'$, the one corresponding to the largest eigenvalue of $\mathbb{C}'$. This means that Algorithm 7 finds the principal component of zero-mean data, and it also implies that the algorithm maximises $\langle y^2 \rangle$ over all $\boldsymbol{w}$ with $|\boldsymbol{w}| = 1$, see Section 6.3.1. Note that $\langle y \rangle = 0$ for zero-mean input data.

Now consider inputs with non-zero mean. In this case Algorithm 7 still finds the maximal eigenvalue direction of $\mathbb{C}'$. But for inputs with non-zero mean, this direction is different from the maximal principal direction (Section 6.3). Figure 10.3 illustrates this difference. The Figure shows three data points in a two-dimensional input plane. The elements of $\mathbb{C}' = \langle \boldsymbol{x}\boldsymbol{x}^\mathsf{T} \rangle$ are

$$\mathbb{C}' = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \tag{10.11}$$

with eigenvalues and eigenvectors

$$\lambda_1 = 1\,, \quad \boldsymbol{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \lambda_2 = \tfrac{1}{3}\,, \quad \boldsymbol{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{10.12}$$

So the the maximal eigenvalue direction of $\mathbb{C}'$ is $\boldsymbol{u}_1$. To compute the principal direction of the data we must determine the data-covariance matrix $\mathbb{C}$, Equation (6.26). Its maximal-eigenvalue direction is $\boldsymbol{u}_2$, and this is the maximal principal component of the data shown in Figure 10.3.

Last but not least note that Oja's rule can be generalised to determine M principal components of zero-mean input data using M output neurons that compute $y_i =$

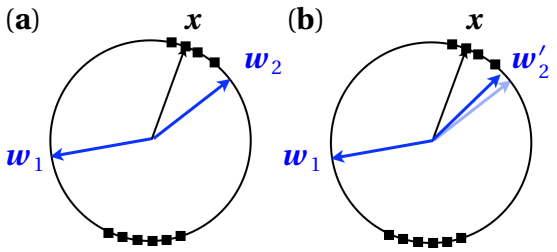


Figure 10.4: Detection of clusters by unsupervised learning. (**a**) Distribution of input patterns on the unit circle and two unit-length weight vectors initialised to random angles. The winning neuron for pattern $\boldsymbol{x}$ is the one with weight vector $\boldsymbol{w}_2$. (**b**) Updating $\boldsymbol{w}'_2 = \boldsymbol{w}_2 + \delta\boldsymbol{w}$ moves this weight vector closer to $\boldsymbol{x}$.

$\boldsymbol{w}_i \cdot \boldsymbol{x}$ for $i = 1,\ldots,M$:

$$\delta w_{ij} = \eta\, y_i\Big(x_j - \sum_{k=1}^{M} y_k\, w_{kj}\Big) \tag{10.13}$$

This is called Oja's M-rule. For $M = 1$ it simplifies to Oja's rule.

10.2 Competitive learning

Oja's M-rule (10.13) results in neurons that are activated simultaneously. Any input usually causes several outputs to assume non-zero values $y_i \neq 0$. In Sections 4.5 and 7.1 we encountered the notion of a winning neuron where the weights are trainedd in such a way that each pattern activates only a single neuron, and different patterns activate different winning neurons. This allows to represent a distribution of input patterns in a neural net.

Unsupervised learning algorithms can categorise or cluster input data in this way: similar inputs are classified to belong to the same category, and activate the same winning neuron (*competitive learning*). Figure 10.4(**a**) shows an example, input patterns on the unit circle that cluster into two distinct clusters. The idea is to find weight vectors $\boldsymbol{w}_i$ that point into the direction of the clusters. To this end we take M linear output units i with weight vectors $\boldsymbol{w}_i$, $i = 1,\ldots,M$. We feed a pattern $\boldsymbol{x}$

Algorithm 8 competitive learning (Figure 10.4)

1: initialise weights to vectors with random angles and norm $|\boldsymbol{w}_i| = 1$;
2: **for** $t = 1,\ldots,T$ **do**
3: draw a pattern $\boldsymbol{x}$ from $P_{\text{data}}(\boldsymbol{x})$ and feed it to the network:
4: find the winning neuron i_0 (smallest angle between $\boldsymbol{w}_{i_0}$ and $\boldsymbol{x}$);
5: update only the winning neuron $\delta\boldsymbol{w}_{i_0} = \eta(\boldsymbol{x} - \boldsymbol{w}_{i_0})$;
6: **end for**

from the distribution $P_{\text{data}}(\boldsymbol{x})$ and *define* the *winning neuron* i_0 as the one that has minimal angle between its weight and the pattern vector $\boldsymbol{x}$. This is illustrated in Figure 10.4(**b**), where $i_0 = 2$. Then only this weight vector is updated by adding a little bit of the difference $\boldsymbol{x} - \boldsymbol{w}_{i_0}$ between the pattern vector and the weight of the winning neuron. The other weights remain unchanged:

$$\delta \boldsymbol{w}_i = \begin{cases} \eta(\boldsymbol{x} - \boldsymbol{w}_i) & \text{for} \quad i = i_0(\boldsymbol{x}, \boldsymbol{w}_1 \ldots \boldsymbol{w}_M), \\ 0 & \text{otherwise}. \end{cases} \tag{10.14}$$

In other words, only the winning neuron is updated, $\boldsymbol{w}'_{i_0} = \boldsymbol{w}_{i_0} + \delta \boldsymbol{w}_{i_0}$. Equation (10.14) is called *competitive-learning* rule.

The learning rule (10.14) has the following geometrical interpretation: the weight of the winning neuron is drawn towards the pattern $\boldsymbol{x}$. Upon iterating (10.14), the weight vectors are drawn to *clusters* of inputs. So if the patterns are normalised as in Figure 10.4, the weights end up normalised on average, even though $|\boldsymbol{w}_{i_0}| = 1$ does not imply that $|\boldsymbol{w}_{i_0} + \delta \boldsymbol{w}_{i_0}| = 1$, in general. The algorithm for competitive learning is summarised in Algorithm 8. When weight and input vectors are normalised, then the winning neuron i_0 is the one with the largest scalar product $\boldsymbol{w}_i \cdot \boldsymbol{x}$. For linear output units $y_i = \boldsymbol{w}_i \cdot \boldsymbol{x}$ (Figure 10.1) this is simply the unit with the largest output. Equivalently, the winning neuron is the one with the smallest distance $|\boldsymbol{w}_i - \boldsymbol{x}|$. Output units with $\boldsymbol{w}_i$ that are very far away from any pattern may never be updated (*dead units*). There are several strategies to avoid this problem [1]. One possibility is to initialise the weights to directions found in the inputs.

Finally, consider the relation between the competitive learning rule (10.14) and Oja's rule (10.13). If we define

$$y_i = \delta_{i i_0} = \begin{cases} 1 & \text{for } i = i_0 \\ 0 & \text{otherwise} \end{cases} \tag{10.15}$$

then the rule (10.14) can be written in the form of Oja's M-rule:

$$\delta w_{ij} = \eta y_i \Big(x_j - \sum_{k=1}^{M} y_k w_{kj} \Big). \tag{10.16}$$

Equation (10.16) is reminiscent of Hebb's rule (Chapter 2) with weight decay.

10.3 Self-organising maps

In order to analyse high-dimensional data it is often useful to map the input patterns to a low-dimensional output space, to obtain a low-dimensional representation of

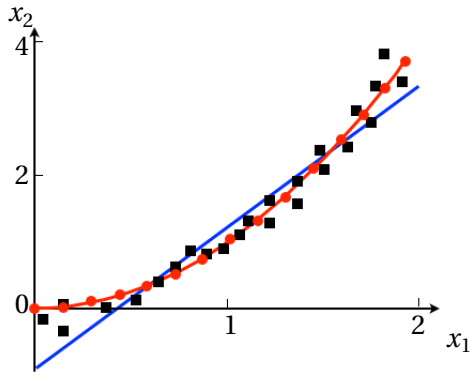


Figure 10.5: Principal-component analysis (Section 6.3) finds the linear principal direction of the data (blue line). A self-organising map can instead find the *principal manifold*, a non-linear approximation to the data (red line).

the input distribution. Principal-component analysis (6.3.1) does just that. However, it does not necessarily preserve distance. To visualise clusters or other arrangements of the input patterns, patterns that are similar or close in input space should be mapped to nearby points in output space, and patterns that are far apart should be mapped to outputs that are far from each other. Such maps are called *semantic* or *topgraphic* maps. Moreover, principal-component analysis is a linear method. As explained in Section 6.3.1, it projects the data to the space spanned by the leading eigenvectors of the correlation matrix. In many cases, however, the data may not lie in a linear subspace, as illustrated in Figure 10.5. In order to project the data onto the non-linear *principal manifold* (red line), a non-linear map is needed.

In neuroscience, the term topographic map refers to the relation between the spatial arrangement of stimuli and the activation patterns in certain parts of the mammalian brain. Similar patterns of visual stimuli on the retina, for instance, activate closeby regions in the visual cortex [129]. Similar maps represent other cognitive stimuli, auditory and sensory. The complex neural nets in the mammalian cortices hosts large numbers of such maps, arranged in a hierarchical fashion. They represent sensory stimuli in terms of spatially localised neural activation? How did this complex structure arise? One possibility is that the mappings are coded in the genetic sequence, that the connections are hard wired, so to speak. However, it is observed that such maps can change over time [130], leading to the hypothesis that the are learned, and that the genetic sequence merely encodes a set of fairly simple *learning rules*.

This motivated Kohonen [130, 131] and others to propose and analyse learning rules for topographic maps. The term *self-organising map* [132, 133] emphasises that the mapping develops in response to the stimuli it maps, that it learns in an unsupervised fashion. Kohonen's model for a non-linear self-organising map relies

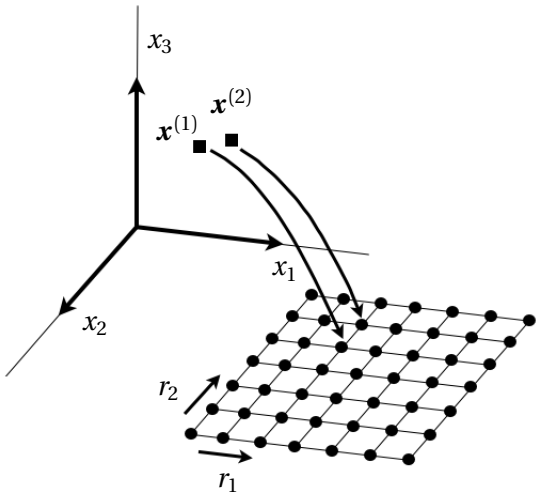


Figure 10.6: Kohonen's self-organising map. If patterns $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ are close in input space, then the two patterns activate neighbouring winning neurons in the output array (with coordinates $\boldsymbol{r} = [r_1, r_2]^{\mathsf{T}}$). Often the dimension of the output array is much lower than that of input space.

on a an ordered array of output neurons, as illustrated in Figure 10.6. The map learns to activate nearby output neurons for similar inputs. This is achieved using a competitive learning rule (10.14), similar to the learning rule described in the previous Section. In order to represent the proximity or similarity of inputs, the rule is endowed with the notion of distance in the output array, by updating not only the winning neuron, but also those that are neighbours in the output array. To this end one replaces the competitive-learning rule (10.14) by

$$\delta \boldsymbol{w}_i = \eta h(i, i_0)(\boldsymbol{x} - \boldsymbol{w}_i), \tag{10.17}$$

where $i_0(\boldsymbol{x}, \boldsymbol{w}_1 \ldots \boldsymbol{w}_M)$ is the index of the winning neuron, the one with weight vector closest to the input $\boldsymbol{x}$. The *neighbourhood function* function $h(i, i_0)$ depends on the distance of the neurons i and i_0 in the output array. The neighbourhood function has a maximum at $i = i_0$ and decreases as the distance between i and i_0 increases. One possibility is to assign decreasing values to $h(i, i_0)$ for nearest neighbours, next-nearest neighbours, and so forth. Another possibility is to use a Gaussian function of the Euclidean distance $|\boldsymbol{r}_i - \boldsymbol{r}_{i_0}|$ in the output array [1]:

$$h(i, i_0) = \exp\Big(-\tfrac{1}{2\sigma^2}|\boldsymbol{r}_i - \boldsymbol{r}_{i_0}|^2\Big). \tag{10.18}$$

Here $\boldsymbol{r}_i$ is the position of neuron i in the output array (Figure 10.6). Different normalisations of the Gaussian [2] can be subsumed in different learning rates.

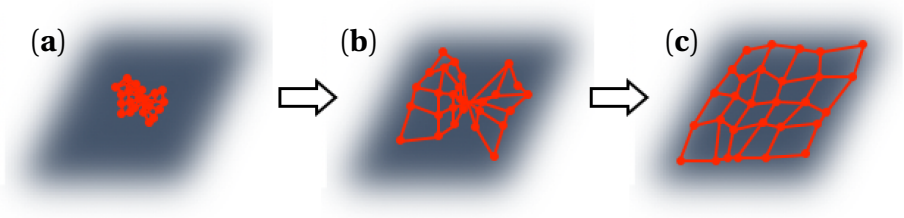


Figure 10.7: Learning a distribution $P_{\text{data}}(\boldsymbol{x})$ (gray) of two-dimensional real-valued inputs $\boldsymbol{x}$ with Kohonen's algorithm. Illustration of the dynamics of the self-organising map in terms of an *elastic net*. (**a**) Initial condition. (**b**) Intermediate stage (note the *kink*). (**c**) In the steady-state the elastic net resembles the shape defined by the input distribution $P_{\text{data}}(\boldsymbol{x})$.

Kohonen's rule has two parameters: the learning rate η, and the width σ of the neighbourhood function. Usually one adjusts these parameters as the learning proceeds. Typically one begins with large values for η and σ (*ordering phase*), and then reduces these parameters as the elastic net evolves (*convergence phase*): quickly at first and then in smaller steps, until the algorithm converges [1, 2, 130].

According to Equations (10.17) and (10.18), similar patterns activate nearby neurons in output space, and their weight vectors change in similar ways. Kohonen's rule drags the winning weight vector $\boldsymbol{w}_{i_0}$ towards $\boldsymbol{x}$, just as the competitive learning rule (10.14), but it also drags the neighbouring weight vectors along. Figure 10.7 illustrates a geometrical interpretation of Kohonen's rule [132]. We can think of the weight vectors as pointing to the nodes of an *elastic net* that has the same layout as the output array. As one feeds patterns from the input distribution, the weights are updated, causing the nodes of the network to move. This changes the shape of the elastic net until it resembles the shape defined by the distribution of input patterns. Figure 10.5 shows another example where the dimensionality of the output array (one-dimensional) is lower than that of the input space (two-dimensional). The algorithm finds a non-linear approximation to the data, the *principal manifold*. As opposed to the principal direction in principal-component analysis, the principal manifold need not be linear. Therefore it can approximate the data more precisely, leading to a smaller residual variance (Exercise 10.8).

In summary, Kohonen's algorithm learns by distributing the weight vectors of the output neurons to reflect the distribution of input patterns. In general this works well, but problems occur at the boundaries. Why this happens is quite clear (Figure 10.7): since the density of patterns outside the parallelogram is low, the elastic net cannot be drawn very close to the boundary. To analyse how the boundaries affect

learning for Kohonen's rule, consider the steady-state condition

$$\langle \delta \boldsymbol{w}_i \rangle_{\boldsymbol{w}_i^*} = \frac{\eta}{T} \sum_{t=1}^{T} h(i, i_0) \big(\boldsymbol{x}^{(t)} - \boldsymbol{w}_i^* \big) = 0 \,. \tag{10.19}$$

This condition is more complicated than it looks at first sight, because i_0 depends on the weights and on the patterns, as mentioned above. The steady-state condition (10.19) is very difficult to analyse in general. One of the reasons is that global geometric information is difficult to learn. It is usually much easier to learn local structures. This is particularly true in the *continuum limit* where we can analyse local learning progress using Taylor expansions.

The analysis of condition (10.19) in the continuum limit is due to Ritter and Schulten [134], and it is described in detail by Hertz, Krogh, and Palmer [1]. One assumes that there is a very dense net of weights, so that one can approximate $i \to \boldsymbol{r}$, $i_0 \to \boldsymbol{r}_0$, $\boldsymbol{w}_i \to \boldsymbol{w}(\boldsymbol{r})$, $h(i, i_0) \to h\big(\boldsymbol{r} - \boldsymbol{r}_0(\boldsymbol{x})\big)$, and $\frac{1}{T} \sum_t \to \int \mathrm{d}\boldsymbol{x} P_{\text{data}}(\boldsymbol{x})$. In this *continuum limit*, Equation (10.19) reads

$$\int \mathrm{d}\boldsymbol{x} \, P_{\text{data}}(\boldsymbol{x}) \, h\big(\boldsymbol{r} - \boldsymbol{r}_0(\boldsymbol{x})\big) \big[\boldsymbol{x} - \boldsymbol{w}^*(\boldsymbol{r}) \big] = 0 \,. \tag{10.20}$$

This is a condition for the steady-state learning outcome, the function $\boldsymbol{w}^*(\boldsymbol{r})$.

In the continuum limit the position $\boldsymbol{r}_0(\boldsymbol{x})$ of the winning neuron in the output array for pattern $\boldsymbol{x}$ is given by

$$\boldsymbol{w}^*(\boldsymbol{r}_0) = \boldsymbol{x} \,. \tag{10.21}$$

We use this to write Equation (10.22) as:

$$\int \mathrm{d}\boldsymbol{x} \, P_{\text{data}}(\boldsymbol{x}) \, h\big(\boldsymbol{r} - \boldsymbol{r}_0(\boldsymbol{x})\big) \big[\boldsymbol{w}^*(\boldsymbol{r}_0(\boldsymbol{x})) - \boldsymbol{w}^*(\boldsymbol{r}) \big] = 0 \,. \tag{10.22}$$

Equation (10.21) defines a mapping $\boldsymbol{r}_0(\boldsymbol{x})$ from input space to output space, the self-organised map. Assuming that this mapping is one-to-one, we change integration variable from $\boldsymbol{x}$ to $\boldsymbol{r}_0$: The neighbourhood function is sharply peaked at $\boldsymbol{r} = \boldsymbol{r}_0(\boldsymbol{x})$, and this makes it possible to evaluate the steady-state condition (10.22) approximately. The first step is to change integration variable from $\boldsymbol{x}$ to $\delta \boldsymbol{r} = \boldsymbol{r}_0(\boldsymbol{x}) - \boldsymbol{r}$:

$$\int \mathrm{d}\boldsymbol{r}_0 \, |\det \mathbb{J}| \, Q(\boldsymbol{r}_0) \, h(\boldsymbol{r} - \boldsymbol{r}_0) \big[\boldsymbol{w}^*(\boldsymbol{r}_0) - \boldsymbol{w}^*(\boldsymbol{r}) \big] = 0 \,, \tag{10.23}$$

where defined $Q(\boldsymbol{r}_0) = P_{\text{data}}\big(\boldsymbol{x}(\boldsymbol{r}_0)\big)$, and where the determinant represents the volume element of the variable transformation. Using Equation (10.21), the Jacobian $\mathbb{J}$ of the transformation has elements

$$J_{ij} = \frac{\partial \, w_i(\boldsymbol{r}_0)}{\partial \, r_j} \,. \tag{10.24}$$

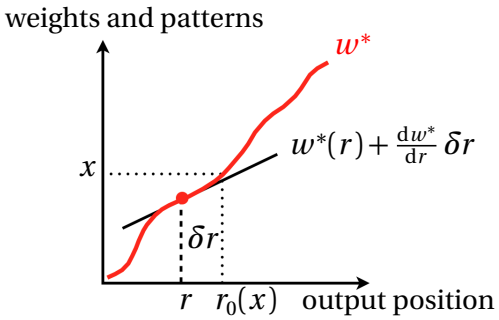


Figure 10.8: To find out how the steady-state map $w^*(r)$ varies near r, one expands w^* in δr around r, $w^*(r+\delta r)=w^*(r)+\frac{\mathrm{d}w^*}{\mathrm{d}r}\,\delta r+\frac{1}{2}\frac{\mathrm{d}^2w^*}{\mathrm{d}r^2}\delta r^2+\ldots$.

The neighbourhood function is sharply peaked at $\boldsymbol{r}=\boldsymbol{r}_0$, and this makes it possible to evaluate the steady-state condition (10.22) approximately, expanding the integrand in $\delta\boldsymbol{r}=\boldsymbol{r}_0-\boldsymbol{r}$, assuming that $\boldsymbol{w}^*(\boldsymbol{r})$ is a smooth function. This is illustrated in Figure 10.8, for one-dimensional inputs and outputs. We consider this special case in the following, not only to simplify the notation, but also because it is one of the few cases that admits mathematical analysis (Exercise 10.10). Expanding $w^*(r+\delta r)$ as shown in Figure 10.8 yields

$$w^*(r+\delta r)-w^*(r)=\frac{\mathrm{d}}{\mathrm{d}r}w^*(r)\delta r+\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}r^2}w^*(r)\delta r^2+\ldots. \tag{10.25}$$

The other factors in Equation (10.23) are expanded in a similar way:

$$J(r+\delta r)=\frac{\mathrm{d}w^*}{\mathrm{d}r}+\frac{\mathrm{d}^2w^*}{\mathrm{d}r^2}\delta r+\ldots, \tag{10.26a}$$

$$Q(r+\delta r)=P_{\mathrm{data}}(w^*)+\delta r\frac{\mathrm{d}w^*}{\mathrm{d}r}\frac{\mathrm{d}}{\mathrm{d}w}P_{\mathrm{data}}(w). \tag{10.26b}$$

Inserting these expressions into Equation (10.22), discarding terms of order higher than δr^2, and changing the integration variable to δr, one finds

$$0=w'[\tfrac{3}{2}w''P_{\mathrm{data}}(w)+(w')^2\tfrac{\mathrm{d}}{\mathrm{d}w}P_{\mathrm{data}}(w)]\int_{-\infty}^{\infty}\mathrm{d}\delta r\,\delta r^2 h(\delta r) \tag{10.27}$$

where we introduced the short-hand notation $w'=\frac{\mathrm{d}}{\mathrm{d}r}w^*(r)$, and we used that the neighbourhood function (10.18) is symmetric, $h(-\delta r)=h(\delta r)$. Since the integral in Equation (10.27) is non-zero, we must either have

$$w'=0 \quad \text{or} \quad \frac{3}{2}w''P_{\mathrm{data}}(w)+(w')^2\frac{\mathrm{d}}{\mathrm{d}w}P_{\mathrm{data}}(w)=0. \tag{10.28}$$

The first solution can be excluded because it corresponds to a singular weight distribution [see Equation (10.30)] that does not contain any geometrical information about the input distribution P_{data}. The second solution gives

$$\frac{w''}{w'}=-\frac{2}{3}\frac{w'\frac{\mathrm{d}}{\mathrm{d}w}P_{\mathrm{data}}(w)}{P_{\mathrm{data}}(w)} \tag{10.29}$$

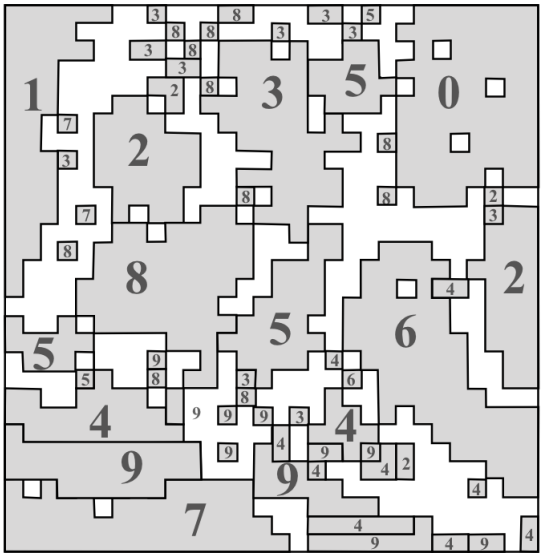


Figure 10.9: Clustering of hand-written digits (MNIST data set) with a self-organising map with a 30×30 output array. In the shaded regions the outputs are quite certain: here the winning neurons are activated by the indicated digit in 80% of the cases. The white regions correspond to outputs where the majority digit appears in less than 80% of the cases, or to outputs that are never activated, or only once. Schematic, based on simulations performed by Juan Arango.

In other words, $\frac{\mathrm{d}}{\mathrm{d}r}\log|w'| = -\frac{2}{3}\frac{\mathrm{d}}{\mathrm{d}r}\log P_{\text{data}}(w)$. This means that $|w'| \propto [P_{\text{data}}(w)]^{-\frac{2}{3}}$. So the distribution ϱ of output weights is

$$\varrho(w) \equiv \left|\frac{\mathrm{d}r}{\mathrm{d}w}\right| = \frac{1}{|w'|} = [P_{\text{data}}(w)]^{\frac{2}{3}}. \tag{10.30}$$

This tells us that the self-organising map learns the input distribution in the following way: the distribution of output weights in the steady state reflects the distribution of input patterns. Equation (10.30) tells us that the two distributions are not equal (equality would have been a perfect outcome). The distribution of weights is instead proportional to $\left[P_{\text{data}}(w)\right]^{\frac{2}{3}}$. Little is known in higher dimensions, but the general idea is that the elastic net has difficulties reaching the corners and edges of the domain where the input distribution is non-zero.

The output of a self-organising map can be interpreted in different ways. For a low-dimensional inputs and outputs, one can simply plot the map $w^*(\boldsymbol{r})$, as in Figure 10.6. Dense regions of weights point to regions in input space with a high density of inputs. Often the output dimension is taken to be much lower than the dimension of input space. In this case the self-organising map performs nonlinear *dimensionality reduction*, and it can be used to find clusters in high-dimensional input data [135].

The analysis proceeds in two steps. First, one runs Kohonen's algorithm until the map has converged to a steady state. Second, one feeds all inputs into the net, and for each input one determines the location of the winning neuron in the output array. The spatial activation patterns in the output array represent clusters of similar inputs. This is illustrated in Figure 10.9, which shows how a self-organising map represents handwritten digits from the MNIST data set. To reveal the semantic map, the Figure labels clusters of outputs that correspond to the same digits (as determined by the labels in the training set). We see that the self-organised map groups the same digits together, but it has some difficulty distinguishing the digits 3 and 8, and also 4 and 9.

10.4 K-means clustering*

Sections 10.2 and 10.3 described different ways of finding clusters in input data. In particular, it was shown how self-organising maps can find clusters in high-dimensional input data, and represent them in a low-dimensional, non-linear projection. A frequently used alternative unsupervised-learning algorithm for this purpose is K-means clustering. Let us compare and contrast this algorithm with the self-organising map. The goal is to cluster p N-dimensional inputs $\boldsymbol{x}^{(\mu)}$ into K clusters, usually $K \ll p, N$. A solution of this task is a mapping $k(\mu)$ that associates each input $\boldsymbol{x}^{(\mu)}$ with one of the clusters $k = 1, \ldots, K$. The function $k(\mu)$ is determined by minimising the energy function

$$H(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) = \frac{1}{2} \sum_{k=1}^{K} \Big(\sum_{\mu \,|\, k(\mu)=k} |\boldsymbol{x}^{(\mu)} - \boldsymbol{w}_k|^2 \Big). \tag{10.31}$$

The second sum is over all values of μ that satisfy $k(\mu) = k$. The vector $\boldsymbol{w}_k$ is the mean of all pattern vectors in cluster k, and the expression in the parentheses is the variance associated with this cluster:

$$\sigma_k^2 = \sum_{\mu \,|\, k(\mu)=k} |\boldsymbol{x}^{(\mu)} - \boldsymbol{w}_k|^2 . \tag{10.32}$$

In other words, H measures the sum of the cluster variances σ_k^2. An optimal solution to the clustering problem corresponds to a local minimum of H. To determine the cluster vectors $\boldsymbol{w}_k$ and the corresponding variances σ_k^2 one starts from an initial guess for the encoding, $k(\mu)$. For each cluster, one adjusts $\boldsymbol{w}_k$ to minimise the cluster variance

$$\underset{\boldsymbol{w}_k}{\arg\min} = \sum_{\mu | k(\mu)=k} |\boldsymbol{x}^{(\mu)} - \boldsymbol{w}_k|^2 . \tag{10.33}$$

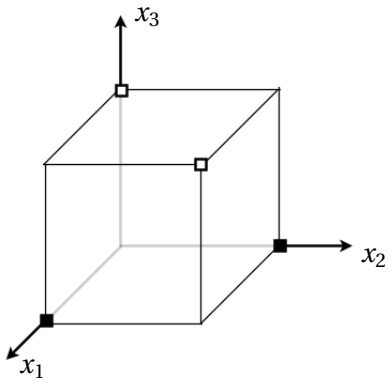


Figure 10.10: The XOR problem can be solved by embedding into a three-dimensional input space.

Given the vectors $\boldsymbol{w}_k$ one optimises the encoding function

$$k(\mu) = \arg\min_{1\le k\le K} |\boldsymbol{x}^{(\mu)} - \boldsymbol{w}_k|^2 \tag{10.34}$$

These steps are repeated until a satisfactory solution is found. The solution is not unique, usually one tries different random initialisations and takes the solution with minimal H.

All three algorithms, competitive learning, the self-organised map, and K-means clustering move weight vectors towards clusters in input space. A difference between the self-organised map and the other two algorithms is that the self-organised map uses a neighbourhood function, so that similar inputs activate closeby neurons in the output array, and update their weight vectors in similar fashion. In this way, a self-organised map with a large output array can find a smooth parameterisation of the principal manifold. Self-organised maps with only a small number of output neurons are not so different from K-means clustering.

10.5 Radial basis functions

Problems that are not linearly separable can be solved by perceptrons with hidden layers, as we saw in Chapter 5. Figure 5.12, for example, shows a piecewise linear decision boundary that can be parameterised by hidden neurons.

Another approach is to *map* the coordinates of input space so that the problem becomes linearly separable. It is usually easier to separate patterns in higher dimensions. To see this consider the XOR problem. It is not linearly separable in two-dimensional input space. The problem becomes separable when we *embed* the points in three-dimensional space, for instance by assigning $x_3 = 0$ to the $t = +1$

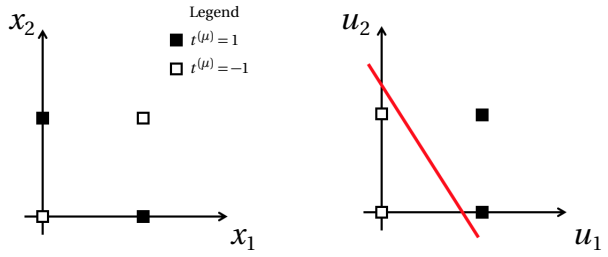


Figure 10.11: Left: input plane for the XOR function (Figure 5.8). The problem is not linearly separable. Right: in the u_1-u_2 plane the problem is linearly separable.

patterns and $x_3 = 1$ to the $t = -1$ patterns (Figure 10.10). This example illustrates why it is often helpful to map input space to a higher-dimensional space – because it is more likely that the resulting problem is linearly separable.

It may also be possible to achieve separability by a non-linear transformation of input space to a space of the same dimension. Figure 10.11 shows how the XOR problem can be transformed into a linearly separable problem by the transformation

$$u_1(\boldsymbol{x}) = (x_2 - x_1)^2 \quad \text{and} \quad u_2(\boldsymbol{x}) = x_2\,. \tag{10.35}$$

The Figure shows the non-separable problem in the x_1-x_2 plane, and in the new coordinates u_1 and u_2. Since the problem is linearly separable in the u_1-u_2 plane we can solve it by a single McCulloch-Pitts neuron with weights $\boldsymbol{W}$ and threshold Θ, parameterising the decision boundary as $\boldsymbol{W} \cdot \boldsymbol{u}(\boldsymbol{x}) = \Theta$. In fact, one does not need the threshold Θ because the function $\boldsymbol{u}$ can have a constant part. For instance, we could choose $u_1(\boldsymbol{x}) = 2(x_2 - x_1)^2 - 1$. In the following we therefore set $\Theta = 0$.

We expect that it should be easier to achieve linear separability the higher the *embedding dimension* is. This statement is quantified by Cover's theorem, discussed in Section 5.4. Consider a set $\boldsymbol{u}(\boldsymbol{x}) = [u_1(\boldsymbol{x}), \ldots, u_m(\boldsymbol{x})]^{\mathsf{T}}$ of m polynomial functions of finite order that embed N-dimensional input space in an m-dimensional space. Then the probability that a problem with p points $\boldsymbol{x}^{(\mu)}$ in N-dimensional input space is separable by a polynomial decision boundary is given by $P(p, m)$ [Equation (5.28)] [2, 55]. Note that this probability is independent of the dimension N of input space.

The question is of course how to find the non-linear mapping $\boldsymbol{u}(\boldsymbol{x})$. One possibility is to use radial basis functions. The idea behind radial basis-function nets is to parameterise the functions $u_j(\boldsymbol{x})$ in terms of weight vectors $\boldsymbol{w}_j$, and to use an unsupervised-learning algorithm (Chapter 10) to find weights that separate the input data. A common choice [2] are *radial basis functions* of the form:

$$u_j(\boldsymbol{x}) = \exp\Big(-\frac{1}{2s_j^2}\left|\boldsymbol{x} - \boldsymbol{w}_j\right|^2\Big). \tag{10.36}$$

Note that these functions are not of the finite-order polynomial form that was as-

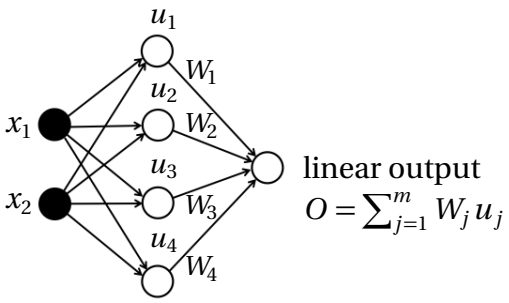


Figure 10.12: Radial basis-function network for $N = 2$ inputs and $m = 4$ radial basis functions (10.36). The output neuron has a linear activation function, weights $\boldsymbol{W}$, and zero threshold.

sumed above. So strictly speaking we cannot invoke Cover's theorem. In practicethe mapping $u_j(\boldsymbol{x})$ works nevertheless quite well. The parameters s_j parameterise the *widths* of the radial basis functions. In the simplest version of the algorithm they are set to unity. When they are allowed to vary they reflect different widths of the radial basis functions. Hertz, Krogh, and Palmer [1] discuss radial basis-function nets with *normalised* radial basis functions

$$u_j(\boldsymbol{x}) = \frac{\exp\left(-\frac{1}{2s_j^2}\left|\boldsymbol{x}-\boldsymbol{w}_j\right|^2\right)}{\sum_{k=1}^{m}\exp\left(-\frac{1}{2s_j^2}\left|\boldsymbol{x}-\boldsymbol{w}_j\right|^2\right)}. \tag{10.37}$$

Other choices for radial basis functions are given by Haykin [2].

Figure 10.12 shows a radial basis-function network for $N = 2$ and $m = 4$. The four neurons in the hidden layer stand for the four radial basis functions (10.36) that map the inputs to four-dimensional $\boldsymbol{u}$-space. The network looks like a perceptron (Chapter 5). But here the hidden layers work in a different way. Perceptrons have hidden McCulloch-Pitts neurons that compute *non-local* outputs $\sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} - \theta)$. The output of radial basis functions $u_j(\boldsymbol{x})$, by contrast, is *localised* in input space [Figure

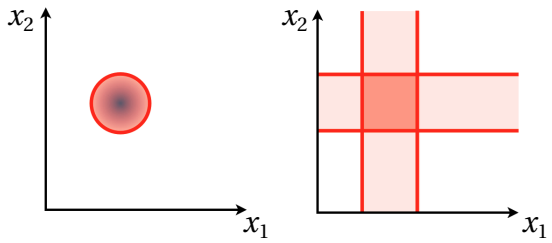


Figure 10.13: Comparison between radial-basis function network and perceptron. Left: the output of a radial basis function is localised in input space. Right: to achieve a localised output with sigmoid units one needs two hidden layers (Figure 7.5).

10.13 (left)]. We saw in Section 7.1 how to make localised basis functions out of McCulloch-Pitts neurons with sigmoid activation functions $\sigma(b)$, but one needs two hidden layers to do that [Figure 10.13 (right)].

Radial basis functions produce localised outputs with a single hidden layer, and this makes it possible to divide up input space into localised regions, each corresponding to one radial basis function. Imagine for a moment that we have as many radial basis functions as input patterns. In this case we can simply take $\boldsymbol{w}_\nu = \boldsymbol{x}^{(\nu)}$ for $\nu = 1, \ldots, p$. Then the classification problem can be written as

$$\sum_\mu U_{\nu\mu} W_\mu = t^{(\nu)}, \tag{10.38}$$

with $U_{\nu\mu} = u_\nu(\boldsymbol{x}^{(\mu)})$. Equation (10.38) uses that the output unit is linear and computes $O^{(\mu)} = \boldsymbol{W} \cdot \boldsymbol{u}(\boldsymbol{x}^{(\mu)})$ (Figure 10.12). If all patterns are pairwise different, $\boldsymbol{x}^{(\mu)} \neq \boldsymbol{x}^{(\nu)}$ for $\mu \neq \nu$, then the matrix $\mathbb{U}$ is invertible [2]. In this case the solution of the classification problem reads

$$W_\mu = \sum_\nu [\mathbb{U}^{-1}]_{\mu\nu} t^{(\nu)}, \tag{10.39}$$

where $\mathbb{U}$ is the symmetric $p \times p$ matrix

In practice one can get away with fewer radial basis functions by choosing their weights to point in the directions of clusters of input data. To this end one can use unsupervised competitive learning (Algorithm 9), where the *winning neuron* is defined to be the one with largest u_j. How are the widths s_j determined? The width s_j of radial basis function $u_j(\boldsymbol{x})$ is taken to be equal to the minimum distance between $\boldsymbol{w}_j$ and the centers of the surrounding radial basis functions. Once weights and widths of the radial basis functions are found, the weights of the output neuron are determined by minimising

$$H = \frac{1}{2} \sum_\mu \left(t^{(\mu)} - O^{(\mu)} \right)^2 \tag{10.40}$$

with respect to $\boldsymbol{W}$. An approximate solution can be obtained by stochastic gradient descent on H keeping the parameters of the radial basis functions fixed. Cover's theorem indicates that the problem is more likely to be separable if the embedding dimension m is higher.

In summary, radial basis-function nets are similar to the perceptrons described in Chapters 5 to 7, in that they are feed-forward nets designed to solve classification problems. A fundamental difference is that the parameters of the radial basis functions are determined by unsupervised learning, whereas perceptrons are

Algorithm 9 radial basis functions

initialise the weights w_{jk} independently randomly from $[-1,1]$;
set all widths to $s_j = 0$;
for $t = 1,\ldots,T$ **do**
    feed randomly chosen pattern $\boldsymbol{x}^{(\mu)}$;
    determine winning neuron j_0: $u_{j_0} \geq u_j$ for all values of j;
    update widths: $s_j = \min_{j\neq k} |\boldsymbol{w}_j - \boldsymbol{w}_k|$;
    update only winning neuron: $\delta\boldsymbol{w}_{j_0} = \eta(\boldsymbol{x}^{(\mu)} - \boldsymbol{w}_{j_0})$;
end for

trained using supervised learning for *all* units. While McCulloch-Pitts neurons compute weights to minimise their output from given targets, the radial basis functions compute weights by maximising u_j as a function of j. The algorithm for finding the weights of the radial basis functions is summarised in Algorithm 9. Further, as opposed to the deep nets from Chapter 7, radial basis-function nets have only one hidden layer, and a linear output neuron. Radial-basis function nets learn using a *hybrid* scheme: unsupervised learning for the parameters of the radial basis functions, and supervised learning for the weights of the output neuron.

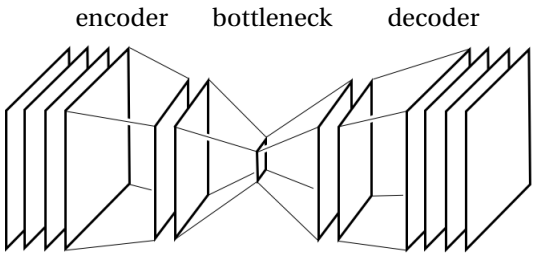


Figure 10.14: Autoencoder (schematic). Both encoder and decoder consist of a number of fully connected or convolutional layers (depicted as squares). In the layout shown, the bottleneck consists of a layer with very few neurons. Sparse autoencoders have bottlenecks with many neurons, but only few are activated.

10.6 Autoencoders*

Multi-layer perceptrons, layered feed-forward networks, were developed for supervised learning, as described in Part II. More recently, such layouts have been used for unsupervised learning. Examples are *autoencoders* and *generative adversarial nets*.

Autoencoders employ layered feed-forward networks for unsupervised learning of an unlabeled data set of input patterns, using the inputs as targets $\boldsymbol{t}^{(\mu)} = \boldsymbol{x}^{(\mu)}$. In the simplest case there is one hidden layer and an output layer (Figure 10.14). One adjusts weights and thresholds by backpropagation until the network learns to approximate the inputs, $\boldsymbol{O}(\boldsymbol{x}^{(\mu)}) = \boldsymbol{x}^{(\mu)}$. The idea is that the states of the hidden layer may encode interesting properties of the patterns. Usually the number of neurons is much smaller than the number of pattern bits, in this case the states of the hidden neurons learn a low-dimensional (*compressed*) representation of the input. Since the activation functions are non-linear, autoencoders can perform non-linear dimensionality reduction, like self-organised maps (Section 10.3).

Sparse autoencoders have a large number of neurons in the bottleneck, possibly more than the number of pattern bits, but only a small number of bottleneck neurons are allowed to be active at the same time. The idea is that sparse representations of input data are more robust than dense ones, and generalise more reliably. At least sparse, high-dimensional representations of binary classification problems are more likely to be linearly separable (Section 5.4). There are different ways of enforcing sparsity, for instance using L_1- or L_2-regularisation (Section 7.7.1). An alternative is to use the Kullback-Leibler divergence [136] to ensure that the average activation of each bottleneck neuron (with sigmoid activation function),

$$f_j = \frac{1}{p}\sum_{\mu=1}^{p} \sigma(b_j^{(\mu)}), \tag{10.41}$$

remains small, $f_j = f \ll 1$. Here $f > 0$ is the sparsity parameter. To ensure spar-

sity, one adds $\lambda D_{\rm KL}(f, f_j)$ to the energy function, where $D_{\rm KL}$ is the Kullback-Leibler divergence (Section 4.5), and λ is a Lagrange multiplier.

Variational autoencoders [137, 138] have layouts similar to the one shown schematically in Figure 10.14, but their purpose is quite different from that of the autoencoders described above, which are used for non-linear dimensional reduction. Variational autoencoders are generative models. Just like restricted Boltzmann machines (Section 4.5) they can be used approximate a data distribution of inputs $P_{\rm data}(\boldsymbol{x})$, and to sample from it. As an example consider the MNIST data set of handwritten digits. The patterns define a data distribution that encodes the properties of the digits in terms of covariances and higher-order correlationts. The question is how to generate new digits from this distribution, different from those in the data set, yet sharing their defining properties. In other words, how can a machine learn to generate images that look like handwritten digits?

The idea of variational autoencoders is to represent the data distribution in terms of a Gaussian distribution of *latent variables* $\boldsymbol{z}$, using the fact that one can approximate any given data distribution $P_{\rm data}(\boldsymbol{x})$ in terms of $P_{\rm L}(\boldsymbol{z})$ by a suitable non-linear transformation $\boldsymbol{f}(\boldsymbol{z})$. This transformation is learned by a multilayer perceptron by backpropagation. An essential difference to the algorithms described in Part II is that the network learns the mean and the variance of a Gaussian distribution. When the variance tends to zero, the algorithm reduces to stochastic gradient descent (Algorithm 3). Given the Gaussian distribution $P_{\rm L}(\boldsymbol{z})$ of the latent variables the goal is to maximise the log-likelihood (Section 4.5):

$$\mathscr{L} = \log P(\boldsymbol{x}) = \log \int \mathrm{d}\boldsymbol{z}\; P(\boldsymbol{x}|\boldsymbol{z}) P_{\rm L}(\boldsymbol{z}) \,. \tag{10.42}$$

Here $P(\boldsymbol{x}|\boldsymbol{z})$ is the probability to generate $\boldsymbol{x}$ given $\boldsymbol{z}$. This distribution is assumed to be Gaussian with mean $\boldsymbol{\mu}_P(\boldsymbol{z}) = \boldsymbol{f}(\boldsymbol{z})$, and correlation matrix $\mathbb{C}_P$.

The encoder represents $P(\boldsymbol{x}|\boldsymbol{z})$ in terms of a multi-layer perceptron. Its weights and thresholds to are determined to maximise $\mathscr{L}$ by gradient ascent. To this end we must find an efficient way of computing $\mathscr{L}$ and its gradients. One possibility is Monte-Carlo sampling, but this is not very efficient because most values of $\boldsymbol{z}$ drawn from $P_{\rm L}(\boldsymbol{z})$ result in unlikely patterns $\boldsymbol{x}$, with only negligible contributions to $\mathscr{L}$. To get around this problem one needs to know, which values of $\boldsymbol{z}$ are likely to produce a given pattern $\boldsymbol{x}$. The idea is to learn a second distribution $Q(\boldsymbol{z}|\boldsymbol{x})$ of values likely to produce $\boldsymbol{z}$ given $\boldsymbol{x}$. In other words, we need to minimise the difference between $Q(\boldsymbol{z}|\boldsymbol{x})$ and the unknown exact distribution $P(\boldsymbol{z}|\boldsymbol{x})$,

$$D_{\rm KL}[Q(\boldsymbol{z},\boldsymbol{x}), P(\boldsymbol{z}|\boldsymbol{x})] = \langle \log Q - \log P(\boldsymbol{z}|\boldsymbol{x}) \rangle_Q \,. \tag{10.43}$$

The trick is to rewrite this expression using $P(\boldsymbol{z}|\boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{z}) P_{\rm L}(\boldsymbol{z}) / P(\boldsymbol{x})$:

$$\mathscr{L} - D_{\rm KL}[Q(\boldsymbol{z}|\boldsymbol{x})|P(\boldsymbol{z}|\boldsymbol{x})] = \langle \log P(\boldsymbol{x}|\boldsymbol{z}) - D_{\rm KL}[Q(\boldsymbol{z}|\boldsymbol{x})|P_{\rm L}(\boldsymbol{z})] \rangle_Q \,, \tag{10.44}$$

and to recognise that the l.h.s of Equation (10.44) is a suitable target function to maximise. We want to maximise $\mathscr{L}$ subject to the constraint that the unknown function $Q(\boldsymbol{z}|\boldsymbol{x})$ approximates the probability of $\boldsymbol{z}$ encoding the pattern $\boldsymbol{x}$. Usually one takes $Q(\boldsymbol{z}|\boldsymbol{x})$ to be a Gaussian with mean $\boldsymbol{\mu}_Q$ and correlation matrix $\mathbb{C}_Q$. The task is then to determine the parameters $\boldsymbol{\mu}_P, \mathbb{C}_P, \boldsymbol{\mu}_Q$, and $\mathbb{C}_Q$ of the Gaussian distributions $P(\boldsymbol{x}|\boldsymbol{z})$ and $Q(\boldsymbol{z}|\boldsymbol{x})$. This is not as straightforward as it may seem, because the target function (10.44) involves an average over random latent variables. The stochastic-gradient algorithm cannot be used in the form described in Algorithm 3, because it is not designed to deal with stochastic output (Exercise 11.1).

The solution is to use *stochastic backpropagation* [138], a generalisation of Algorithm 3 for target functions of the form (10.44), expressed as average over Gaussians. Stochastic backpropagation allows to learn the parameters of these Gaussians, using the relation [138]

$$\frac{\partial}{\partial w_{mn}} \langle F(\boldsymbol{z}) \rangle_Q = \langle \boldsymbol{b} \cdot \tfrac{\partial}{\partial w_{mn}} \boldsymbol{\mu}_Q + \frac{1}{2} \mathrm{tr} \mathbb{A} \tfrac{\partial}{\partial w_{mn}} \mathbb{C} \rangle_Q \,, \tag{10.45}$$

where $\boldsymbol{b}$ and $\mathbb{A}$ are the gradient and the Hessian of the function $F(\boldsymbol{z})$. When the correlation matrix $\mathbb{C}$ is constant, this rule is equivalent to standard backpropagation, Algorithm 3. A further difficulty is that Equation (10.45) requires the gradients and the Hessian of the target function. How to deal with this problem is described in Ref. [138]. Once the parameters of the Gaussian distributions P and Q are determined, one can sample from the distribution $P_{\mathrm{l}}(\boldsymbol{z})$ and and apply the *decoder* $P(\boldsymbol{x}|\boldsymbol{z})$.

Variational autoencoders are used for different purposes. Ref. [139] suggests to employ a variational autoencoder for active learning (Section 7.9). The idea is to represent the input distribution in terms of lower-dimensional latent variables, and to use K-means clustering (Section 10.4) to find identify groups of patterns that should be labeled. Variational autoencoders have also been used for outlier detection [140] and language generation [141].

Generative adversarial networks [142] are generative models based on learning rules similar to that described above for variational autoencoders, but there are some differences in detail. Generative adversarial networks consist of two multilayer perceptrons, a generator and a discriminator. The generative network produces new outputs from a given data distribution (*fakes*), and the task of the disriminator is to classify these outputs into two classes: real or fake data. Generator and discriminator are trained together. The weights of the generator are adjusted to maximise the classification error of the discriminator, while those of the discriminator are trained to minimise this error [143].

10.7 Summary

The unsupervised-learning algorithms described in Sections 10.1 and 10.2 are based on Hebb's rule. These algorithms can learn different features of unlabeled input data: they can detect the familiarity of inputs, perform principal-component analysis, and identify clusters in the input data. Self-organised maps also rely on Hebb's rule. An important difference is that the outputs are arranged in an array, and that output neurons that are closeby in the output array are updated in similar ways. Self-organised maps can therefore represent topographic and semantic maps, where closeby or similar inputs are mapped to nearby outputs. When the dimension of the output array is much lower than the input dimension, self-organised maps perform non-linear dimensional reduction. Radial basis-function nets are classifiers, just like multilayer perceptrons. Their output neurons are trained in the same way, using labeled input data. However, the decision boundaries of radial basis-function nets are polynomial functions (not just hyperplanes), and their parameters are determined by unsupervised learning. Autoencoders are multilayer perceptrons. They can learn to encode non-linear features of unlabeled input data by using the input patterns as targets. Finally, generative adversarial nets do not require labeled inputs, so they can be considered unsupervised-learning machines. They are used to generate synthetic data in order to expand training sets for supervised learning, and pose an ethical dilemma because they can be used to generate *deepfakes* [144], manipulated videos that put someone else's words into a well-known person's mouth.

Similar and sometimes equivalent algorithms are used in Mathematical Statistics (K-means clustering) and Bioinformatics (`structure` [145]) where large data sets must be analysed, such as Human sequence data (HGDP) [146]. At any rate, the simple algorithms described in this Chapter provide a proof of concept: how machines can learn without labels. In addition there is one significant application of unsupervised learning, or rather semi-supervised learning: *reinforcement learning* allows a machine to learn from partial feedback on its output, in the form of a penalty or reward. This is discussed in the next Chapter.

10.8 Further reading

The primary source for Sections 10.1 and 10.2 is the book by Hertz, Krogh, and Palmer [1]. A good reference for self-organising maps is Kohonen's book [131]. Radial-basis function nets are discussed by Haykin in Chapter 5 of his book [2]. It has been argued that radial-basis function nets do not generalise as well as perceptrons do [147]. To solve this problem, Poggio and Girosi [148] suggested to determine the parameters $\boldsymbol{w}_j$ of the radial basis function by supervised learning, using stochastic gradient

descent.

10.9 Exercises

10.1 Continuous Oja's rule. Using the ansatz $\boldsymbol{w} = \boldsymbol{q}/|\boldsymbol{q}|$ show that Equations (10.4) and (10.5) describe the same angular dynamics. The difference is just that $\boldsymbol{w}$ remains normalised to unity, whereas the norm of $\boldsymbol{q}$ may increase of decrease. See Ref. [128].

10.2 Data-covariance matrix. Determine the data-covariance matrix and the principal direction for the data shown in Figure 10.3.

10.3 Oja's rule. The aim of unsupervised learning is to construct a network that learns the properties of a distribution $P_{\text{data}}(\boldsymbol{x})$ of input patterns $\boldsymbol{x} = [x_1, \ldots, x_N]^\mathsf{T}$. Consider a network with one linear output that computes $y = \sum_{j=1}^{N} w_j x_j$. Show that Oja's learning rule $\delta w_j = \eta y(x_j - y\, w_j)$ has the stable steady state $\boldsymbol{w}^*$ corresponding to the leading eigenvector of the matrix $\mathbb{C}'$ with elements $C'_{ij} = \langle x_i x_j \rangle$. Here $\langle \cdots \rangle$ denotes the average over $P_{\text{data}}(\boldsymbol{x})$.

10.4 Linear stability analysis for Oja's rule. Iterate the stochastic dynamics (10.3) near a fixed point $\boldsymbol{w}^*$, linearise, and average the result over a random sequence of patterns $\boldsymbol{x}$. Expand the result to leading order in the learning rate η to show that the linear stability of $\boldsymbol{w}^*$ to this order is determined by Equation (10.9).

10.5 Competitive learning for binary patterns. A competitive learning rule for binary patterns with 0/1 bits reads $\delta w_{ij} = \eta(x_j / \sum_{k=1}^{N} x_k - w_{ij})$ for the winning neuron $i = i_0$, and $\delta w_{ij} = 0$ otherwise. Show that the steady-state weight vectors $\boldsymbol{w}_{i_0}$ have positive components and are normalised as $\sum_{k=1}^{N} w_{i_0,k} = 1$.

10.6 Self organising map. Explain the meaning of the parameter σ in the neighbourhood function in Kohonen's learning rule, Equations (10.17) and (10.18). Discuss the nature of the update rule in the limit of $\sigma \to 0$. Discuss and explain the implementation of Kohonen's algorithm in a computer program. In the discussion, refer to and explain the following terms: *output array, neighbourhood function, ordering phase, convergence phase, kinks.*

10.7 Self-organising map. Write a computer program that implements Kohonen's algorithm with a two-dimensional output array, to learn the properties of a two-dimensional input distribution that is uniform inside an equilateral triangle with sides of unit length, and zero outside. *Hint*: to generate this distribution, sam-

ple at least 1000 points uniformly distributed over the smallest square that contains the triangle, and then accept only points that fall inside the triangle. Increase the number of weights and study how the two-dimensional density of weights near the boundary depends on the distance from the boundary.

10.8 Principal manifolds. Create a data set like the one shown in Figure 10.5, using $x_2 = x_1^2 + r$ where r is a Gaussian random number with mean zero and variance $\sigma_r^2 = 0.01$. Determine the principal component (blue line) of the data set (Section 6.3). Use Kohonen's algorithm to find a better approximation to the data, the *principal manifold* (red line). For both cases determine the variance of data that remains unexplained.

10.9 Iris data set. Write a computer program that combines a two-dimensional self-organising map with a simple classifier to classify the Iris data set (Figure 5.1).

10.10 Steady state of two-dimensional Kohonen algorithm. Repeat the analysis of Equation (10.23) for a two-dimensional self-organising map. (a) Derive the equivalent of Equation (10.27) and determine a relation between the weight density ϱ and P_{data} assuming that the data distribution factorises $P_{\text{data}}(\boldsymbol{w}) = f(w_1)g(w)_2$ [134]. (b) Assume that $\boldsymbol{w}(\boldsymbol{r}) = u + iv$ can be written as an analytic function of $\boldsymbol{r} = x + iy$ and derive a relation between ρ and P_{data}.

10.11 Radial basis functions for XOR. Show that the two-dimensional Boolean XOR problem with 0/1 inputs can be solved using the two radial basis functions $u_1(\boldsymbol{x}^{(\mu)}) = \exp(-|\boldsymbol{x}^{(\mu)} - \boldsymbol{w}_1|^2)$ and $u_2(\boldsymbol{x}^{(\mu)}) = \exp(-|\boldsymbol{x}^{(\mu)} - \boldsymbol{w}_2|^2)$ with $\boldsymbol{w}_1 = (1,1)^\mathsf{T}$ and $\boldsymbol{w}_2 = [0,0]^\mathsf{T}$. Draw the positions of the four input patterns in the transformed space with coordinates u_1 and u_2.

10.12 Radial basis functions. Table 10.1 describes a classification problem. Show that this problem can be solved as follows. Transform the inputs x_1, x_2, x_3 to two-dimensional coordinates u_1, u_2 using radial basis functions:

$$u_1 = \exp(-\tfrac{1}{4}|\boldsymbol{x} - \boldsymbol{w}_1|^2), \text{ with } \boldsymbol{w}_1 = [\text{-}1,1,1]^\mathsf{T}, \tag{10.46}$$

$$u_2 = \exp(-\tfrac{1}{4}|\boldsymbol{x} - \boldsymbol{w}_2|^2), \text{ with } \boldsymbol{w}_2 = [1,1,\text{-}1]^\mathsf{T}, \tag{10.47}$$

with $\boldsymbol{x} = [x_1, x_2, x_3]^\mathsf{T}$. Plot the positions of the eight input patterns in the u_1-u_2-plane. *Hint*: to compute u_j use the following approximations: exp(-1) ≈ 0.37, exp(-2) ≈ 0.14, exp(-3) ≈ 0.05. The transformed data is used as input to a simple perceptron (no hidden layers) with one output unit $O^{(\mu)} = \text{sgn}\big(\sum_{i=1}^{2} W_j u_j{}^{(\mu)} - \Theta\big)$. Draw a decision boundary in the u_1-u_2-plane and determine the corresponding weight vector $\boldsymbol{W}$ and the threshold Θ.

x_1	x_2	x_3	t
-1	-1	-1	1
-1	-1	1	1
-1	1	-1	1
-1	1	1	-1
1	-1	-1	1
1	-1	1	1
1	1	-1	-1
1	1	1	1

Table 10.1: Inputs and targets for Exercise 10.12.

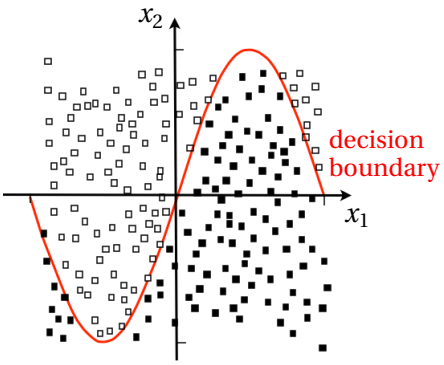


Figure 10.15: Non-linear decision boundary $x_2 = \sin(2\pi x_1)$ for a non-linearly separable binary classification problem defined on the square $-1 \leq x_1 \leq 1$ and $-1 \leq x_2 \leq 1$. Exercise 10.13.

10.13 A two-dimensional binary classification problem. Figure 10.15 illustrates a binary classification problem defined on the square $-1 \leq x_1 \leq 1$ and $-1 \leq x_2 \leq 1$ with decision boundary $x_2 = \sin(2\pi x_1)$. Make your own input data set by distributing 1000 inputs in the two regions shown, half of them with target $t = +1$ (■), the other half with $t = -1$ (□). Find approximate decision boundaries using a radial-basis function network with m radial basis functions, for $m = 5, 10, 20$ and 100. Plot the decision boundaries in the input plane. Determine the corresponding classification errors.

11 Reinforcement learning*

Supervised learning requires labeled data, where each input comes with a target pattern that the network is supposed to learn. Unsupervised learning, by contrast, does not require labeled data. *Reinforcement learning* lies between these extremes. The term reinforcement describes the principle of learning with only incomplete feedback, in the form of *penalty* or *reward*. For a neural net with a vector of outputs, for instance, the feedback may consist only of a single bit of information: *reward* (all outputs correct) or *penalty* (some or all bits are wrong):

$$r = \begin{cases} +1, & \text{reward}\,, \\ -1, & \text{penalty}\,. \end{cases} \tag{11.1}$$

The goal is to learn to produce outputs that receive positive feedback (reward) more frequently than those that trigger a penalty. We say that rewarded outputs are *reinforced*. More generally, the feedback may be random, given by a distribution initially unknown to the network.

One distinguishes two different types of reinforcement problems, *associative* and *non-associative* problems [149]. An example for a non-associative task is the N-armed bandit problem [149]. Imagine N slot machines with different reward distributions, initially unknown to the player. Given a finite amount of money, the question is in which order to play the machines so as to maximise the overall profit. The dilemma is whether to stick with a machine that yields a decent reward, or whether to try out other machines that may yield a low reward initially, but could give much higher rewards eventually (exploit versus explore dilemma). In this type of problems the player receives only the reinforcement signal, no other inputs. In associative tasks, by contrast, the agent receives inputs (*stimuli*) and it should learn to *associate* with each stimulus the output that yields the highest reward. Such tasks occur for instance in behavioural psychology, where the problem is to discriminate between different stimuli, and to associate the right behaviour with each stimulus.

In general, such associative tasks can be described as *sequential decision processes* (Figure 11.1), where an *agent* explores a sequence of states $\boldsymbol{s}_0, \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots$ through a sequence of actions $\boldsymbol{a}_0, \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots$. Consider for instance a motile microorganism in the turbulent ocean that should swim to the water surface as quickly as possible [150]. It determines its state by observing the local environment. The microorganism, for example, might measure local strain and vorticity of the flow. The enviroment provides a reinforcement signal (the distance to the surface for example), and the agent determines which action to take, given its state and the reinforcement signal. Should it turn, stop to swim, or accelerate? The agent learns to associate actions with certain states that maximise the reward. This sounds quite similar to associating

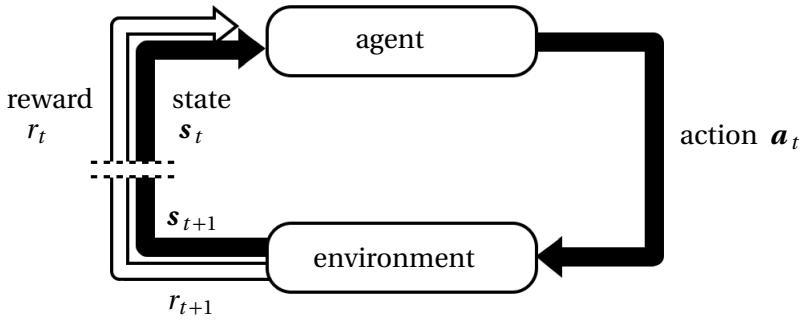


Figure 11.1: Sequential decision process (schematic). Adapted from Figure 3.1 in Ref. [149].

optimal outputs with stimuli. A conceptual difference is that the action of the agent modifies the enviroment: its actions take it to a different place in the turbulent flow, with different vorticity and different strain. The goal is generally to optimise the expected future reward, given the information collected so far. One distinguishes two different kinds of tasks, and *continuous* and *episodic* tasks.

In continuous tasks, the intertwined sequences of states and actions have no natural end, so that one must either terminate the sequence in an *ad-hoc* fashion or introduce a weighting factor to ensure that the expected future reward remains finite. In episodic tasks, by contrast, the learning is divided into episodes that terminate after a certain number of steps. An example is the task to learn a strategy for winning a board game, each episode corresponds to a round of the game, and the reward (draw/won/lost) is incurred at the end of each episode. Often the number of steps per round (episode length T) varies from round to round. In order to estimate the expected reward one usually averages over many episodes. A second, very simple example is the stimulus problem described above, where the states (stimuli) are independent from the actions. Each episode consists of only one step: in response to a randomly chosen state $\boldsymbol{s}_0$ the agent learns to perform the action $\boldsymbol{a}_0$ that maximises the immediate reward. This can be achieved by the associative reward-penalty algorithm (Section 11.1). It uses stochastic neurons with weights that are trained by gradient ascent to maximise the expected immediate reward.

To estimate the future expected reward one must use a different method, usually *temporal difference learning*. It allows to estimate the expected future reward, after T steps say, by breaking up the learning into time steps $t = 1, \ldots, T$. The idea is that it is better to adjust the prediction of a future reward as one iterates, rather than waiting for T iterations before updating the prediction, expressing the reward at time T in terms of differences at time steps $t+1$ and t (*telescoping sum*) [151]. The

algorithm builds up a lookup table that summarises the best actions for each state, the *Q-table*. The elements of the Q-table contain estimates of the expected future reward for each state-action pair. In general it is difficult to prove convergence of algorithms for temporal difference learning. A simplified scheme is *Q-learning*, an approximation to the temporal difference algorithm. In Q-learning, the Q-table is updated assuming that the agent always follows the greedy policy, even though it might follow a different policy. Q-learning allows agents to learn to play strategic games [6]. A simple example is the game of tic-tac-toe (Section 11.3). Games such as chess or go require to keep track of a very large number of states, so large that Q-learning in its simplest form becomes impractical (*curse of dimension*). An alternative is to represent the state-action mapping by a deep neural net [6].

11.1 Associative reward-penalty algorithm

The associative reward-penalty algorithm uses *stochastic neurons* which are trained to maximise the average immediate reward. In Chapters 5 to 9 the output neurons were deterministic functions of their inputs. For reinforcement learning, by contrast, it is better to use stochastic neurons. The idea is the same as in Chapters 3 and 4: stochastic neurons can explore a wider range of possible states which may in the end lead to a better solution. The state y_i of neuron i is given by the *stochastic update rule* (3.1):

$$y_i = \begin{cases} +1, & \text{with probability } p(b_i), \\ -1, & \text{with probability } 1-p(b_i), \end{cases} \tag{11.2}$$

where $b_i = \boldsymbol{w}_i \cdot \boldsymbol{x}$ is the local field (no thresholds), and $p(b) = (1+e^{-2\beta b})^{-1}$ as before. Recall that the parameter β^{-1} is the noise level. Since the output can assume only two values, $y_i = \pm 1$, Equation (11.2) describes a *binary* stochastic neuron.

To illustrate the associative reward-penalty algorithm for a single binary stochastic neuron, consider an agent experiencing different stimuli $\boldsymbol{x}$ drawn with equal probability from a distribution of inputs. Given a stimulus $\boldsymbol{x}$, the stochastic neuron outputs either $y = 1$ or $y = -1$. The environment provides a stochastic *reinforcement signal* $r(\boldsymbol{x}, y) = \pm 1$, a reward ($r = 1$) with probability $p_{\text{reward}}(\boldsymbol{x}, y)$, and a penalty ($r = -1$) with probability $1 - p_{\text{reward}}(\boldsymbol{x}, y)$. The goal is to adjust the weights so that the neuron produces outputs that are rewarded with high probability. Figure 11.2(**a**) shows an example with just two stimuli, $\boldsymbol{x}_1 = [1,0]^{\mathsf{T}}$ and $\boldsymbol{x}_2 = [1,1]^{\mathsf{T}}$. The numerical values of $p_{\text{reward}}(\boldsymbol{x}, y)$ indicate that the expected reward is maximised when the neuron outputs $y = 1$ in response to $\boldsymbol{x}_1$, and $y = -1$ in response to $\boldsymbol{x}_2$, and in this case it attains the value

$$r_{\max} = p_{\text{reward}}(\boldsymbol{x}_1, +1) + p_{\text{reward}}(\boldsymbol{x}_2, -1) - 1 = 0.1\,. \tag{11.3}$$

(**a**)

p_{reward}	$y=-1$	$y=+1$
$\boldsymbol{x}_1=[1,0]^{\mathsf{T}}$	0.6	0.8
$\boldsymbol{x}_2=[1,1]^{\mathsf{T}}$	0.3	0.1

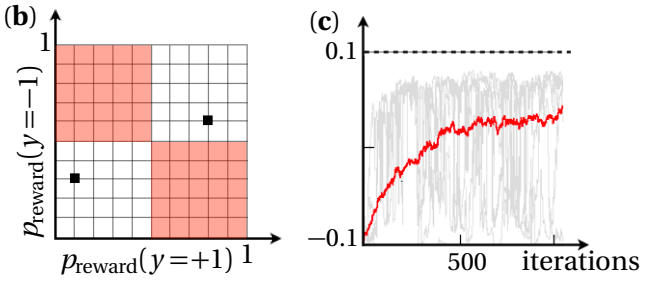


Figure 11.2: Conditioning by reward [152]. A stochastic neuron responds to stimuli $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ with different outputs, $y=\pm 1$, receives a reward $r=+1$ with probability $p_{\text{reward}}(\boldsymbol{x},y)$, and a penalty $r=-1$ with probability $1-p_{\text{reward}}(\boldsymbol{x},y)$. The goal is to always respond with the output that maximises the expected reward. (**a**) Reward probability. (**b**) Contingency space of the problem, see text. (**c**) Reward versus iteration number of the associative reward-penalty rule (11.9) with $\delta=0.05$: individual realisations (gray), ensemble average (red). Simulations by Phillip Graefensteiner.

Here we used that $\langle r(\boldsymbol{x},y)\rangle = p_{\text{reward}}(\boldsymbol{x},y)-[1-p_{\text{reward}}(\boldsymbol{x},y)]$, and that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ appear with equal propabilities.

Figure 11.2 (**b**) shows the *contingency space* of the problem, representing the inputs $\boldsymbol{x}$ in a plane with coordinates $p_{\text{reward}}(\boldsymbol{x},+1)$ and $p_{\text{reward}}(\boldsymbol{x},-1)$. It is easier to learn to associate the correct output with inputs that lie in the red regions where $p_{\text{reward}}(\boldsymbol{x},+1)>\frac{1}{2}$ and $p_{\text{reward}}(\boldsymbol{x},-1)<\frac{1}{2}$, or vice versa. In this case one can solve the problem by fixing $y=+1$ and sampling $p_{\text{reward}}(\boldsymbol{x},+1)$ for all $\boldsymbol{x}$. If $p_{\text{reward}}(\boldsymbol{x},+1)>\frac{1}{2}$ then $y=+1$ is the optimal output for $\boldsymbol{x}$, otherwise it is $y=-1$. This strategy cannot be used outside the red region. For example, if both reward probabilities are larger than one half, it is necessary to sample both $p_{\text{reward}}(\boldsymbol{x},-1)$ and $p_{\text{reward}}(\boldsymbol{x},+1)$ sufficiently often in order to determine which one is larger, one must find *the greater of two goods* according to Barto [152]. This illustrates the fundamental dilemma of reinforcement learning: an output that appears at first to yield a high reward may not be the optimal one in the long run. To find the optimal output it is necessary to estimate both reward probabilities precisely. This means that one must try all possible outputs frequently, not only the one that appears to be optimal at the moment.

To derive a learning rule we need a cost function. The associative reward-penalty algorithm uses the average of the immediate reward

$$\langle r\rangle = \frac{1}{T}\sum_{t=1}^{T}\sum_{y_1=\pm q,\ldots,y_M=\pm 1}\Big\langle r(\boldsymbol{x}(t),\boldsymbol{y})P(\boldsymbol{y};\boldsymbol{x}(t),\{w_{ij}\})\Big\rangle_{\text{reward}}. \tag{11.4}$$

Here

$$P(\boldsymbol{y};\boldsymbol{x},\{w_{ij}\})=\prod_{i=1}^{M}\begin{cases}p(b_i) & \text{if} \quad y_i=1\,,\\ 1-p(b_i) & \text{if} \quad y_i=-1\end{cases} \tag{11.5}$$

is the probability that the network produces the output $\boldsymbol{y}$ given weights w_{ij} and input $\boldsymbol{x}$, and $\langle\cdots\rangle_{\text{reward}}$ is an average over the response of the environment, determined by the reward distribution $p_{\text{reward}}(\boldsymbol{x},\boldsymbol{y})$. It is assumed that the reward distribution is stationary, just like the distribution of inputs.

To find the maximum of $\langle r\rangle$ one uses gradient ascent, analogous to maximising the log-likelihood for Boltzmann machines (Section 4.4), and to gradient descent on the energy function for perceptrons in supervised learning (Chapter 6). Accordingly, the weight increments are given by:

$$\delta w_{mn}=\eta\frac{\partial\langle r\rangle}{\partial w_{mn}} \tag{11.6}$$

with learning rate $\eta>0$. The gradient is computed by applying the chain rule, as usual. The calculation is similar to the one for Boltzmann machines (Chapter 4). Using $\mathrm{d}p(b_m)/\mathrm{d}b_m=\frac{1}{2}[1+\tanh(\beta b_m)]$ one finds:

$$\frac{\partial P(\boldsymbol{y};\boldsymbol{x},\{w_{ij}\})}{\partial w_{mn}}=\beta[y_m-\tanh(\beta b_m)]x_n\,. \tag{11.7}$$

This yields the learning rule

$$\delta w_{mn}=\alpha r\big[y_m-\tanh(b_m)\big]x_n \tag{11.8}$$

with $\alpha=\beta\eta$ and $b_m=\sum_j w_{mj}x_j$. It is plausible that this rule maximises the average immediate reward because it tends to increase this quantity on average, and because the weight increments approach zero as the network learns to produce the output $\max_{\boldsymbol{y}}\{p_{\text{reward}}(\boldsymbol{x},\boldsymbol{y})\}$, independent of $\boldsymbol{y}$ so that $\boldsymbol{y}-\langle\boldsymbol{y}\rangle$ averages to zero.

In practice it is better to use an asymmetric learning rule [152], the associative reward-penalty rule:

$$\delta w_{mn}=\alpha\begin{cases}\big[y_m-\tanh(b_m)\big]x_n & \text{for} \quad r=+1\,,\\ \delta\big[-y_m-\tanh(b_m)\big]x_n & \text{for} \quad r=-1\,,\end{cases} \tag{11.9}$$

with $0<\delta\ll 1$. For $r=1$, the learning rules (11.8) and (11.9) give the same weight increment. But the weight increments differ for $r=-1$. With rule (11.9) the agent learns primarily from positive feedback. One advantage of this asymmetric rule is that it can be proven to converge in the limit of $\delta\to 0$ [152]. In general, however, the convergence becomes quite slow when δ is small. Figure 11.2(**c**) shows simulation

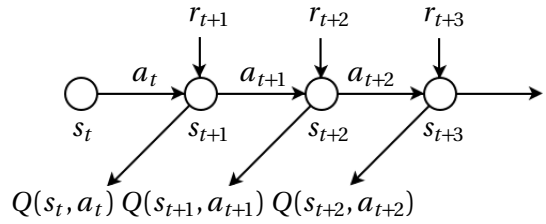


Figure 11.3: Sequence of states s_t in sequential reinforcement learning. The action a_t leads from s_t to s_{t+1} where the agent receives reinforcement r_{t+1}. The Q-table with elements $Q(s_t, a_t)$ estimates the future discounted reward.

results for the immediate reward as function of iteration number of the rule (11.9) for a quite small value of δ. Shown are individual realisations of the learning curve (gray), as well as an average over 100 realisations. The algorithm appears to converge to a steady state, but the steady-state average of the immediate reward is smaller than $r_{\max} = 0.1$. Comparing one single realisation of a learning curve with the ensemble average, we see that there are significant fluctuations. Furthermore, the convergence proof assumes that the input patterns are linearly independent. This means that the number of patterns cannot exceed N. Associative reinforcement problems with linearly dependent inputs can be solved by embedding the input patterns in a higher-dimensional input space (Section 10.5).

The associative reward-penalty rule illustrates how an agent can use a reinforcement signal to learn to maximise the expected immediate reward. It is a model for how an animal may learn to respond in different ways to different stimuli.

Often the reward is not immediate. When we play chess, for example, the reward comes at the end of the game, $r = +1$ if we won, $r = -1$ if we lost, and $r = 0$ if the game ended in a draw. More generally, an agent navigating an complex environment should not only consider which immediate reward a certain action gives, but also how this action affects possible future rewards. One way of estimating future rewards for such tasks is temporal difference learning, discussed next.

11.2 Temporal difference learning

How does temporal difference learning allow an agent to learn to optimise its expected future reward? Given an episodic task, the agent visits a finite sequence of T states $\boldsymbol{s}_0, \ldots, \boldsymbol{s}_{T-1}$ during an episode, and collects the rewards $r_1, \ldots, r_T$. The future expected reward is defined as

$$R_t = \sum_{\tau=t}^{T-1} r_{\tau+1} \,. \tag{11.10}$$

Continuous tasks, by contrast, do not have a defined end point. Since the sum in (11.10) might diverge as $T \to \infty$, it is customary to introduce a weighting factor $0 < \gamma \leq 1$ in the sum over rewards:

$$R_t = \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_{\tau+1} . \tag{11.11}$$

The weighting factor reduces the contribution of the far future to the estimate. Smaller values of γ give more weight to the immediate future. The sum in Equation (11.11) is called *future discounted reward*.

We use a neural net with input $\boldsymbol{s}_t$ to estimate R_t. In general, the network output is a non-linear function of the inputs, parameterised by weights that could be arranged into several layers of hidden neurons (Chapter 6). The simplest choice is to use a linear unit, just like in Equation (5.19):

$$O(\boldsymbol{s}_t, \boldsymbol{w}) = \boldsymbol{w} \cdot \boldsymbol{s}_t . \tag{11.12}$$

The components w_j of the weight vector $\boldsymbol{w}$ are determined so that the network output $O(\boldsymbol{s}_t, \boldsymbol{w})$ approximates R_t. This can be achieved by minimising the energy function

$$H = \tfrac{1}{2} \sum_{t=0}^{T-1} [R_t - O(\boldsymbol{s}_t)]^2 \tag{11.13}$$

using gradient descent. The corresponding learning rule reads:

$$\delta w_m = \alpha \sum_{t=0}^{T-1} [R_t - O(\boldsymbol{s}_t)] \frac{\partial O}{\partial w_m} . \tag{11.14}$$

The idea of temporal difference learning [151] is to express the error $R_t - O(\boldsymbol{s}_t)$ as a sum of temporal differences, using $R_t - O(\boldsymbol{s}_t) = \sum_{\tau=t}^{T-1} [r_{\tau+1} - O(\boldsymbol{s}_{\tau+1}) - O(\boldsymbol{s}_\tau)]$. Using the gradient-descent rule (11.14) one obtains

$$\delta \boldsymbol{w}_t = \alpha [r_{t+1} + O(\boldsymbol{s}_{t+1}) - O(\boldsymbol{s}_t)] \boldsymbol{s}_t . \tag{11.15}$$

This learning rule corresponds to the following update rule for O:

$$O_{t+1}(\boldsymbol{s}_t) = O_t(\boldsymbol{s}_t) + \alpha [r_{t+1} + O_t(\boldsymbol{s}_{t+1}) - O_t(\boldsymbol{s}_t)] . \tag{11.16}$$

The notation O_t emphasises that O is updated recursively. The update rule (11.16) applies to estimating the future reward (11.10) for episodic tasks. If the environment is stationary, one may average over many consecutive episodes, using the final weights from episode k as initial weight values for episode $k+1$. For continuous

tasks, the corresponding rule for estimating the future discounted reward (11.11) reads:

$$O_{t+1}(\boldsymbol{s}_t) = O_t(\boldsymbol{s}_t) + \alpha[r_{t+1} + \gamma O_t(\boldsymbol{s}_{t+1}) - O_t(\boldsymbol{s}_t)]. \tag{11.17}$$

Returning to the problem outlined in the beginning of this Chapter, consider an agent exploring a complex environment. The task might be to get from location A to location B as quickly as possible, or expending as little energy as possible. At time t the agent is at position $\boldsymbol{x}_t$, it may have velocity $\boldsymbol{v}_t$. These variables as well as the local state of the environment are summarised in the state vector $\boldsymbol{s}_t$. Given $\boldsymbol{s}_t$, the agent can act in certain ways: it might slow down, speed up, or turn for example. These possible actions are summarised in a vector $\boldsymbol{a}_t$. At each time step, the agent take the action $\boldsymbol{a}_t$ that optimises the expected future discounted reward (11.11), given its present state $\boldsymbol{s}_t$. The estimated expected future reward for any given state-action pair is summarised in a table, the Q*-table* with elements $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$, the analogue of O. Different rows of the Q-table correspond to different states, and different columns to different actions. The temporal-difference learning rule for Q reads:

$$Q_{t+1}(\boldsymbol{s}_t, \boldsymbol{a}_t) = Q_t(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha_t\big[r_{t+1} + \gamma Q_t(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) - Q_t(\boldsymbol{s}_t, \boldsymbol{a}_t)\big]. \tag{11.18}$$

This algorithm is called SARSA, because one needs $\boldsymbol{s}_t, \boldsymbol{a}_t, r_{t+1}, \boldsymbol{s}_{t+1}$, and $\boldsymbol{a}_{t+1}$ to update the Q-table (Figure 11.3). A difficulty with the rule y (11.18) is that it depends not only on the present state-action pair $[\boldsymbol{s}_t, \boldsymbol{a}_t]$, but also on the next action $\boldsymbol{a}_{t+1}$, and thus indirectly upon the policy. Sometimes this is indicated by writing Q_π for the Q-table given policy π.

A common policy is the *greedy* policy where the agent chooses the action that maximises the expected future reward. The ε-greedy policy does mainly that, but with a small probability ε it takes a suboptimal action. This allows the agent to explore potentially better alternatives. In general the policy can change as the algorithm is iterated. It is customary, for instance, to reduce the parameter ε as the agent learns.

11.3 Q-learning

The one-step Q-learning rule [153] is an approximation to Eq. (11.18) that does not depend on $\boldsymbol{a}_{t+1}$. Instead one assumes that the next action, $\boldsymbol{a}_{t+1}$, is the optimal one:

$$Q_{t+1}(\boldsymbol{s}_t, \boldsymbol{a}_t) = Q_t(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha_t\big[r_{t+1} + \gamma \max_{\boldsymbol{a}} Q_t(\boldsymbol{s}_{t+1}, \boldsymbol{a}) - Q_t(\boldsymbol{s}_t, \boldsymbol{a}_t)\big], \tag{11.19}$$

Algorithm 10 Q-learning for episodic task

initialise $\boldsymbol{s}_0$ and Q;
for $k = 1, \ldots, K$ **do**
  for $t = 0, \ldots, T_k - 1$ **do**
    choose $\boldsymbol{a}_t$ from $Q(\boldsymbol{a}_t, \boldsymbol{s}_t)$ according to ε-greedy policy;
    compute $\boldsymbol{s}_{t+1}$ and record r_{t+1};
    update $Q(\boldsymbol{s}_t, \boldsymbol{a}_t) \leftarrow Q(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha[r_{t+1} + \max_{\boldsymbol{a}} Q(\boldsymbol{s}_{t+1}, \boldsymbol{a}) - Q(\boldsymbol{s}_t, \boldsymbol{a}_t)]$;
  end for
end for

regardless of the policy that is currently followed. Although Equation (11.19) does not refer to any policy, the learning outcome nevertheless depends on it, because the policy determines the sequence $[\boldsymbol{s}_t, \boldsymbol{a}_t]$. For the greedy policy, Eq. (11.19) is equivalent to (11.18), but in general the two algorithms differ. For episodic tasks one puts $\gamma = 1$ in (11.19), and averages over many episodes using the outcome Q_{T_k} from episode k as initial condition for episode $k+1$. In general, Q-learning learns the optimal strategy as well as SARSA, but it may give lower rewards than SARSA.

The Q-learning algorithm is summarised in Algorithm 10. Usually one sets the initial entries in the Q-table to large positive values (*optimistic* initialisation), because this prompts the agent to explore many different actions, at least in the beginning. If the agent is in state $\boldsymbol{s}_t$, it chooses the action $\boldsymbol{a}_t$ from $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ according to the given policy. For the ε-greedy policy, for example, the agent picks a random action from the corresponding row of the Q-table with probability ε. With probability $1-\varepsilon$ it chooses the action $\boldsymbol{a}_t$ that yields the largest[1] $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ given $\boldsymbol{s}_t$. The choice of action $\boldsymbol{a}_t$ determines the next state $\boldsymbol{s}_{t+1}$, and this in turn allows to update the Q-table using Eq. (11.19).

When the sequence $\boldsymbol{s}_0, \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots$ is a Markov chain (Section 4.2), then the Q-learning algorithm can be shown [154] to converge if one uses a time-dependent learning rate α_t that satisfies

$$\sum_{t=0}^{\infty} \alpha_t = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty \,. \tag{11.20}$$

It is important to note that the learning outcome depends on the reward function imposed. In general it is instructive to analyse how the optimal strategy changes as one varies the reward function. Sometimes we are faced with the inverse problem: suppose you obsere how a microorganism swimming in the turbulent ocean

[1] If several elements in the relevant row have the same maximal value then any one of them is chosen with equal probability.

responds to different stimuli. How was this behaviour shaped by genetic evolution? In other words, which quantity is optimised? Is it most important to reduce the energy cost for moving? Or is it more important to avoid predation?

Most often, Q-learning is implemented in combination with the ε-greedy policy. This policy shares an important property with the associative reward-penalty algorithm with stochastic neurons: stochasticity allows for a wider range of responses some of which may turn out beneficial in the long run. When ε is very small, the agent picks the action that appears optimal. As a consequence, sub-optimal Q-elements are sampled less frequently and therefore subject to larger errors. Therefore it is advantageous to begin with a relatively large value of ε. It is customary to decrease ε as the algorithm is iterated.

Another challenge is to determine suitable states and actions. An agent navigating a complex environment may have a continous range of positions and velocities, and may experience real-valued signals from the environment. To represent the corresponding states in a Q-table it is necessary to discretise. To this one must determine suitable ranges and resolutions of these variables, and for the actions. If there are too many states and actions, Q-learning becomes unwieldy. This is sometimes referred to as the *curse of dimension*.

Let us see how Q-learning works for a very simple example, for the associative task described in Fig. 11.2. One episode corresponds to computing the output of the neuron given its initial state $\boldsymbol{s}_0$, so $T=1$. There is no sequence of states, and the task is to estimate the immediate reward. In this case the update rule (11.19) simplifies to

$$\delta Q(\boldsymbol{s},a)=\alpha\big[r(\boldsymbol{s},a)-Q(\boldsymbol{s},a)\big]. \tag{11.21}$$

The term $\max_{\boldsymbol{a}} Q(\boldsymbol{s}_{t+1},\boldsymbol{a})$ in Equation (11.19) does not appear in (11.21) since Q estimates the immediate reward. There are only two states in this problem $\boldsymbol{s}=\boldsymbol{x}_1$ and $\boldsymbol{s}=\boldsymbol{x}_2$, and two actions, $a=\pm 1$. In each round, one of the states is chosen randomly, with equal probability. The action is determined from the current estimate of the immediate reward as $\operatorname{argmax}_a Q(\boldsymbol{s},a)$ with probability $1-\varepsilon$, and uniformly randomly otherwise. These steps are iterated over many iterations (episodes), using the outcome of episode k as initial condition for episode $k+1$.

The rule (11.21) describes exponential relaxation to the target, as we can see from the solution of the stochastic differential equation $\frac{\mathrm{d}}{\mathrm{d}t}Q(\boldsymbol{s},a)=\alpha[r(\boldsymbol{s},a)-Q(\boldsymbol{s},a)]$. Averaging over the reward distribution gives:

$$\langle Q_t(\boldsymbol{s},a)\rangle=\langle Q_0(\boldsymbol{s},\boldsymbol{a})\rangle\,\mathrm{e}^{-\alpha t}+\alpha\int_0^t \mathrm{d}t'\,\langle r(\boldsymbol{s},\boldsymbol{a})\rangle\mathrm{e}^{\alpha(t'-t)}. \tag{11.22}$$

Using the fact that $\langle r(\boldsymbol{s},a)\rangle=2p_{\text{reward}}(\boldsymbol{s},a)-1$, we see that $Q_t(\boldsymbol{s},\boldsymbol{a})$ converges on

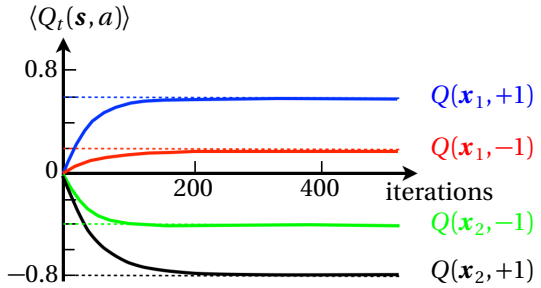


Figure 11.4: Q-learning for the task described in Figure 11.2. Shown are the entries of the Q-table versus the number of iterations of Equation (11.21) for $\alpha = 0.1$ and $\varepsilon = 0.3$. Schematic, based on simulations by Navid Mousavi, averaged over 5000 independent realisations of the learning curve.

average to

$$\begin{bmatrix} Q^*(\boldsymbol{x}_1,-1) & Q^*(\boldsymbol{x}_1,+1) \\ Q^*(\boldsymbol{x}_2,-1) & Q^*(\boldsymbol{x}_2,+1) \end{bmatrix} = \begin{bmatrix} 0.2 & 0.6 \\ -0.4 & -0.8 \end{bmatrix}, \tag{11.23}$$

for the reward probabilities in Figure 11.2. Figure 11.4 demonstrates how the learning rule (11.21) approaches the maximal immediate reward $r_{\max} = \frac{1}{2}[Q^*(\boldsymbol{x}_1,+1) + Q^*(\boldsymbol{x}_2,-1)] = 0.1$. Shown is the average over many realisations of the random process (11.21).

A second example is illustrated in Figure 11.5, the board game tic-tac-toe. It is a very simple two-player game who take turns in placing their pieces on a 3×3 board. The player who manages to first obtain three pieces in a row, column, or diagonal wins and receives the reward $r = +1$. A draw gives $r = 0$, and the player receives $r = -1$ when the round is lost. The goal is to win as often as possible, to maximise the future expected reward. However, there is a strategy for both players to ensure that they do not lose. If both players try to maximise their expected future reward, they end up following this strategy, so that every game must end in a draw [155]. As a result, the game is quite boring.

Nevertheless it is instructive to ask how the players can learn to find this strategy using Q-learning with the ε-greedy policy. To this end we let two agents play many rounds against each other. The state space is the collection of all board configurations. Player $\times$ starts, and thus always sees a board with an even number of pieces, while the number of pieces is odd for player $\circ$. Since the players encounter different sets of states, each must keep track of their own Q-table. The task is episodic, and the number T of steps may vary from round to round. Feedback is only obtained at the end of each round.

We use Equation (11.19) with a constant learning rate α. We can set $\gamma = 1$ since the number of steps in each round is finite. Since the number of states is large, it is

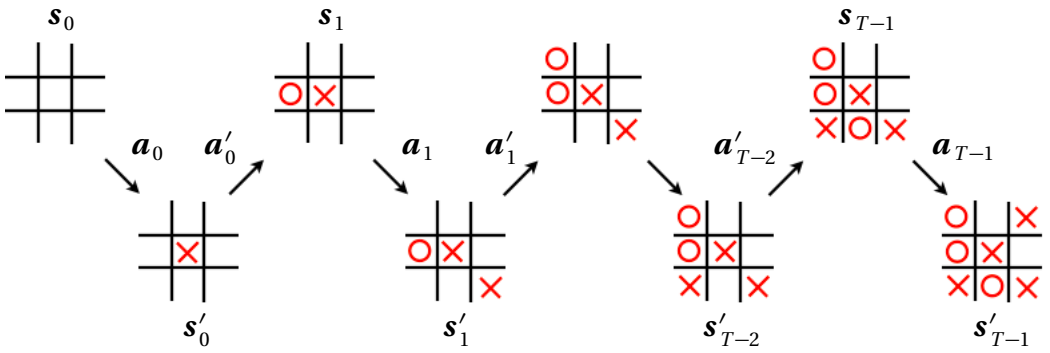


Figure 11.5: Tic-tac-toe. Two players, × and ∘, take turns in placing a piece on an empty field of 3×3 board. The goal is to be the first to complete a row, column, or diagonal consisting of three of one's own pieces. In the example shown, player × starts and ends up winning the game. The states encountered by player × are denoted by $\boldsymbol{s}_t$, those encountered by player ∘ by $\boldsymbol{s}'_t$. Their actions are denoted by $\boldsymbol{a}_t$ and $\boldsymbol{a}'_t$.

better to initialise the elements of the Q-table only when they are first encountered. We can think of the Q-table as $2 \times n$ table where each entry is a 3×3 array. Here n is the number of states the player encountered so far. The first row lists the states, each a 3×3 array with entries -1 (∘), 1 (×), or 0 (empty). The second row contains the Q-values. Given a certain state of the board, a player can play a piece on any empty field. The corresponding estimate of the future expected reward is stored in the corresponding 3×3 array in the second row. Since one cannot place a piece onto an occupied field, the corresponding entries in the Q-table are assigned NaN.

During a round, the Q-matrices of the players are updated in turns, always the one of the player who places a piece. After a round, the Q-matrices of both players are updated. During the first round the elements of Q-matrices encountered are initialised to zero and remain zero. The first change to Q occurs in the last step of the round. If player × wins, the element $Q(\boldsymbol{s}_{T-1}, a_{T-1})$ corresponding to the state-action pair that led to the final state $\boldsymbol{s}'_{T-1}$ is updated for the winning player, and $Q(\boldsymbol{s}'_{T-2}, \boldsymbol{a}'_{T-2})$ is set to -1 for player ∘.

Both players follow the ε-greedy policy. With probability $1-\varepsilon$ they take the optimal move (if the maximal Q-element in the relevant row is degenerate, then one of the maximal elements is chosen randomly). With probability ε, a random action is chosen. As the players continue to play rounds against each other, the rewards slowly spread to other elements of Q. Suppose that the state $\boldsymbol{s}_{T-1}$ is encountered once more, the one that allowed player × to the first round with $\boldsymbol{a}_{T-1}$. Then the term $\max_a Q(\boldsymbol{s}_{T-1}, \boldsymbol{a})$ causes a Q-element for the previous state to change, the one from which $\boldsymbol{s}_{T-1}$ was reached the second time. However, as times goes on this process slows down because later updates are multiplied with higher powers of the learning rate α. This is a reason for choosing a constant learning rate, independent of t.

Figure 11.6 illustrates how the players learn, after playing many rounds against

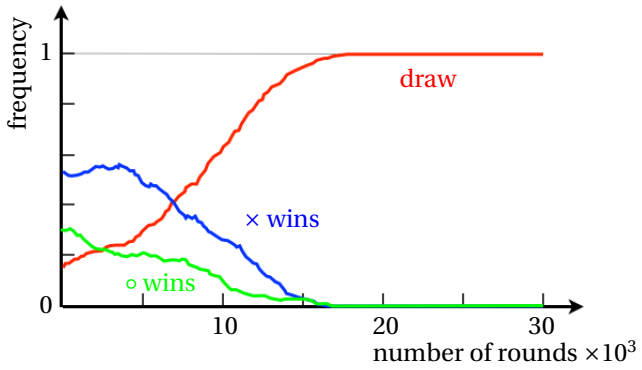


Figure 11.6: Learning curves for two players learning to play tic-tac-toe with Q-learning and the ε-greedy policy. Shown are the frequencies that the game ends in a draw (red), that player × wins (blue), and that player ∘ wins (green). Similar curves are obtained using a learning rate $\alpha = 0.1$. The parameter ε was equal to unity for the first 10^4 rounds, and then decreased by a factor of 0.9 after each 100 rounds, and averaging each curve over a running window of 30 rounds. Schematic, based on simulations performed by Navid Mousavi.

each other. Since both players try to maximise their expected future reward, all games end in a draw in the steady state of the Q-learning algorithm. The corresponding Q-tables contain the strategies each player should adopt to maximise their reward. Suppose for example player ∘ places the first piece as shown below:

$$\begin{bmatrix} 1.00 & 1.00 & 1.00 \\ 0.34 & \text{NaN} & \text{NaN} \\ 1.00 & 1.00 & 1.00 \end{bmatrix} \quad \begin{bmatrix} \text{NaN} & -0.57 & 1.00 \\ -0.69 & \text{NaN} & \text{NaN} \\ -0.59 & -0.73 & \text{NaN} \end{bmatrix} \quad \begin{bmatrix} \text{NaN} & 1.00 & \text{NaN} \\ -0.86 & \text{NaN} & \text{NaN} \\ \text{NaN} & 0.027 & \text{NaN} \end{bmatrix}. \tag{11.24}$$

How does the game continue? There are several different ways for player × to win. The left Q-table in Equation (11.24) shows that one possibility is to place the piece in a corner because this creates the opportunity of creating a *bridge* in the next move and thus of winning the game. The right Q-table shows that player × could still lose or end up with a draw if he makes the wrong move. The corresponding Q-entries have not quite converged to -1 and 0, respectively. Q-elements corresponding to suboptimal states are not estimated as precisely because they are visited less frequently. Here $\varepsilon = 0.3$ was chosen quite large. Smaller values give less precise estimates than those in Equation (11.24).

As pointed out above, the learning outcome depends on the reward function. If one increases reward for winning, to $r = +2$ for instances, the optimal strategy appears to be to take turns in winning. The same learning outcome is expected if

one imposes a penalty for a draw, $r = +1$ (win), $r = -1$ (draw, lose), Exercise 11.8. More examples for reinforcement-learning problems in robotics and in the natural sciences are described in Ref. [156].

11.4 Summary

Reinforcement learning lies between unsupervised learning (Chapter 10) and supervised learning (Chapters 5 to 9). The learning is not based on labeled data sets, but the neural net or agent learns by feedback from the environment in the form of a reward or a penalty. The goal is to find a strategy that maximises the expected reward.

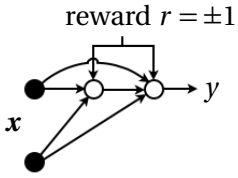
Reinforcement learning is applied in a wide range of fields, from psychology to mechanical engineering, using a large variety of algorithms. The associative reward-penalty algorithm and many versions of time-difference learning were originally formulated using neural nets. Q-learning, an approximation to time-differerence learning for sequential sequential decision processes, does not rely on neural nets. The Q-learning algorithm is quite efficient when the number of states and actions is quite small. For large Q-tables, by contrast, the algorithm becomes inefficient.

11.5 Further reading

The standard reference for reinforcement learning is the book by Sutton and Barto [149]. The original reference for the convergence of the Q-learning algorithm is given in Ref. []. A more mathematical introduction is given in Ref. Szepesvari2010. Examples for reinforcement learning in statistical and non-linear physics are given in Ref. [156].

A widely discussed application of reinforcement learning is AlphaGo, the algorithm that learnt to play the game of Go. How it works is described in this blog. An important point is that the Q-table represented by a Q-function that maps states to actions, in terms of a convolutional neural nets. This makes it possible to use Q-learning despite the fact that the number of states is enormous.

An open question is when and how symmetries can be exploited to simplify a reinforcement problem. For a small microorganism learning to navigate a turbulent flow, some aspects are discussed in Ref. [**?**], but little is known in general.



p_{reward}	$y=-1$	$y=+1$
$\boldsymbol{x}_1 = [0,0]^{\mathsf{T}}$	0.8	0.1
$\boldsymbol{x}_2 = [0,1]^{\mathsf{T}}$	0.1	0.8
$\boldsymbol{x}_3 = [1,0]^{\mathsf{T}}$	0.1	0.8
$\boldsymbol{x}_4 = [1,1]^{\mathsf{T}}$	0.8	0.1

Figure 11.7: Associative penalty-reward algorithm with one hidden neuron for a stochastic XOR problem [152].

11.6 Exercises

11.1 Binary stochastic neurons. A layer has binary stochastic neurons y_i with update rule $y_i = +1$ with probability $p(b_i)$ and $y_i = -1$ otherwise. Here $b_i = \sum_j w_{ij} x_j$, and $p(b) = (1 + \mathrm{e}^{-2\beta b})^{-1}$. The parameter β^{-1} is the noise level, x_j are inputs, and w_{ij} are weights. Consider the the energy function $H = \frac{1}{2}\sum_{i\mu}\left(t_i^{(\mu)} - y_i^{(\mu)}\right)^2$ with targets $t_i^{(\mu)} = \pm 1$. Stochastic neurons can be trained by gradient descent on the energy function $H' = \frac{1}{2}\sum_{i\mu}\left(t_i^{(\mu)} - \langle y_i^{(\mu)}\rangle\right)^2$, defined in terms of the average outputs $\langle y_i^{(\mu)}\rangle$. The error $\delta_m^{(\mu)}$ is defined by $\delta w_{mn} \equiv -\eta \frac{\partial H'}{\partial w_{mn}} = \eta \sum_\mu \delta_m^{(\mu)} x_n^{(\mu)}$. Show that $\delta_m^{(\mu)} = (t_m^{(\mu)} - \langle y_m^{(\mu)}\rangle)\beta(1 - \langle y_m^{(\mu)}\rangle^2)$. Show that this rule does not necessarily minimise $\langle H\rangle$.

11.2 Gradient of average reward. Derive Equation (11.7).

11.3 Klopf's self-interested neuron. Klopf's self-interested neuron [157] is a binary stochastic neuron with outputs 0 and 1. Derive a learning rule that is equivalent to Eq. (11.8).

11.4 Associate reward-penalty algorithm. Barto [152] explains how to solve association tasks with linearly dependent inputs using hidden neurons. One of his examples is the XOR problem, illustrated in Fig. 11.7. Using the network shown in Fig. 11.7, implement the associate reward-penalty algorithm and analyse its convergence for the XOR task. Then verify that the task can be solved by a single stochastic neuron if you embed the input data in four-dimensional input space.

11.5 Three-armed bandit problem. Implement the Q-learning algorithm for the three-armed bandit problem with reward distributions shown in Fig. 11.8. Analyse the convergence of Q-learning for different values of ε. Discuss the exploitation-exploration dilemma. Illustrate with results of your computer simulations.

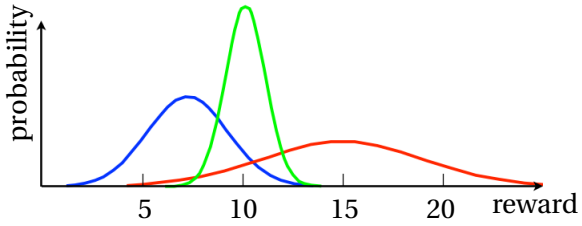


Figure 11.8: Three-armed bandit problem [158]. Three slot machines have Gaussian reward distributions shown, with means and standard deviations $\mu_1 = 7.5, \sigma_1 = 2$, $\mu_2 = 10, \sigma_2 = 1$, and $\mu_3 = 15, \sigma_3 = 5$.

11.6 Psychology of rock-paper-scissors. Learn to exploit idiosynchrasies in the strategy of your opponent. Suppose that you opponent tends to repeat his action if he won the previous round[159], but changes his action if he lost. Let us say that your opponent randomly chooses between two strategies. With probability p he repeats his action if he won, but chooses one of the other two actions if he lost. With probability $1-p$ he picks one of the three actions, rock with probability $1-2q$, and paper or scissors with probability q (your opponent has a preference for rock, so $q < \frac{1}{3}$), and this is also the strategy for the first round. Implement the Q-learning algorithm to determine the optimal strategy as a function of q and p after N rounds.

11.7 Tic-tac-toe. Implement the Q-learning algorithm for two agents learning to play tic-tac-toe (Figure 11.5), with reward function $r = +1$ (win), $r = 0$ (draw), and $r = -1$ (lose). Crowley and Siegler [155] described how a perfect player should play to never lose. Their Table 1 summarises how to play given a certain configuration of the board. Determine the Q-table of a perfect player, to verify Table 1 of Crowley and Siegler.

11.8 Different reward function for tic-tac-toe. For the tic-tac-toe problem (Figure 11.5) Investigate, with Q learning, how the optimal strategy depends on the reward function. Determine the optimal strategy for $r = +2$ (win), $r = 0$ (draw), $r = -1$ (lose), and for $r = +1$ (win), $r = -1$ (draw, lose).

11.9 Connect four. Implement the Q-learning algorithm for two agents learning to play connect four (Figure 11.9) on a 6×6 board. Show that the second player can always achieve a draw against a perfect player [160].

11.10 Eat or save the chocolate. Suppose you get a piece of chocolate every morning. Either you save your chocolate for the next day, or you eat all of it during the day, including all pieces you may have saved from previous days. Each day, before you go to bed, you receive a reward: for each piece of chocolate you ate during the day

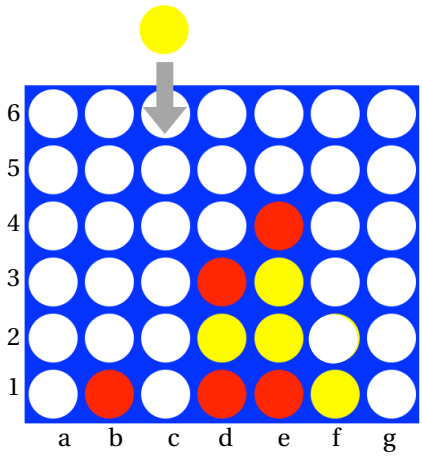


Figure 11.9: Connect four is a game for two players who take turns dropping their pieces into one of k columns of height l ($k = 7$ and $l = 6$ in the Figure). The player first completing a horizontal, vertical, or diagonal row of four pieces wins. The red player started.

you get +2. If you save the chocolate instead you get +1 for each piece of chocolate in stock. But your brother likes chocolate too, and he searches for your stock while you are asleep. Suppose he finds it with probability p, and eats all the chocolate. What is your strategy to optimise the future reward over N days?

Bibliography

[1] HERTZ, J, KROGH, A & PALMER, R 1991 *Introduction to the Theory of Neural Computation*. Addison-Wesley.

[2] HAYKIN, S 1999 *Neural Networks: a comprehensive foundation*, 2nd edn. New Jersey: Prentice Hall.

[3] HORNER, H, Neuronale Netze, [Online; accessed 8-November-2018].

[4] GOODFELLOW, I, BENGIO, Y & COURVILLE, A, Deep Learning, [Online; accessed 5-September-2018].

[5] NIELSEN, M, Neural Networks and Deep Learning, [Online; accessed 13-August-2018].

[6] DEEPMIND, AlphaGo, [Online; accessed: 20-August-2018].

[7] 2020, N. M. A, The nobel prize in physiology or medicine 1906, NobelPrize.org, [Online; accessed 1-October-2020].

[8] GABBIANI, F & METZNER, W 1999 Encoding and processing of sensory information in neuronal spike trains. *Journal of Experimental Biology* **202** (10), 1267.

[9] MCCULLOCH, W & PITTS, W 1943 A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115.

[10] ROSENBLATT, F 1958 A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**, 386.

[11] ROSENBLATT, F 1958 The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Rev.* **65**, 386.

[12] LITTLE, W 1974 The existence of persistent states in the brain. *Mathematical Biosciences* **19**, 101 – 120.

[13] HOPFIELD, J. J, Hopfield network, [Online; accessed 14-August-2018].

[14] FISCHER, A & IGEL, C 2014 Training restricted Boltzmann machines: An introduction. *Pattern Recognition* **47** (1), 25–39.

[15] LIPPMANN, R. P 1987 An introduction to computing with neural nets **3**, 9–44.

[16] MATHEWS, J & WALKER, R. L 1964 *Mathematical Methods of Physics*. New York: W.A. Benjamin.

[17] WolframMathWorld, [Online; accessed 17-September-2019].

[18] KADANOFF, L. P, More is the same: phase transitions and mean field theories, arxiv:0906.0653, [Online; accessed 3-September-2020].

[19] HOPFIELD, J. J 1982 Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* **79** (8), 2554–2558.

[20] HOPFIELD, J. J 1984 Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences* **81** (10), 3088–3092.

[21] MÜLLER, B, REINHARDT, J & STRICKLAND, M. T 1999 *Neural Networks: An Introduction*. Heidelberg: Springer.

[22] AMIT, D. J, GUTFREUND, H & SOMPOLINSKY, H 1985 Spin-glass models of neural networks. *Phys. Rev. A* **32**, 1007.

[23] GESZTI, T 1990 *Physical models of neural networks*. World Scientific.

[24] AMIT, D. J & GUTFREUND, H 1987 Statistical mechanics of neural networks near saturation. *Ann. Phys.* **173**, 30.

[25] ZIRNBAUER, M 1994 Another critique of the replica trick arxiv:cond-mat/9903338.

[26] STEFFAN, H & KUEHN, R 1994 Replica symmetry breaking in attractor neural network models arxiv:cond-mat/9404036.

[27] VOLK, D 1998 On the phase transition of Hopfield networks – another Monte Carlo study. *Int. J. Mod. Phys. C* **9**, 693.

[28] LÖWE, M 1998 On the storage capcaity of Hopfield models with correlated patterns. *Ann. Prob.* **8**, 1216.

[29] N.METROPOLIS, ROSENBLUTH, A. W, ROSENBLUTH, M. N, TELLER, M & TELLER, E 1953 Equation of state calculations by very fast computing machine. *Journal of Chemical Physics* **21**, 1087–1092.

[30] GUBERNATIS, J. E 2005 Marshal Rosenbluth and the Metropolis algorithm. *Physics of Plasmas* **12**, 057303.

[31] HINTON, G. E & SEJNOWSKI, T. J 1986 *Learning and Relearning in Boltzmann Machines*. MIT Press.

[32] HINTON, G. E, Boltzmann machine, [Online; accessed 21-September-2019].

[33] HINTON, G. E 2014 *Boltzmann machines*. Springer.

[34] HINTON, G. E, A Practical Guide to Training Restricted Boltzmann Machines, [Online; accessed 18-September-2019].

[35] MACKAY, D. J. C 2003 *Information Theory, Inference and Learning Algorithms*. New Jersey: Cambridge University Press.

[36] KAMPEN, N. V 2007 *Stochastic processes in physics and chemistry*. North Holland.

[37] MEHLIG, B, HEERMANN, D. W & FORREST, B. M 1992 Hybrid Monte Carlo method for condensed-matter systems. *Phys. Rev. B* **45**, 679–685.

[38] KIRKPATRICK, S, GELATT, C. D & VECCHI, M. P 1983 Optimization by simulated annealing. *Science* **220**, 671–680.

[39] POTVIN, J. Y & SMITH, K. A 1999 Artificial neural networks for combinatorial optimization. In *Handbook of Metaheuristics* (ed. G F. & G Kochenberger). Heidelberg: Springer.

[40] MANDZIUK, J 2002 Neural networks for the n-Queens problem: a review. *Control and Cybernetics* **31**, 217, Special issue on neural networks for optimization and control.

[41] WATERMAN, M 1995 *Introduction to Bioinformatics*. Prentice Hall.

[42] HOPFIELD, J. J & TANK, D. W 1985 Neural computation of decisions in optimisation problems. *Biol. Cybern.* **52**, 141.

[43] SMOLENSKY, P 1987 *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, pp. 194–281. MITP.

[44] CARLEO, G & TROYER, M 2017 Solving the quantum many-body problem with artificial neural networks. *Science* **355** (6325), 602–606.

[45] MONTUFAR, G, GHAZI-ZAHEDI, K & AY, N 2014 A theory of cheap control in embodied systems. *PLoS Computational Biology* **11**, e1004427.

[46] BINDER, K, ed. 1986 *Monte Carlo Methods in Statistical Physics*. Heidelberg: Springer.

[47] SOKAL, A 1997 *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms*. Boston: Springer.

[48] MURPHY, K. P 2012 *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts: MIT Press.

[49] FISCHER, A & IGEL, C 2012 An introduction to restricted Boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (ed. L Alvarez, M Mejail, L Gomez & J Jacobo), pp. 14–36. Berlin, Heidelberg: Springer Berlin Heidelberg.

[50] MONTUFAR, G. F, RAUH, J & AY, N 2011 Expressive power and approximation errors of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems 11*.

[51] BENGIO, Y 2009 Learning deep architectures for AI. *Foundations and Trends in Machine Learning* **2**, 1–127.

[52] MACHINE LEARNING REPOSITORY UNIVERSITY OF CALIFORNIA IRVINE, archive.ics.uci.edu/ml, [Online; accessed 18-August-2018].

[53] FISHER, R. A 1936 The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179.

[54] MINSKY, M & PAPERT, S 1969 *Perceptrons. An Introduction to Computational Geometry*. MIT Press.

[55] COVER, T. M 1965 Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. on electronic computers* p. 326.

[56] KANAL, L. N 2001 Perceptrons. In *International Encyclopedia of the Social and Behavioral Sciences*.

[57] SOMPOLINSKY, H, Introduction: the perceptron, The perceptron, [Online; accessed 9-October-2018].

[58] GREUB, W 1981 *Linear Algebra*. New York: Springer.

[59] LECUN, Y, BOTTOU, L, ORR, G. B & MÜLLER, K.-R 1998 Efficient back prop. In *Neural networks: tricks of the trade* (ed. G. B Orr & K.-R Müller). Springer.

[60] NESTEROV, Y 1983 A method of solving a convex programming problem with convergence rate o($1/k^2$). *Soviet Mathematics Doklady* **27**, 372.

[61] SUTSKEVER, I 2013 Training recurrent neural networks. PhD thesis, University of Toronto, [Online; accessed 27-October-2018].

[62] RUMELHART, D. E, HINTON, G. E & WILLIAMS, R. J 1986 Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition* (ed. D. E Rumelhart & J. L MCClelland).

[63] HORNIK, K, STINCHOME, M & WHITE, H 1989 Neural networks are universal approximators. *Neural Networks* **2**, 359.

[64] LAPEDES, A. S & FARBER, R. M 1987 How neural nets work. In *Neural Information Processing Systems (NIPS 1987)*, pp. 442–456.

[65] CVITANOVIC, P, ARTUSO, G, MAINIERI, R, TANNER, G & VATTAY, G, chaosbook.org/version15 (Niels Bohr Institute, Copenhagen 2015), Lyapunov exponents, [Online; accessed 30-September-2018].

[66] PENNINGTON, J, SCHOENHOLZ, S. S & GANGULI, S 2017 Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems 30*.

[67] GLOROT, X, BORDES, A & BENGIO, Y 2011 Deep sparse rectifier neural networks. In *Proceedings of Machine Learning Research*.

[68] HE, K, ZHANG, X, REN, S & SUN, J 2015 Deep residual learning for image recognition arxiv:1512.03385.

[69] SUTSKEVER, I, MARTENS, J, DAHL, G & HINTON, G 2013 On the importance of initialization and momentum in deep learning ACM Digital Library.

[70] GLOROT, X & BENGIO, Y 2010 Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of Machine Learning Research*.

[71] SRIVASTAV, N, HINTON, G, KRISHEVSKY, A, SUTSKEVER, I & SALKHUTDINOV, R 2014 Dropout: A simple way to prevent neural networks from overfitting Journal of Machine Learning Research.

[72] HANSON, S. J & PRATT, L. Y 1989 Comparing biases for minimal network construction with backpropagation. In *Advances in Neural Information Processing Systems 1*.

[73] HASSIBI, B & G.STORK, D 1993 Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5*.

[74] LECUN, Y, DENKTER, J. S & SOLLA, S 1990 Optimal brain damage. In *Advances in Neural Information Processing Systems 2* (ed. D. S Touretzky), p. 598.

[75] FRANKLE, J & CARBIN, M 2018 The lottery ticket hypothesis: Finding small, trainable neural networks arxiv:1803.03635.

[76] IMAGENET, image-net.org, [Online; accessed 3-September-2018].

[77] IOFFE, S & SZEGEDY, C 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift arxiv:1502.03167.

[78] SANTURKAR, S, TSIPRAS, D, ILYAS, A & MADRY, A 2018 How does batch normalization help optimization? (No, it is not about internal covariate shift) arxiv:1805.11604.

[79] KIRKPATRICK, J, PASCANU, R, RABINOWITZ, N, VENESS, J, DESJARDINS, G, RUSU, A. A, MILAN, K, QUAN, J, RAMALHO, T, GRABSKA-BARWINSKA, A, HASSABIS, D, CLOPATH, C, KUMARAN, D & HADSELL, R 2016 Overcoming catastrophic forgetting in neural networks arxiv:1612.00796.

[80] SETTLES, B 2009 Active Learning Literature Survey. *Tech. Rep.* 1648. University of Wisconsin–Madison.

[81] CHOROMANSKA, A, HENAFF, M, MATHIEU, M, BEN AROUS, G & LECUN, Y 2014 The loss surfaces of multilayer networks arxiv:1412.0233.

[82] KRIZHEVSKY, A, SUTSKEVER, I & HINTON, G. E 2012 Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*.

[83] TENSORFLOW, tensorflow.org, [Online; accessed 3-September-2018].

[84] LECUN, Y & CORTES, C, MNIST, [Online; accessed 3-September-2018].

[85] SMITH, L. N 2015 Cyclical learning rates for training neural networks arxiv:1506.01186.

[86] CIRESAN, D, MEIER, U & SCHMIDHUBER, J 2012 Multi-column deep neural networks for image classification arxiv:1202.2745.

[87] PICASSO, J. P, Pre-processing before digit recognition for NN and CNN trained with MNIST dataset, stackexchange, [Online; accessed 26-September-2018].

[88] KOZIELSKI, M, FORSTER, J & NEY, H 2012 Moment-based image normalization for handwritten text recognition. In *Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition*.

[89] PASCAL VOC DATA SET, Pascal VOC, [Online; accessed 6-September-2018].

[90] REDMON, J, DIVVALA, S, GIRSHICK, R & FARHADI, A, You only look once: Unified, real-time object detection, arxiv:1506.02640.

[91] RUSSAKOVSKY, O, DENG, J, SU, H, KRAUSE, J, SATHEESH, S, MA, S, HUANG, Z, KARPATHY, A, KHOSLA, A, BERNSTEIN, M, BERG, A. C & FEI-FEI, L 2014 Imagenet large scale visual recognition challenge arxiv:1409.0575.

[92] LI, F, JOHNSON, J & YEUNG, S, CNN architectures, CNN architectures, [Online; accessed 23-September-2018].

[93] HU, J, SHEN, L & SUN, G 2018 Squeeze-and-excitation networks arxiv:1709.01507.

[94] SEIF, G, Deep learning for image recognition: why it's challenging, where we've been, and what's next, Towards Data Science, [Online; accessed 26-September-2018].

[95] Tensor processing unit, github.com/tensorflow/tpu, [Online; accessed 23-September-2018].

[96] SZEGEDY, C, LIU, W, JIA, Y, SERMANET, P, REED, S, ANGUELOV, D, ERHAN, D, VANHOUCKE, V & RABINOVICH, A 2014 Going deeper with convolutions arxiv:1409.4842.

[97] ZENG, X, OUYANG, W, YAN, J, LI, H, XIAO, T, WANG, K, LIU, Y, ZHOU, Y, YANG, B, WANG, Z, ZHOU, H & WANG, X 2016 Crafting GBD-net for object detection arxiv:1610.02579.

[98] HERN, A, Computers now better than humans at recognising and sorting images, The Guardian, 13 May 2015, [Online; accessed 26-September-2018].

[99] KARPATHY, A, What I learned from competing against a convnet on imagenet, blog, [Online; accessed 26-September-2018].

[100] KHURSHUDOV, A, Suddenly, a leopard print sofa appears, [Online; accessed 23-August-2018].

[101] KARPATHY, A, ILRSVC labeling interface, [Online; accessed 26-September-2018].

[102] GEIRHOS, R, TEMME, C. R. M, RAUBER, J, SCHÜTT, H. H, BETHGE, M & WICHMANN, F. A 2018 Generalisation in Humans and deep neural networks arxiv:1808.08750.

[103] SZEGEDY, C, ZAREMBA, W, SUTSKEVER, I, BRUNA, J, ERBAN, D. A ND GOODFELLOW, I & FERGUS, R 2013 Intriguing properties of neural networks arxiv:1312.6199.

[104] NGUYEN, A, YOSINSKI, J & CLUNE, J 2015 Deep neural networks are easily fooled: high confidence predictions for unrecognisable images arxiv:1412.1897.

[105] YOSINSKI, J, CLUNE, J, NGUYEN, A, FUCHS, T & LIPSON, H 2015 Understanding neural networks through deep visualization arxiv:1506.06579.

[106] OTT, E 2002 *Chaos in Dynamical Systems*, 2nd edn. Cambridge University Press.

[107] STROGATZ, S. H 2000 *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press.

[108] SUTSKEVER, I, VINYALS, O & LE, Q. V 2014 Sequence to sequence learning with neural networks arxiv:1409.3215.

[109] LIPTON, Z. C, BERKOWITZ, J & ELKAN, C 2015 A critical review of recurrent neural networks for sequence learning arxiv:1506.00019.

[110] PASCANU, R, MIKOLOV, T & BENGIO, Y 2012 On the difficulty of training recurrent neural networks arxiv:1211.5063.

[111] HOCHREITER, S & SCHMIDHUBER, J 1997 Long short-term memory. *Neural Computation* **9**, 1735.

[112] OLAH, C, Understanding LSTM Networks, [Online; accessed 30-September-2020].

[113] CHO, K, VAN MERRIENBOER, B, GULCEHRE, C, BAHDANAU, D, BOUGARES, F, SCHWENK, H & BENGIO, Y 2014 Learning phrase representations using rnn encoder-decoder for statistical machine translation arXiv:1406.1078.

[114] HECK, J & SALEM, F. M 2017 Simplified minimal gated unit variations for recurrent neural networks arxiv:1701.03452.

[115] MACHEREY, W, KRIKUN, M, CAO, Y, GAO, Q, MACHEREY, K, KLINGNER, J, SHAH, A, JOHNSON, M, LIU, X, KAISER, L, GOUWS, S, KATO, Y, KUDO, T, KAZAWA, H, STEVENS, K, KURIAN, G, PATIL, N, WANG, W, YOUNG, C, SMITH, J, RIESA, J, RUDNICK, A, VINYALS, O, CORRADO, G, HUGHES, M & DEAN, J 2016 Google's neural machine translation system: bridging the gap between Human and machine translation arxiv:1609.08144.

[116] PAPINENI, K, ROUKOS, S, WARD, T & ZHU, W.-J 2002 Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311.

[117] LUKOSEVICIUS, M & JAEGER, H 2009 Reservoir computing approaches to recurrent neural network training. *Computer Science Review* **3**, 127.

[118] PATHAK, J, HUNT, B, GIRVAN, M, LU, Z & OTT, E 2018 Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Phys. Rev. Lett.* **120**, 024102.

[119] JAEGER, H & HAAS, H 2004 Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* **304**, 78–80.

[120] LIM, S. H, GIORGINI, L. T, MOON, W & WETTLAUFER, J, Predicting critical transitions in multiscale dynamical systems using reservoir computing, arxiv:1908.03771.

[121] LUKOSEVICIUS, M 2012 *A practical guide to applying echo state networks*. Berlin,Heidelberg: Springer.

[122] TANAKA, G, YAMANE, T, HÉROUX, J. B, NAKANE, R, KANAZAWA, N, TAKEDA, S, NUMATA, H, NAKANO, D & HIROSE, A 2019 Recent advances in physical reservoir computing: A review. *Neural Networks* **115**, 100 – 123.

[123] DOYA, K 1993 Bifurcations of recurrent neural networks in gradient descent learning. *IEEE Transactions on Neural Networks* **1**, 75.

[124] WILLIAMS, R. J & ZIPSER, D 1989 A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* **1**, 270–280.

[125] DOYA, K 1995 *Recurrent networks: supervised learning*, pp. 796–799. Cambridge MA: MIT Press.

[126] KARPATHY, A, The unreasonable effectiveness of recurrent neural networks, webpage, [Online; accessed 4-October-2018].

[127] OJA, E 1982 A simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15**, 267.

[128] WILKINSON, M, BEZUGLYY, V & MEHLIG, B 2009 Fingerprints of random flows? *Phys. Fluids* **21**, 043304.

[129] WELIKY, M, BOSKING, W. H & FITZPATRICK, D 1996 A systematic map of direction preference in primary visual cortex. *Nature* **379**, 1476–4687.

[130] KOHONEN, T 2013 Essentials of the self-organizing map. *Neural Networks* **37**, 52 – 65.

[131] KOHONEN, T 1995 *Self-Organizing Maps*. Berlin: Springer.

[132] KOHONEN, T 1990 The self-organizing map. *Proceedings of the IEEE* **78**, 1464–1480.

[133] MARTIN, R & OBERMAYER, K 2009 Self-organizing maps. In *Encyclopedia of Neuroscience* (ed. L. R Squire), p. 551. Oxford: Academic Press.

[134] RITTER, H & SCHULTEN, K 1986 On the stationary state of kohonen's self-organizing sensory mapping. *Biological Cybernetics* **54**, 99–106.

[135] SNYDER, W, NISSMAN, D, VAN DEN BOUT, D & BILGRO, G 1990 Kohonen networks and clustering: Comparative performance in color clustering. In *Advances in Neural Information Processing Systems 3*.

[136] NG, A, Sparse autoencoder, [Online; accessed 13-October-2020].

[137] DOERSCH, C 2016 Tutorial on variational autoencoders arxiv:1606.05908.

[138] REZENDE, D. J, MOHAMED, S & WIERSTRA, D 2014 Stochastic backpropagation and approximate inference in deep generative models arxiv:1401.4082.

[139] POURKAMALI-ANARAKI, F & WAKIN, M. B 2019 The effectiveness of variational autoencoders for active learning arxiv:1911.07716.

[140] EDUARDO, S, NAZABAL, A, WILLIAMS, C. K. I & SUTTON, C 2019 Robust variational autoencoders for outlier detection and repair of mixed-type data arxiv:1907.06671.

[141] LI, C, GAO, X, LI, Y, PENG, B, LI, X, ZHANG, Y & GAO, J 2020 The effectiveness of variational autoencoders for active learning arxiv:2004.04092.

[142] GOODFELLOW, I. J, POUGET-ABADIE, J, MIRZA, M, XU, B, WARDE-FARLEY, D, OZAIR, S, COURVILLE, A & BENGIO, Y 2014 Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*.

[143] ROCCA, J, Understanding Generative Adversarial Networks, [Online; accessed 15-October-2020].

[144] SAMPLE, I, What are deepfakes – and how can you spot them?, The Guardian, 13 Jan 2020, [Online; accessed 30-September-2020].

[145] PRITCHARD, J. K, STEPHENS, M & DONNELLY, P 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**, 945.

[146] Human genome diversity project, HGDP, [Online; accessed 10-October-2018].

[147] WETTSCHERECK, D & DIETTERICH, T 1992 Improving the performance of radial basis function networks by learning center locations. In *Advances in Neural Information Processing Systems 4*, pp. 1133–1140. Morgan Kaufmann.

[148] POGGIO, T & GIROSI, F 1990 Networks for approximation and learning. In *Proceedings of the IEEE*, , vol. 78, p. 1481.

[149] SUTTON, R. S & BARTO, A. G 2018 *Reinforcement Learning: An Introduction*, 2nd edn. The MIT Press.

[150] COLABRESE, S, GUSTAVSSON, K, CELANI, A & BIFERALE, L 2017 Flow navigation by smart microswimmers via reinforcement learning. *Phys. Rev. Lett.* **118**, 158004.

[151] SUTTON, R. S 1988 Learning to predict by the methods of temporal differences. *Machine Learning* **3**, 9–44.

[152] BARTO, A. G 1985 Learning by statistical cooperation of self-interested neuron-like computing elements. *Hum. Neurobiol.* **4**, 229–56.

[153] WATKINS, C. J. C. H 1989 Learning from delayed rewards. PhD thesis, University of Cambridge, [Online; accessed 25-December-2019].

[154] SZEPESVARI, C 2010 Algorithms for reinforcement learning. In *Synthesis Lectures on Artificial Intelligence and Machine Learning* (ed. R. J Brachmann & T Dietterich). Morgan and Claypool Publishers.

[155] CROWLEY, K & SIEGLER, R. S 1993 Flexible strategy use in young children's tic-tac-toe. *Cognitive Science* **17**, 531–561.

[156] CICHOS, F, GUSTAVSSON, K, MEHLIG, B & VOLPE, G 2020 Machine learning for active matter. *Nature Machine Intelligence* **2**, 94–103.

[157] KLOPF, A. H 1982 *The Hedonistic Neuron: Theory of Memory, Learning and Intelligence*. Taylor and Francis.

[158] SUTTON, R. S, ed. 1992 *A special issue of machine learning on reinforcement learning, reprinted from Machine Learning Vol. 8, Nos. 3–4*. Springer.

[159] MORGAN, J, How to win at rock-paper-scissors, BBC News, 2 May 2014, [Online; accessed 7-September-2020].

[160] ALLIS, V, A knowledge-based approach of connect-four, Report IR-163, Faculty of Mathematics and Computer Science at the Vrije Universiteit Amsterdam, [Online; accessed 9-September-2020].

Index