Nada är en gemensam institution mellan
Kungliga Tekniska högskolan och Stockholms universitet.

Some computational aspects of attractor memory

MARTIN REHN

Avhandling som med tillstånd av Kungliga Tekniska högskolan
framlägges till offentlig granskning för avläggande av teknologie licentiatexamen
tisdagen den 15 mars 2005 kl 13.15
i sal E32, Lindstedtsvägen 3,
Kungliga Tekniska högskolan, Stockholm.

ISBN 91-7283-983
TRITA-NA-0509
ISSN 0348 2952
ISRN KTH/NA/R--05/09--SE

Abstract

In this thesis I present novel mechanisms for certain computational capabilities of the cerebral cortex, building on the established notion of attractor memory. A sparse binary coding network for generating efficient representation of sensory input is presented. It is demonstrated that this network model well reproduces receptive field shapes seen in primary visual cortex and that its representations are efficient with respect to storage in associative memory. I show how an autoassociative memory, augmented with dynamical synapses, can function as a general sequence learning network. I demonstrate how an abstract attractor memory system may be realized on the microcircuit level – and how it may be analyzed using similar tools as used experimentally. I demonstrate some predictions from the hypothesis that the macroscopic connectivity of the cortex is optimized for attractor memory function. I also discuss methodological aspects of modelling in computational neuroscience.

Keywords: Attractor memory, cerebral cortex, neural networks, sequence learning, generative models, serial order, computational neuroscience, connectionism.

ISBN 91-7283-983 • TRITA-NA-0509 • ISSN 0348 2952 • ISRN KTH/NA/R--05/09--SE

Acknowledgements

Thanks to Anders Lansner and the others in the SANS group at KTH for creating a unique environment, where brains meet computer science. It has been great fun to work with all of you; Christopher Johansson also read this manuscript. Many thanks also to Fritz Sommer and the others at the Redwood Neuroscience Institute, in particular Bruno Olshausen and Tony Bell, for a great collaboration and many interesting discussions.

Table of Contents

Chapter 1

Introduction

Attractor memory is an essential model for understanding information processing and computation in the cerebral cortex of the primate brain. Based on a very simple, yet elegant, mathematical construction, an attractor memory is a natural way of embedding arbitrary information content in the vast network of neurons and synapses that is the brain. Today we have a solid theory of the basic properties of attractor memory. There is also a mounting body of evidence that cortical circuitry does indeed implement attractor memory function, even though details remain to be determined. But a memory mechanism in itself is of little evolutionary value, unless coupled with the ability to interact with the external world. Sensory impressions need to be processed into a form suitable for storage. The information stored must be retrievable in a controlled fashion and ultimately used for producing motor output.

In this thesis I will investigate some questions related to how an attractor memory system in the cerebral cortex may work and interact with its environment. Under the constraints of known properties of the cortex, and certain optimality conditions, this will lead to very specific propositions about cortex as an attractor memory and about these interactions. For the domains under study, these can be thought of as part of the set of primitive computational operations available to the brain.

1.1 Structure of the thesis

The thesis begins with a critical look at some of the methodology employed in the field of computational neuroscience, appearing in chapter 2. In particular, I discuss the use of realistic models, when underconstrained by empirical data. In chapter 3, I review some relevant aspects of the cerebral cortex and explain how and why it may implement an attractor memory system. This chapter also contains an introduction to attractor memory models and the tools used to analyze them. The thesis then, in chapters 4–5, discusses how attractor memory fits into a larger picture of the brain; how it receives its input and how it may be extended with the ability to perform temporal tasks, including cognitive processing and the production of motor output. Next, in chapter 6, I describe a biophysically detailed attractor memory model; its differences and similarities to the abstract models so far employed and how to translate between high level models and more detailed ones. Moving from the small to the big, chapter 7 describes global aspects of attractor memory functionality; it deals with scaling such a model over several orders of magnitude, to the size of the human brain, and what constraints this imposes on the topology of the system. The thesis concludes in chapter 8 with a summary and discussion of the results and an outlook on further work.

1.2 Contributions

- I critically examine the methodology of modelling as used in theoretical neuroscience.

- I present the sparse assembly coding network (SACN); a novel principle for representational learning in a cortical region.

- I show that the SACN is more efficient in generating a binary sparse code than a pruned graded model.

- I demonstrate that receptive fields produced by the SACN match experimental observations fields in V1 better than previous models.

- I show that the SACN generates a code that is efficiently processed by an associative memory.

- I present a model for sequence learning (SL), based on attractor memory.

- I demonstrate that the SL model reproduces data from free recall experiments.

- I show how the SL model may form a basis for generic sequence processing in the cortex.
- I analyse a realistic model of a cortical circuit, extracting measures directly comparable to experimental data.
- I demonstrate topological constraints on scalable attractor memory.
- I mathematically analyse an attractor memory model, demonstrating the advantage of a certain form of connection structure ("patchy connectivity").

1.3 Articles

This thesis is based on the following articles, which also appear as appendices, numbered as below.

I. Martin Rehn, Anders Lansner, "Sequence memory with dynamical synapses", *Neurocomputing* 58–60 (2004), 271–278.

II. Martin Rehn, Friedrich T. Sommer, "Early sensory representation in cortex optimizes information content in small neural assemblies".

III. Martin Rehn, Friedrich T. Sommer, "A network model for the rapid formation of binary sparse representations of sensory inputs".

IV. Christopher Johansson, Martin Rehn, Anders Lansner, "Attractor Neural Networks with Patchy Connectivity". *(ESANN 2005, oral presentation)*

   - Christopher Johansson primarily contributed to the simulation experiments.
   - I primarily contributed to the mathematical analysis.

V. Mikael Lundqvist, Martin Rehn, Anders Lansner, "Attractor dynamics in a large-scale modular network model of neocortical layers 2/3".

   - I primarily contributed to the model analysis.

Chapter 2

On modelling

Several areas of research strive to increase our understanding of the operational principles of the brain and the nervous system. Some theoretical approaches, in order of increasing level of abstraction are neurophysiological modeling, computational neuroscience and artificial neural networks. But why study the brain? One general reason is that the brain has impressive abilities, unmatched by any other system; natural or artificial. Three specific goals, or objectives, for studying the brain, all motivated by this observation, are:

1. Understanding the biological system
2. Investigating its mathematical properties
3. Incorporating its principles into technological systems.

These three objectives correspond to scientific, mathematical and engineering approaches, respectively. We expect different methods and different success criteria to characterise the pursuit of these goals. The work described in this thesis falls under the first objective; the goal is to understand the brain, not to prove theorems or to build machines. Even though the overall goal is thus clear, it turns out that subgoals and methods from the other approaches also figure. In this chapter I will further clarify methodological issues concerning the work presented here.

One tool that is used regardless of the ultimate research goal is modelling. Specifically, one may model parts of the system under study, at different scales. In the case of the brain, models range from subparts of individual neurons, such as ion channels, to networks of neurons, to psychological models of the whole organism. Some models may be structurally very unlike what they model; for instance a nerve cell may be modeled using an information theoretic construct (Bell and Sejnowski, 1995). On the other hand, many models in this field share key structural aspects of what they are modelling; they will incorporate the geometry of a nerve cell or the topology of a network of cells. Such models are known in the methodological literature as *realistic* (Lloyd, 1998). In computational neuroscience, the term realistic is sometimes used in a narrower sense, meaning *biophysically detailed* (Durstewitz et al., 2000).

2.1 Modelling objectives

Three separate *objectives* for studying brain-like models were listed above; one scientific, one mathematical and one technological, each having its own motivation. Of course, this classification is in part an artificial one; one researcher may well be motivated by more than one cause and contribute to more than one objective. Nevertheless I will introduce the different motivations in their pure forms below.

2.1.1 Scientific objective

Under the first of the three objectives, the *scientific* goal of understanding the brain, we will judge a model according to how well it ”corresponds to” biology. Models are motivated by empirical data and should be experimentally testable. Occam's razor applies in that when two models explain known facts equally well, the simpler one (such as the one with fewer parameters) should be preferred, as it most likely has the strongest predictive power. Sometimes simplicity will conflict with realism; making a model share structural properties of biology may incur a cost of additional complication, so we may face a tradeoff between structure and parsimony.

In practice, for realistic models, the immediate value of constructing the model is often that it provides hints on which further experiments might be the most crucial, namely the ones that would decide on uncertainties in the model. In this sense, these models become heuristic tools (Ekeberg, 1992, p. 19).

Determining ”correspondence” between a model and reality is an interesting problem in itself. We do not have immediate access to the inner workings of nature; all we can do is to perform experiments and observe the outcomes. For a model that we ourselves have constructed, on the other hand, we do know the details of its inner structure. This presents us with an apparent asymmetry between model and reality; for the former we can directly read off parameters; for the latter experiments must be performed. In the case of realistic models, this is sometimes indeed the end of the story; some of the model parameters directly correspond to ion channel parameters, physical constants or other things that we believe also to be basic aspects of reality. In general, however, matters are not so simple. More on this latter case in section 2.2.1.

2.1.2 Mathematical objective

The second purpose of studying models based on the brain, such as artificial neural networks, is to discover their mathematical properties. In mathematics we may choose to study whatever formal systems we prefer; there is no requirement that it be rooted in reality. Nevertheless, the power of the biological brain may indicate that interesting mathematics lies hidden therein, motivating the study of related models. But once a model has been chosen, it is treated by the mathematician as a purely abstract entity (Whitehead and Russell, 1927). The mathematical approach to the brain has been a fruitful one; very general results have been obtained for some of the most abstract models, regarding e.g. representational ability, learning properties and memory capacity (Amit et al., 1987; Minsky and Papert, 1988; Hertz et al., 1991).[1]

From a mathematical point of view, simple models are often the best to work with, since proving general theorems may otherwise be prohibitively difficult. Theoretical work surrounding a model is judged according to conventional criteria for mathematics or indeed science in general; a good theory makes specific statements about a general class of circumstances (Kuhn, 1970). If we want to put the mathematics to use, another criterion should be added; that the mathematical model be applicable to a real world problem. In fact, one does not apply mathematical tools directly to reality, but to yet another *model*, derived from biology or from an engineering application. It is to this model that the mathematical theorems developed should, at least approximately, be applicable. Therefore, this will impose a competing constraint on the choice of a mathematical model; while it still needs to be simple in or-

[1] Important contributions have been made by physicists and computer scientists, as well as mathematicians. I refer to as mathematical all efforts that do not lean on empirical data.

der to be tractable, it also must be similar enough to the more detailed model that results carry over.

2.1.3 Engineering objective

The third reason why one might be interested in building models imitating the brain is to create technological artifacts. Noting the superiority that natural systems hold over current technology in many fields, such as pattern recognition and interaction with natural environments, the hope is to reduce this gap by borrowing some operational principles from the brain. There is no reason to copy all of the brain, nor to discard traditional engineering approaches altogether. Therefore, the brain-like models chosen need not be biologically plausible, in the sense that they could reasonably be implemented in the brain itself. Thus un-biological features, such the ”backpropagation” learning rule (that is unlikely to be realized in the brain) may well be used in the brain-like construction. That in turn becomes just one of several engineering building blocks. The ultimate goal is a well functioning technical system; that may perform data analysis, control a robot or carry out some other task. At least in the short run, therefore, the engineering and scientific projects may well not make substantial contributions to each other. The models preferred for first purpose do not qualify as models of the brain – and the models painstakingly developed to be faithful to the brain are often outperformed by existing engineering software, optimized for serial computers.

2.1.4 Craftsmanship

Constructing a neural network model is often a difficult task; one that requires substantial experience. Regardless of what originally motivated constructing such a model, a lot of time and energy will devoted to the *craft* of making it work. Both for models intended to model the brain and those intended to perform an engineering function, parameters need to be tuned and the implementation optimized, before they can produce anything useful at all. A substantial investment of time and resources goes into these tasks, an investment that may shift attention from the original research question to problems specific to model building itself; an individual scientist or a group of scientists may spend years working on the same model. The following motivation may therefore become every bit as important:

Objective	***Object***	***Truth criterion***	***Goal***
Science	nature	hypothesis testing	explanation of phenomena
Mathematics	system of axioms	deduction	insight on logical relations
Technology	artifacts	function	ability to perform tasks
Craft	model	aesthetics	personal skill

Table 2.1: Four different projects, all working with models of the brain, are contrasted with regard to methodological issues.

4. Modelling as a craft.

In contrast to the other objectives, the *craftmanship objective*, does not have objective success criteria. Rather, the goal might be for the individual or the group to improve their ability to create artifacts, which are in turn judged according to aesthetic criteria.

2.1.5 Mixing of objectives

In practice, as was hinted in the introduction to this section, the objectives here described are often mixed together. Individual researchers may pursue more than one of them at a time, not necessarily keeping them separate in the daily work. From one perspective this is a good thing; the different projects have much to learn from each other. If, on the other hand, one does not make clear which project is currently being pursued, then potentially dangerous ground is trod. It is then unclear what *object* is being studied, which *truth criterion* is used to evaluate results, and what the ultimate *goal* of the research is. In table 2.1 the different projects are contrasted with respect to these points.

2.2 Modelling and realism

2.2.1 A different kind of empirical science

Traditionally, science has distinguished two forms of activity; empirical investigations and theoretical work. The former is concerned more directly with reality, which is probed using experiments and other investigations, in order to ultimately produce data. Theoretical activity deals with stating hypotheses, analysing them and deriving predictions from them. Computer models fit nicely into this framework as special forms of hypotheses. They are sources of predictions for the experimental work.

Producing predictions from a model may sometimes be done using purely mathematical tools. For instance, conditions for some measure of optimality may be derived for a neural model, leading to a prediction that these conditions are satisfied in biology. For realistic models, direct comparison of model parameters to reality is sometimes possible. But in general, one must simulate the model's behaviour, perhaps under several different conditions, and produce measurements from the simulations. This is similar to conducting an empirical study. Experiments are performed on the model using methodology similar to that used in the laboratory, only they are being carried out in the ”alternate universe” wherein the model lives. If a discrepancy is then found between model behaviour and the outcome of a laboratory experiment, neither the model nor the ”real” experiment is necessarily at fault; it is also possible that the ”virtual” experiment in the model universe was flawed. The symmetric relationship between model and reality is illustrated in figure 2.1. One consequence of this is that one should stick to one and the same model for long enough to experimentally probe its properties, or one will be chasing a moving target. Also note that from a methodological perspective, there is nothing wrong with a model that must be subject to computer simulation to be useful; all that we must ask is that there is some specified procedure to produce predictions from the model.

2.2.2 Realism

The diverse class of models that are collectively known as ”artificial neural networks” borrow their fundamental structure from the nervous system. They are constructed from a number of nodes, corresponding to biological neurons, which are linked by connections, corresponding to synapses. This prompts us to consider them realistic models. One advantage of realism in modelling is that it makes it relatively straightforward to translate between model and biological system; some observation that pertains to the nodes of an artificial neural network model should also be applicable to the neurons of the brain. But in fact, there are a number of qualifications to this identification. A node

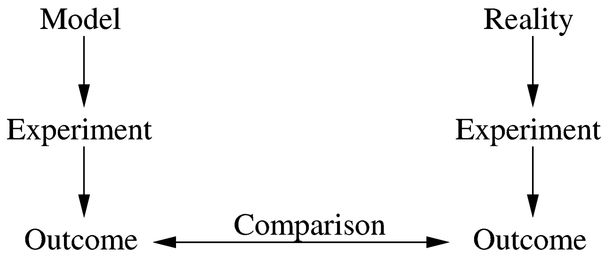


Figure 2.1: Comparison of a model to reality.

in an artificial neural network may include more computational machinery than a biological neuron does. For instance, while a biological neuron is generally restricted to being either excitatory or inhibitory[2], nodes in artificial networks are often allowed to transmit both types of signals. In such a case, one node in the model should rather be identified with a group of neurons working together, containing both excitatory and inhibitory neurons (see chapter 6). But there may also be a one-to-many mapping in the other direction; perhaps the simple, atomic nodes of an artificial neural network should really be thought of as lower level components of biological neurons, the latter being capable of performing parallell processing in their dendritic trees (Taylor et al., 2000; Hausser et al., 2000).

Taking these objections seriously and ruling out a one-to-one mapping between natural and artificial neurons makes it nontrivial to translate model results to predictions about nature. What then separates neural network models, as applied to e.g. psychological data, from purely phenomenological models, is that their underlying assumptions are ”realistic” or ”analogous” to what is known about biological networks (Lee et al., 1998). Specifically, one assumption may be that any information processing in the model is localized. In the words of Milton Friedman, we implicitly claim that ”the conformity of these ’assumptions’ to ’reality’ is a test of the validity of the hypothesis *different from* or *additional to* the test by implications.” Friedman however, claims that this view is fundamentally flawed and detrimental to hypothesis driven sci-

[2] This rule was long thought to be a fundamental principle of neural interactions, but lately some exceptions have been discovered.

Level	Subcellular	Neuron	Network	System
Methods	pharmacological, genetic knock-out	patch clamp recording	multicell recording, anatomy	imaging, psychology, psychophysics

Table 2.2: Some examples of experimental techniques, ordered according to the the level of understanding to which they contribute most directly.

ence (Friedman, 1966). This criticism should be moderated, however, by noting that realism may often be a prerequisite for producing predictions in the first place. A physiologically detail model, for instance, has a wider interface to experimental work, in that certain model features, such as ion channel properties, can sometimes be tested in a very direct way, perhaps by experimentally blocking that ion channel by an antagonist substance.

2.2.3 Model determination

Why are realistic models used in computational neuroscience if there is a risk, according to Friedman's view, that they undermine the rigor of the field? One simple reason is the relative paucity of constraints on models from biology. To illustrate the basic picture, table 2.2 lists some current experimental techniques. They are ordered according to the scale at which they probe the biological system, from local to global. It is such methods that are the source of empirical data against which we compare model predictions and hence constrain certain models and rule out others. On the network level, which is at the focus of this thesis, one of the most impressive experimental techniques is multi-electrode recording. Using a large number of electrodes, this technique makes it possible to record spike events from hundreds or more individual neurons (Nicolelis and Ribeiro, 2002; Pellizzer et al., 1995). While this is spectacular in itself, it samples just a minuscule fraction of the billions of neurons in the neocortex. The information thus obtained is somewhat anecdotal in character; some reports detail the response characteristics of individual neurons. Further constraints are thus needed.

2.2.4 Additional objectives

One of the studies reported in this thesis deals with sequence learning in the cerebral cortex (chapter 5). I will here use that study as an example case. The aim of the study is to learn about how the brain solves tasks with prominent temporal aspects; think of learning to sing a song. In order to understand this type of behaviour, we would like to build a model that performs similar tasks and to demonstrate that the model shares important properties with the brain. For instance, we would constrain the model to using mechanisms known from biological neurons (subcellular and neuron levels in table 2.2) and we would compare its network activity to multi neuron recordings (network level).

Human subjects have performed various sequence learning tasks, their performance being recorded (system level) (Koch and Hoffmann, 2000; Avons, 1998). We might ask of our model to perform similarly to humans on sequence learning tasks. This would mean that tasks that prove hard for people should also be hard for the model. If possible, the failure modes of the model should be similar to the mistakes humans make. Note how this differs from the engineering objective of making the system perform as good as possible in all cases. These two requirements impose constraints on the model, the second one is the more specific, since it has a higher dimensionality. Continuing to constrain the model, we may next add several more requirements:

- Efficiency
- Scalability
- Robustness.

These requirements mean that we prefer models that can store long sequences, to those that can only store shorter ones. We prefer models whose storage capacities increase as fast as their resources for storage, usually the number of plastic synapses. We prefer models that are little affected by noise and variability in the model components. All of this can be motivated from an evolutionary perspective; the cortex is likely to have an efficient design, to be similar to its evolutionary predecessors and to not easily break down. One objection to these heuristic criteria could be that they are not directly based on experimental evidence. Although they are also likely realized by an evolutionary process, are they not in fact engineering objectives, the pursuit of which might lead to solutions that perform on par with the brain, but bears little resemblance to it? Certainly, it is hard to see how these assumptions could experimentally be put to test. On the other hand it is generally the case that each scientific paradigm has at its core certain unproven assumptions; we might put the above hypotheses of optimality into this category. This is perfectly acceptable, as long as productive science is produced in the community (Kuhn, 1970).

2.3 Conclusions

A common view is that good science should deal with producing hypotheses that are in some sense testable; *verifiable* or *falsifiable* (Popper, 1959). Another criterion is that it should postulate *mechanisms for* or *explanations to* phenomena (Bechtel and Abrahamsen, ress; Woodward, 2003). This second criterion is seemingly fulfilled in theoretical neuroscience; the heuristic criteria for model selection, and the preference for realistic models, generally lead to explanations in mechanistic terms. In principle, such models are also falsifiable, but there may be practical problems. Models are often nebulous in nature, having a wealth of parameters and variants (Collins and Pinch, 1998). Furthermore, a prescription is needed for how to generate predictions from a model. Thus it may be harder to test a model in neuroscience than some other scientific hypotheses. The conclusion to draw from this is that we should strive to make models easily falsifiable. This is achieved by contrasting core claims in a model from the free parameters and by clearly stating how to generate predictions.

Chapter 3

The cortex as an attractor memory

3.1 The cerebral cortex

3.1.1 Development

One of the early steps in the development of the vertebrate embryo is the gastrula phase; so called because what is to become the digestive tract is then first apparent (Larsen, 2001). During this phase, a bulge is formed in the ectoderm, the outermost of just three cell layers in the embryo at this stage, the one which later also develops into the skin. From this bulge, known as the *neural plate*, located in the back of the embryo, the neural system in its entirety is formed (Streit and Stern, 1999; Xanthos et al., 2002). As the cells in the neural plate divide, its center begins to protrude inwards, forming a *neural groove* and finally closing in on itself into a *neural tube* (O'Rahilly and Muller, 1994). The part towards the tail of the neural tube (the *caudal* part) will develop into the spinal cord, a relatively homogeneous structure, differentiated primarily along a dorsal-ventral axis, with neurons receiving and processing sensory information located *dorsally* (towards the back) and motor processing located *ventrally* (towards the belly) (Kandel et al., 2000, Figure 17-3, p. 319).

The part of the neural tube towards the nose (the *rostral* part) is on the other hand destined for greater glory. It undergoes a series of differentiations. First it develops into three clearly distinguishable, rounded cavities: the forebrain, midbrain and hindbrain (Larsen, 2001). Towards the forebrain, the dorsoventral division between sensory and motor regions, prominent in the spinal cord, becomes less distinct and moves ventrally, meaning that the forebrain is likely developed entirely from the dorsal, sensory side. After this first differentiation, the forebrain and hindbrain cavities are further divided, whereas the midbrain is not. The forebrain's anterior part is now called the telencephalon and the posterior part the diencephalon. The rostral part of the hindbrain forms the metencephalon, the caudal part the myelencephalon. At this point in development we therefore have six anatomically separate parts in the nervous system; the spinal cord, the myelencephalon (which will develop into the lower part of the brain stem), the metencephalon (the upper brain stem and the cerebellum), the midbrain or mesencephalon (e.g. the tectum), the diencephalon (the thalamus and the hypothalamus) and finally the telencephalon (the basal ganglia, the hippocampus and the cerebral cortex). Now the cerebral cortex and other structures begin to grow rapidly outwards. The cerebral hemispheres emerge as two prominent bulges, that will eventually, in the human brain, encase and cover all of the midbrain structures.

3.1.2 Anatomy and function

The cerebral cortex forms the surface of the cerebral hemispheres, the largest structure in the human brain. Underneath the cortex itself, where cell bodies, synapses and short range connections are located, is the white matter, containing long range axons (Kandel et al., 2000, p. 322). Most of these connect different parts of the cortical surface; when referring to the cortex as a functional system, these corticocortical connections through the white matter are often implicitly included. The cerebral hemispheres also contain the hippocampus, the basal ganglia and the amygdala. The hippocampus is located ventrally of the lateral ventricle, medially (on the inside) of the temporal lobe. It is sometimes thought of as an unusually agile version of the cortex itself; it is involved with intermediate storage of memories, before they are committed to long term memory (Treves and Rolls, 1992, 1994; Rolls and Treves, 1997). The basal ganglia are located in the center of the cerebral hemispheres, near the lateral ventricle. They play an important role in executive functions; decision making with regard to initiating and stopping movements and are most likely also involved with cognitive decisions (Mink, 1996). The amygdala is a set of nuclei in the temporal lobe that is involved with innate and learned responses related to fear and other emotions (Gallagher and Chiba, 1996).

The first thing to note about the large scale connectivity of the cortex is that it mostly listens to itself. The majority of the long range connections arriving at any cortical area originate in other areas of the cortex (Abeles, 1991). This is true even for the primary sensory cortices; which may otherwise be thought of as the input areas of the cortex (Lamme et al., 1998; Salin and Bullier, 1995). Sensory inputs from all modalities but olfaction arrive at these areas via nuclei in the thalamus. But the thalamus is not a pure input processor either; it receives most of its input from the cortex (Kandel et al., 2000; Guillery and Sherman, 2002a, pp. 341–343). The thalamocortical system thus forms a loop, where information is passed from the thalamus to the cortex, is processed by the cortex and then passed back to another part of the thalamus. While anatomically it is part of the diencephalon (rather than the telencephalon) the thalamus is therefore tightly integrated with the cortex from a functional perspective.

Each hemisphere in the cerebral cortex is divided into four lobes; occipital, parietal, temporal and frontal. These lobes, which form the visible area of the cortex, are clearly separated by deep sulci. But there is also a hidden part; in particular the lateral sulcus, separating the temporal lobe from the parietal and frontal lobes, hides a substantial cortical surface area, called the insular cortex. Also hidden is the cingulate cortex, located between the two hemispheres.

3.1.3 Evolution

In its present, much expanded form, the cerebral cortex is a relatively young structure, but already in the amphibian brain, there are analogues to the different parts of the human telencephalon; the hippocampus, the pyriform olfactory cortex and the isocortex (Northcutt and Kaas, 1995). The likely precursor to the latter is known as the pallium. Like the isocortex, it receives inputs from different sensory modalities and it displays reentrant connectivity, in that connections are often reciprocal rather than feed forward. Unlike the isocortex, the pallium does not have specific areas for different modalities and it lacks any trace of the homogeneous, layered structure of the isocortex (Herrick, 1948). Its structure is even ”inside-out” compared to the mammalian cortex; in the pallium the cell bodies are on the inside, near the ventricles, axons are on the surface (Super and Uylings, 2001). Moving forward to the reptilian brain, we find the pallium much enlarged, but still lacking layered structure (Aboitiz et al., 2002, 2003). In contrast, the reptilian pyriform cortex, which is involved with olfactory processing, has already formed the three-layered structure seen in mammals.

In all living mammals, the isocortex displays a homogeneous, layered structure; hence the name. For this reason, we may assume that this was the case already for the earliest mammals (Northcutt and Kaas, 1995). What is remarkable about the evolution of the cortex since then is not innovation, but the lack of it. A great growth of the cortical surface area has taken place without changing the fundamental design – and with little additional differentiation between cortical areas. It thus seems that the cortical design was exceptionally well suited for scaling. This scaling of the cortex has taken place by increasing the size of individual areas, but even more so by adding to the number of areas. This process has been interpreted as new areas being added to perform higher order processing. Speaking for that interpretation is the fact that in the primate cortex, the prefrontal areas, associated with abstract planning activities, have seen the most rapid expansion (Fuster, 2002).

3.1.4 Microcircuitry

Local circuit

There are two main classes of neurons in the cortex; pyramidal cells and interneurons. The pyramidal cells, shaped like little pyramids, pointing outwards, are responsible for all non-local communication. Any information that leaves a local cortical circuit is passed through the axon of a pyramidal cell. The pyramidal cells are excitatory cells, communicating mainly by the neurotransmitter glutamate. Depending on the layer to which they belong, pyramidals either tend to project laterally, within the cortex, or send their axons into the white matter, projecting to other cortical areas or subcortically. Both is often the case; axons typically branch and project to more than one destination. Pyramidal cells receive their input through two sets of dendrites. The apical dendrite extends like an antenna tower from the top (the apex) of the pyramid, ascending to the cortical surface. The basal dendrites are found near the base of the pyramid and extend horizontally. There are numerous types of interneurons. The spiny stellate cells are glutamatergic and also share other properties with pyramidal cells (Staiger et al., 2004). They are prominent in layer IV of primary sensory cortices, where thalamic input is received. But most interneurons are inhibitory, their main emitted transmittor substance being GABA. They are collectively known as non-spiny stellate or granule cells; small, locally projecting, inhibitory cells that come in an assortment of shapes.

One major class are the basket cells, so named because their axons tend to encase the cell bodies of the pyramidal cells on which they synapse. Most basket cell dendrites are local, some extending about 100 μm, whereas their axons can reach upwards of 1 mm (Wang et al., 2002). Another class is the double bouquet cell. They project to dendrites of pyramidal cells inside a narrow (<100 μm) vertical column (Markram et al., 2004). Chandelier cells project to the axons of pyramidal cells; they have been found to reliably fire when the overall activity in a local circuit is high and have therefore been hypothesized to moderate excitatory signals in that condition (Zhu et al., 2004).

The isocortex is organized in six anatomical layers. Of particular interest to us will be the cortical layers II/III. The pyramidal cells in these layers project mostly locally, within the cortex. In the rat visual cortex, about 70% of the projections terminate within 300 μm of the cell body, but there are also lateral connections of longer range (Nicoll and Blakemore, 1993). Thus they form a local, highly recurrent network that may form a substrate for attractor memory.

Projections

Having already mentioned the local circuitry of layer II/III, I will now relate functional aspects of the other cortical layers. Layer IV may be thought of as the input layer. It receives input from the thalamus and from ”earlier” or ”lower” cortical areas. Layers I and VI also receive input, but in the form of feedback from ”later” or ”higher” cortical areas. By a lower area is understood one that is closer to a primary sensory input stream; a signal pathway that enters the thalamocortical loop for the first time. Because cortical areas are interconnected in an intricate web and most connections are reciprocal, the notion of a hierarchy could not be extended much beyond the first few steps; if it were not for the different roles of the cortical layers. As it turns out, one can construct a fairly consistent hierarchy of cortical areas using the rule that projections to layer IV are forward connections and connections to layers I and VI are backward projections (Felleman and Van Essen, 1991).

Layers I-III are the source of feedforward corticocortical connections, those that terminate in layer IV. The feedback connections originate in layers V and VI. Layer v is the output layer of cortex. In the primary motor areas this is literally true; here large pyramidal cells send off axons that control movement, either directly (corticospinal fibers) or through motor centers in the brain stem (corticobulbar) (Kandel et al., 2000, p. 671f). In the case of distal muscles, the former class of pyramidal cells are just one synapse away from directly controlling the muscles; they project to the large ”alpha” motor neurons in the spinal cord, that in turn innervate muscle fibres (de Noordhout et al., 1999; Ziemann et al., 2004). With regard to the hierarchical terminology, the dual role of layer V as a source of feedback and motor output may seem to be a paradox; should not the motor output be in the forward direction? But the primary motor cortex, where the motor output is produced, is at the bottom of the motor hierarchy; it contains the lowest abstraction level, just like the bottom rungs of the sensory hierarchy, because it must operate on a concrete representation of motor actions, in terms of simple motor programs and individual muscle movements (though primary motor cortex also processes more complex information) (Scott, 2003). Noting this, one can also label the forward projections, from the superficial layers to layer IV, a sensory stream and the backwards projections, from layer V to layers I and VI a motor stream. The thalamus is not only the sensory gateway to the cortex; also the cortical motor output impinges on the thalamus, which receives branches of outgoing motor axons originating in layer V, together with modulatory input from layer VI (Guillery and Sherman, 2002b).

Columnar structure

A standard technique for investigating the brain is to cut it in thin slices and observe them under a microscope. To do this, the preparation must be stained with an agent that colors just one component of it. For instance, the Nissl stain colors the ”Nissl substance”, related to protein synthesis, therefore selectively staining neural cell bodies (Simmons and Swanson, 1993). When this staining was applied to vertical slices of the cortex, cell bodies were found to be localized in narrow columns, 35-60 µm wide in the human cortex. Each such column contains about 80-100 neurons; both pyramidal cells and interneurons (Buxhoeveden and Casanova, 2002). It has been suggested, based on vertical connectivity within the minicolumn, that it acts as a functional unit, meaning that all neurons in the column receive much the same input (Peters and Yilmaz, 1993). In primary sensory cortices it indeed has been found that neurons within a minicolumn share receptive field properties; i.e. they respond to the same stimuli (Hubel and Wiesel, 1977; Favorov and Kelly, 1994; Sugimoto et al., 1997). An attractor memory built on minicolumns rather than the neurons as functional units allows for more dense connectivity and hence higher storage capacity (Fransén and Lansner, 1998).

Columnar structures on a larger scale have been found in several cortical areas. The term ”hypercolumn” was originally used to describe the finding of localized areas in visual cortex that include representations of the full range of a visual variable; ocular dominance or orientation preference. It was nevertheless proposed as a general organizational principle for the cortex and indeed similar structure has since been found elsewhere (Mountcastle, 1997). In the context of attractor memory, hypercolumns are interesting because they imply a normalization property over minicolumnar activity. If one hypercolumn covers the full range of some variable, and if the activity distribution over the constituent minicolumns code for the brain's estimate of that variable, then the total activity should sum to one, or to some measure of confidence in the estimate (Carandini et al., 1997).

3.1.5 Synaptic plasticity

In 1949, Donald O. Hebb proposed the following rule for how the connection between two neurons should be modified, depending on activity (Hebb, 1949):

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

Hebb then pointed out that no such mechanism was known at the time; only much later it was found that such a mechanism does indeed exist. It was found that synaptic transmission, under certain conditions, could become more efficient after repeated stimulation. Long term potentiation (LTP), as the phenomenon was called, was first found in the hippocampus, not surprisingly, since that is probably one of the most plastic part of the brain (Lomo, 1968, 1971). Mechanistic explanations of several forms of plasticity followed (Fox and Lloyd, 2002). One important component of learning and synaptic plasticity is the NMDA receptor (Newcomer and Krystal, 2001). These receptors, found postsynaptically, are gated by glutamate, directly regulating membrane ion channels.[1] NMDA receptors are special when compared to other glu-

[1] The receptors are named for a substance, N-methyl-D-aspartate, that is used to selectively control them in the laboratory.

tamate gated receptors, in two respects. Firstly, the ion channels they gate are normally blocked by magnesium ions; only when the cell membrane is depolarized is the blocking removed and the channel becomes effective. Secondly, the channels are permeable to calcium ions, which are known to trigger and regulate a large number of intracellular functions. Specifically, calcium levels in dendrites have been shown to correlate with LTP (Ismailov et al., 2004). To become active, the receptor requires simultaneously a presynaptic spike (to provide glutamate gating) and postsynaptic depolarization (to remove the magnesium blocking). When both conditions are met, there is a calcium influx that lead to potentiation of the synapse. The NMDA receptor therefore serves as a coincidence detector between pre- and postsynaptic activity, triggering synaptic plasticity and realizing Hebb's learning rule (Colbert, 2001).

3.2 Attractor memory

The term ”attractor memory” denotes a fairly large class of abstract neural network models. What they have in common is that they provide a means of embedding information as attractors in a dynamical system. One example of a *dynamical system* is just an interconnected network of neurons; the ”dynamics” being the equations describing how the state of neurons and synapses evolve in time. As the word implies, an *attractor* is something that will tend to attract the evolution of such a system, if the system state comes sufficiently close to the attractor. In fact, for a deterministic system, the initial condition alone determines the attractor in which it will end up, and once there the system will then never again depart. The attractor itself may be a point in state space, a limit cycle or an irregular-looking ”strange” attractor.

3.2.1 The sparse Hopfield network

The autoassociative Hopfield network is a very pure and simple associative memory. The network consists of just a single pool containing N artificial neurons, all interconnected through N^2 artificial synapses, collected in a matrix J_{ij}. The network is capable of storing static patterns taking the form of binary vectors $\xi = [\xi_1, \xi_2, \dots, \xi_N]$ where $\xi_i \in \{0, 1\}$. It operates, in its learning phase, by a version of Hebb's rule; basically the reciprocal synaptic connection between a pair of neurons is strengthend when they are both active. More precisely, for storing sparse patterns, where units are inactive most of the time, the optimal form of the learning rule is based on the deviation of unit activity from its mean (Okada, 1996; Frolov et al., 1997). When storing a set of p patterns $\{\xi_i^1 \dots \xi_i^p\}$ the learned synaptic weights become by this rule:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \left(\xi_i^\mu - \langle \xi_i \rangle\right) \left(\xi_j^\mu - \langle \xi_j \rangle\right).$$

Once the network has been trained in this way (the J_{ij} matrix filled in) the network may be used to retrieve the stored patterns. To do this, the network is cued with a partial, noisy or otherwise distorted version of one of the patterns; $x^0 = \tilde{\xi}^\mu$. This means that the state vector of the network neurons $[x_1 \dots x_N]$ is initialized to the cue. The current state is then propagated through the synaptic matrix. Either one neural state is updated at a time, as was the case in the original Hopfield model, or the whole network is updated in parallell (Hopfield, 1982). In the latter case, the the network dynamical equations become:

$$x_i^{t+1} = \Theta\left(\sum_{j \neq i}^{N} J_{ij} x_j^t - \theta\right).$$

Here $\Theta(\cdot)$ is the Heaviside step function and θ is an activity threshold. If the cue x^0 is sufficiently close to one of the patterns ξ^μ, and not too many patterns have been stored, the new state x^1 will be closer still to ξ^μ and after a few iterations the full pattern will be retrieved. Because the retrieved pattern is the same as the one that generated the cue, this Hopfield type network is referred to as autoassociative; it associates a pattern with itself.

3.2.2 Learning paradigm

In the above presentation of the Hopfield network, we saw a learning paradigm that is in fact the most common way to use an attractor memory system. The *pattern completion* paradigm consists of three phases:

1. Store

   A number of patterns are presented to the learning system.

2. Cue

   A fragment of one pattern is input to the system. The cue may be distorted in two ways; parts of the pattern may be missing and parts of the information may be incorrect.

3. Recall

   The autoassociative memory fills in and corrects the cue, converging on the full and undistorted cued pattern.

The usefulness of this paradigm is that patterns may be retrieved using content addressing; there is no need to specify an index number or any other special key to find a particular pattern, we only need to present a part of the pattern itself – whatever is known about the information that one wants to retrieve. When it comes to testing the performance of an attractor memory, the pattern completion paradigm is also suitable, because there are straightforward performance measures for it.

3.2.3 Performance measures

When a partial or noisy pattern is processed by an autoassociative network, the desired effect is that some or all of the missing information is filled in. The network operation can be regarded as a mapping from the cue to an output pattern; a natural performance measure is then how much closer the latter is to the original pattern. One measure of how close a distorted pattern is to the original is the amount of information needed to correct the errors. Depending on context, there may be arbitrarily clever ways of encoding that information, but we will here settle for a simple code: First the indexes of the units that should be switched on are listed, then those that should be turned off (Schwenker et al., 1996).[2] Let as before N be the pattern length and let a^μ be the number of active units in pattern ξ^μ. In the distorted pattern $\tilde{\xi}^\mu$ there are $e^\mu_- = \sum_{j=1}^N \Theta(\xi^\mu - \tilde{\xi}^\mu)$ inactive units that should be active and $e^\mu_+ = \sum_{j=1}^N \Theta(\tilde{\xi}^\mu - \xi^\mu)$ active units that should in fact be inactive. (This implies that the activity level of the distorted pattern ξ^μ is $\hat{a}^\mu = a^\mu + e^\mu_+ - e^\mu_-$.) The information needed to correct the distorted pattern becomes:

$$r(\tilde{\xi}^\mu, \xi^\mu) = \sum_{j=0}^{e^\mu_- - 1} \log(N - \tilde{\xi}^\mu - j) + \sum_{j=0}^{e^\mu_+ - 1} \log(\tilde{a}^\mu - j).$$

For small distortions, this becomes approximately $r(\tilde{\xi}^\mu, \xi^\mu) \approx e^\mu_- \log(N - a^\mu) + e^\mu_+ \log(a^\mu)$. For sparse activity, the measure may be further approximated as $r(\tilde{\xi}^\mu, \xi^\mu) \approx e^\mu_- \log(N) + e^\mu_+ \log(a^\mu)$.

[2] If we are dealing with sparse patterns, the former corrections will be more expensive, because there are more zeros than ones to choose from.

We may now describe the information gain from operating on a noisy patterns by the associative network as the decrease in missing information. If the network transforms the pattern according to $\widetilde{\xi^\mu} \to \hat{\xi^\mu}$, the information gain is $r(\tilde{\xi}^\mu, \xi^\mu) - r(\hat{\xi}^\mu, \xi^\mu)$. We can take the total decrease in missing information (over all the stored patterns) as a measure of the useful information content in the associative memory:

$$I_c = \sum_{\mu=1}^{p} \left(r(\tilde{\xi}^\mu, \xi^\mu) - r(\hat{\xi}^\mu, \xi^\mu) \right) .$$

The information content divided by the number of synapses is a measure of the efficacy of the network, as it relates the useful information to the total information embedded in the system. Naturally, this efficacy measure is conditional on the distortion applied to the patterns; the mapping $\xi^\mu \to \hat{\xi}^\mu$. If no distortion at all is applied, that is $\hat{\xi}^\mu = \xi^\mu$, then of course $r(\hat{\xi}^\mu, \xi^\mu)$ will already be zero and the best we could hope for would be zero information gain, meaning that the stored patterns are stable under the dynamics of the associative memory. Conversely, if too severe distortions are applied, making the patterns indistinguishable, any associative memory will fail to produce a noticeable information gain. The type of distortion to apply should therefore be reasonable in severity and its type chosen according to the intended application of the associative memory.

Chapter 4

Input processing

Attractor memories deal with storing and retrieving cortical representations for different kinds of behaviourally relevant entities; sensory impressions, words and concepts as well as motor programs. But the external world; its shapes, colors and sounds, does not present itself in a form suitable for processing by an attractor memory. Sensory impressions must therefore be pre-processed into a form digestable by such a system. For this representation to be useful, two conflicting requirements should additionally be satisfied: The code should preserve similarity of inputs, in order that the cortex may generalize between similar situations, but it should otherwise make efficient use of the available code space, in order for storage capacity to be maximized.

Efficient coding has previously been proposed as a computational objective for sensory regions (Barlow, 1983), but in this chapter the objective will be more specific; the representation should be efficient in the sense that it may be *efficiently processed by an attractor memory*. The critical resource for associative memory is the capacity of plastic synapses. Making efficient use of Hebbian synaptic memory has been shown to require a set of active units that is a small fraction of the whole network (Willshaw et al., 1969; Palm, 1980; Palm and Sommer, 1995). As recently emphasised, another critical resource is metabolic energy consumption limiting the number of permissible spikes (Laughlin and Sejnowski, 2003). This limit has been estimated to be a few action potentials per neuron per second (Lennie, 2003). Thus, the efficient sensory code is sparse and binary, a conclusion also argued for previously (Földiák, 1990).

Generative models are a traditional approach to create coding maps in sensory areas. A generative model optimizes the ability to reconstruct the input data from the code vectors. Setting this objective gives a local performance measure, allowing for unsupervised learning. A very successfull approach to understanding cortical sensory coding is based on linear generative models (Olshausen and Field, 1996; Bell and Sejnowski, 1997). Using InfoMax in a linear coding model, which is also known as independent component analysis (ICA), the receptive fields become localized, as seen experimentally in primary visual areas; though the actual receptive field shapes generated are unlike those seen in V1 simple cells. In this chapter, I introduce a generative model that yields sparse population patterns, or *small neural assemblies*, as sensory representations. As it turns out, the model itself may be represented as an attractor memory structure.

4.1 The gated-linear generative model

The focus of this chapter will be a generative model where a piece of data $x \in R^n$, e.g. an image patch, is reconstructed in terms of a nonorthogonal and overcomplete set of basis vectors $\{\Psi_{ij} : i = 1, ..., m; j = 1, ..., n; m > n\}$ as

$$\hat{x}_i = \sum_{l=1}^{m} a_l y_l \Psi_{li} \in R^n. \tag{4.1}$$

In this *gated-linear generative model* one factor in the bilinear expansion is a binary gating variable y_l, the other a real-valued coefficient a_l. I will refer to $y \in \{0, 1\}^m$ as the *gating vector,* since it determines what set of basis vectors is used. The vector $a \in R^m$ is referred to as the *coefficient vector*. What the model does, in other words, is to encode an input image in the form of one analogue and one digital vector, in such a way that the original image can be reconstructed from this representation. The gain from this is that the new representation, in particular the digital vector, may be better suited for e.g. storing in associative memory. The model is further described by an objective function to be minimized:

$$E(x, y, a) = E_{rec}(x, y, a) + E_{sp}(y), \tag{4.2}$$

where E_{rec} is a reconstruction term describing the deviation of the reconstruction $\hat{x}$ from the data x. The binary sparseness term E_{sp} penalizes one-entries in the gating vector as

$$E_{sp}(y) = \theta \sum_i y_i, \tag{4.3}$$

where θ is a threshold parameter controlling the reconstruction benefit required for inclusion of a nonzero y_i in the gating vector. The gating vector can also be expressed as a projection matrix in coefficient space

$$P^y := \delta_{lm} y_l \in R^m \times R^m. \tag{4.4}$$

Thus, we can rewrite (4.1) in matrix notation, $\hat{x} = (P^y \Psi)^T a$, and write the objective function as

$$E(x, y, a) = \frac{1}{2} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 = \frac{1}{2} \langle a, P^y C P^y a \rangle - \langle a, P^y c \rangle + \frac{\langle x, x \rangle}{2} + \theta \langle y, y \rangle . \tag{4.5}$$

where $\langle x, y \rangle := \sum_i x_i y_i$ is the inner product between the vectors x and y. We may now move from the image coordinate space to the inner product space, writing the overlaps between basis vectors $C := \Psi \Psi^T \in R^m \times R^m$ and the filter outputs $c := \Psi x \in R^m$. The term $\langle x, x \rangle$ is constant for a given x and will further on be omitted. Note that for fixed y the model becomes an ordinary linear generative model. However, we will here be interested in the dynamics of the gating vector, which will ultimately be the representation of input seen by the associative memory – and each change in y involves an optimization of the coefficients a.

4.1.1 Optimizing the coefficients

For every fixed y the optimal coefficient vector a^* has to minimize the energy (4.5), so

$$a^* = \text{argmin}_a \left[E(y, c) \right] = \text{argmin}_a \left[\frac{1}{2} \langle a, P^y C P^y a \rangle - \langle a, P^y c \rangle \right] . \tag{4.6}$$

If the operator $C^P := P^y C P^y$ were invertible, we could solve for a^* by taking the derivative and inverting that matrix. As this is generally not the case, we must settle for the pseudoinverse solution:

$$a^* = [C^P]^+ P^y c \in R^m. \tag{4.7}$$

This leaves the $\{a_i^* : y_i = 0\}$ undetermined; we may arbitrarily set them to zero. To optimize the gating vector, a^* from (4.7) is inserted into the applicable part of (4.5);

$$E_{rec}(y, c) = -\frac{1}{2} \langle P^y c, [C^P]^+ c \rangle . \tag{4.8}$$

The above equation allows for direct optimization of the y variables. A local search method to optimize y is to use (4.8) in a sequential update scheme; given a y vector the single bit flip is chosen that yields the maximum energy decrease,

$$\Delta E(y \to \bar{y}) = \frac{1}{2}\left\langle P^y c, ([C^P]^+ - [C^y]^+)c \right\rangle + \theta\left(\langle \bar{y}, \bar{y}\rangle - \langle y, y\rangle\right), \tag{4.9}$$

where $\bar{y}$ is one of the m vectors that can be obtained from y by a single bit flip. However, in general the optimization is computationally expensive because of the calculation of a matrix pseudoinverse required for every change in the y-vector. One way to greatly speed up this computation by an approximative method follows (section 4.2).

4.1.2 Learning the basis vectors

To learn the set of basis vectors, the optimal input reconstruction using the current set of basis vectors is first determined, yielding a and y. Then we follow the gradient of the energy function (4.2) with respect to the basis vector components, which yields a local "delta" learning rule:

$$\frac{\partial E}{\partial \Psi_{ij}} = (\hat{x}_i - x_i)\, a_j y_j. \tag{4.10}$$

At every update step the normalization of the basis vectors is maintained by re-normalizing.

4.2 The small assembly coding network

The matrix inversion in the objective function (4.8) prevents direct network implementation of the optimization of the gating vector y. To derive an approximative solutions to this optimization problem, the operator C^P is written as a product of an operator that is, in most cases, full rank and a projection operator:

$$C^P = P^y C P^y = [P^y(C-1)P^y + 1]P^y =: C^y P^y$$

Substituting the newly introduced C^y into the energy equation yields:

$$E(y) = -\frac{1}{2}\langle c, [C^y]^+ P^y c\rangle$$

The operator C^y is full rank and can thus be inverted using the ordinary inverse, rather than the pseudoinverse, if the set of basis vectors selected by the nonzero y_i are linearly independent. We may then use the power series expansion

$$[C^y]^{-1} = 1 - P^y(C-1)P^y + [P^y(C-1)P^y]^2 - ... \tag{4.11}$$

which converges as long as $||P^y(C-1)P^y|| << 1$. This condition can be ensured by a) making the y sparse through adding an appropriate sparseness constraint in the energy function and b) by ensuring that y does not contain basis vectors with any mutual C_{ij} close to one.

Using the expansion (4.11) up to the first order yields the first-order approximation to equation (4.2)

$$\begin{aligned} E^{FO}(c,y) &= -\frac{1}{2}\langle c, P^y c\rangle + \frac{1}{2}\langle c, P^y(C-1)P^y c\rangle + E_{sp}(y) && (4.12)\\ &= \frac{1}{2}\langle c, P^y(C-2)P^y c\rangle + E_{sp}(y). && (4.13)\end{aligned}$$

The interpretation of E^{FO} is most obvious in the form (4.12), where the first term describes the support of basis vectors by filter inputs c and the second term accounts for "explaining away", competition between basis vectors with high overlap. At the same time the second term directs the choice of y towards regions where the first-order approximation is valid. Equation (4.13) is the energy function of a recurrent neural network;

$$E^{FO}(y) = \frac{1}{2}\sum_{l,m} T_{lm} y_l y_m + \theta \sum_l y_l. \tag{4.14}$$

This is formally the energy function of a Hopfield associative memory network, only the connection matrix depends on the stimulus;

$$T_{ij} := c_i C_{ij} c_j - 2\delta_{ij} c_i^2. \tag{4.15}$$

In the first order approximation, the continuous variables, $a^{FO} = [1 - P^y(C-1)P^y]c$, are then computed as

$$a_i = c_i - y_i \sum_{j \neq i} C_{ij} c_j y_j, \tag{4.16}$$

which completes the description of the *small assembly coding network* (SACN).

4.3 Results from simulation experiments

The SACN model was first tested on patches of natural scene images. The images have undergone a ”whitening” preprocessing to equalize the spatial frequency distribution (Olshausen and Field, 1996). The number of basis vectors was three times overcomplete, with $m = 192$. Unless stated otherwise, the patch size used was $8 \times 8 = 64$ pixels.

The SACN model at a given sparseness level – as adjusted by choosing θ in equation (4.3) – is compared with a *graded sparse generative model*. In that model, the reconstruction map is of the same form as in the SACN; $\hat{x}_i = \sum_{l=1}^{m} a_l \Psi_{li}$ (Olshausen and Field, 1997). However, in that model a different energy function is used, aiming for a graded form of sparseness. Using one such a sparseness function, we have:

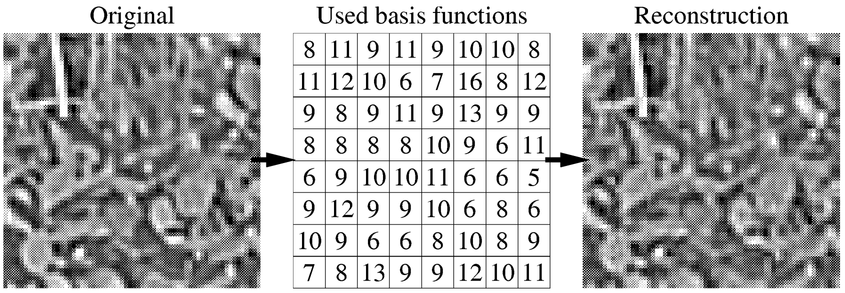
$$E(x, a) = E_{rec}(x, a) + \beta \sum_i |a_i|. \tag{4.17}$$

Sparseness is varied in the graded model by changing the parameter β. To make the models comparable, the output of the graded model is optimally *pruned*, keeping just the largest of the coefficients, such that the energy function 4.2 is optimized.

4.3.1 Reconstruction quality

An example of an original image and its reconstruction is shown in Fig. 4.1. The choice $\theta = 5.6 \cdot 10^{-2}$ in equation (4.14) sets the sparseness of the mean usage to $\langle y \rangle \approx 4.8$. The number of basis vectors used to reconstruct each of the 8×8 patches is displayed in the grid on the right hand panel.

By systematically varying θ in equation (4.14) we may investigate how constraining binary sparseness effects reconstruction. Fig. 4.2 (left) shows that with growing sparseness (decreasing mean usage) the reconstruction error increases. Approaching a mean usage $\langle y \rangle = 1$ means to enter the regime of vector quantization, where the reconstruction quality is limited by our fixed number of basis vectors (corresponding to a small dictionary size). Comparing reconstruction achieved with the SACN model and with the pruned linear model, the residual errors of the former are significantly lower.



8	11	9	11	9	10	10	8
11	12	10	6	7	16	8	12
9	8	9	11	9	13	9	9
8	8	8	8	10	9	6	11
6	9	10	10	11	6	6	5
9	12	9	9	10	6	8	6
10	9	6	6	8	10	8	9
7	8	13	9	9	12	10	11

Figure 4.1: Image, reconstruction and map indicating the usage in individual patches. This particular image part was chosen because there is some interesting structure in it. As a consequence, the SACN has assigned a higher average usage number than the overall $\langle y \rangle \approx 4.8$.

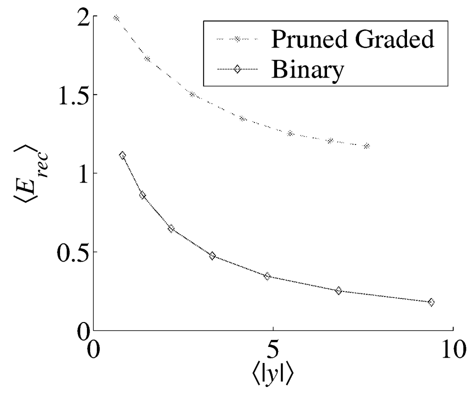


Figure 4.2: Reconstruction error as a function of mean usage. The curve of the pruned linear model ends at about $\langle y \rangle = 10$ because even with setting $\beta = 0$ in the linear coding model, this is the maximum usage that survives the the pruning process described above.

4.3.2 Model basis vectors vs. biological receptive fields

To explore the shapes of the basis vectors resulting from the learning with natural scenes more quantitatively, they are fitted with model functions. In the sparse regime, each basis vector can be well fitted with two-dimensional Gabor function in the image coordinates u, v;

$$h(u', v') = A \ \exp\left[-\left(\frac{u'}{\sqrt{2}\sigma_{u'}}\right)^2 - \left(\frac{v'}{\sqrt{2}\sigma_{v'}}\right)^2\right] \cos\left(2\pi f u' + \Phi\right), \quad (4.18)$$

where u' and v' are translated and rotated image coordinates, $\sigma_{u'}$ and $\sigma_{v'}$ represent the widths of the Gaussian envelope, f and Φ are the spatial frequency and phase of the sinusoidal grating.

The parameters of the Gabor fits allow a much more compact description of the shapes of basis vectors. In figure 4.3 (left), a two dimensional display is used, that has previously been employed for experimentally determined receptive fields (Ringach, 2002). On the horizontal and vertical axes, are mapped respectively $n_u = \sigma_u f$ and $n_v = \sigma_v f$; the size of the Gaussian envelope measured in units of the period of the sinusoidal grating. Center surround geometries are located near the origin, slim edge-detector type geometries are at large n_u and small n_v values, geometries with multiple subfields are at large n_u and n_v values. The right diagram is a histogram of the carrier phase Φ with respect to the center of the envelope. The properties of the basis vectors of the generative models are compared to receptive fields recorded in the primary visual cortex of monkeys (Ringach, 2002).

As can be assessed in figure 4.3, the basis vectors of the SACN model resemble the spatial structure of biological receptive fields much more closely than those of the linear coding model.

4.4 Conclusions

The SACN model works by ”explaining away” parts of the input using a set of basis vectors, paying a penalty for each one used. The interaction between basis vectors is mediated by the lateral weights C, similar as in linear generative models (Olshausen and Field, 1996; Lee and Seung, 1997) and in a heuristically derived binary sparse coding network (Földiák, 1990). The network computation in the SACN differs from these models; the SACN units receive a product of the bottom-up input times the difference between bottom-up input and the input over the lateral connections, in a Hopfield type model. The representations

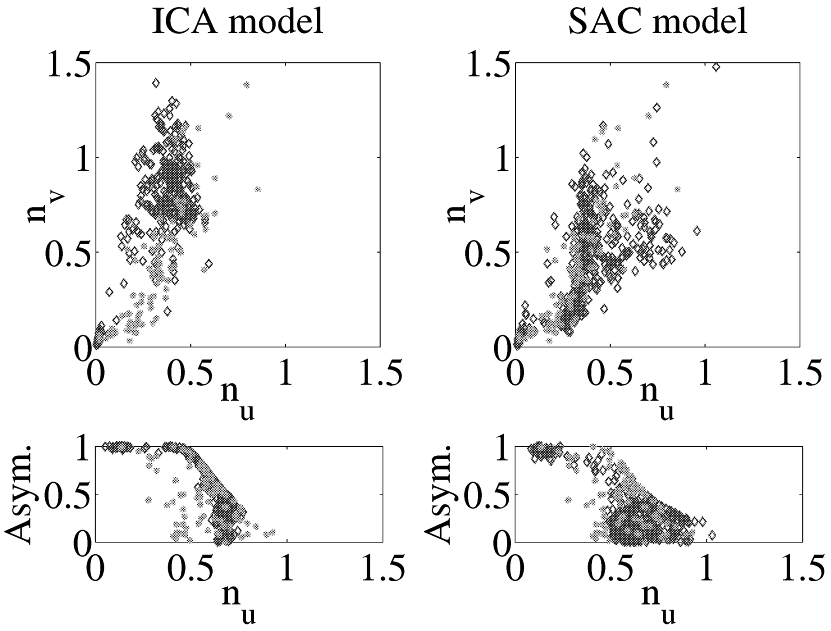


Figure 4.3: This figure displays shape properties of the basis vectors and compares them to receptive fields in V1 of monkey (Ringach, 2002). In the top diagram the n_u and n_v on the axes are the relation between carrier width and oscillatory period. The bottom diagram relates the asymmetry of basis functions/receptive fields.

produced are sparse and binary, suitable for storage in an attractor memory.

The simulation experiments presented reveal an interesting difference, introduced by the gating nonlinearity of the overcomplete SACN: The basis vectors develop more diverse profiles, there are not only oriented shapes as found in the linear model, but also center surround shapes (Olshausen and Field, 1996). When compared to the pruned linear model the SACN model finds more parsimonious representations; better reconstruction by same number of basis vectors.

Chapter 5

Sequence memory

Abstract models of attractor memory, such as the Hopfield network, were originally though of as static systems. In terms borrowed from psychology, no concepts of timing or serial order are connected to them. This means, among other things, that it does not matter to the memory system if a millisecond or century passes between two operations. Also, what is stored in the memory is nothing but a set of memory snapshots, with no internal ordering.

5.1 Timing and serial order

The brain of any living organism, or for that matter a computer that interacts with its environment, can in the most general case be thought of as performing a mapping from sensory perceptions and time, to motor commands; $(x, t) \to y$. I will here initially be concerned with learning and reproduction of stereotypic spatiotemporal patterns, in which case we may drop the input part and be left with a mapping $t \to y$. Furthermore, my primary interest will be in spatiotemporal patterns that change only at a finite number of points in time. That is, there are no gradual changes, only discrete transitions between spatial patterns. I will refer to this class of spatiotemporal patterns as *sequences*.

A sequence can be decomposed into two parts; the points in time when transitions take place and the ordered list of patterns exhibited. The two parts are known in the psychological literature respectively as *timing* and *serial order*. It has been demonstrated that the two concepts are learned and processed independently. If a subject has first learnt either component (timing or serial order) of a sequence, learning the full sequence then proceeds faster. Furthermore, separate brain areas are involved in processing timing and serial order (Ullén et al., 2003). This separation hints that if we would like to faithfully model the brain, the two should be kept separate. I will here deal with serial order, mentioning timing only in passing.

5.2 Sequences and the nervous system

A single cell organism that swims in the direction of a nutrient, but away from harmful substances, exhibits a behaviour that can be considered a static mapping from input to output. No history of past experience, and no learning, is required to produce the behaviour. More complex behaviours, on the other hand, are based not just on the sensory input, but also on the internal state of the organism. This is the case already for only slightly more complex behaviours. The motor commands for walking, generated in the spinal cord, involve a stateful system, incorporating (among other components) internal oscillators, pacing the movements. Of course, sensory input also plays a role, but without internal state the system will know neither when to start swinging a leg forwards, nor that a movement of the left leg should be followed by one of the right (Wadden and Ekeberg, 1998).

Associative memories certainly have internal state. When cued with a ”stimulus”, their response – the completed pattern – will depend on which patterns have been previously stored in the memory. In their basic form however, we may conclude that they don’t have the capability of performing even a task like walking. Two things are again missing: timing and serial order.

Moving straight from the ancient circuitry in the spinal cord, to the highest level of neocortex, let me next define *cognitive sequences*. A cognitive sequence is a series of events taking place on the highest level in the information processing hierarchy of the nervous system. This is where processing of abstract concepts; such as words, numbers or other symbolic constructs takes place, sometimes under direct conscious control. Simple examples of cognitive sequences are the acts of counting or singing. The act of thinking itself may be speculated to be a more sophisticated form; more on that in the conclusion of this chapter (section 5.4). I have described motor acts and cognitive sequences, although seemingly vastly different, in the same context, for a reason. Experiments have demonstrated that motor acts are truly sequences, in the sense defined above. That is, they are based on a high level plan that is discrete in nature, consisting of a number of well-defined intermediate steps. Actual movement seems to be a form of interpolation based on the discrete plan (Johansson et al., 2001).

The cerebral cortex is not alone in producing sequence output. Some reports indicate that the basal ganglia may be even more important for serial order production (Cromwell and Berridge, 1996). The other component, timing, seems to be dependent on the cerebellum; a structure which is much smaller in size than the cerebral cortex, but which contains more neurons than all the other parts of the brain. Motor tasks can be performed without a cerebellum, but the fine structure and timing is lost (Sakai et al., 2002; Diener et al., 1993). This anatomical separation of timing from serial order again motivates treating the two aspects separately.

5.3 Sequence processors

Let us now return to the simplest case of sequence processing, that of reproducing a learned sequence. While the ability to sing a song or to count may not be as impressive as the skills of an advanced chess player, I will argue that more advanced skills emerge from a combination of a basic sequence machinery, combined with associative memory. The requirement that an associative memory structure be maintained will then be an important constraint. Already Hopfield, a pioneer of associative memory, discovered that this is a nontrivial requirement. He experimented with adding a sequence producing term, $J_{ij} = \sum_{\mu=1}^{P} \xi_i^{\mu} \xi_j^{\mu+1}$ to his autoassociative memory. This synaptic matrix term stores a series of patterns by chaining them together; each pattern in the sequence being linked to its successor by Hebbian association.[1] This does not work well in the asynchronously updated Hopfield network, but it does work in a synchronously updated version, provided the autoassociative part of the memory matrix is dropped (Hopfield, 1982; Düring et al., 1998). The resulting heteroassociative network has but one recall attractor state[2]; an endless reproduction of the learned sequence; this model will be incapable of dwelling in any of the pattern states, but will rather always march steadily on.

[1] Here we may make the sequence of patterns cyclic by defining $\xi_i^{P+1} := \xi_i^1$.

[2] In addition, there may be spurious attractors, but those are in general not viewed as contributing to the computational power of the model (Amit, 1989).

5.3.1 Learning paradigm

For evaluating sequence learning networks, we slightly modify the pattern completion paradigm, to read as follows:

1. Store

   A sequence of patterns are presented to the learning system.

2. Cue

   A fragment of one sequence is input to the system.

3. Recall

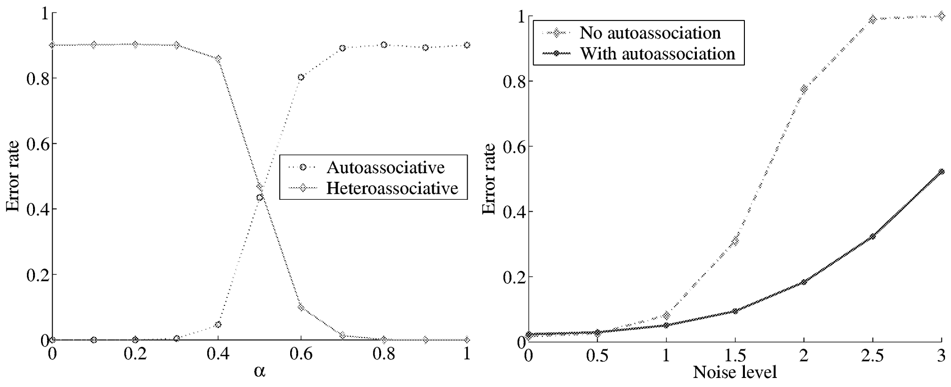
   The system continues sequence recall from the cue onwards.

The fragment presented in the cue phase may be a subsequence of the original sequence; consisting of a few patterns. The distortion of the cue may take additional forms, compared to the static case: apart from distortion of individual patterns, there may be distortions with respect to order (omitting, inserting or exchanging patterns) and with respect to timing.

5.3.2 Naive mixing

Can the auto- and heteroassociative Hopfield networks be combined, adding sequence recall to the autoassociative memory function? One way to do this would be to follow Hopfield's lead and simply add the two learning equations together; $J_{ij} = \sum_{\mu=1}^{N} \xi_i^\mu((1-\alpha)\xi_j^\mu + \alpha\xi_j^{\mu+1})$. Here α is a parameter regulating the amount of heteroassociation (Hertz et al., 1991). We may understand the poor performance observed for the asynchronously updated version of this network from considering the performance of a synchronous network trained in this way. The heteroassociative part of the weight matrix then acts as as a noise term for the autoassociative recall and vice versa, as seen in figure 5.1(a).

Another way of combining auto- and heteroassociation in a synchronously updated network is to interleave them, using an autoassociative step to "clean up" each new pattern before applying the heteroassociative step. In the simple case of noise free sequences, we should however not expect better performance from such a network than from a purely heteroassociative network. The reason for this is that the heteroassociative Hopfield network has a larger memory capacity than the autoassociative version (Düring et al., 1998). Looking



(a) Naive mixing of auto- and heteroassociation. Shown is unit-wise error rate for one-step retrieval.

(b) Autoassociation in sequence retrieval. Noise was added to each retrieval step. Shown is the error rate for retrieval of a 10 item sequence.

Figure 5.1: Mixing of auto- and heteroassociation. This has benefits only under certain circumstances. Both plots refer to a k-winner-take-all Hopfield network with 100 neurons and 10% activity level.

at the case for zero noise in figure 5.1(b) we see that performance is in fact slightly worse for a network where we include an autoassociative step than for a purely heteroassociative one. We also see that the interleaved model will pay off at higher noise levels. More importantly, processing in the brain, mediated by spikes and with no central pacing, is very unlikely to be instantaneous and perfectly synchronized, as required by the purely heteroassociative model.

5.3.3 A sequence learning model

I have argued that the ability to perform temporal tasks should be built on top of an autoassociative memory, so that the system retains the ability to represent each memory state individually. But as was also explained, naively adding additional mechanisms to an autoassociative memory may break it. I will next outline a clean and simple model where sequence learning has been added to an associative memory, while maintaining a high memory capacity. The model is a fully connected, single layer, k-winner-take-all network with synchronous updating. The k-winner-take-all rule means that at each moment in time, exactly k units are active; the units with the strongest support (Kown and Zervakis, 1995). Thus we need not concern ourselves with activity control, the problem of keeping the number of active units at an appropriate level (O'Reilly, 1998). The neural units themselves are leaky integrators; they have a short-lasting memory of their input history.

The key feature of the model is that the synapses are dynamical; their efficacies vary depending on past use. Each synapse possesses a finite pool of "resources" used for transmitting information. This is in analogy with real synapses, where the presynaptic neuron maintains a number of vesicles with transmittor substance, ready for release. A fraction of the available resources are used each time the synapse is activated, which happens each time the presynaptic cell fires (Tsodyks and Markram, 1997). Thus, a synapse that has recently been activated a few times will be less influential than one that has not been activate for a while. This turns out to be highly advantageous for sequence processing in that it helps separating autoassociation from forward association. The two types of signal are separated in time, such that autoassociation dominates when a pattern is newly activated but gives way to heteroassociation after some time; determined by how fast resources are expended. The network equations are as follows:

$$\begin{aligned}
h_i(t+1) &= (1-\mu_{mem})h_i(t) + \sum_{j=1}^{N} u_{ij} r_{ij}(t) s_j(t) \\
s_i(t) &= \begin{cases} 1 & \text{if } i \in \text{n-argmax}_j \left(h_j(t) + n_j(t)\right) \\ 0 & \text{otherwise} \end{cases} \\
n_i(t) &\in N(0,\sigma) \\
r_{ij}(t+1) &= \left(1 - u_{ij} s_j(t)\right) r_{ij}(t) + \mu_{rec}\left(1 - r_{ij}(t)\right)
\end{aligned}$$

Here N is the number of units in the network, n is the number of active units at any one time. The model uses synchronous time updating, which is equivalent to assuming that all spiking activity is synchronized to a 50 Hz gamma rhythm. Thus other time constants in the model are related to a 20 ms gamma time interval. The key variables of the model are the integrated support vector h, the spike vector s and the synaptic resource pool matrix r. The latter represents the synaptic resources that are temporarily expended when used; they recover with a time constant of 800 ms. (Tsodyks and Markram, 1997).

The synapses in the model implement a simple Hebbian learning rule. The pre- and postsynaptic sides of a synapse each maintains a short-term memory of past activity, the combination of which is used to update the synaptic release parameters u_{ij}. A synapse thus potentiated will release more ”vesicles” per transmission event, expending its resources faster. While there is no explicit synaptic ”strength”, or conductance parameter, a potentiated synapse will still have an increased overall efficacy, since resources recover at a speed proportional to the amount that has been spent.

Model properties

The sequence recall behaviour of the model is illustrated in figure 5.2, where support levels of four stored patterns are shown. The four patterns take turn in being active. Importantly, the model is empirically found to be efficient, meaning that its storage capacity is just a constant factor below the theoretical limit, determined by the information content in the synapses (see figure 5.3). An interesting aspect of the model is that the sequence recall behaviour is emergent from the combination of a non-pointwise Hebbian learning rule and dynamical synapses. This means that recall goes in the forward direction when the presynaptic time constant is larger then the postsynaptic one,

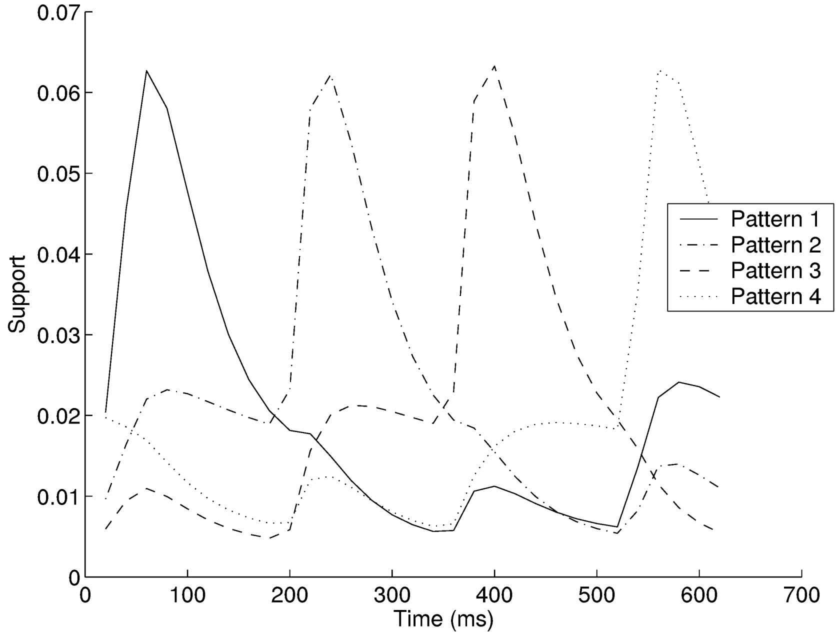


Figure 5.2: Sequence recall in the model network. Each curve shows the mean support level $\langle h_j \rangle$ for the units belonging to four stored patterns. Synaptic depression weakens the active pattern, eventually allowing the next pattern in the sequence to take over.

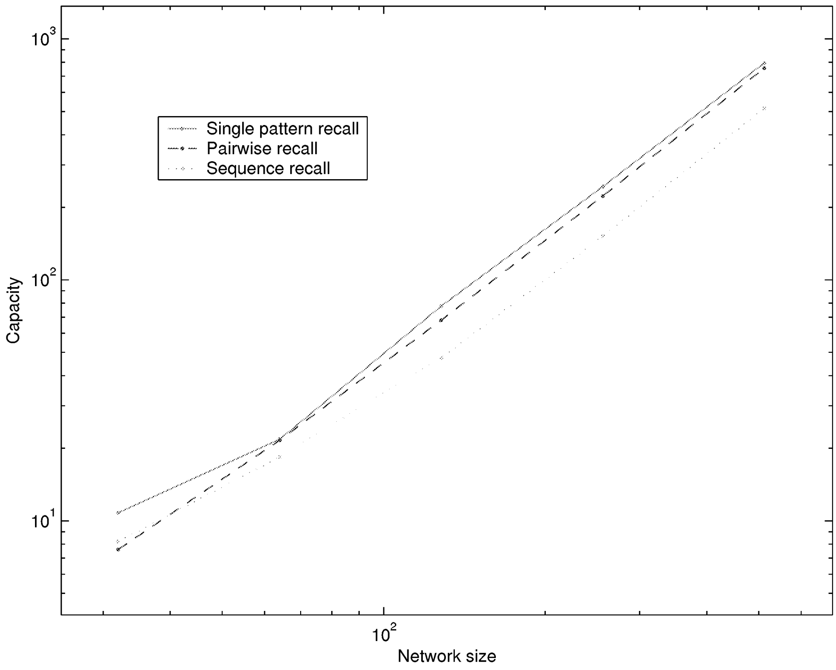


Figure 5.3: Scaling of memory capacity in the sequence learning network. Shown is storage capacity in patterns, based on performace criteria for autoassociation (single pattern), one step heteroassciation (pairwise) and full requence recall.

which is consistent with what has been observed in biological synapses. In terms of sequence learning methods, as related in the psychological literature, this model falls into the category of associative chaining, meaning that each pattern is associated by the next one in the sequence. This association is provided by the memory traces built into the synapses.

In the model as described so far, timing is hard wired. More precisely, how long a pattern can remain active depends on the rate at which an autoassociatively potentiated synapse expends resources. This determines the number of spikes that can be fired before these synapses are too exhausted to support the pattern, something that will always happen eventually, since only the release rate is modified by the learning. However, by slightly modifying the model, such that in addition to release probability also synaptic efficacy is modified in the learning phase, pattern recall will be stable. An active pattern will still be weakened after a short while, as resources are expended, but lateral inhibition (as represented by the k-winner-take-all rule in the model) will prevent another pattern from becoming active. If the lateral inhibition is now relaxed for a moment (the number k temporarily increased), units from other patterns are also activated in parallell with the previous pattern. Once inhibition is restored, the previously active, weakened, pattern will lose in the competition with one of the newly activated ones. The end result is that momentarily reduced inhibition triggers a change of active pattern. If the network is trained with a sequence, it will switch to the next pattern in line. In figure 5.4 such a behaviour is shown. Which pattern is next activated need not be specified by the intrinsic associative chaining property, but could be specified by any external system, separating timing and serial order. The important thing to note is that the system retains its autoassociative properties, performing pattern completion when presented with a distorted version of one of the stored patterns. Meanwhile, autoassociation gets out of the way once its work is done, not interfering with pattern transitions.

5.4 Sequences and cognition

From computer science we know that a very simple sequence processing capability, that of the Turing machine, is all that is required to perform a vast class of computation (Turing, 1936). The idea that the brain is just a serial computer, implemented in neurons and synapses, was

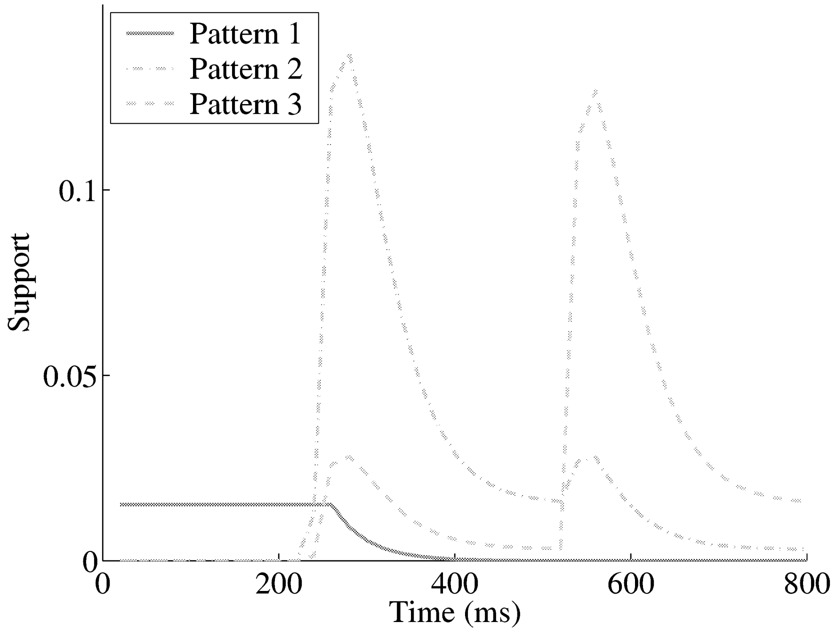


Figure 5.4: Disinhibition triggers pattern transitions. In this version of the sequence learning network, patterns are stable , but destabilize as activity control is relaxed. At t=240 and t=540 ms the k-value is momentarily doubled, leading to transitions first from pattern 1 → 2, then from pattern 2 → 3. Note that the steady state support value from weakened synapses is much smaller than the transient signals that appear at pattern transitions.

the tenet of the most radical proponents of traditional artificial intelligence. Traditional AI has not failed; it has spawned many techniques that are fundamental to computing today, such as high level programming languages and relational databases. On the other hand, it has not been successful in mimicking human intelligence. This is illustrated by the fact that some tasks that are easy for symbolic AI systems, such as tasks requiring deep recursion, are difficult for humans, whereas other tasks, such as image recognition, are easy for us but have not been successfully approached using symbolic methods. In this respect, the approaches of connectionism, neural networks and parallel distributed processing yield results more similar to biological intelligence. The associative memory systems that are the focal point of this thesis belong to this family of models. Still, it is apparent that we do use a serial approach to some mental tasks, for example when planning a route, playing chess or performing arithmetic. Doing these things our working memory passes through a series of intermediate states, which we may identify with places along the route, positions on the chess board or partial results of a calculation. Unlike the Turing machine, the mental path followed in such tasks may require something more complex than following a simple list of rules. Presumably, our ability for pattern recognition and other forms of parallell processing comes heavily into play. An understanding of higher cognitive function therefore seems to require a combination of parallell processing – associative memory – and sequence processing.

Chapter 6

A biophysically detailed attractor memory model

I have so far in this thesis analysed abstract models of attractor memory. The network units have been assumed to operate on real valued inputs and to produce binary outputs. In contrast, the basic unit of the brain, the neuron, communicates through spike trains; series of all-or-nothing impulses occurring asynchronously. Real neurons overall have a much richer set of behaviours than the artificial neurons so far employed. On the other hand, some features that seem natural and are easily implemented in an abstract model may be non-trivial in a more realistic setting. In the previous, I have relied on two such features, so far without justification.

1. I have assumed that there is no limit on the number of cells to which a given neuron can send its output. In reality, a pyramidal cells project to a finite number of other neurons.

2. In the chapter on sequence memory, a k-winner-take-all rule was included in the network dynamics to maintain a tight control over the network activity level; to prevent runaway activity or termination of the network's activity. In fact, activity control is a hard problem in neural modelling.

In this chapter I will remedy these omissions, by introducing a much more detailed, biologically plausible network model. I will show how properties of the abstract model, including those mentioned above, emerge from the detailed model, validating their use in the abstract models.

6.1 The cortical network model

The model I present in this chapter is based on the minicolumnar hypothesis of cortical function. In this view, the minicolumns are taken as the computational units of the network. In the model, and presumably in the real cortex, excitatory coupling within a minicolumn means that it will be either active or inactive as a whole, the activities of the individual neurons need not be detailed. A minicolumn naturally has a richer repertoire than a single neuron; for instance it turns out that it can exert inhibitory as well as excitatory influence over another minicolumn far away, even though such non local interactions are always mediated by nominally excitatory pyramidal cells. In addition to the tightly coupled minicolumn, there is another level of structure in the model; a number of minicolumns are grouped together into a hypercolumn. While the minicolumn was internally dominated by excitation, inhibition is the dominant interaction in between them, rendering the hypercolumn a winner-take-all unit. This is, as we have already seen, a very stable and efficient way of controlling the activity level in an associative memory, but this time it will be emergent from underlying mechanisms.

The neurons in the simulated network are conductance-based models, meaning that they include a simplified model for the fundamental electrical properties of the neuron. This includes ion currents passing through channels in the cell membrane and electrical currents travelling between different parts of the neuron. The internal currents are modelled by compartmentalizing the neurons; their geometries are approximated by joining together a finite number of building blocks; spheres and cylinders. The interactions between neurons take place by simulating the opening and closing of ion channels at the appropriate compartment of the postsynaptic neuron. The synaptic dynamics, determining precisely how this takes place, may be arbitrarily complex both on the pre- and postsynaptic sides.

In the present model, there are three types of neurons. The excitatory pyramidal cells project both locally, within a minicolumn, and globally to other hypercolumns; the local connections serving to forge the minicolumnar unit, the long range connections corresponding to the weight matrix of the associative memory. Here the minicolumnar organization allows for negative weights in the matrix. This is realized by pyramidal cells projecting to an inhibitory ”regular spiking non-pyramidal” (RSNP) class of cells, including e.g. double bouquet cells. This changes the sign of the effective projection, while maintaining its specificity, as the projection of these cells in turn is very localized; projecting largely to a single minicolumn. The third cell type is the basket cell. Like the RSNP cells these are inhibitory, local interneurons, but their projections are spread out over the local hypercolumn. They serve to maintain activity control in the network. Their influence in the model has been tuned such that it is unlikely that two minicolumns in the same hypercolumn would be simultaneously active.

6.2 Network properties

I apply several methods to analyze the characteristics of the model network. Where possible, I have followed methodology from experimental neuroscience, in order to make results obtained from the model easily comparable to the experimental literature. The following properties of the network activity have been studied:

1. Pattern completion and rivalry
2. Up/down states
3. Artificial EEG
4. Unitary events
5. Temporal correlations.

6.2.1 Pattern completion and rivalry

Pattern completion is the basic element of associative memory. This is tested by stimulating just some of the minicolumns participating in a pattern, by simulated layer IV input. Equally important is pattern rivalry; competition between patterns. If two or more patterns are simultaneously stimulated, one would expect the pattern that receives the strongest stimulation to be activated, which was tested by injecting ambiguous layer IV input. It turns out that pattern completion is very fast; after just one burst from the stimulated minicolumns, the others are activated. Similarly, for the rivalry experiment the assembly with more stimulation is found to become active; the other activity is terminated. This is decided by the network in a matter of tens of milliseconds.

6.2.2 Up/down states

It has recently been found that pyramidal cells in the cortex have a discrete nature to their behaviour; they alternate between a ”down” state of low firing rate and an ”up” state where the firing rate is high. It is also seen that the soma potential is elevated when a neuron is in the up state (Cossart et al., 2003). Looking at the model network activity, we clearly see a similar phenomenon. The patterns take turns in being active, which for the individual neurons, that are coupled to many other neurons participating in the same patterns, means an input that varying accordingly. This manifests itself both as a raised membrane potential and as an accompanying variation in spike rate, as seen in figure 6.1(C-D). In a down state, the average soma potential is 65 mV and spike frequency is 0.2 Hz. In an up state they are 57 mV and 8 Hz, respectively. The up states last for some 700 ms, the rise of soma potential at their onset happening during about 50 ms. There is no exponential decay after termination of an up state (as has been reported experimentally); this is overshadowed by inputs from the next active pattern, but there is indeed more activity in the beginning of up states.

The attractor dynamics of the model is dependent on the long-range connections. When these were scaled down ten times, no attractor dynamics was observed. Again increased three times from this level, some pattern completion could be evoked, but only noisy after-activity and no spontaneous attractor dynamics. At four times increase there was after-activity and some noisy spontaneous attractor activity and with connection strength at half that used normally the dynamics was essentially the same as for the standard model.

6.2.3 Artificial EEG

I create an artificial EEG trace from the network simulation, intended to be similar to the actual EEG that would be recorded if the model were embedded in a human cortex. This allows for relating network activity to the wealth of empirical EEG data that has been recorded from human subjects under many different conditions. The main source of the EEG signal is believed to be currents in the apical dendrites of pyramidal cells, but the exact mechanisms are likely rather complicated (Alexander D. Protopapas and Bower, 1998). Rather than striving for complete realism in the model, I therefore settle for a simplified model, wherein the dendritic current in pyramidal cells is represented by the derivative of the soma potential. This signal, aggregated from

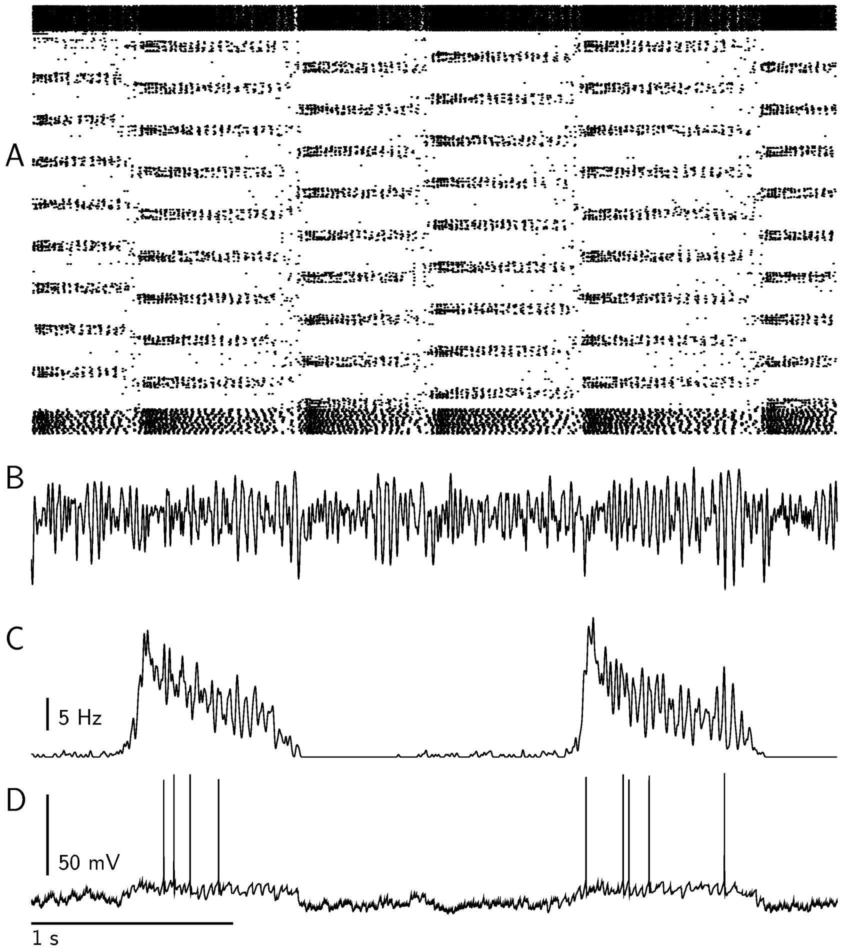


Figure 6.1: Up and down states in the biophysically detailed network model. *A:* Raster plot of spike activity in the entire network; the topmost, rapidly firing cells are the RSNP cells, then follow the pyramidal cells (sorted by hypercolumn and minicolumn) and finally the basket cells. Simulated time is 4 seconds and different patterns take turns being active, with short transitional periods in between. *B:* Local field potential; note asynchronous spindles at the up state onsets. *C:* Mean spike frequency of the pyramidal cells in one of the patterns, showing up and down states, the spike frequency being highest at the beginning of up states. *D:* Soma potential of one of the pyramidal cells in the same pattern. Membrane potential and spike rate are clearly elevated in the up state.

all pyramidal neurons in the model, is then low-pass filtered to generate the simulated signal, as it would be recorded by an EEG apparatus.

Following practice in the field, the EEG signal is divided into shorter segments, for each of which a cosine-shaped Hanning filter is applied to regularize the signal; smoothly bringing it to zero at the beginning and end of the segments (Dressler et al., 2004). For each subpart, the power spectrum is calculated, and these spectra are added together. A gamma-like oscillation with a frequency around 25–30 Hz is robustly identified, as is evident from figure 6.3(A). The signal is almost exclusively generated by the activity in the up states, consistent with experimental data, relating gamma patterns in human EEG to memory matches (Tallon-Baudry et al., 1998).

6.2.4 Unitary events

To analyse the fine structure of network activity, the *unitary event* technique was applied. The method starts by binning the set of neural spike trains into small time bins, a typical bin size being 5 ms. For each neuron, a bin is marked ”0” if the neuron emitted no spike inside of the time interval associated with the bin and ”1” otherwise. The method then proceeds to calculate the probability of obtaining the particular vectors of ”0”:s and ”1”:s that describe each bin, based on individual spike frequencies for the neurons. A measure of ”surprise”, based on the number of occurrences of each such vector, compared to the expected number of occurrences, is calculated and bins carrying ”surprise” exceeding some threshold are flagged as unitary events (Grün et al., 2002a). There is also a version of the method that proceeds without binning, but this would incur prohibitive computational costs in our case, where we record from hundreds of simulated neurons; instead we verify validity by slightly varying the bin size (Grün et al., 1999, 2002a).

Since activity in our network is certainly nonstationary, each minicolumn alternating between bursting and quiescent modes, we calculate firing frequency using a windowed average. The window size is chosen such that the estimated firing frequencies for individual neurons track the instantaneous firing frequency for the full pattern to which they belong. The window size turns out to be an important parameter in our case; as the spike rate in the network varies on a short time scale, a small window size is required (Grün et al., 2002b).

Because of the large number of neurons in the network, we may expect that, at least in the high activity, bursting phase, there will be few or no repeats of one and the same spike pattern. We must therefore distinguish unitary events during the bursting phase without repeated spike patterns, setting the surprise threshold such that individual spike events will be flagged as appropriate. If a fixed surprise threshold were used, we would flag roughly the bins in which the most neurons fire; they all occur during peak network activity. To better control the sensitivity, I use a moving threshold that for each bin calculates the ”expected surprise”, based on the instantaneous firing frequencies, and similarly also the ”expected surprise deviation”. I then choose the threshold for flagging a bin as the expected surprise plus a multiple of the deviation, chosen such that the expected number of unitary events would be less than one, if there were no structure at all in the data; thereby the variations in network activity are compensated for.

With the unitary event method, it is found that there is significant substructure in the network bursts; it is not exhaustive to describe the minicolumns in the network as having one active and one quiescent state. On the other hand, when the frequency calculation is performed using a sufficiently small window size, just a handful of unitary events are identified. These occur in the beginning and towards the end of a burst period (figure 6.2). Unitary events have been associated with behaviourally salient points in time. In monkeys they have been observed prominently near the end of the delay period in a delayed response task, when the animal is preparing to execute a movement and also in relation to external stimuli (Riehle et al., 2000, 1997).

6.2.5 Temporal correlations

Using an autocorrelogram measure, the temporal structure of the patterns is analyzed on short and long time scales. The measure used is equivalent to the average crosscorrelogram between pairs of neurons. Two such measures are produced, one within the minicolumn and one averaged over pairs of minicolumns participating in the same pattern. A correction, similar to the shift predictor is applied, by subtracting from the raw correlations the correlogram for low pass filtered spike trains, removing gross temporal dynamics but preserving fine structure (Gerstein and Perkel, 1969). The averaged crosscorrelogram between units shows a well defined central peak, with the peaks corresponding to one oscillatory period at about 40 ms clearly visible and peaks for two periods at 80 ms being less prominent; figure 6.3(C). This synchronization is the source of the gamma frequency apparent in the EEG. In figure 6.3(D) we see that synchronization is much weaker be-

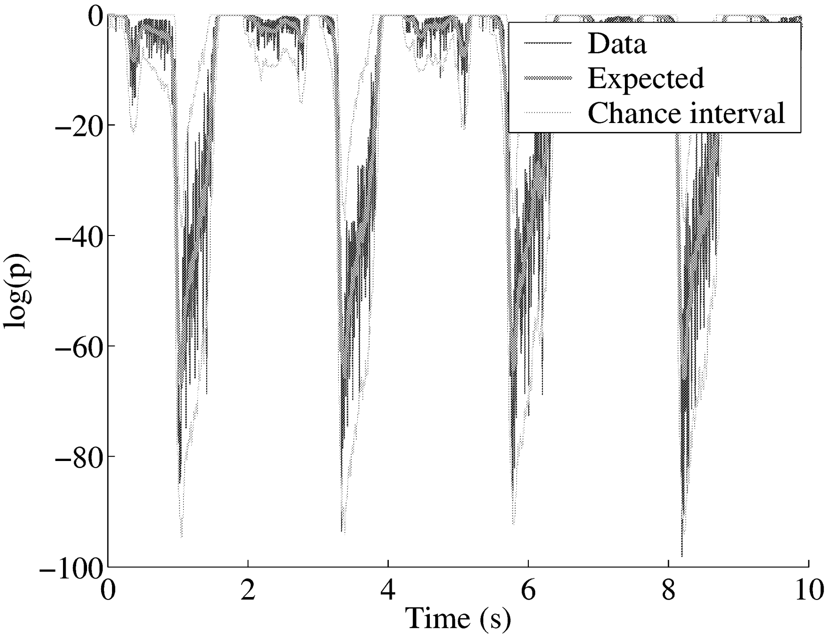


Figure 6.2: Unitary events calculated using a moving surprise threshold and a small spike rate averaging window. Only in the beginning and towards the end of up states are unitary events detected.

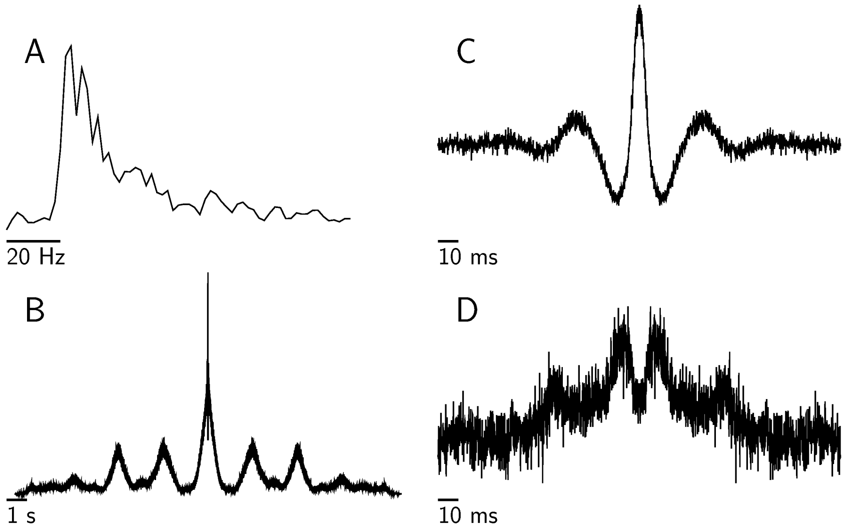


Figure 6.3: Temporal fine structure of network activity. *A:* Frequency spectrum of artificial EEG, generated by pyramidal cell currents. There is a peak energy around 25–30 Hz. *B:* Average autocorrelation between spiking activity within a minicolumn. UP state duration is reflected in peak width, pattern recurrence in the side peak spacings. *C:* Autocorrelation within a minicolumn on a short time scale, corrected for slow dynamics. The synchronization giving rise to the 25–30 Hz oscillation is evident. *D:* Average crosscorrelation between different minicolumns, belonging to the same pattern. An imprecise synchronization is evident, but with a tendency towards inverting the patterns seen within a minicolumn on the 10 ms time scale.

tween different minicolumns, even those belonging to the same cell assembly (Steriade et al., 1996). Similar oscillations have been found in local field potentials of awake behaving monkeys (Brovelli et al., 2004). On a longer time scale, there is also a non-zero correlation appearing after about a second and then again at longer time intervals, corresponding to a pattern being reactivated after a quiet period and can be seen in figure 6.3(B).

6.3 Conclusions

In constructing this detailed network model, there have been influences both from the top down model and from detailed empirical data on cortical structure. The model clearly demonstrates that attractor memory systems are realizable in the cortical substrate. Too much top down influence probably went into constructing the model though, to draw the stronger conclusion that the cortex *must* implement an attractor memory. Of particular interest is also the methodology of mimicking real experiments in the model universe, thereby extending the interface of observables between simulation and experiment.

Chapter 7

The large scale cortex

The models discussed so far have dealt with models of attractor memory and related concepts that each can be realized using just a tiny part of the cerebral cortex, as evidenced by the modest scale of the cortical network presented in the previous chapter. In this final main chapter of the thesis, I will deal with the cortex on a global scale. More accurately, I will deal with both the cortex, and the underlying white matter, which in the human brain occupies a volume about the same as the cortex itself (Pakkenberg and Gundersen, 1997). The corticocortical connections through the white matter become a very prominent feature as we consider the cortex on a larger scale.

The large scale structure of the cortex and the white matter has been approached with purely mathematical tools, including dimensionality analysis, that do not take functional aspects into consideration. Here I will stay true to the theme of my thesis and approach the structure of the cortex from a functional perspective, based on the hypothesis that the cortex largely functions as an attractor memory and has been optimized to operate as such. This hypothesis then leads to structural predictions.

7.1 Connectivity of the cortex

The cortex is sparsely connected on a global scale. If it were not, the amount of wiring would grow dramatically as the number of neurons was scaled up; going from rodent to primate to human the brain would come to consist almost exclusively of white matter. Suppose that the cortex were built from N computational units (individual neurons or minicolumns) that each occupy a fixed amount of gray matter volume

ν_g, for a total volume of $V_g = N\nu_g$. In this volume we include cell bodies, synapses and local connections, but not long range connections. If connectivity were global, here taken to mean that for each pair of units there is a fixed probability γ that they are connected, the number of connections becomes γN^2 (Braitenberg, 2001). If we now approximate the cerebral hemispheres, consisting of the cortex and the underlying white matter, as a sphere of volume V, the average length of a single connection, assuming it can be constructed as a straight line, would be proportional to the sphere radius. We can then write its volume as $a_w V^{1/3}$, bringing the total white matter volume to $V_w = \gamma N^2 a_w V^{1/3}$. The constant a_w incorporates the axon area and a dimensionless geometric constant. The volume of the model brain can be expressed as the sum of gray and white matter volumes; $V = V_g + V_w = \nu_g N + \gamma a_w V^{1/3} N^2$. Solving for N we find that $N = \frac{\sqrt{\nu_g^2 + 4\gamma a_w V^{4/3}} - \nu_g}{2\gamma a_w \sqrt[3]{V}}$; the gray matter fraction of the total volume becomes $\frac{V_g}{V} = \frac{\nu_g \sqrt{\nu_g^2 + 4\gamma a_w V^{4/3}} - \nu_g^2}{2\gamma a_w V^{4/3}}$. For large V we see that the gray matter fraction vanishes as $\frac{V_g}{V} \sim \frac{1}{V^{2/3}}$.

A mouse brain contains about $V_g^m = 0.1\,\mathrm{cm}^2$ gray and $V_w^m = 0.01\,\mathrm{cm}^2$ white matter (it contains about $1.6 \cdot 10^7$ pyramidal cells or $1.6 \cdot 10^5$ minicolumns) (Zhang and Sejnowski, 2000). Using these numbers, we may calculate ν_g and a_w and then plot how the white and gray matter volumes grow as the brain is scaled up by adding more neurons. In figure 7.1 it is shown that white matter completely dominates our hypothetical brain as it is scaled up. This will happen eventually, and with the same asymptotic power law ($\frac{V_g}{V} \sim \frac{1}{V^{2/3}}$), regardless of the following:

- The details of the gray matter computational unit and the white matter wiring.
- The average length of global connections, as long as it is a fixed fraction of brain size.
- The fraction of possible connections that are realized.

7.2 Connectivity structure

The above scaling argument, while simplistic, conclusively rules out random connectivity in the large scale cortex. This means that there must be some form of localized structure in corticocortical connectivity. Two types of such models are small world- and scale free networks.

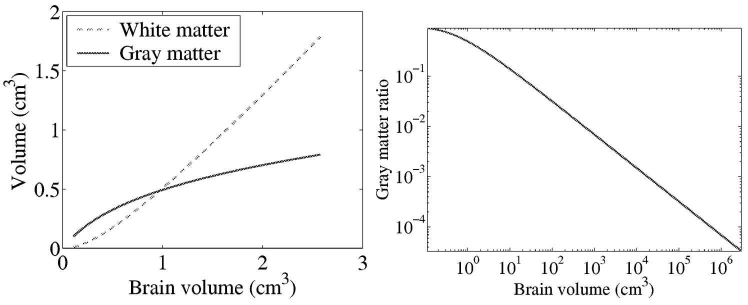


Figure 7.1: Scaling of white and gray matter in a hypothetical brain.

A small world network can be represented as a graph, where vertices typically have a large number of local connections to nearby vertices, but may also have one or a few global ones. Whether this architecture falls under the above scaling argument depends on how the global connections are set up; I will not here look into how to scalably wire the cortex based on a small world prescription. The prevalence of local connections leads to a high *clustering*, whereas the smaller number of global connections suffice to yield a short *average path length*. In the remainder of this chapter one particular small-world type network that I have already dealt with will be investigated; the hypercolumnar attractor network. This network has high local connectivity in the hypercolumns, but fewer global connections in between them.

7.2.1 Patchy connectivity

When viewed on an intermediate scale, the connectivity of the cerebral cortex shows a peculiar structure. It is patchy, meaning that two areas on cortical surface the scale of cortical hypercolumns tend to be either extensively interconnected, or not interconnected at all (Boyd and Matsubara, 1991). In this section I describe a model designed to understand this patchy connectivity, based on a hypercolumnar network, structurally similar to the biophysically detailed model described in chapter 6.

First we consider two extremes; one ”fully patchy” case where each hypercolumn communicates with all minicolumns in another hypercolumn, or with none. This means that the network connectivity is defined strictly on the hypercolumnar level, as opposed to the minicolumnar one, and would manifest itself as highly patch connectivity if we were to trace the lower level connectivity. At the opposite ”non-patchy” extreme, connections between minicolumns are independently established, meaning that the connectivity matrix is defined on the minicolumnar level. We investigate three versions of this kinds of patchiness; one which applies to the incoming connections, one which applies to the outgoing ones and finally a reciprocal case. The storage capacity of the network is evaluated as a function of patchiness, using numerical simulations as well as mathematical analysis.

7.3 Analysis of patchy connectivity

We analyse the network in two versions, the only difference being the learning rule for the synaptic weight matrix. The first version uses the sparse Hopfield learning rule previously described, the second one uses the Willshaw rule, which is similar, but with all weight values clipped at zero or one. For both models, the stability of the learned patterns is analysed. A pattern is stable if for each hypercolumn the ”correct” (already active) minicolumn receives the highest support value. For both models I describe how to calculate an approximate probability distribution for the support values, for both the ”correct” and ”incorrect” minicolumns and thereby the probability of pattern stability. We describe the clustering of the hypercolumnar network by the parameter C, running from 1 in the fully clustered network to 0 in the independently wired one. The task is then to find the optimal clustering value, for the highest storage capacity.

Further notation will be U for the number of minicolumns (meaning that $1/U$ is the activity level of the network), H for the number of hypercolumns and P for the number of stored patterns. The connectivity is described by K; for the fully clustered network this is the number of hypercolumns that are connected to each other hypercolumn.

7.3.1 Hopfield learning rule

The Hopfield learning rule is linear in the sense that each stored pattern adds to the synaptic matrix. The weight increment for each pattern can be described by a stochastic variable X_{ij};

$$X^{\mu}_{ij} = \left\{ \begin{array}{cc} (1-1/U)^2 & 1/U^2 \\ (1-U)/U^2 & 2(U-1)/U^2 \\ 1/U^2 & (U-1)^2/U^2 \end{array} \right.$$

These are *potential* synaptic weights in the sense that they will be realized only for those $\{ij\}$ where a connection actually exists. The realized synaptic weights are the sums over the contributions to X^{μ}_{ij} from each pattern, gated with a binary variable Y_{ij} that indicates whether the connection exists; $W_{ij} = Y_{ij} \sum_{\mu=1}^{P} X_{ij}$. The support for any minicolumn, when the pattern ξ^{ν}_j is active, is the sum of the synaptic weights linking it to the active minicolumns;

$$S^{\nu}_i = \frac{1}{N} \sum_k \xi^{\nu}_k W_{ij} = \frac{1}{N} \sum_k \xi^{\nu}_k Y_{ik} \sum_{\mu=1}^{P} X^{\mu}_{ik} \,.$$

This sum may be partitioned across the "block" hypercolumns that are, depending on the clustering parameter C, more or less completely connected to the hypercolumn where our minicolumn resides and the "non-block" hypercolumns that are not connected at all in the fully clustered case, but otherwise have a non-zero probability of being connected. The central limit theorem is applied to approximate the sum of the support variables as a normal distribution. We have the average $E(X_{ij}) = 0$ and variance $V(X_{ij}) = (U-1)^2/U^4$. We then treat the minicolumns that are part of the active pattern separately from those that are not. For the latter, the ξ, Y and X are independent and we calculate the average and variance of their support values to be $E(S^{-}_i) = 0$ and $V(S^{-}_i) = KPV(X_{ij})/N^2$. In the variance calculations, the Y_{ik} variables contribute, through an addition formula for second order moments, but in this case they have no effect because the average weight is zero. For the S^{+}_i, the situation is similar, except for the connections to other units in the same pattern. For those, the ξ^{ν}_k is known to be active and the weights will have a non-zero mean since this pattern was one of the learned ones. The Y_{ij} may still be zero however and this is where the difference between the patchy and non-patchy networks come in. The mean $E(S^{+}_i) = K(1-1/U)^2$ is independent on patchiness, but the variance;

$$\begin{aligned} V(S^{+}_i) &= \frac{K}{N^2(H-1)} \\ &* \ \big((1-1/U)^4(H(1-C^2)+(K+1)C^2-K-1)+(H-1)V(X^{\mu}_{ij})\big) \end{aligned}$$

becomes greater for the less patchy cases (smaller C) because of how the second order moments are added.

7.3.2 Willshaw learning rule

The probability that an entry in the synapse matrix of the Willshaw memory is used, when storing a single pattern is $p_o = 1/n^2$. After storing P patterns, the density of ones in the memory matrix is

$$p_1 = 1 - (1 - p_0)^P.$$

We may think of the fully clustered (patchy) extreme as a starting point, generating the less patchy networks by randomly relocating minicolumn connections. In the initial, fully clustered, network, the input to a particular minicolumn unit comes from the K "block" hypercolumns that are fully connected to the unit's hypercolumn. For a given clustering measure C, the ratio of units that have been relocated from this starting point is $1-C$. When a connection is relocated, it may be moved either to another "block" hypercolumn, or to a "non-block" hypercolumn. The probability of the former case is $\frac{K}{H}$. Therefore, the probabilities that a connection is present are, in the block and non-block cases respectively;

$$\begin{aligned} p_b &= C + (1 - C)\frac{K}{H} \\ p_n &= (1 - C)\frac{K}{H}. \end{aligned}$$

We now consider the stability of patterns. To this end, we first consider one hypercolumn. We calculate the support level of the unit that is part of the active pattern (S^+) and that of the other units (S^-):

$$\begin{aligned} S^+ &= \sum_{i=1}^{K} B_i^+ + \sum_{i=1}^{H-K} N_i^+ \\ S^- &= \sum_{i=1}^{K} B_i^- + \sum_{i=1}^{H-K} N_i^-, \end{aligned}$$

where the variables B_i are the contributions from the block parts and the N_i are from the non-block parts. Each of these stochastic variables take the value 1 precisely when a) there is a connection from the active unit in the other hypercolumn to the unit under consideration, and b) this synapse entry is set to one. The latter is always true for the "+" units, in the active pattern. Assuming independence between the B_i:s and the N_i:s the sums become binomial distributions;

$$\begin{aligned} S^+ &\in \mathrm{Bin}(K,\, p_b) + \mathrm{Bin}(H-K-1,\, p_n) \\ S^- &\in \mathrm{Bin}(K,\, p_b p_1) + \mathrm{Bin}(H-K-1,\, p_n p_1). \end{aligned}$$

The probability for the pattern unit having strictly larger support than any one other unit becomes

$$p_{\text{unit}} = P(S^+ > S^-). \tag{7.1}$$

The probability for stable recall in one hypercolumn becomes $p_{\text{hyper}} = p_{\text{unit}}^{U}$ and the probability that all hypercolumns are thus stable is $p_{\text{pattern}} = p_{\text{hyper}}^{H} = p_{\text{unit}}^{UH}$. This is also the expected ratio of stable patterns for a given memory load;

$$r = P(s_+ > s_-)^{UH}.$$

We can now determine the storage capacity, given our performance criterion. The summation over the possible outcomes for the stochastic variables $S^{\pm}$ implicit in equation 7.1 can be carried out exactly for any reasonably small network size, or be approached by a normal approximation for very large networks.

7.4 Results

In addition to the above analysis, we simulated the network storage capacity using a computer implementation. We determined storage capacity in both the analytical approach and the simulations as the largest number of patterns that could be retrieved, varying the number of stored patterns to achieve the optimum number.

Networks with patchy connectivity were found to have higher storage capacity than non-patchy networks (figure 7.2). This is due to a larger variability in the non-patchy networks. Simply stated, there is a risk that important, signal-carrying connections may be missing, when connectivity is less structured. This may explain the connectivity structure seen in the brain. Further, the analysis shows that the most important type of clustering is "sender side" clustering; meaning that all minicolumns should be represented when a hypercolumn projects to another hypercolumn, leading to a prediction that this type of clustering should be strongly represented in the brain.

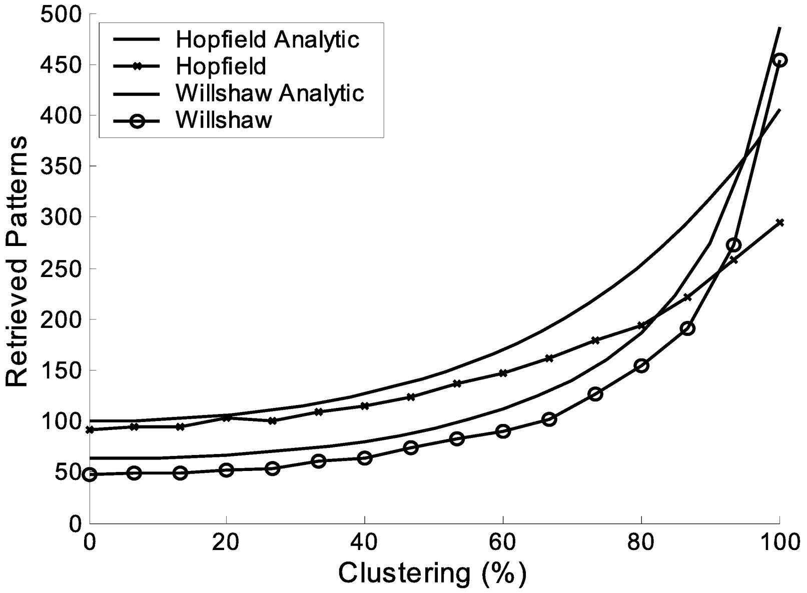


Figure 7.2: Storage capacity in the patchy hypercolumnar network, as a function of clustering. Shown are analytical and simulation results for networks with Hopfield and Willshaw learning rules. High clustering leads to a higher storage capacity.

Chapter 8

Conclusions

I have in this thesis studied some aspects in an overall picture of the cortex. The central piece is attractor memory function which allows for storage and retrieval of arbitrary memory patterns. Attractor memory systems are notable for their content addressing capabilities. This allows for generalizing experiences to novel situations; very rapidly accessing past experiences similar to the present situation. Using such similarity-based computation, many challenging tasks may prove to be nothing more than memory retrieval, but there is a catch. To access similar memories, there must be a metric or a measure of what ”similar” means. To an attractor memory system, the measure of similarity is embedded in the representation – the codes used by the system for different inputs, objects or situations. Constructing such a code is highly non-trivial; for it must entail all the high level features and abstract concepts that one desires the system to handle.

I have suggested a first step in the complex process of generating representations for sensory inputs; a way to produce efficient codes from natural image data. The codes generated by the proposed model are well suited for storage in an attractor memory; they are binary, sparse and have little internal correlations. In experiments I have found that transmission and storage of such codes is more efficient than using previously suggested representations that do not optimize a discrete objective function.

When dealing with artificial attractor memory systems, we often measure just a single aspect of their performance, notably storage capacity. It is important to realize that other characteristics are necessary for attractor memories to be good citizens in the computational community of the brain. Working memory has been suggested to operate as an attractor memory and long term memory is certain to be fundamentally involved in cognitive tasks. Such tasks certainly involve a temporal aspect. I model this kind of task using a sequence learning model and argue that in such a situation heteroassociative attractor memory does best to perform its retrieval and then *get out of the way*. My model sequence learning network demonstrates how that can be accomplished by way of synaptic dynamics, without sacrificing storage capacity.

Brain theories must be solidly rooted in empirical knowledge. I show, in a methodological review, that model realism is not sufficient to ensure this. But by constructing similar models at various levels of abstraction, high level models may be more tightly coupled to the real brain. For a biophysically detailed attractor memory model, I show how virtual experiments, e.g. EEG measurements, may be performed on the model and directly compared to actual experimental results. Using an abstract model of the same type on the other hand, a possible functional explanation for the experimentally observed phenomenon of reciprocally patchy connectivity in the cortex is put forth.

By linking abstract and detailed models, by analyzing constraints imposed on cortical architecture from scaling requirements and by considering experimental results from a functional perspective, we are slowly increasing our knowledge of the computational functions hidden in the cerebral cortex. In this thesis I have sampled but a few pieces of that intriguing puzzle.

Bibliography

Abeles, M. (1991). *Corticonics: Neural Circuits of the Cerebral Cortex.* Cambridge University Press, Cambridge, UK.

Aboitiz, F., Montiel, J., and Lopez, J. (2002). Critical steps in the early evolution of the isocortex: insights from developmental biology. *Braz J Med Biol Res*, 35(12):1455–1472.

Aboitiz, F., Morales, D., and Montiel, J. (2003). The evolutionary origin of the mammalian isocortex: towards an integrated developmental and functional approach. *Behav Brain Sci*, 26(5):535–52; discussion 552–85.

Alexander D. Protopapas, M. V. and Bower, J. M. (1998). *Methods in Neuronal Modeling: From Ions to Networks*, chapter 12: "Simulating Large Networks of Neurons". MIT Press.

Amit, D. J. (1989). *Modelling Brain Function.* Cambridge University Press, Cambridge.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173:30–67.

Avons, S. E. (1998). Serial report and item recognition of novel visual patterns. *Br J Psychol*, 89 (Pt 2):285–308.

Barlow, H. B. (1983). *Understanding natural vision.* Springer-Verlag, Berlin.

Bechtel, W. and Abrahamsen, A. (in press). Explanation: A mechanistic alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences.*

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Bell, A. J. and Sejnowski, T. J. (1997). The independent components of natural images are edge filters. *Vision Research*, 37:3327–3338.

Boyd, J. and Matsubara, J. (1991). Intrinsic connections in cat visual cortex: a combined anterograde and retrograde tracing study. *Brain Res*, 560(1-2):207–215.

Braitenberg, V. (2001). Brain size and number of neurons: an exercise in synthetic neuroanatomy. *J Comput Neurosci*, 10(1):71–77.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality. *Proc Natl Acad Sci U S A*, 101(26):9849–9854.

Buxhoeveden, D. P. and Casanova, M. F. (2002). The minicolumn hypothesis in neuroscience. *Brain*, 125(Pt 5):935–951.

Carandini, M., Heeger, D. J., and Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *J Neurosci*, 17(21):8621–8644.

Colbert, C. M. (2001). Back-propagating action potentials in pyramidal neurons: a putative signaling mechanism for the induction of hebbian synaptic plasticity. *Restor Neurol Neurosci*, 19(3-4):199–211.

Collins, H. M. and Pinch, T. (1998). *The Golem : What You Should Know about Science*. Cambridge University Press, 2nd edition.

Cossart, R., Aronov, D., and Yuste, R. (2003). Attractor dynamics of network up states in the neocortex. *Nature*, 423(6937):283–288.

Cromwell, H. C. and Berridge, K. C. (1996). Implementation of action sequences by a neostriatal site: a lesion mapping study of grooming syntax. *J Neurosci*, 16(10):3444–3458.

de Noordhout, A. M., Rapisarda, G., Bogacz, D., Gerard, P., Pasqua, V. D., Pennisi, G., and Delwaide, P. J. (1999). Corticomotoneuronal synaptic connections in normal man: an electrophysiological study. *Brain*, 122 (Pt 7):1327–1340.

Diener, H. C., Hore, J., Ivry, R., and Dichgans, J. (1993). Cerebellar dysfunction of movement and perception. *Can J Neurol Sci*, 20 Suppl 3:S62–S69.

Dressler, O., Schneider, G., Stockmanns, G., and Kochs, E. F. (2004). Awareness and the EEG power spectrum: analysis of frequencies. *Br J Anaesth*, 93(6):806–809.

Düring, A., Coolen, A. C. C., and Sherrington, D. (1998). Phase diagram and storage capacity of sequence processing neural networks. *submitted to J. Phys. A: Math. Gen.*

Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nat Neurosci*, 3 Suppl:1184–1191.

Ekeberg, Ö. (1992). *Computer Simulation Techniques in the Study of Neural Systems*. PhD thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden. TRITA-NA-P9232.

Favorov, O. V. and Kelly, D. G. (1994). Minicolumnar organization within somatosensory cortical segregates: II. emergent functional properties. *Cereb Cortex*, 4(4):428–442.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1 – 47.

Földiák, P. (1990). Forming sparse representaions by local anti–Hebbian learning. *Biological Cybernetics*, 64:165–170.

Fox, L. E. and Lloyd, P. E. (2002). Mechanisms involved in persistent facilitation of neuromuscular synapses in aplysia. *J Neurophysiol*, 87(4):2018–2030.

Fransén, E. and Lansner, A. (1998). A model of cortical associative memory based on a horizontal network of connected columns. *Network: Computation in Neural Systems*, 9(2):235–264.

Friedman, M. (1966). The methodology of positive economics. *Essays In Positive Economics*, pages 3–16, 30–43. University of Chicago Press.

Frolov, A. A., Husek, D., and Muraviev, I. P. (1997). Informational capacity and recall quality in sparsely encoded Hopfield-like neural network: Analytical approaches and computer simulation. *Neural Networks*, 10(5):845–855.

Fuster, J. M. (2002). Frontal lobe and cognitive development. *J Neurocytol*, 31(3-5):373–385.

Gallagher, M. and Chiba, A. A. (1996). The amygdala and emotion. *Curr Opin Neurobiol*, 6(2):221–227.

Gerstein, G. L. and Perkel, D. H. (1969). Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science*, 164(881):828–830.

Grün, S., Diesmann, M., and Aertsen, A. (2002a). Unitary events in multiple single-neuron spiking activity: I. detection and significance. *Neural Comput*, 14(1):43–80.

Grün, S., Diesmann, M., and Aertsen, A. (2002b). Unitary events in multiple single-neuron spiking activity: II. nonstationary data. *Neural Comput*, 14(1):81–119.

Grün, S., Diesmann, M., Grammont, F., Riehle, A., and Aertsen, A. (1999). Detecting unitary events without discretization of time. *J Neurosci Methods*, 94(1):67–79.

Guillery, R. W. and Sherman, S. M. (2002a). Thalamic relay functions and their role in corticocortical communication: generalizations from the visual system. *Neuron*, 33(2):163–175.

Guillery, R. W. and Sherman, S. M. (2002b). The thalamus as a monitor of motor outputs. *Philos Trans R Soc Lond B Biol Sci*, 357(1428):1809–1821.

Hausser, M., Spruston, N., and Stuart, G. J. (2000). Diversity and dynamics of dendritic signaling. *Science*, 290(5492):739–744.

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley & Sons, New York.

Herrick, J. C. (1948). *The brain of the tiger salamander*. Univ. Chicago Press.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison Wesley, New York.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Acadamy of Science (USA)*, 79:2554–2558.

Hubel, D. H. and Wiesel, T. N. (1977). Ferrier lecture. functional architecture of macaque monkey visual cortex. *Proc R Soc Lond B Biol Sci*, 198(1130):1–59.

Ismailov, I., Kalikulov, D., Inoue, T., and Friedlander, M. J. (2004). The kinetic profile of intracellular calcium predicts long-term potentiation and long-term depression. *J Neurosci*, 24(44):9847–9861.

Johansson, R. S., Westling, G., Backstrom, A., and Flanagan, J. R. (2001). Eye-hand coordination in object manipulation. *J Neurosci*, 21(17):6917–6932.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (2000). *Principles of Neural Science*. McGraw-Hill Companies, 4th edition.

Koch, I. and Hoffmann, J. (2000). Patterns, chunks, and hierarchies in serial reaction-time tasks. *Psychol Res*, 63(1):22–35.

Kown, T. M. and Zervakis, M. (1995). KWTA networks and their applications. *Multidimensional Syst. Signal Process.*, 6(4):333–346.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*. Chicago University Press, 2nd edition.

Lamme, V. A., Super, H., and Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Curr Opin Neurobiol*, 8(4):529–535.

Larsen, W. J. (2001). *Human Embryology*. Churchill Livingstone, 3rd edition.

Laughlin, S. B. and Sejnowski, T. J. (2003). Communication in neuronal networks. *Science*, 301(5641):1870–1874.

Lee, C., van Heuveln, B., Morrison, C. T., and Dietrich, E. (1998). Why connectionist nets are good models. *Psycoloquy*.

Lee, D. D. and Seung, H. S. (1997). Information, prediction, and query by committee. In Hanson, S. J., Cowan, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 515–521. Morgan Kaufmann, San Mateo, CA.

Lennie, P. (2003). The cost of cortical computation. *Curr Biol*, 13(6):493–497.

Lloyd, E. A. (1998). *Routledge Encyclopedia of Philosophy*, volume 6, chapter Models, pages 443–447.

Lomo, T. (1968). Potentiation of monosynaptic epsps in cortical cells by single and repetitive afferent volleys. *J Physiol*, 194(2):84–5P.

Lomo, T. (1971). Potentiation of monosynaptic EPSPs in the perforant path-dentate granule cell synapse. *Exp Brain Res*, 12(1):46–63.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., and Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nat Rev Neurosci*, 5(10):793–807.

Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Prog Neurobiol*, 50(4):381–425.

Minsky, M. L. and Papert, S. (1988). *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, Massachusetts, expanded edition.

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain*, 120 (Pt 4):701–722.

Newcomer, J. W. and Krystal, J. H. (2001). Nmda receptor regulation of memory and behavior in humans. *Hippocampus*, 11(5):529–542.

Nicolelis, M. A. and Ribeiro, S. (2002). Multielectrode recordings: the next steps. *Curr Opin Neurobiol*, 12(5):602–606.

Nicoll, A. and Blakemore, C. (1993). Single-fibre EPSPs in layer 5 of rat visual cortex in vitro. *Neuroreport*, 4(2):167–170.

Northcutt, R. G. and Kaas, J. H. (1995). The emergence and evolution of mammalian neocortex. *Trends Neurosci*, 18(9):373–379.

Okada, M. (1996). Notions of associative memory and sparse coding. *Neural Netw*, 9(8):1429–1458. (ENG).

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–608.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed in V1. *Vision Research*, 37:3311–3325.

O'Rahilly, R. and Muller, F. (1994). Neurulation in the normal human embryo. *Ciba Found Symp*, 181:70–82; discussion 82–9.

O'Reilly, R. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2:455–462.

Pakkenberg, B. and Gundersen, H. J. (1997). Neocortical neuron number in humans: effect of sex and age. *J Comp Neurol*, 384(2):312–320.

Palm, G. (1980). On associative memory. *Biological Cybernetics*, 36:19–31.

Palm, G. and Sommer, F. T. (1995). Associative data storage and retrieval in neural networks. In E. Domany, J. L. van Hemmen, K. S., editor, *Models of Neural Networks III*, pages 79 –118. Springer, New York.

Pellizzer, G., Sargent, P., and Georgopoulos, A. P. (1995). Motor cortical activity in a context-recall task. *Science*, 269:702–705.

Peters, A. and Yilmaz, E. (1993). Neuronal organization in area 17 of cat visual cortex. *Cereb Cortex*, 3(1):49–68.

Popper, K. (1959). *The logic of scientific discovery*. Hutchinson.

Riehle, A., Grammont, F., Diesmann, M., and Grün, S. (2000). Dynamical changes and temporal precision of synchronized spiking activity in monkey motor cortex during movement preparation. *J Physiol Paris*, 94(5-6):569–582.

Riehle, A., Grün, S., Diesmann, M., and Aertsen, A. (1997). Spike Synchronization and Rate Modulation Differentially Involved in Motor Cortical Function. *Science*, 278(5345):1950–1953.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J Neurophysiol*, 88(1):455–463.

Rolls, E. T. and Treves, A. (1997). *Neural Networks and Brain Function*. Oxford University Press.

Sakai, K., Ramnani, N., and Passingham, R. E. (2002). Learning of sequences of finger movements and timing: frontal lobe and action-oriented representation. *J Neurophysiol*, 88(4):2035–2046.

Salin, P. A. and Bullier, J. (1995). Corticocortical connections in the visual system: structure and function. *Physiol Rev*, 75(1):107–154.

Schwenker, F., Sommer, F. T., and Palm, G. (1996). Iterative retrieval of sparsely coded associative memory patterns. *Neural Networks*, 9(3):445–455.

Scott, S. H. (2003). The role of primary motor cortex in goal-directed movements: insights from neurophysiological studies on non-human primates. *Curr Opin Neurobiol*, 13(6):671–677.

Simmons, D. M. and Swanson, L. W. (1993). *Neuroscience Protocols*, chapter The Nissl stain, pages 1–7. Elsevier Science.

Staiger, J. F., Flagmeyer, I., Schubert, D., Zilles, K., Kotter, R., and Luhmann, H. J. (2004). Functional diversity of layer IV spiny neurons in rat somatosensory cortex: quantitative morphology of electrophysiologically characterized and biocytin labeled cells. *Cereb Cortex*, 14(6):690–701.

Steriade, M., Contreras, D., Amzica, F., and Timofeev, I. (1996). Synchronization of fast (30-40 Hz) spontaneous oscillations in intrathalamic and thalamocortical networks. *J Neurosci*, 16(8):2788–2808.

Streit, A. and Stern, C. D. (1999). Establishment and maintenance of the border of the neural plate in the chick: involvement of FGF and BMP activity. *Mech Dev*, 82(1-2):51–66.

Sugimoto, S., Sakurada, M., Horikawa, J., and Taniguchi, I. (1997). The columnar and layer-specific response properties of neurons in the primary auditory cortex of mongolian gerbils. *Hear Res*, 112(1-2):175–185.

Super, H. and Uylings, H. (2001). The Early Differentiation of the Neocortex: a Hypothesis on Neocortical Evolution. *Cereb. Cortex*, 11(12):1101–1109.

Tallon-Baudry, C., Bertrand, O., Peronnet, F., and Pernier, J. (1998). Induced gamma-band activity during the delay of a visual short-term memory task in humans. *J Neurosci*, 18(11):4244–4254.

Taylor, W. R., He, S., Levick, W. R., and Vaney, D. I. (2000). Dendritic computation of direction selectivity by retinal ganglion cells. *Science*, 289(5488):2347–2350.

Treves, A. and Rolls, E. T. (1992). Computational analysis of the operation of a real neuronal network in the brain: the role of the hippocampus in memory. volume 2, pages 891–898. Elsevier Science Publishers B.V.

Treves, A. and Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391.

Tsodyks, M. V. and Markram, H. (1997). The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc. Natl. Acad. Sci. USA*, 94:719–723.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proc. London Math. Soc.*, 2(42):230–265.

Ullén, F., Forssberg, H., and Ehrsson, H. H. (2003). Neural networks for the coordination of the hands in time. *J Neurophysiol*, 89(2):1126–1135.

Wadden, T. and Ekeberg, Ö. (1998). A neuro-mechanical model of legged locomotion: Single leg control. 79:161–173.

Wang, Y., Gupta, A., Toledo-Rodriguez, M., Wu, C. Z., and Markram, H. (2002). Anatomical, physiological, molecular and circuit properties of nest basket cells in the developing somatosensory cortex. *Cereb Cortex*, 12(4):395–410.

Whitehead, A. N. and Russell, B. (1925-1927). *Principia mathematica*. Cambridge University Press.

Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, 222:960–962.

Woodward, J. (Summer 2003). Scientific explanation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Xanthos, J. B., Kofron, M., Tao, Q., Schaible, K., Wylie, C., and Heasman, J. (2002). The roles of three signaling pathways in the formation and function of the spemann organizer. *Development*, 129(17):4027–4043.

Zhang, K. and Sejnowski, T. J. (2000). A universal scaling law between gray matter and white matter of cerebral cortex. *Proc Natl Acad Sci U S A*, 97(10):5621–5626.

Zhu, Y., Stornetta, R. L., and Zhu, J. J. (2004). Chandelier cells control excessive cortical excitation: characteristics of whisker-evoked synaptic responses of layer 2/3 nonpyramidal and pyramidal neurons. *J Neurosci*, 24(22):5101–5108.

Ziemann, U., Ilic, T. V., Alle, H., and Meintzschel, F. (2004). Estimated magnitude and interactions of cortico-motoneuronal and Ia afferent input to spinal motoneurones of the human hand. *Neurosci Lett*, 364(1):48–52.