KUNGL
TEKNISKA
HÖGSKOLAN

Bayesian Attractor Neural Network Models of Memory

Anders Sandberg

Stockholm 2003

Doctoral Dissertation

Stockholms universitet

Institutionen för numerisk analys och datalogi

Akademisk avhandling som med tillstånd av Stockholms universitet framlägges till offentlig granskning för avläggande av filosofie doktorsexamen 6 Juni 2003 13:15 i D1, Lindstedtsvägen 17.

ISBN 91-7265-684-0, pp 1–96,125–155
TRITA-NA-0310
ISSN 0348-2952
ISRN KTH/NA/R--03/10--SE

Universitetsservice US AB, Stockholm 2003

Abstract

The work presented in this thesis deals with neural network models of human memory based on the Bayesian Confidence Propagation Neural Network (BCPNN). The focus is to explore how a model derived from a statistical framework can link more abstract top-down cognitive models with biologically plausible cortex models. Of special interest is whether it there exists necessary architectural differences between different memory systems or whether they can all be achieved within the same neural architecture for different parameter values.

The main contributions are:

A derivation of an attractor neural network based on the BCPNN. The BCPNN framework consists of interpreting network unit activations as probability estimates ("confidence") in the presence of features or states of the world, and the update dynamics as Bayesian inference producing posterior estimates from the initially known information. The process converges to a self-consistent estimate of the state of the world given the prior known information. Learning consists of updating probability estimates, and by using exponential smoothing the network becomes able to learn incrementally with a learning time constant that can be regulated. The network exhibits palimpsest properties, avoiding catastrophic forgetting by forgetting old patterns at a rate dependent on the learning time constant. By changing the time constant the network can act both as a fast-learning, fast forgetting working memory and a slowly learning, slowly forgetting long term memory. The statistical derivation introduces a degree of modularity with similarities to cortical hypercolumns. The capacity is found to scale optimally with network size. The convergence time to an attractor state is dependent on training set size and the age of the state, suggesting a simple parallel model of the Sternberg reaction time effect.

Regulation of the learning time constant enables selective enhancement or inhibition of learning new information. This regulation acts by changing the relative sizes of basins of attraction, which can produce retroactive inhibition. Using this regulation an account of the isolate effect based on plasticity modulation is developed and compared to a model based on pattern decorrelation.

By adding a phenomenological model of cellular and synaptic adaptation the attractor states become unstable over time, producing a dynamics of fast convergence to a quasi-attractor where the network state dwells until slower adaptation forces it to move to another attractor. This dynamics enables intrinsically driven free recall or reinstatement that preserves information of basin of attraction size, external control over dynamics including second-best match and on-line learning during recall.

The network was applied to the delayed response oculomotor working memory task based on a fast Hebbian plasticity hypothesis of working memory. It reproduced the behaviour of other simulations and generated a range of experimental predictions.

Based on the assumption of a plasticity-limited memory capacity and the need for maximising stored information at reproductive age a model of cognitive aging is derived where the learning rate decreases in an optimal manner. A continuation of learning rate decrease past reproductive age produces age-dependent memory impairments. This model produces plausible autobiographical memory curves with infantile amnesia and the autobiographical bump.

ISBN 91-7265-684-0, pp 1–96,125–155 • TRITA-NA-0310 • ISSN 0348-2952 • ISRN KTH/NA/R--03/10--SE

Sammanfattning

Denna avhandling behandlar neuronnätsmodeller av mänskligt minne baserade på bayesianska konfidenspropageringsneuronnät (BCPNN). Fokus ligger på att utforska hur en modell härledd från ett statistiskt ramverk kan knyta samman abstrakta "top-down" kognitiva modeller med biologiskt plausibla "bottom-up" modeller av hjärnbarken. Av särskilt intresse är frågan huruvida det existerar nödvändiga arkitektoniska skillnader mellan olika minnesystem eller om deras funktionella specialiseringar kan uppnås med hjälp av samma neurala arkitektur med olika parametervärden.

De huvudsakliga bidragen är:

En härledning av ett attraktorneuronnät baserat på BCPNN. BCPNN-ramverket tolkar nätverksenheternas aktivering som sannolikhetsuppskattningar (konfidens) av olika kännetecken eller tillstånd i omvärlden och uppdateringsdynamiken som Bayesiansk inferens av *a posteriori* sannolikheter från den ursprungliga informationen. Processen konvergerar till en uppskattning av världens tillstånd konsistent med den tidigare kända informationen. Inlärning utgörs av att uppdatera sannolikhetsuppskattningar, och genom att använda exponentiell utjämning blir nätverket kapabelt att lära inkrementellt och med en regleringsbar tidskonstant. Nätverket uppvisar palimpsestegenskaper och undviker katastrofal glömska genom att glömma gamla mönster med en hastighet beroende på inlärningstidskonstanten. Genom att ändra tidskonstanten kan nätverket fungera både som ett snabbt lärande och glömmande arbetsminne eller ett långsamt lärande och glömmande långtidsminne. Den statistiska härledningen introducerar en modulär struktur som påminner om kortikala hyperkolumner. Kapaciteten visar sig skala optimalt med nätverksstorleken. Konvergenstiden till ett attraktortillstånd beror på träningsmängdens storlek och tillståndets ålder, vilket kan användas för en enkel parallell model av Sternbergs reaktionstidseffekt.

Reglering av inlärningstidskonstanten möjliggör selektiv förstärkning eller försvagning av hur ny information inlärs. Denna reglering verkar genom att förändra den relativa storleken av attraktionsbassänger. En modell av isolationseffekten baserad på plasticitetsreglering jämförs med en baserad på dekorrelerade mönster.

Genom att addera en fenomenologisk modell av cellulär och synaptisk adaptation blir attraktortillstånden instabila över tiden, vilket producerar en dynamik med snabb konvergens till en kvasiattraktor där nätverkets tillstånd förblir tills den långsamma adaptationen tvingar det till en annan kvasiattraktor. Denna dynamik möjliggör internt driven fri erinran eller återkallande som bevarar information om storlekar på attraktionsbassänger, extern kontroll över dynamiken inklusive second-best match och inlärning parallellt med erinran.

Nätverket tillämpas också på en okulomotorisk arbetsminnesuppgift utifrån en alternativ hypotes om arbetsminne baserat på snabb Hebbsk plasticitet. Det reproducerar beteendet av andra simulationer och genererar en uppsättning experimentella prediktioner.

Baserat på antagandet om en plasticitetsbegränsad minneskapacitet och det evolutionära behovet av att maximera lagrad information vid reproduktiv ålder, härleddes en modell av kognitivt åldrande där inlärningshastigheten minskar på ett optimalt sätt. En fortsättning av inlärningshastighetsminskningen bortom reproduktiv ålder producerar åldersberoende minnestörningar. Denna modell reproducerar plausibla självbiografiska minneskurvor med infantil amnesi och den självbiografiska kullen.

ISBN 91-7265-684-0, pp 1–96,125–155 • TRITA-NA-0310 • ISSN 0348-2952 • ISRN KTH/NA/R--03/10--SE

Contents

1 Introduction **5**
- 1.1 Thesis structure . . . 7
- 1.2 Articles . . . 8
- 1.3 Abbreviations and Notation . . . 9

2 Memory and the Brain **13**
- 2.1 Memory: A Cognitive Neuroscience View . . . 13
  - 2.1.1 Types of Memory . . . 13
  - 2.1.2 Memory Systems and Brain Systems . . . 17
- 2.2 Memory: A Cortical View . . . 22
  - 2.2.1 Cortical Circuitry and Modularity . . . 23
  - 2.2.2 Neural Plasticity . . . 25
  - 2.2.3 Cell Assemblies . . . 26
- 2.3 Memory: A Neural Network View . . . 31
  - 2.3.1 Computational Memory Models . . . 31
  - 2.3.2 Attractor Neural Networks . . . 34
  - 2.3.3 Biologically Inspired Attractor Memory Models . . . 35
  - 2.3.4 Memory Models and this Thesis . . . 37

3 Bayesian Confidence Propagation Neural Networks (BCPNN) **39**
- 3.1 Introduction . . . 39
  - 3.1.1 Palimpsest Memories . . . 40
- 3.2 BCPNN . . . 42
- 3.3 Heuristic Derivation of Network Architecture and Learning Rule . . . 43
  - 3.3.1 Naive Bayesian Classifier BCPNN . . . 43
  - 3.3.2 Discrete Valued Attribute Network . . . 44
  - 3.3.3 Recurrent BCPNN . . . 46
  - 3.3.4 The Prior State . . . 47
  - 3.3.5 Summing Bayesian Learning . . . 48
  - 3.3.6 Incremental Bayesian Learning . . . 48
- 3.4 Network Learning and Dynamics . . . 50
  - 3.4.1 Behaviour of Single Connection Weights . . . 50

3.4.2 Learning and Forgetting 52
3.4.3 Storage Capacity 53
3.4.4 Convergence Speed 58
3.4.5 Free Recall with Noise 61
3.4.6 Comparison with Clipped Weights 61
3.4.7 Different Learning and Forgetting Rates 63
3.5 Discussion 63

4 Memory Modulation **69**
4.1 Relevance Modulation 69
4.2 Effects of Learning Time Constant Modulation 71
4.2.1 Performance Model 73
4.3 Correlated Patterns 76
4.4 Discussion 77

5 Adaptation **81**
5.1 Synaptic and Cellular Adaptivity 81
5.2 Phenomenological Adaptation and Reinstatement Model 82
5.3 Network 84
5.4 Quasi-Attractor Dynamics 85
5.5 Dwell times and Pattern Overlaps 89
5.6 Second-Best Match 90
5.7 Online Learning 91
5.8 Discussion 92

6 Working Memory **97**
6.1 Introduction 97
6.2 The network simulation model 98
6.3 Setup of the delayed oculomotor task 99
6.4 Results 100
6.4.1 Adaptation 105
6.4.2 Non-hebbian plasticity 106
6.5 Discussion 107

7 Mental Ageing **111**
7.1 Age-Related Memory Impairment 111
7.2 Episodic and Autobiographical Memory 113
7.3 Cognitive Aging Models 114
7.4 Evolutionary Neuroscience 115
7.5 Optimal Learning Rate 116
7.5.1 Convex f 117
7.5.2 Simulation Model 118
7.5.3 Simulation Results 119
7.6 Discussion 121

8 Discussion and Conclusions **125**
8.1 Encoding and Retrieval 125
8.2 BCPNN as a Memory Model 127
8.3 Current Work 128
8.3.1 Networks of Networks 129
8.4 Open Issues and Further Research 130

To my dear mother.

Socrates: They say the cause of these variations is as follows: When the wax in the soul of a man is deep and abundant and smooth and properly kneaded, the images that come through the perceptions are imprinted upon this heart of the soul – as Homer calls it in allusion to its similarity to wax –; when this is the case, and in such men, the imprints, being clear and of sufficient depth, are also lasting. And men of this kind are in the first place quick to learn, and secondly they have retentive memories, and moreover they do not interchange the imprints of their perceptions, but they have true opinions. For the imprints are clear and have plenty of room, so that such men quickly assign them to their several moulds, which are called realities; and these men, then, are called wise. Or do you not agree?

Theaetetus: Most emphatically.

Socrates: Now when the heart of anyone is shaggy (a condition which the all-wise poet commends), or when it is unclean or of impure wax, or very soft or hard, those whose wax is soft are quick to learn, but forgetful, and those in whom it is hard are the reverse. But those in whom it is shaggy and rough and stony, infected with earth or dung which is mixed in it, receive indistinct imprints from the moulds. So also do those whose wax is hard; for the imprints lack depth. And imprints in soft wax are also indistinct, because they melt together and quickly become blurred; but if besides all this they are crowded upon one another through lack of room, in some mean little soul, they are still more indistinct. So all these men are likely to have false opinions. For when they see or hear or think of anything, they cannot quickly assign things to the right imprints, but are slow about it, and because they assign them wrongly they usually see and hear and think amiss. These men, in turn, are accordingly said to be deceived about realities and ignorant.

–Plato, Theatetus 194d and following (Fowler trans.)

Acknowledgements

I wish to thank:

- Anders Lansner, my advisor, for his patience and support. Time after time he has shown great intuition for what the networks would do and why. Had I listened to him more closely I would have been finished in half of the time.
- Karl Magnus Petterson for introducing me to the cognitive neuroscience of memory and pushing me onwards.
- Erik Fransén for stimulating discussions of methodology and what we really can learn from neural network simulations, as well as his excellent thesis.
- Anders Holst for showing in his thesis how to clearly explain BCPNN.
- Örjan Ekeberg for many wise insights and a much help in getting Debian to work.
- Christopher Johansson for much invaluable simulation and theoretical work.
- Jesper Tegnér for much assistance with working memory, especially pointing out what results were important.
- Mikael Djurfeldt for many useful comments and for introducing me to the music of Philip Glass.
- Sverker Sikström for inspiring remarks about the isolate effect and consolidation.
- David Eriksson for many interesting discussions about the nature of clustering within a BCPNN.
- Bengt Ström for demonstrating how well simple normalisation can work.
- Axel Liljenkrantz for exploring the consolidation model and forcing me to clarify my thinking on adaptation.

- The entire SANS group for providing a stimulating and creative environment: Jeanette Hellgren Kotaleski, Alexander Kozlov, Peter Raicevic, Charlotte Eriksson, Mikael Huss, Martin Rehn, Pål Westermark, Anders Fagergren and Einar Larsson.
- Roland Orre for showing how to really write good acknowledgements.
- Martin Ingvar for much help and inspiration.
- Håkan Andersson for feeding me on unexpected occasions.
- My gaming group for patience when their gamemaster spent weekends working instead of playing.
- The members of the Eudoxa think tank for patience when their colleague spent weekends writing instead of working.
- And many more...

This work was supported by TFR grant No. 97274 and MFR grant No. 12716, and a doctoral position at Stockholm University.

Chapter 1

Introduction

Memory at its most general can be defined as the ability to retain and reuse past experienced information. It is a necessary ability for all agents that need to adapt to a changing complex world, be they biological or artificial. Understanding the factors that enable adaptive behaviour is both important for constructing and applying complex artificial systems, and in deepening the understanding of biology, medicine and psychology.

This thesis will focus on neural memory with a particular emphasis on certain forms of human memory. The main issue is to study what memory phenomena can be implemented within a basic neural network framework. Can different memory systems be modelled using the same basic neural architecture? How much of the properties of different memory systems can be explained in terms of different parameters rather than structural differences? Another issue paralleling this is the development of a flexible building block for more complex memory models.

Memory models span the range from detailed computational neuroscience models of synapses and their internal biochemical networks over more simplified cell models to connectionist networks and functional models of memory that contain little or no biological detail. In moving towards higher levels of abstraction larger systems can be studied, enabling linking to quantitative behavioural data but often they are structurally or functionally underconstrained. Detailed models on the other hand are constrained by the rapidly increasing amount of complex biological data, making it hard to see the forest for the trees. The combination of large voids in our knowledge of certain parameters and phenomena and almost excessive information about other aspects makes it imperative to base exploration on simple, robust hypotheses and models that allow directed exploration without being too limited by lacking data.

Memory models also range from qualitative to quantitative. Many conceptual or qualitative memory models have been proposed at varying levels of specificity and connection to biological detail (see chapter 2 for a brief review). They have been influential in guiding research, but have the disadvantage of often being under-

constrained. Quantitative modelling is necessary in order to tie models to empirical data and generate testable predictions. It enables falsifiable theoretical constructs as well as connecting different levels of detail.

This thesis is neither intended to describe a neurophysiological model nor a purely functional/cognitive model. Rather, it is positioned in the intermediate level between structural bottom-up models where memory is modelled in terms of more or less realistic neurons and cell assemblies, and functional top-down models where memory is modelled in terms of correlations and high-level constructs. Ideally this kind of mid-level model enables translation and integration of observations or theories between the levels.

Important issues that this kind of model could help explore involve: Are there any necessary architectural differences between different memory systems such as working memory, short-term memory and long-term memory, or can they all be achieved within the same neural architecture? Can this architecture be consistent with cognitive and neurobiological data? What memory effects can modulation of parameters achieve? How can interacting simple systems produce adaptive effects?

In this thesis I will to some extent attempt to bridge the work described in the previous thesis of Erik Fransén (Fransén, 1996) and the thesis of Anders Holst (Holst, 1997). Fransén's thesis deals with biophysically detailed models of cortical memory, studying how Hebbian cell assemblies could be implemented within a biologically realistic neural network. Holst's thesis studies artificial neural networks derived from Bayesian inference and their applications. This thesis attempts to use models derived from the framework of Holst and earlier work to study memory systems on a larger scale than would be possible with biophysical models, while attempting to retain a reasonable level of abstraction as well as a connection to the biologically plausible models of Fransén et al. Ideally these models should allow the construction of more complex "networks of networks" for the study of interacting memory systems.

The methodology here is in a sense negative. Since it is hard to prove at this stage that the brain actually does implement a particular neural or functional architecture, it may be worthwhile to study how simple models can demonstrate the same properties. This is especially important in evaluating other models. A complex model can in general fit any observed facts well by adapting its multiple degrees of freedom, but is less satisfactory than a more parsimonious model as a tool for qualitative understanding. If an empirical finding can be modelled by an elaborate model, the demonstration of a less elaborate model that can replicate the finding is a step forward. If the simpler model provides the same results as is observed then the more complex model should be tentatively rejected according to Occam's razor. If the simpler model does not replicate all but a significant amount of the empirical data, then the discrepancy between the models provides a potential avenue of deeper understanding of what aspects of the studied system are generic in the space of models and what aspects can be used to chose between models.

The approach here is to model memory in terms of a family of neural networks within a probabilistic framework where learning consists of updating probability

estimates and decisions/retrieval consist of deducing a likely state of the world from prior information and the current, uncertain information at hand. Learning is controlled by time constants which can be externally modulated and affect the behaviour of the network. The statistical derivation implies a natural interpretation of memory and network behaviour in terms of inference, decision making and confidence estimation.

This statistical framework can be viewed either as a phenomenological model of neural activity similar to other artificial neural networks or a hypothesis of brain information processing. This thesis does not claim that the brain is implementing the attractor network that is studied or that the only natural interpretation of cortical activity is as confidence estimates. Rather, it attempts to show that such a simple model can account for a wide range of memory phenomena observed in humans and other animals.

There is a often a trade-off between mathematical analysability and expressiveness in models. The framework used in this thesis is derived from a mathematically analysable origin, but the inclusion of new assumptions and functionality makes it relatively intractable to conventional analysis. Hence the approach used will be more empirical than analytical.

The main contributions of this thesis are:

- A derivation and analysis of an autoassociative Bayesian confidence propagation neural network that performs as an efficient palimpsest memory.
- A demonstration and investigation of how learning can be controlled by a time constant that has a more natural biological interpretation than the control parameters of most previous palimpsest models. Modulation of this learning time constant acts as a print-now signal, with implications to emotional modulation of memory and the isolate effect.
- A phenomenological model of synaptic and cellular adaptation is introduced and shown to exhibit the same properties as more detailed and biologically realistic models. It enables the network to visit multiple stored quasi-stable states as part of free recall, to produce second best match to an input and to maintain a working memory of multiple items.
- A model of the delayed response oculomotor task with bump states is demonstrated, giving an alternative account for the formation of spatial working memory.
- A model of the effects of lifespan modulation of brain plasticity, replicating autobiographical memory curves.

1.1 Thesis structure

The basic structure of the thesis is:

- An initial review of information from cognitive science, neuroscience and the theory of neural networks relevant for the thesis (Chapter 2).
- A derivation and examination of the properties of the BCPNN family of neural networks, with particular emphasis on issues of memory capacity, convergence speed, cued and free recall (Chapter 3).
- An analysis of modulation of learning rate in the network and its relationship to neuromodulation, emotional memory and the von Restorff effect (Chapter 4).
- An extension of the basic network to include a phenomenological model of cellular and synaptic adaptation, enabling a more complex intrinsic dynamic (Chapter 5).
- An application of the framework to working memory in the form of a model of the delayed response oculomotor task (Chapter 6).
- An application of the framework to the issue of autobiographic learning and lifespan changes in neuromodulation (Chapter 7).

1.2 Articles

This thesis is based on the following reports and articles:

1. A. Sandberg, A. Lansner, K.M. Petersson and Ö. Ekeberg, *An incremental Bayesian learning rule*, NADA Tech. report TRITA-NA-P9908 1999.
2. A. Sandberg, A. Lansner, K.M. Petersson and Ö. Ekeberg, A Palimpsest Memory based on an Incremental Bayesian Learning Rule, *Neurocomputing* 32-33 (2000) 987-994.
3. A. Sandberg, A. Lansner, K.M. Petersson, Selective Enhancement of Recall through Plasticity Modulation in an Autoassociative Memory. *Neurocomputing* 38-40 (2001) 867-873.
4. A. Sandberg and A. Lansner, Synaptic Depression as an Intrinsic Driver of Reinstatement Dynamics in an Attractor Network, *Neurocomputing* 44-46, June 2002, 615–622
5. A. Sandberg, A. Lansner, K.M. Petersson and Ö. Ekeberg, A Bayesian attractor network with incremental learning *Network: Comput. Neural Syst.* 13 (May 2002) 179-194
6. C. Johansson, A. Sandberg and A. Lansner, A Neural Network with Hypercolumns, ICANN 2002 International Conference, Madrid, Spain, Proceedings, Lecture Notes in Computer Science 2415, Springer-Verlag 2002.

7. A. Sandberg, J. Tegnér and A. Lansner, A Working Memory Model Based on Fast Hebbian Learning. Submitted to Network: Computation in Neural Systems.

8. A. Lansner, E. Fransén and A. Sandberg. Cell assembly dynamics in detailed and abstract attractor models of cortical associative memory, *Theory in Biosciences*. In press 2003.

1.3 Abbreviations and Notation

Abbreviations

ACh Acetylcholine

ANN Artificial Neural Network

BCPNN Bayesian Confidence Propagation Neural Network

CF Catastrophic Forgetting

EPSP Excitatory Postsynaptic Potential

LTD Long-Term Depression

LTM Long-term Memory

LTP Long-Term Potentiation

MTL Medial Temporal Lobe

NBC Naive Bayesian Classifier

PFC Prefrontal Cortex

PPC Posterior Parietal Cortex

STDP Spike Timing Dependent Plasticity

STM Short-term Memory

WM Working Memory

Symbols Used

In the following log will be used to denote the natural logarithm and $\log_2$ the base-two logarithm.

For compatibility with the notation in Holst (1997) the double index ($x_{ii'}$) notation for unit i' in hypercolumn i will be used in chapter 3. However, being cumbersome it will be replaced with the standard single-index notation (x_i) in subsequent chapters, where i is assumed to be an index ranging over all units.

Some symbols are used with slightly different meaning in different parts of the thesis, e.g. the synaptic weight w which is both used to denote weights in the BCPNN memory and in other memories. In the table below are listed the sections of the main definitions of the quantities. Only symbols used elsewhere than where they are defined are included.

Symbol	Meaning	Defined in section
ξ	A learning pattern, the current learning pattern	
ξ^p	Learning pattern p	
N	Number of units in the network	
H	Number of hypercolumns in the network	
V	Presentation time for each pattern	3.4.3
z	Number of training patterns	3.3.5
$c(x)$	Clipping function	3.1.1
$\pi_{ii'}$	$P(x_{ii'}\|\mathbf{x})$, the probability conditioned on known information	3.3.1
$o_{ii'}$	Indicator variable representing known information	3.3.1
τ_c	Membrane time constant	7.5.1
τ_L	Learning time constant	3.3.6
τ_A	Adaptation time constant	5.2
τ_S	Learning rate decrease time constant	7.5.1
α	Learning rate, $1/\tau_L$	3.3.6
$h_{ii'}$	Support of unit ii'	3.3.2
$\hat{\pi}_{ii'}$	Activation of unit ii'	3.3.2
$\Lambda_{ii'}$	Rate estimate of unit ii'	3.3.6
$\Lambda_{ii'jj'}$	Rate estimate of connection unit jj' to unit ii'	3.3.6
λ_0	Underflow protection, intrinsic noise rate	3.3.6
$\mu_{ii'}$	Rate estimate of adaptation for unit ii'	5.3
$\mu_{ii'jj'}$	Rate estimate of adaptation for connection unit jj' to unit ii'	5.3
$\gamma_{ii'}$	Cellular adaptation of unit ii'	5.3

$\beta_{ii'}$	Bias of unit ii'	3.3.2
$w_{ii'jj'}$	Weight between unit jj' and unit ii'. Note the difference between the usage in section 3.3.1 and section 3.3.2; when both are used the expression in 3.3.1 is denoted log-weight.	3.3.2
$v_{ii'jj'}$	Synaptic depression/facilitation between unit jj' and unit ii'	5.3
g_A	Gain of adaptation	5.3
g_L	Gain of associative projection	5.3
g_I	Gain of input projection	5.3
$\kappa(t)$	Print-now modulation of τ_L	4.2
κ_i	Modulation of isolate pattern	4.2
$H(i)$	The set of units in the same hypercolumn as unit i	
M_i	Number of units in hypercolumn i / number if possible valus of attribute i	3.3.2

Chapter 2

Memory and the Brain

2.1 Memory: A Cognitive Neuroscience View

2.1.1 Types of Memory

Memory has been of interest to philosophers long before it became a field of psychological and biological study. Aristotle developed a theory of mind where the senses receive the form of sensible objects without their matter, leaving imprints in the mental matter in the same way as a signet ring leaves imprints in wax. Memory is seen as the persistence of sense impressions, and recall is defined as an act which causes an imagination of an earlier impression to become an actual sensation. Recollection is governed by laws which regulate how an initial imagination associates with other potential imaginations based on similarity, contrast or continuity (Aristotle, 350a,b). This was in many ways the first general model of memory, not just an analogy like the wax and aviary analogies in Platon's Theaetetus (Plato, 360) where the main issue was how impressions could be stored and why erroneous recall was possible. The Aristotelian model expands on this, distinguishing between an inactive storage state and an active experience as well as adding a theory of association and of the processes accessing memory.

Despite (or perhaps because of) the elaboration of the Aristotelian memory model the simpler model of memory as a passive warehouse was to hold sway within psychology for nearly two millennia, and still remains a common metaphor in folk psychology. The role of the Aristotelian model was rather to influence empirical philosophers such as Locke outside the field of psychology.

The fundamental insight of Aristotle was that memory can be viewed as a function rather than storage; it does not consist solely of imprints of past experiences, associations between them and the underlying physiological basis but also the processes that enables it to affect the behaviour of the organism. As Eichenbaum and Cohen (2001) put it, memory can be "conceived of as a fundamental *property* of brain systems and a natural *outcome* of the brain's various processing activities,

rather than an *entity* stored in the brain".

The definition of memory I will use in the thesis is the ability to retain and reuse prior experienced information. Retrieval of an active representation is just as important as proper storage. The active representation is necessary for mental processing and the storage for transferring information across time. The processes that introduce new information into the storage and activate useful representations are necessary for memory as a function.

Phyletic Memory

Platon suggested the existence of innate memories that had no basis in previous experience. While this theory has been criticised on philosophical and psychological grounds, the basic architecture of the brain and mind is genetically predetermined and this provides an advantage in dealing with a structured environment. A completely general learning system would require far more training data to cope with an environment than one with predetermined suitable biases (Geman et al., 1992), and without prior selection of the system to the environment there is no advantage between different systems in general (Wolpert and Macready, 1995, 1997). Fuster coined the term phyletic memory for the inherited basic structure of the nervous system and default connections of the brain. In his words, it is the memory of the species, which has been accumulated through an evolutionary learning process (Fuster, 1995).

While memory in this sense fits with the working definition of memory I suggested above, it is beyond this thesis. However, the importance of a predefined context for each memory system or model cannot be overstated. Prior information relevant to the general function of the whole system can be encoded in its basic structure, and individual learning serves to improve the function beyond this by allowing adaptation to a variable environment.

A mental ability, a brain structure or a neural network in isolation does not make any sense or fulfil any purpose: it is only as part of an adaptive perception-action system it gains meaning. It is through how well a particular implementation functions in order to make this system achieve its goals that it can be judged.

Memory Systems and Their Taxonomies

The earliest theories of memory treated it as a unitary system. The main interests of study were how learning occurred and the rules underlying association. Association can be defined as linkage of information with other information, making the experience or retrieval of one piece increase the likelihood of retrieval of associated pieces.

Another early distinction was between recall and recognition. Recall is the ability to retrieve an item that was previously learned when given an appropriate cue (cued recall) or spontaneously (free recall). Recognition is the ability to successfully acknowledge that a certain item has or has not appeared in previous experience.

William James introduced a dichotomy between primary (short-term, STM) and secondary (long-term, LTM) memory, which also represents the start of analysing memory as a non-unitary system (James, 1890). He described primary memory as that which is held only for a moment in our conscious mind and secondary memory as unconscious but permanent information. Note the similarity to Aristotelian imagination and impressions.

Empirical evidence for the short-term and long-term account emerged with experiments such as the ones performed by Brown (1958) and Peterson and Peterson (1959), where even small amounts of information given to test subjects was rapidly forgotten when active rehearsal was prevented. Even more convincing was neuropsychological evidence of brain lesions causing impaired LTM but preserved STM (Scoville and Milner, 1957) and impaired STM with preserved LTM (Shallice and Warrington, 1970).

The evidence pointed to the existence of at least two memory systems, one with durable long-term storage, unlimited capacity, a slow rate of acquisition and a tendency to encode items according to meaning, and another with rapid dissipation and rapid acquisition, a limited capacity and encoding sensitive to phonetic similarity and other surface characteristics (Waughn and Norman, 1965; Baddeley, 1966b,a). Information from the first short-term system was assumed to be consolidated into a more permanent and robust form in the long-term system.

Sperling (1960) also showed that there appeared to exist a very short-term perceptual memory, an iconic sensory store. These sensory stores were sensitive to similar sounding or looking stimuli and had a very limited capacity and memory span. The Atkinson and Shiffrin (1971) multi-store memory model expanded on this into a model where input entered the short-term sensory store, where it was available for entering into short-term memory. From short-term memory information was in turn transferred to long-term memory through repetition, and could be retrieved back into working memory if needed to generate response output. However, as shown by Craik and Watkins (1973) memory encoding is not directly related to the time the information is kept in working memory. Instead the level of processing, how much information is processed and associated, might affect encoding by creating a richer and more durable memory trace (Craik and Lockhart, 1972; Baddeley, 1999).

As more studies accumulated more elaborate models of the information flow between sensory buffers, working memory and long-term memory were developed (Baddeley and Hitch, 1974; Baddeley, 2000). As in the multi-store model sensory information is first received in sensory stores (the visuo-spatial scratch pad and the phonological loop). Working memory consists of these stores and a central executive function regulating information flow and usage, which also connects to long-term memory and other cognitive functions. Later the model has been extended with an episodic buffer, a limited capacity system that temporarily binds together information from the other stores in a multimodal code, which the central executive functions can manipulate (Baddeley, 2000).

Beside the fractioning of short-term memory, long-term memory was also found

to exhibit different subcomponents. Cohen and Squire introduced the distinction between declarative (explicit) and non-declarative (implicit or procedural) memory (Cohen and Squire, 1980; Squire et al., 1993). Declarative memory was defined by conscious recollection: memory content such as facts and events can be recalled to consciousness. Non-declarative memory causes behavioural changes (such as the acquisition of skills, habituation or priming) but the memory content remains inaccessible (Zola-Morgan and Squire, 1993). As expressed by Tulving (1999) a declarative memory act results in a product that can be held in mind, while non-declarative/procedural memory acts do not. Studies in amnesic patients revealed intact learning abilities for motor skills, classical conditioning and priming despite lack of conscious recollection of the learning events (Cohen and Squire, 1980). Over time declarative memory has become increasingly defined as a brain systems construct (see section 2.1.2) rather than tied directly to consciousness.

Another division within declarative memory was suggested by Tulving (1972), between episodic memory and semantic memory. Episodic memories are memories of past experienced events. They are particular, covering a specific learning experience with strong autobiographical aspects (although the exact time of the learning experience in relation to personal history may be hard or impossible to recall). Semantic memory represents world-knowledge. Semantic memories contain meaning and relationships between objects, people, places and concepts often without a recallable source or autobiographical content. The relationship between episodic and semantic memory has been debated (Graham et al., 2000). One suggestion has been that semantic memory is the result of the merging of many episodic memories based on their commonalities rather than their individual character (McClelland, 1994; Baddeley, 1999). However, some evidence suggests that semantic knowledge can be acquired as one-shot learning and during impaired episodic memory (Nadel and Moscovitch, 1998).

The richness of types of memory in the literature suggested the need for taxonomies. Schachter and Tulving (1994) defined memory systems based on their psychological characteristics. A memory system is a system that is necessary for tasks from a large class of tasks that have the same functional features, such as working memory tasks or skill learning. Each memory system exhibits unique functional properties, and can be distinguished from another through dissociations such as different effects of lesions.

The Schachter and Tulving definition of memory systems in terms of observable psychological effects led to a memory taxonomy with five major systems and 11 "subsystems": procedural memory (motor skills, cognitive skills, simple conditioning, simple associative conditioning), perceptual representation (visual word form, auditory word form, structural description), semantic memory (spatial, relational), primary memory (visual, auditory) and episodic memory.

Squire and Zola-Morgan (1991) proposed another taxonomy, only slightly overlapping with the Schachter-Tulving taxonomy. They divided memory into declarative and procedural memory, with declarative memory divided into episodic and

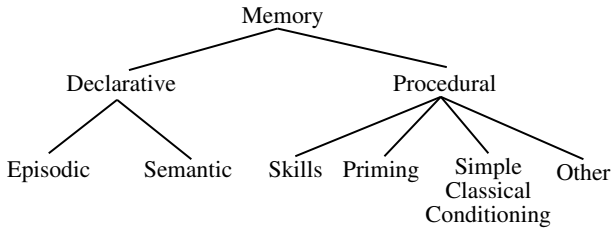


Figure 2.1. Memory taxonomy of Squire and Zola-Morgan (Squire and Zola-Morgan, 1991).

semantic memory and procedural memory divided into skills, priming, simple classical conditioning and other categories (Figure 2.1).

These two taxonomies are based on psychological data, only using brain data as confirming evidence. There is no consensus on how many memory systems exist or along what cognitive dimensions to organise them (Eichenbaum and Cohen, 2001). Psychologically identifiable memory systems are linked to biological memory systems in some way; the challenge to cognitive neuropsychology is to determine the exact relationships between the behavioural/functional macroscopic description in terms of cognition and the microscopic description in terms of synapses, neurons, networks, brain structure and brain functions.

2.1.2 Memory Systems and Brain Systems

Declarative Long-Term Memory

The most successful demonstration of a link between a psychological memory system and a brain system has been the link between declarative long-term memory and the medial temporal lobe (see Zola-Morgan and Squire (1993); Eichenbaum (2000) for reviews).

The patient known as HM is the paradigmatic case. HM underwent surgery in 1953 to treat intractable epilepsy, in which the medial temporal lobes (MTL) were removed bilaterally. The MTL encompasses the hippocampal formation, the entorhinal, parahippocampal and perirhinal cortex; in HM's case the amygdala and adjacent cortex was also removed. After surgery HM was severely impaired in learning new material as measured with recall and recognition tests (anterograde amnesia). No new episodic or semantic memories could be laid down after this event, but other forms of learning such as working memory, motor skill learning and repetition priming were normal (Scoville and Milner, 1957; Cohen and Squire, 1980). HM's case demonstrated that declarative memory could be dissociated from non-declarative memory and suggested a link to the medial temporal lobe.

Other patients with focal lesions in the MTL exhibit the same anterograde amnesia, implying that the MTL at least plays an important role in laying down

declarative long-term memory. Since patients with anterograde amnesia have functional memory for information acquired sufficiently long before the lesion, the MTL memory system is not necessary for maintaining or recalling long-term memories of this type.

Patients with MTL lesions also often exhibit retogerade amnesia, amnesia of the time period before the lesion (Ribot, 1882). This amnesia can exhibit a temporal grading, with the most remote memories less affected than more recent memories. One interpretation of this is in terms of a gradual consolidation process where experiences first exist as a MTL-dependent representation and then are consolidated into a non-MTL-dependent representation (Squire and Zola-Morgan, 1991; Zola-Morgan and Squire, 1993; Squire and Alvarez, 1995), likely in the neocortex (Fuster, 1995).

Further evidence of the role of the MTL in declarative memory have been found through functional imaging methods such as PET and fMRI studies in humans. The MTL shows increased activity when retrieving a less practised memory state compared to retrieving a well practised memory (Petersson et al., 1997, 1999). Especially notable is that a higher level of activity can be observed within the hippocampus during retrieval accompanied with a conscious recollection of the learning episode compared to merely recognising a learned item as familiar (Eldridge et al., 2000).

Animal lesion studies have largely confirmed this evidence and narrowed down the brain regions necessary for acquisition of memories in delay conditioning, place learning and contextual fear conditioning (Zola-Morgan and Squire, 1993; Quillfeldt et al., 1996; Packard and Teather, 1998). Drug infusion in specific brain areas of the rat shows that the hippocampus and amygdala are involved in memory expression for a few days after the learning occasion, entorhinal cortex for more than 31 days but less than 60 days and parietal cortex (presumably the final storage of the learned information) for more than 60 days (Quillfeldt et al., 1996). The MTL can also be dissociated from procedural learning, as demonstrated by the differential effects of selectively inactivating the hippocampus and caudate nucleus (Packard and McGaugh, 1996).

The converging evidence from lesions and functional neuroimaging points towards a set of structures in the MTL, particularly the hippocampus, entorhinal, parahippocampal and perirhinal cortex, and parts of the diencephalon as being essential for declarative long-term memory (Squire, 1992).

The MTL is a convergence area of polymodal information, well suited to form representations of a current context which can then be consolidated. It has been implicated in contextual learning of different kinds, such as classical context conditioning (Holland and Bouton, 1999; Anagnostaras et al., 1999).

The hippocampus and related systems appear to have strong ties to navigation in both rodents and humans (Maguire et al., 1998). In rats place cells fire when the rat is in a particular location in the environment (O'Keefe and Dostrovsky, 1971; Wilson and McNaughton, 1993). While it has been argued that the hippocampus is mainly about spatial maps, there also exist cells sensitive to other features, such as

olfactory cues (Wood et al., 1999) and cells in humans sensitive to facial expression feature combinations (Fried et al., 1997). Hence it appears likely that this system does represent general cognitive maps of stimulus features.

Morris (1996) argues that having a learning system with fast plasticity (such as the hippocampus) would enable it to do one trial learning of the unique and random events of everyday life, while slower learning would tend to represent more invariant or consistent features of the world. Episodic and semantic memories are abstracted from the hippocampal recording of experience.

An important hypothesis relating to encoding of memory is MTL-dependent consolidation through repeated reinstatement of the neocortical representations. According to this hypothesis the MTL stores a "snapshot" or "index" of experience through fast learning, while the neocortex is a slow learner. Over time this trace is reactivated, causing the reinstatement of the activity caused by past experience. This reactivation strengthens the neural interconnections between parts of the representations so that eventually the neocortical memory network can support declarative memory retrieval on its own without support from the MTL (Marr, 1971; Alvarez and Squire, 1994; McClelland, 1994; McClelland et al., 1995; Squire and Alvarez, 1995). Different versions of this hypothesis exists, differing in the nature of the MTL representation and how the reinstatement occurs (Squire, 1992; Bibbig et al., 1995; Murre, 1996; Robins, 1996; McClelland and Goddard, 1997). Neurophysiological evidence supports the view that this reinstatement occurs during sleep or inactivity (Wilson and McNaughton, 1994; Buzsáki and Solt, 1995).

fMRI imaging of the MTL during a famous face remote memory test was temporally graded (Haist et al., 2001), supporting the consolidation account. The hippocampus proper exhibited a mixed response that was interpreted as evidence for that it participated in consolidation for only a few years, while the entorhinal cortex exhibited temporally graded changes extending up to 20 years.

While both episodic and semantic memories are affected in classical anterograde amnesia, they can be dissociated in semantic dementia. Semantic dementia is a disorder of progressive deterioration of semantic knowledge about people, objects, facts and words that is not accompanied by declines in other cognitive skills such as visuospatial ability, nonverbal problem solving and working memory (Graham et al., 1999). Learning of nonverbal information can be relatively intact (Graham et al., 1999, 2000). Semantic dementia is typically caused by atrophy of the inferolateral temporal neocortex that has spared the hippocampal complex (Graham et al., 1999). This suggests that at least some aspect of semantic memory is closely linked to this cortical region.

Working memory

Short-term memory/working memory involves the retention or maintenance of information for short periods of time, usually linked with an ongoing behavioural task. It is similar to declarative long-term memory in that it is associated with awareness and mental representations, but it is short-lived and not affected by LTM lesions.

The region most well linked to working memory tasks is the prefrontal cortex (PFC), especially the dorsolateral PFC. Prefrontal lesions disrupt the ability to perform tasks requiring recall over brief periods of time, while sparing the ability to perform tasks where the sensory information remains available (Jacobsen, 1935; Pribram et al., 1952; Passingham, 1975). Patients with frontal and parietal lesions are impaired in working memory tasks but have normal declarative memory (Shallice and Warrington, 1970; Freedman and Oscar-Berman, 1986).

Functional neuroimaging studies also show heightened activation of dorsolateral PFC during working memory tasks as well as activation of the inferior parietal cortex, with a possible role of the cingulate cortex for motor action (Klingberg, 1997).

Single-cell electrophysiological recordings in the prefrontal cortex and cingulate gyrus have demonstrated neurons that sustain firing during the delays in working memory tasks (Fuster and Alexander, 1971; Fuster et al., 1982; Funahashi et al., 1989, 1993). The paradigmatic experiment has been the delayed response task (Goldman-Rakic, 1995; Goldman-Rakic et al., 1999). In the oculomotor version of this task (Funahashi et al., 1989), the location in visual space has to be remembered over a few seconds after which a suitable response should be generated. Neurons with sustained delay period firing are found in a circumscribed part of PFC with columnar units sharing sensitivity for the same visuospatial coordinates. If the firing of a recorded neuron is not maintained throughout the delay period the animal is likely to make an error (Funahashi et al., 1989), and the temporary inactivation of a module of cortex results in the loss of memory for particular target locations (Sawaguchi and Goldman-Rakic, 1991). The delay activity appears to encode sensory attributes of a remembered stimulus regardless of their task relevance, suggesting that they are not solely preparation for a response (Constantinidis et al., 2001).

There may be a separation of memory domains in the PFC, with different areas dedicated to different contents. In monkeys the dorsolateral area around the principal sulcus and the anterior arcuate may be important for spatial working memory, while the inferior convexity appears involved in working memory of nonspatial visual information (Goldman-Rakic et al., 1999; Undergleider et al., 1998).

Similar delay activity can be found in other cortical areas such as the parietal lobe (spatial memory tasks) or the inferotemporal lobe (object memory tasks). It has been suggested that the prefrontal and posterior sensory areas are connected by cortico-cortical reciprocal connections that exhibit reverberating activity during the delays (Fuster, 1995).

A model of working memory will be further studied in chapter 6.

Non-declarative memory

Non-declarative or procedural memory is to an extent a catch-all term for many different memory functions and should not be expected to be a unitary brain system but rather a collection of different systems (Willingham and Preuss, 1995).

The basal ganglia have been suggested to be involved in motor, habit and skill learning (Mishkin et al., 1984). Patients with Parkinson's and Huntington's disease often suffer from lapses of procedural memory, without any explicit memory deficits. However, the impairments seem to vary between different groups of patients and tasks (Vakil and Herishanu-Naaman, 1998). Inactivation of the caudate nucleus impairs response learning but not place learning which instead is subserved by the hippocampus (Packard and McGaugh, 1996).

Priming, the increased probability of retrieving a recently observed memory, may be a neocortical phenomenon. PET studies have demonstrated decreases in activation in different cortical areas during priming tasks (Yasuno et al., 2000). This has been interpreted as a facilitation of processing the second time a stimulus is shown (Squire et al., 1992).

The cerebellum has been suggested as a site of motor learning (Marr, 1969; Thach, 1996), especially for timing of motor responses and it appears to be involved in different forms of classical conditioning (Medina et al., 2002). Another brain structure implicated in conditioning is the amygdala. It appears tied to "emotional memory", particularly fear conditioning (see next section).

Modulation of Long-Term Memory

The ability to retrieve earlier experience is strongly influenced by factors at the time of encoding in addition to retrieval-time influences.

Part of this influence is from meaning-based, context and relational processing and factors like emotional significance and attentional allocation (for reviews see e.g. Buckner et al. (1999); Wagner et al. (1999)). In many cases the activity observed during encoding predicts whether subsequent retrieval attempts will be successful (Buckner et al., 1999).

Endogenous processes activated by experience can modulate memory strength in terms of recall probability (McGaugh, 2000). For example, emotionally arousing (Christianson, 1992) or humorous (Schmidt, 1994) experiences are generally better remembered than less affective experiences. Successful recognition of a task-relevant stimulus can give rise to enhanced learning of task-irrelevant stimuli, suggesting that there exists a reinforcement signal linked to perceived task-relevance that also reinforces learning of other stimuli (Seitz and Watanabe, 2003).

The novelty of a stimulus also plays an important role. The von Restorff or isolation effect consists of improved recall or recognition of an item (the isolate) that is distinct or different from the others in a set, while the other items are less well recalled (retroactive and proactive inhibition) (von Restorff, 1933). While this has mainly been studied in human list recall, a similar effect has been observed in rats (Reed and Richards, 1996) and monkeys (Parker et al., 1998).

On the neurochemical level hormones and neuromodulators can affect how strongly experiences are retained (Martinez et al., 1991). The pharmacology of memory enhancers generally demonstrate that drugs that stimulate the dopamine,

noradrenaline and acetylcholine modulator systems or act as agonists have memory enhancing effects (Ennaceur and Delacour, 1987; Hasselmo. et al., 1992; Hasselmo et al., 1996; Levin and Simon, 1998; Clark et al., 1999). Even nutrient levels can affect the strength of memory traces (Winder and Borril, 1998; Boccia et al., 1999), possibly by triggering the release of neuromodulators such as acetylcholine (Ragozzino et al., 1996).

The amygdala has been implicated in many forms of emotional memory. While originally believed to be an integral part of the MTL declarative memory system, animal lesion studies have demonstrated that it is not necessary for declarative memory (see Zola-Morgan and Squire (1993) for a review). Rather it appears to act as the site of Pavlovian fear conditioning, learning associations between stimuli and hippocampal contexts to produce fear responses (Maren and Fanselow, 1996; Anagnostaras et al., 1999; Fanselow and LeDoux, 1999; Medina et al., 2002; Moita et al., 2003) as well as modulating the consolidation of memory in other brain regions (Cahill and McGaugh, 1998; McGaugh, 2000).

These mechanisms appear linked to a small number of "print-now" signals that changes overall plasticity in a memory system in a fairly global manner, e.g. dopamine (Wickens and Kötter, 1995).

There thus appears to exist several groups of memory modulating effects. One group consists of plasticity-modulating influences, regulated by emotional and cognitive factors. Another is intrinsic properties of the brain's network structure making certain types of data more likely to be retained than others due to more efficient or orthogonal representations. Activation of larger neural networks with more extensive connections would make the traces more likely to be stimulated by a cue and more resistant to interference, similar to the levels of processing theory of Craik and Lockhart (1972). A better or multimodal representation of the information would also improve storage, such as in memory arts (Patten, 1990). Many memory enhancement techniques rely on using spatial representations, which may be stored more efficiently in the hippocampus than other representations (Nadel and Moscovitch, 1998). A closely related group consists of attentional gating of what and how deeply information is processed, which can be hard to disambiguate from direct changes in plasticity (Warburton et al., 1992).

An important issue is how to distinguish these factors from each other. We will return to the subject of models of memory modulation in chapter 4.

2.2 Memory: A Cortical View

Cognitive psychology and neuropsychology provides a top-down view of memory, where the actual behavioural effects of memory are the most easily observable phenomena and ever finer subdivisions of systems the result of a reductionistic research program. In the end the hope is to ground these systems in the actual neurobiology of the brain, a top-down deductive approach. The cortical neurophysiology perspective instead addresses the issues of memory in a bottom-up constructive approach

where neural plasticity, microcircuits, global networks and the formation of representations is seen as the physiological substrate of memory functions. The hope here is to explain higher and higher memory functions in terms of simpler neural correlates, eventually reaching the level of the constructs of cognitive psychology.

2.2.1 Cortical Circuitry and Modularity

The basic cortical structure appears to be relatively regular with local variations of the same underlying modular and layering theme (Braitenberg and Schüz, 1991; Shepherd, 1998; Arbib et al., 1998). The neurons are distributed between distinct cortical layers, where each layer receives and sends projections in a specific manner to other cortical and subcortical sites, suggesting a different functional role for each layer (Thomson and Bannister, 2003). The layers are similar across the cortex although their relative size varies. The cortex consists of ≈80% pyramidal cells, a relatively homogeneous group of excitatory neurons (with some variation between cortical layers). The remaining ≈20% inhibitory interneurons are more heterogeneous, with several subclasses of cells (Gupta et al., 2000).

Neurons in the cerebral cortex receive both thalamic afferents and extensive horizontal recurrent connections from nearby cells and more remote cortical areas. Intracortical connections are relatively dense locally and sparse on the global level (Palm, 1982; Gilbert et al., 1990), and horizontal connections provide up to 80% of synapses onto some pyramidal cells (LeVay and Gilbert, 1976). Superficial pyramidal cells connect extensively to neighbouring cells. Roughly 70% of the excitatory synapses on such pyramidal cells come from cells less than 0.3 mm away (Calvin, 1995). Connections between cortical areas are extensive and in general reciprocal (Felleman and van Essen, 1991; Scannell et al., 1995). This produces a densely interconnected system well suited to associate information both intra- and inter-modally.

Mountcastle introduced the concept of "cortical columns" as the basic functional module of the neocortex based on physiological experiments with cat somatosensory cortex (Mountcastle, 1957, 1998). Cells with similar functional properties were found to be located in narrow columns 40-50 μm across comprising some 80–100 cells. This is similar to the earlier observations of de Nó (1938); de Nó (1938) of synaptically linked chains of neurons in rodent cortex spanning the cortical layers, which he theorised was an elementary unit of neocortex which could sustain reverberating activity in the form of impulses circulating along closed loops. However, it is not yet verified that the functional columns observed by Mountcastle actually correspond to the geometrical minicolumns of the brain (Arbib et al., 1998).

According to Mountcastle (1978) the minicolumn is the basic functional module for input-output processing. It comprises a set of cells with heavy vertical interconnections and more sparse horizontal interconnections. Pericolumnar inhibition isolates neighbouring minicolumns from each other, while they are assumed to have specific long-range connections. While the general activity of cells belonging to the same column is similar, different cell types differ in their responses. The variables

represented by the minicolumns (direction, frequency, colour etc.) from different cortical areas differ, and are determined by the nature of the thalamic input and intracortical processing.

The functional columns appear to be organised into larger patterns. For instance, minicolumns are organised in broader groups, 300-500 μm hypercolumns consisting of 50–80 minicolumns with a common input and extensive connections between the pyramidal and inhibitory interneurons. Long-range intracortical projections link columns with similar functional properties. Minicolumn size does not appear to be very sensitive to brain size (Bugbee and Goldman-Rakic, 1983), but varies between cortical areas.

Columnar organisation allows an intermittently recursive mapping of two or more variables onto the cortical surface (Mountcastle, 1998). The canonical example is the coding of edge orientation in the primary visual cortex (Hubel and Wiesel, 1977), where each orientation minicolumn responds selectively to a range of orientations and spatial frequencies (Issa et al., 2000). The hypercolumn contains orientation columns covering all angles as well as two different ocular dominance columns, and thus represents the local edge orientation pertinent to a given point in visual space. A similar modular arrangement is found in blobs and in many other cortical areas, e.g. whisker barrels in rodent barrel cortex (Purves et al., 1992) and in auditory cortex (Imig and Adrian, 1977; Middlebrooks et al., 1980). Modular geometry has also been found in high-level association cortex such as the entorhinal cortex (Hevner and Wong-Riley, 1992). The exact geometry appears to be variable between brain areas, species and possibly even individually (Adams and Horton, 2003). It seems reasonable to assume that while modularity is a common feature, the self-organising processes bringing it about can produce many different physical geometries but with the same basic circuitry and functional structure.

The functional role of hypercolumns is largely unknown. In the visual cortex normalisation models propose that the total activity of all the cells within a hypercolumn is normalised, possibly by shunting inhibition via basket cells (Carandini et al., 1997).

Collateral axons often travel a characteristic lateral distance before terminating in a patchy way, with patch sizes coinciding with minicolumn sizes (Gilbert, 1993; Lund et al., 1993; Schüz, 1994; Arbib et al., 1998). Cells with the same orientation selectivity in the primary visual cortex are often connected Hirsch and Gilbert (1991). This suggests that connectivity is selective for minicolumns, i.e. "patchy". Even arborisation of afferents arriving over macroscopic distances appears to be modular Goldman-Rakic and Schwartz (1982); Arbib et al. (1998).

From scaling and volume considerations Braitenberg finds that dividing cortical circuitry into compartments of $\sqrt{N}$ neurons where each neuron in a compartment sends an axon to another compartment (and has local dense connectivity) produces compartments with the same size as Hubel and Wiesel's hypercolumns and fits data on dendritic spread and white matter volume (Braitenberg, 2001; Braitenberg and Schüz, 1991). Perelmouter proposed a hierarchy of k levels of compartments with $N^{1/k}$ compartments of the next level, producing realistic volume values for $k = 4$

(Braitenberg, 2001).

This kind of volume consideration is one reason to expect modularity in the connections of the cortex. The advantages of co-localising closely related information for normalisation and other shared operations is another. The independence and capacity considerations in chapter 3 provides a third reason.

2.2.2 Neural Plasticity

The idea that memory is based on a cellular process where the synaptic connections between neurons is changed was initially suggested by Tanzi in 1893 and independently by Cajal in 1894 (Fuster, 1995). While a common theme in most psychological and neural theories of memory, it took until 1973 before a form of synaptic change that could form a plausible basis for such memory storage was found experimentally.

Bliss and Lømo (1973) and Bliss and Gardner-Medwin (1973) discovered that high frequency tetanic stimulation of afferents to hippocampal neurons caused a subsequent steeper rise time in the excitatory postsynaptic potential, as well as recruitment of spike activity from more cells in the population. These changes in single impulse stimuli remained for hours after the original tetanic stimulation, giving the name Long Term Potentiation. Further studies have characterised LTP to a greater degree and described it in many brain areas (Baudry and Davis, 1996).

The underlying biochemical basis for LTP has been studied in great detail, see Bliss and Collingridge (1993). The proposed underlying mechanisms involve the combination of presynaptic input and postsynaptic depolarisation causing NMDA receptors to open, producing further depolarisation and increased calcium influx from voltage dependent channels as well as the NMDA receptors. The calcium influx into dendritic spines induce a cascade of kinases that activate second-messenger systems and affect the synapse, e.g. by recruiting previously inactive AMPA receptors. LTP also depends upon protein synthesis and the activation of transcriptional regulatory elements (Bailey et al., 1996; Alberini, 1999); the second messenger cascade also leads to gene expression of synaptic effector proteins (Frey and Morris, 1997; Bhalla and Iyengar, 1999). Structural reorganisation tied to consolidation may also involve proteolysis and transmembrane proteins that control cell adhesion (Lynch, 1998).

LTP develops rapidly, typically within one minute after a stimulus train arrives, and persists for several hours or more. It is also specific to the synapses activated by the stimuli, and associative in that all synapses participating in input signals during the time when the postsynaptic neuron fires are potentiated (Levy and Steward, 1979). It hence appears to fulfil the requirements of Hebbian learning (section 2.2.3) and may be a plausible neural implementation of it (Gustafsson and Wigström, 1988; Sejnowski, 1989; Bear, 1996).

The temporal resolution of LTP was originally thought to be a few milliseconds, which would limit its ability to bind together associations separated in time. However, "synaptic tagging" has been observed in the hippocampus, in which a stimulus

leaves a "tag" for a few hours. During this period the arrival of another stimulus will cause potentiation both at the synapse of that stimulus and at the synapse of the original tagged stimulus (Frey and Morris, 1997). This enables the association of delayed stimuli.

Another phenomenon that appears to be quite common is spike timing dependent plasticity (STDP), where the exact time relationship between spikes in the presynaptic and postsynaptic neuron can cause both potentiation when the presynaptic neuron fires first, and depression (LTD) when the postsynaptic neuron fires first (Levy and Steward, 1983; Bell et al., 1997; Markram et al., 1997; Bi, 2002). Long-term depression also occurs, a decrease of synaptic efficiency due to the firing of one but not both of the neurons (Christie et al., 1994). This is a saturable phenomenon reversible by LTP-inducing stimulation.

These interactions may be only the first steps in a longer and more complex neural consolidation process that stabilises long-term memory (Abel and Lattal, 2001). During this process different modulatory factors and activity can affect the synapse for at least several hours after the learning experience (Izquierdo and Medina, 1997). Long-term learning may be related to synaptogenesis and other major structural changes (Ramirez-Amaya et al., 2001).

An unusual demonstration of very fast memory consolidation was described by Gleissner et al. (1997). All activity in the left language-dominant hemisphere could be suppressed in patients given an intracarotid injection of amobarbital. Words given during the period of inactivation could not be recalled afterwards, while words given one minute before the injection could be recalled. The lack of retogerade amnesia suggests that a stable representation of the words had time to form in the minute before the inactivation.

Proving that all or some memory is based on LTP has turned out to be complex, since most interventions also interfere with other brain systems such as motor or sensory functions. Significant circumstantial evidence suggests a role for LTP in memory (Morris, 1996; Eichenbaum and Cohen, 2001). Xu et al. (1998) showed that the stimulation-induced early-phase LTP in CA1 was rapidly and persistently reversed by exploration of a new, non-stressful environment, while LTP expression was not affected by exploration of familiar environments.

2.2.3 Cell Assemblies

LTP provides a mechanism where experience can shape neural networks into functional units, cell assemblies. The cell assembly hypothesis of Hebb (1949) represents an influential theoretical attempt at answering the questions of how to represent concepts, behavioural structures and learning in neural tissue. It was introduced before much was known about the microarchitecture of cortex and synaptic plasticity. The theory has inspired much research in these areas, becoming a productive if not uncontroversial framework (Lansner et al., 2003).

Cell assemblies are formed through experience. Hebb suggested that repeated simultaneous stimulation of cells cause changes in the connections between them,

making them more likely to activate each other (Hebb's first postulate). The structural changes of learning were described as:

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

This form of learning has become known as Hebb's rule or Hebbian learning. As Hebb and others have pointed out, there are earlier references to the same idea found in works such as James (1892), McCulloch and Pitts (1943) and von Hayek (1952).

Donald Hebb introduced the concept of the cell assembly as the simplest representation of an image or idea:

> "It is proposed first that a repeated stimulation of specific receptors will lead slowly to the formation and 'assembly' of association-area cells which can act briefly as a closed system after stimulation has ceased; this prolongs the time during which the structural changes of learning can occur and constitutes the simplest instance of a representative process (image or idea)." – (Hebb, 1949, ch. 4)

This is sometimes called Hebb's second postulate.

A cell assembly would consist of a group of cells that are connected to each other by excitatory synapses. Through mutual excitation the cells of the assembly would be able to continue firing after an external stimulus had activated it, implementing an active representation through the reverberatory activity.

In this way the brain might use two separate neural mechanisms for short-term and long-term information storage, with short-term storage based on electrical activation and reverberation within cell assemblies and long-term memory based on changes and growth in the synaptic connections between assemblies. Again this dual active and passive representations correspond well with Aristotele's "imaginations" and "imprints" and William James' "primary memory" and "secondary memory". It should be noted that short-term information storage could in addition be stored in synapses through fast plasticity or adaptation.

Structures in the input to the brain corresponding to objects and categories would cause the formation of corresponding cell assemblies, and these would act as mental representations of the external objects. Connections between assemblies could themselves be assemblies, enabling a conceptual hierarchy. Sequential activation of assemblies could underlie association and sequential thought processes (Hebb's third postulate, "phase sequences"). In the end signals from assemblies could produce motor responses, but unlike many early connectionist models this was not emphasised.

The dense interconnections between neurons within an assembly gives it the ability to perform pattern completion. If only a subset of neurons are activated

(for example due to a weak or partial input) they will stimulate the remaining neurons bringing them to firing until all of the assembly is active.

In Hebb's model the assemblies overlap: cells can belong to several assemblies at the same time. Connections between cells not belonging to the same assembly are assumed to be inhibitory (this was a later addition to the theory (Milner, 1957; Hebb, 1959)). This causes competition/rivalry between the assemblies: if two overlapping assemblies are activated their mutual inhibition will make one of them decrease its activity until only one is fully active. This ensures that only one assembly at a time is active and prevents activity within one assembly from spreading to overlapping assemblies. It also disambiguates a mixed input and prevents randomly activated neurons from activating assemblies if there exists one fully active assembly.

Hebb pointed out that this hypothesis could explain phenomena observed in Gestalt psychology such as perceptual completion and rivalry (Milner, 1957; Freeman, 1991). Reactivation of a cell-assembly would also be a reconstructive process, similar to memory reconstruction where gaps are filled in by general knowledge and expectations.

The exact location of cell assemblies was not strictly specified, but Hebb speculated that they consisted of cells both in the cortex and diencephalon and possibly elsewhere. Assemblies were assumed to include few neurons compared to the whole brain neural network, leading to a sparse coding.

Hebb's theory was in many ways speculative and incomplete, but proved stimulating to research both in synaptic plasticity, the role of early experience in perceptual development, sensory deprivation and many other fields.

Many issues surrounding the neural implementation of assemblies have been hotly debated due to its underconstrained nature. One such issue is the form of neural activity within assemblies. In the simplest case it would either be a low firing rate state (inactive) or a high firing rate (active). Both states would be stable until outside influences caused a shift in activity. Milner (1957) further suggested that inhibition could produce a graded response.

Beyond rate codes, von der Malsburg (1986) suggested that activity within an assembly should be synchronised. In synchronous models assemblies would be connected not so much by a shared high activity state as a state of shared synchronised firing. Synchronous input would sum strongly on target neurons, which would ensure transmission over asynchronous input (Bressler, 1990). This coding has especially been proposed for feature binding since it avoids ambiguities of which feature belongs to which object (von der Malsburg, 1981; Eckhorn et al., 1988; Singer et al., 1997).

The activity of individual assemblies might also be relatively brief as they pass through the phase sequence. Braitenberg (1984) proposed that threshold control could act as "the pump of thoughts", causing a currently activated cell assembly to associate to the assembly most consistent or familiar with the present. This is similar to the suggestion of Palm (1990) of short pulses of fixed point activation under global threshold control.

Analysis in attractor neural networks (see below) have shown that a symmetric connection matrix produces fixed point attractors suitable for supporting cell assemblies. However, the brain may not have the exact reciprocity required in these models. This was analysed in Wickelgren (1992), where it was shown that exact symmetry is not necessary as long as the total number of connections in both direction are equal. Similar conclusions were reached by Fransén and Lansner (1998) with reciprocity between cortical minicolumns but not individual neurons. Also, where there are enough feedback connections Hebbian learning enhances the symmetry of connections between groups of co-active cells since they are both strengthened by the same amount.

This brings up the issue of the location of assemblies. Are they local networks, or are they distributed across several brain areas? Are there global assemblies encompassing the whole brain? Wickelgren (1992) assumes assemblies to be part of a local cell group. Amit (1994) similarly proposes local networks, while Hebb, Fuster as well as Fransén and Lansner assumes that assemblies extend over large cortical areas. Eichenbaum (1993) suggests local subassemblies close to the sensory/motor areas and disperse assemblies at higher areas.

The Wickelgren (1992) argument for locality is interesting in that it is based on issues of connectivity. Randomly connected networks are very unlikely to contain the "webs" of dense mutual connectivity assumed necessary for the reverberatory aspect of cell assemblies. Random networks where connection probabilities favour local connections are likely to have such webs. The Fransén-Lansner model on the other hand is based on densely interconnected minicolumns with sparse neuron-neuron connections to each other. While the average neural connectivity is sparse and mainly local, the network itself acts as a densely connected network of minicolumns where sets of minicolumns form the basic units of cell assemblies rather than individual neurons. In many ways this is a variant of a small world network (Watts and Strogatz, 1998) where a mostly local network still exhibits the same short diameter and large interconnectedness as would be expected by a completely randomly connected network. If the brain is not a random network but rather has a certain measure of structure, it is possible to maintain dense "webs" that have widely distributed parts.

Finding experimental support for cell assemblies has proven complex, since the neurons participating in an assembly may be spread out among countless non-participating neurons. Hence very many recording electrodes would be needed to have a sufficiently high probability to pick up two or more member cells at the same time (Strangman, 1996). One approach to detecting cell assemblies is to search for dynamic shifts in correlations between cells. Cells belonging to the same cell assembly should show correlations in their activity, and a cell belonging to several assemblies would show different correlation patterns depending on which assembly was active. Such changes have been observed in hippocampal cells (Sakurai, 1998).

Ensemble recordings in the hippocampus show changes after introduction to a new environment or the return to a known environment suggestive of the cell assembly account (Wilson and McNaughton, 1993). Similarly, the study of Hoffman and

McNaughton (2002) demonstrated the appearance of correlated activity between widely distributed neurons in a monkey during and after performing a task.

An old hypothesis of memory retrieval (again with roots in Aristotle (350a) and James (1890)) is that retrieval of specific sensory events reactivates the cortical regions that were active during the the encoding of the event. Wheeler et al. (2000) demonstrated that vivid remembering of visual or auditory information causes reactivation of subsets of sensory cortex that were activated during a separate but corresponding perception task. Retrieval of visual words that were paired with sounds causes the activation of auditory brain regions that were active during encoding, even when there is no need for the auditory information (Nyberg et al., 2000). It thus appears likely that holding a concept active in awareness or recalling it corresponds to reinstating at least part of the activity within the cell assembly that would be activated by its sensory occurrence. This fits well with Hebb's predictions, including the formation of multimodal assemblies by the pairing of stimuli.

The cell assembly theory is not the only bottom-up theory of cognition (e.g. . (Abeles and Gerstein, 1988; Gray et al., 1989; Abeles et al., 1990; Singer et al., 1997)) but appears to work well with the known data of synaptic plasticity, cortical modularity and the formation of integrated networks of activity across the brain. Modulatory substances may regulate the formation, reorganisation and dynamics of assemblies depending on behavioural state. This can be said to be a successful basis for a bottom-up hypothesis of how cortical memory works.

What is still lacking in this paradigm are mainly three things. One is the phyletic background regulating afferents, efferents, initial connectivity of brain structures and the nature of the modulatory systems. The gradual formation of perceptual representations during development appears to be controlled by self-organising processes that are partially learning-dependent, but also highly regulated by phyletic factors. Without this context cell assemblies cannot link perception and action in behaviourally relevant ways, and cannot be said to represent anything. This background is on the other hand amenable to traditional neurophysiological study.

The second lacuna is an understanding of the temporal succession of assemblies. This is likely linked to the issue of how sequences of experiences are recognised and behavioural sequences performed in an adaptive manner. Many different proposals have been made, but so far no consensus has emerged and no models can exhibit the full range of temporal abilities of animals.

The final challenge is the link to cognitive phenomena; despite the conceptual successes of the cortical perspective of memory we still lack an understanding of how the different observed varieties of cognitive memory systems (e.g. episodic vs. semantic memory, procedural vs. declarative memory) can be grounded in cell assemblies or other cortical structures. This is what this thesis is about.

2.3 Memory: A Neural Network View

Artificial neural network (ANN) models often bridge the gap between the cortical perspective and the cognitive perspective, since they can model networks on the scales that are currently hard or impossible to study experimentally. They provide both a qualitative language to express hypotheses about memory, a potentially quantitative bridge to experimental data and a means to make predictions and propose new experiments.

2.3.1 Computational Memory Models

The earliest neural networks such as that proposed by McCulloch and Pitts (1943) were more similar to networks of logical gates and stimulus-response psychology than to biological neural networks and memory retrieval. Rosenblatt's perceptron (Rosenblatt, 1962) introduced learning of an input-output mapping, but associative learning was not present.

The first truly associative neural network model was the Lernmatrix/Willshaw model (Steinbuch, 1961; Steinbuch and Piske, 1963; Willshaw et al., 1969; Palm, 1980). The associator mapped a sparse binary input vector to an output vector through a binary weight matrix and a threshold function similar to the one used by McCulloch and Pitts (1943). Each element of the weight matrix can be interpreted as a synapse between two neurons, and the conjunctive learning rule for setting the weight as a Hebbian learning rule. By judicious threshold-setting the capacity and robustness could be optimised even for noisy input and incomplete connectivity (Palm, 1980, 1981; Schwenker et al., 1996; Graham and Willshaw, 1997). This form of sparse binary hetero- or autoassociative memory inspired many strands of memory models. One approach is represented by the Sparse Distributed Memory model of Kanerva (1988) and its extensions.

Another field was matrix memory models (Kohonen, 1972). In these networks an activation vector is multiplied by an association matrix to produce an association, typically a linear mapping. In order to achieve memory storage the association matrix had to be suitably defined. This is a general feature of distributed memory models: the existence of an associative matrix, whose structure is molded by the learning process to enable retrieval of stored patterns and possibly other functions.

The three main approaches to constructing the matrix were pseudoinverse, gradient descent and correlation matrix methods. Of these the correlation matrix memories have been the most important in memory models. The reason is that they employ a local learning rule, where the update is based on information available to a single synapse or neuron, and incremental learning, where learning modifies the old network configuration to memorise new patterns without need to refer to earlier patterns. The pseudoinverse and gradient descent methods require non-local information processing and the pseudoinverse is non-incremental, making them unlikely candidates for cortical memory; nonlocal information transfer and storage of earlier data outside the network does not fit well with what is known or surmised about

biological systems. Correlation matrix memories create the association matrix as the sum of the outer product of input vectors. They are not in general optimal, but can be trained iteratively and are based on local learning similar to the Willshaw model. Through the introduction of nonlinearities in the response in the "Brain State in a Box" (BSB) model of Anderson et al. (1977) the matrix memory models evolved towards modern attractor neural networks.

Within cognitive psychology variants of matrix models continued to be developed (Murdock, 1982; Eich, 1982; Pike, 1984; Humphreys et al., 1989). These models usually employ abstract representations that can not easily be mapped to neurobiology, and various retrieval/decision criteria. The main thrust has been to explain observed results in cognitive experiments, linking them more with the cognitive than the cortical perspective of memory.

Another approach was represented by the continuum models of Wilson and Cowan (1973) of neural sheets, where activity was described as partial differential equations. This strand of exploration were later to become relevant to models of spatially localised states in visual cortex hypercolumns and working memory (Amari, 1977a; Ben-Yishai et al., 1995).

In parallel with the studies of matrix memories neural network models of memory developed during the 70's and early 80's (Little and Shaw, 1975; Grossberg, 1976). Marr (1971) approached biology closely, examining the computational abilities of the hippocampus and suggesting a model of hippocampal function.

The start of the second wave of connectionism was marked by the publication of the volume "Parallel models of associative memory" by Hinton and Anderson (1981) and the appearance of the Hopfield (1982) model (an asynchronous version of the Little (1974) model) which stimulated research into the statistical physics of attractor neural networks (Hertz et al., 1991). These developments gave access to a highly interdisciplinary toolkit for the study of models of neural and synaptic interactions and increased interest in constructive models of cognition that were more closely linked with cortical architecture. The developing field of memory models branched out to encompass a wide spectrum of model abstraction levels and scales of the studied memory systems (Figure 2.2).

At the cellular and synaptic level various detailed models of relevant processes can be found. These include models of synaptic biochemical networks linked to transmission and plasticity, the effects of backpropagating action potentials and STDP. There is a continuum of more reduced models of synaptic plasticity ranging from phenomenological descriptions of STDP and the BCM rule to mathematical simplifications such as the Hopfield learning rule.

Above the synaptic models we find elaborate single cell compartmental models like the Purkinje cell model of De Schutter (1994) and models of the interactions of small networks of such neurons. As larger and larger networks are studied, the models by necessity become more simplified and abstract due to computational demands. The largest biophysical simulations of spinal, neocortical or hippocampal networks involve tens of thousands of neurons with few compartments and millions of synapses (Hammarlund and Ekeberg, 1998; Kozlov et al., 2003).

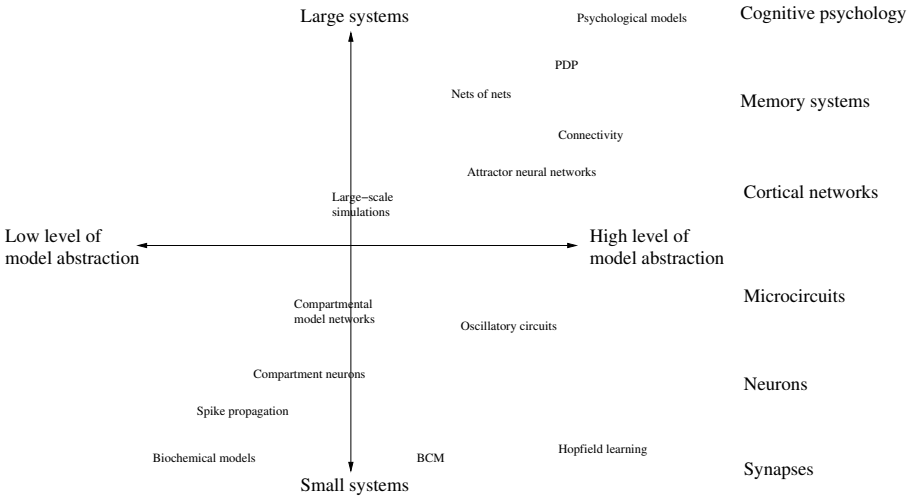


Figure 2.2. Rough overview of memory models, plotted in terms of model abstraction (from attempts of biological emulation to mathematical abstraction) and the scale of the studied systems.

On the abstract side there is a flourishing field of study of microcircuits and larger networks, mostly dominated by attractor neural networks used as generic models of memory or as models of particular brain systems, such as the visual cortex, area MT, IT or CA3. The level of abstraction varies from spiking neuron models to simple rate coding elements. Studies of connectivity constraints can also be found here, where graph theory is used to analyse the representational abilities or capacities of networks. At largest scales networks of networks are studied, where the interactions between simpler participating networks produce memory-related phenomena.

The Parallel Distributed Processing framework can be seen as the manifest of more abstract connectionist models (Rumelhart et al., 1986) that still retain some architectural link with neural systems but have abstracted away much detail of neurons and synapses. Beyond these in abstraction are the psychological matrix models (see above) and other cognitive memory models that do not map their dynamics onto any neurophysiological correlates but rather aim at a phenomenological understanding of cognitive functionality.

While one goal of computational memory modelling can be said to be to cover the entire map, different regions provide different forms of knowledge. Even fairly small biologically realistic models show very complex dynamics depending on little understood or uncertain biological parameters, making generalisation hard. While realistic modelling of small subsystems can constrain the parameters and suggest functions, there is usually a need for feedback from reduced models and data on

larger scales to formulate functional hypotheses. More abstract and conceptual models help support this process, by concentrating on the functional issues and their internal constraints. By formulating more abstract high-level models functional ideas can be tested, and if found worthwhile compared to increasingly detailed models. These abstract models are also easier to grasp conceptually than detailed ones, thus promoting communication between different disciplines.

2.3.2 Attractor Neural Networks

Auto-associative attractor neural networks, from the early binary associative memories and the Hopfield net to networks of spiking neurons, have been proposed as models for biological associative memory (Willshaw et al., 1969; Hopfield, 1982; Amit, 1989, 1994). They can be regarded as formalisations of Hebb's original ideas of how synaptic plasticity could produce the emergence of cell assemblies (Hebb, 1949). A number of psychological memory and Gestalt perception phenomena have been modelled based on such networks, as well as various models of psychiatric disorders (Freeman, 1991; Quinlan, 1991; Ruppin and Yeshurun, 1991; Ruppin, 1995).

The Little-Hopfield network (Hopfield, 1982) has in many ways become the Drosophila melanogaster of attractor neural networks, a simple model that can easily be simulated, analysed, modified and extended. The memory is trained with a set of patterns $\{\boldsymbol{\xi}\}$ where each pattern $\boldsymbol{\xi}^p$ is a vector with bipolar components $\xi_i^p \in \{-1, 1\}$. These are added together using an outer product learning rule $w_{ij} = \sum_p \xi_i^p \xi_j^p$ into a weight matrix (the "mnemonic" equation in the terminology of Caianiello (1961)). An activity state $\boldsymbol{x}$ is updated by the rule $x_i \leftarrow \text{sign}(\sum_j w_{ij} x_j)$ either asynchronously or synchronously (the "neuronic" equation). The basic framework can be extended to continuous time and activation (Hopfield, 1984), binary rather than bipolar patterns, sparse activity and more elaborate learning rules (see section 3.1.1).

A fruitful concept related to the physical interpretation of the Hopfield model is the energy function. The energy function (or Lyapunov function) is defined as a state function that is bounded below and always decreases as the network state evolves according to the neuronic equation (Hopfield, 1982; Cohen and Grossberg, 1983). While mainly technically useful for proving the stability and statistical properties of states, it has also helped conceptualise differences between different memory states (see below).

Attractor neural networks have in general held a similar interpretation of the neural activity state as "primary memory" or working memory, and the weight matrix as long-term memory (Amit, 1994). By uncoupling the neuronic from the mnemonic update equations ("the adiabatic learning hypothesis") analysis can be simplified (Caianiello, 1961, 1989). However, recent models are exploring non-adiabatic issues where the separation in time scales between learning and neural activity has been blurred, such as synaptic adaptation models (Tsodyks et al., 1998) or the working memory model of chapter 6.

Each stored memory in an attractor network corresponds to an attractor state of activity which usually is identical or close to the original activity imposed by the learning experience. This would correspond to an active cell assembly. Activity states similar to the attractor state will converge towards it, producing pattern completion and noise removal. Cued recall in an attractor model can be defined as successful when an initial cue state evolves into a state in the neighbourhood of the stored memory nearest to the cue (Ruppin and Yeshurun, 1991).

The volume of the state space that is attracted to a certain attractor state is called its basin of attraction. In many attractor network models an important issue is the presence of spurious states, i.e. attractor states that do not correspond to stored memories (Amit et al., 1985; Amit, 1989). Such states emerge due to overlaps between stored patterns, and since the number of potential spurious states increase fast with the number of stored patterns they decrease the size of the basins of attraction of the desired memory states and eventually overwhelm them (see section 3.1.1). The probability of ending up in a memory state from a random initial state will decrease as the volume of state space occupied by basins of attraction around spurious states increases.

Recognition can be implemented in an attractor network as the detection of whether the rate of change of activity of the network units becomes lower than some threshold within a certain time of the presentation (Hopfield, 1982; Ruppin and Yeshurun, 1991). Non-memory stable states in general have higher energy and smaller basins of attraction than memory states (Amit et al., 1985; Amit, 1989) and convergence takes longer time (Ruppin and Yeshurun, 1991; Tanaka and Yamada, 1993). The energy of states can hence be seen as a possible measure of memory trace strength.

It should be noted that there exists a great deal of terminological confusion about the meaning of incremental learning (Sarle, 2002). In the following I will use it in the same sense as Storkey (1999) to mean on-line learning: each input pattern is discarded after it has been processed and the weights been updated. This is somewhat different from the suggestion of (Sarle, 2002), where incremental learning denotes updating weights based on one pattern at a time but includes the possibility of multiple repetitions of the pattern set as part of iterative convergence to a final weight matrix (as in the backpropagation rule).

2.3.3 Biologically Inspired Attractor Memory Models

The recurrent architecture of attractor networks corresponds to the large-scale as well as the local patterns of connectivity of the cerebral cortex discussed in section 2.2.1. There one finds a local and medium-range horizontal network as well as long-range feed-forward, feed-back and lateral connectivity between cortical areas (Gilbert and Wiesel, 1989; Shepherd, 1998). The former can be seen as connections within the same network or within a module of the network, the latter as connections between larger, less strongly coupled networks.

One of the earliest applications of neural networks to memory was attempts to test the Hebbian theory by simulating cell assemblies (see (Fransén, 1996, p. 14–15) for a brief review). The earliest attempts of Rochester et al. (1956) using simple spiking neurons with modifiable synapses did not produce assemblies, while continuous output functions enabled assembly formation. Later models moved towards more biologically plausible cell models (MacGregor and Palasek, 1974; MacGregor and McMullen, 1978), where issues such as spike synchronisation and the problem of achieving satisfactory after-activity were studied. Later simulations have demonstrated that after-activity could be achieved by using pyramidal cells rather than motor neurons (Lansner, 1982; Fransén and Lansner, 1990; Lansner and Fransén, 1992).

Networks of cortical pyramidal and basket cells can perform well as an attractor network (Hasselmo. et al., 1992; Haberly and Bower, 1989; Amit and Brunel, 1995; Fransén and Lansner, 1998). Specifically relevant for this thesis, a network using the counting Bayesian learning rule (Lansner and Ekeberg, 1989) (discussed in next chapter) operates as an attractor neural network when implemented with biologically detailed compartmental model neurons with cortical minicolumns as its functional units (Fransén and Lansner, 1998). Moreover it has dense local connectivity that with sparse long-range connectivity, and when scaling up this network model to a patch of 8x8 mm cortex one obtains cell counts and connectivity patterns (including EPSP and IPSP amplitudes) compatible with experimental data on cortical microarchitecture.

One problem for the early assembly models was the saturation of the activity due to self-excitation of an assembly/attractor state, leading to unphysiological firing rates Amit and Tsodyks (1991a,b). By local inhibition (Amit and Brunel, 1997b) or synaptic saturation and slow NMDA kinetics (Fransén and Lansner, 1995; Tegnér et al., 2002) this can be ameliorated.

Neuromodulation is another factor that has been extensively studied, especially in terms of signal-to-noise ratio and plasticity gating (Hasselmo. et al., 1992; Hasselmo et al., 1996, 1997; Brunel and Wang, 2001), as well as the regulation of network after-activity (Lansner and Fransén, 1992; Fransén and Lansner, 1995).

Attractor networks have also been extended beyond discrete attractors to handle parametric storage of stimuli (Seung et al., 2000). An important family consists of ring attractors where the activity is a spatially localised bump (Amari, 1977a), originally proposed as a model of the visual hypercolumn (Ben-Yishai et al., 1995; Hansel and Sompolinsky, 1996) but also used as a model of both head orientation (Skaggs et al., 1995; Redish et al., 1996; Zhang, 1996) and working memory (Wilson and Cowan, 1973; Amit and Brunel, 1997b; Camperi and Wang, 1998).

Attractor working memory models have explored the interplay between neural parameters and network performance in simulations of the delayed oculomotor response task (Camperi and Wang, 1998; Durstewitz et al., 2000; Compte et al., 2000; Wang, 2001; Laing and Chow, 2001; Tegnér et al., 2002), which will be further studied in chapter 6.

2.3.4 Memory Models and this Thesis

The perspectives represented in this necessarily brief and incomplete review serve as a backdrop to the models of this thesis.

In terms of memory systems and cognitive functions I will in the following mainly focus on declarative memory in the form of semantic and episodic long-term memory and working memory. Models of non-declarative memory will be discussed briefly in the final chapter.

This thesis is focused on attractor neural network memories with point-like attractor states. Real memories certainly involve temporal sequence learning, a far more complex issue. It is likely that the general dynamical rules applicable to attractor memories with quasi-stable attractors are applicable to memories where the stored states represent or are themselves temporal sequences.

The models used here are rate coded and the activity is assumed to be sparse and modular. Spiking models are assumed to have similar basic dynamics with additional complexities such as spike synchronisation, which may add further functionality. As perceptual information is processed in the primary and secondary sensory areas, it is decorrelated and becomes sparsely encoded. This means that the input to the models will not in general correspond to a primary sensory input, but rather to the distributed representations of association cortex. The models also mirror ideas of cortical modularity, with units corresponding to minicolumns rather than neurons and groups of units bound together into hypercolumns of normalised total activity containing a full set of the different features represented by the included minicolumns.

The learning rule is a functional model of synaptic plasticity, treating it as a statistical measure of correlation similar to the ideas of Hebbian learning. Neural and synaptic adaptation processes are modelled in an equally abstract form.

The complex details of synaptic consolidation (the transformation of plastic changes from an early and unstable form to a late robust form) will not be modelled. The assumption is that in a network where individual connections correspond to groups of synapses connecting e.g. cortical minicolumns rather than to individual synapses between cells such details will be averaged out.

Chapter 3

Bayesian Confidence Propagation Neural Networks (BCPNN)

3.1 Introduction

In contrast to the arranged learning situation often faced by artificial neural networks, real world learning presents a next to unlimited number of learning examples, potentially surpassing the storage capacity of the learner by orders of magnitude. An additional complication is that the world may be non-stationary, partly due to the changing behaviour of the system and its interaction with the environment. In general, it is essential for a capacity limited real world learning system to give priority to the retention of recent information of a kind that is relevant to its operation. Several different time-scales may be involved, as in for example short-term and long-term memory. These are fundamental constraints for existing biological learning and memory systems and are also critical for advanced artificial learning systems.

This chapter will deal with the derivation and properties of the recurrent incremental Bayesian Confidence Propagation Neural Network (BCPNN), comparing it with other neural networks and memory models. Auto-associative attractor ANNs are an important class of learning systems since they both formalise the ideas of Hebbian cell assemblies and act as neurophysiological models on different levels of abstraction. Hence issues of their learning capacity, representational capability and dynamics have direct bearing on brain theory.

Especially the issue of catastrophic forgetting (CF) appears relevant, since it is not observed in biological memories but standard correlation based learning rules for attractor ANNs suffer from CF. As more and more patterns are presented to the network the ability to recall all the patterns will start to decline. If enough patterns are presented the network will become unable to retrieve any of the stored patterns, often in an abrupt manner (Amit et al., 1985; Amit, 1989; French, 1999). The reason

is that the total storage capacity is finite (proportional to number of synapses, (Amari, 1977b)), and beyond a certain point new information has to overwrite old information. But the most common learning rules such as the Hopfield learning rule or the summing BCPNN learning rule (section 3.3.5) does not distinguish between new and old information, causing all memories to interfere with each other equally. As the interference level rises, all patterns will be disrupted at nearly the same time. CF is a variant of the stability-plasticity dilemma, a general problem in learning systems: how to design a system that is both sensitive to new input but not disrupted by it (Grossberg, 1982, 1987).

3.1.1 Palimpsest Memories

To cope with the problem of catastrophic forgetting Nadal, Toulouse and Changeaux (Nadal et al., 1986) proposed a so called "marginalist" learning paradigm where the acquisition intensity is tuned to the present level of crosstalk "noise" from other patterns. The learning rule is equivalent to an exponential increase in the weight changes $w_{ij}(t) = w_{ij}(t-1) + Ae^{Bt}\xi_i^t\xi_j^t$ for the Hopfield network. This makes the most recently learned pattern the most stable; new patterns are stored on top of older ones, which are gradually overwritten and become inaccessible, a so-called "palimpsest memory"[1]. The network will remember well the p most recent patterns, where p is the palimpsest capacity (Storkey, 1999).

Another smoothly forgetting learning scheme is "learning within bounds" (originally suggested by Hopfield (Hopfield, 1982) and discussed in Nadal et al. (1986); Parisi (1986)), where the connection weights w_{ij} are bounded $-A \leq w_{ij} \leq A$. The learning rule for training patterns ξ^t is

$$w_{ij}(t) = c\left(w_{ij}(t-1) + \frac{\xi_i^t\xi_j^t}{\sqrt{N}}\right)$$

where N is the network size and c is a clipping function

$$c(x) = \begin{cases} -A & x < -A \\ x & |x| < A \\ A & A < x \end{cases} \tag{3.1}$$

For high values of A CF occurs, for low values the network remembers only the last pattern. The optimal capacity $0.05N$ is reached for $A \approx 0.4$ (Parisi, 1986; Bonnaz, 1997). This implies a decrease in storage capacity from $0.137N$ of the standard Hebbian rule; total capacity has been sacrificed for long term stability. It can be shown that all learning rules of the Hopfield network in the general form

$$w_{ij}(t) = \sum_{k=0}^{\infty} \kappa(k)\xi_i^{t-k}\xi_j^{t-k}$$

[1] Palimpsest: writing material (such as parchment) used one or more times after earlier writing has been erased. From Greek palimpsestos, "scraped again".

with a time dependent learning rate $\kappa(k)$ satisfying $\int_0^\infty \kappa^2(u)du = 1$ have less capacity than the original Hopfield rule (Mezard et al., 1986).

More recently a higher capacity learning rule has been introduced that exhibits palimpsest properties (Storkey and Valabregue, 1999; Storkey, 1999). It is based on exploiting local information in order to make a first-order approximation to the pseudo-inverse that is local and incremental (Storkey and Valabregue, 1997):

$$w_{ij}(t) = w_{ij}(t-1) + \frac{1}{N}[\xi_i^t \xi_j^t - \xi_i^t h_j(t) - h_i(t)\xi_j^t]$$

where

$$h_i(t) = \sum_{k=1}^{N} w_{ik}(t-1)\xi_k^t$$

is the local fields. This rule exhibits a capacity greater than $0.25N$, greater than the other palimpsest rules.

It should be noted that forgetting within palimpsest memories is an active process due to the arrival of new information, rather than the decay of earlier information (Nadal et al., 1986). However, the distinction is not always clear. For example, the exponential strengthening of acquisition intensity in the marginalist learning rule can be re-interpreted as a constant acquisition intensity in a network where weights decline at a similar rate $w_{ij}(t) = \lambda[w_{ij}(t-1) + (\epsilon/N)\xi_i^t\xi_j^t]$ where $\lambda = (1+\epsilon^2/N)^{-1/2}$ and ϵ is a constant (Mezard et al., 1986). Similarly, a scheme with weights decaying over time where the rate of decay depends on the presence of new information so that no decay occurs in the absence of new learning would exhibit palimpsest properties.

One group of approaches to the stability-plasticity dilemma involve networks that are extended by new neurons when needed (Grossberg, 1976; Carpenter and Grossberg, 1987; Hamker, 2001). This approach avoids CF by increasing capacity rather than removing old memories. While originally highly speculative from a biological point of view, the recent findings that neurogenesis does occur in the adult brain (Eriksson et al., 1998; Gage, 2002) and even have an influence on memory (Shors et al., 2001) does make this approach less biologically implausible. However, the timescale of neurogenesis (days) is far longer than most learning timescales, suggesting that introduction of new neurons might at most be important for long-term memory and consolidation.

Unlearning theories approach the dilemma by adding active mechanisms of removing spurious states and unwanted information. By allowing the network to converge from random initial states to attractor states (which are likely spurious in a heavily loaded network) and then unlearning them using negative Hebbian learning $w_{ij}(t+1) = w_{ij}(t) - \eta x_i x_j$ spurious states can be removed and the capacity increased (van Hemmen, 1997). This has been a suggested reason for dream sleep (Crick and Mitchison, 1983). The net effect is palimpsest-like, as old memories are occasionally unlearned and gradually weaken. There are conflicting reports on the

capacity; Christos (1996) claims a capacity of $0.05N$ while van Hemmen (1997) claims $\approx 0.5N$.

Geszti and Pázmándi (1987) points out that the energy of learned attractor states in a palimpsest memory decrease significantly as they grow older. This is accompanied with a fast decrease of the probability of convergence to the states from random initial conditions. Their suggestion is that dream sleep consists of random activation and relearning of patterns, which would favour recent, strongly learned patterns and eliminate weak incidental memories. This would act as a filter for the information to be transferred from medium-term memory to long-term memory.

These unlearning theories relate to the adaptation model in chapter 5 and the experiments with free recall in this and the following chapter.

3.2 BCPNN

A neural network architecture and learning rule derived from Bayes' rule (Bayes, 1958), the Bayesian Confidence Propagation Neural Network (BCPNN), has previously been developed (Kononenko, 1989; Lansner and Ekeberg, 1987, 1989; Lansner and Holst, 1996; Holst, 1997). It employs a Hebbian learning rule that reinforces connections between simultaneously active units and weakens or makes connections inhibitory between anti-correlated units. This learning rule is based on a probabilistic view of learning and retrieval, with input and output unit activities representing confidence of feature detection and posterior probabilities of outcomes, respectively. The connection strengths are based on the probabilities of the units firing together, estimated by counting co-occurrences in the training data.

Originally it was applied as a feed-forward network, used for classification tasks (Holst, 1997) and data mining (Orre, 1998). When applied to a recurrent attractor network this learning rule gives a symmetric weight matrix, allowing for fixed point attractor dynamics. It also generates a proper balance between excitation and inhibition, avoiding the need for external means of threshold regulation. The update of the weights in the network resembles what has been proposed as rules for biological synaptic plasticity (Levy and Desmond, 1985; Wahlgren and Lansner, 2001).

It should be noted that the type of Bayesian neural network studied here differs from e.g. those proposed by MacKay (MacKay, 1995) and Sommer and Dayan (Sommer and Dayan, 1998). In MacKay's approach the network is seen as a model of the data and the learning dynamics of the network as an inference of the most likely parameters for approximating the data. Sommer and Dayan study how a Bayesian treatment of noisy initial patterns and weight matrices naturally leads to an iterative retrieval strategy.

The most important difference between our approach and both of the above mentioned ones is that in BCPNN unit activations have a direct interpretation as confidence estimates of attribute values rather than intermediate results in a function approximations or ordinary network states. The way in which we use learning

to estimate the parameters of our model is also quite different from MacKay's way of doing this, as our weights are directly expressed in terms of probability estimates based on training data. BCPNN is more closely related to methods of Bayesian abduction used in AI (Charniak and McDermott, 1985) and the Bayesian neural network model proposed by Kononenko (Kononenko, 1989).

In this chapter, we demonstrate that by estimating the probabilities underlying network biases and weights by moving averages instead of counters as in the previous versions, it is possible to derive a continuous, real-time Bayesian learning rule with the properties of a palimpsest memory. The forgetfulness of the network can conveniently be regulated by the time constant of the moving averages. We evaluate this learning rule in the context of long-term and short-term memory properties of an attractor network with some biologically plausible elements. The consequences of a time dependent energy landscape in terms of convergence speed and plasticity modulation are also investigated.

3.3 Heuristic Derivation of Network Architecture and Learning Rule

The Bayesian Confidence Propagation Neural Network is based on an update rule heuristically derived from Bayes rule and the naive Bayesian classifier (NBC) (Good, 1950).

3.3.1 Naive Bayesian Classifier BCPNN

We start with the NBC, calculating the probabilities of the attributes y_j given a set $\mathbf{x}$ of observed occurences of attributes x_i. Both are assumed to be discrete, and the x_i are assumed to be independent ($P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i)$) and conditionally independent given y_j ($P(x_1, \ldots, x_n|y_j) = \prod_{i=1}^{n} P(x_i|y_j)$). These independence assumptions are used for the derivation but will weakened in the network.

Using Bayes' rule we get

$$\pi_j = P(y_j|\mathbf{x}) = P(y_j) \prod_{i=1}^{n} \frac{P(x_i|y_j)}{P(x_i)} = P(y_j) \prod_{i=1}^{n} \frac{P(x_i, y_j)}{P(y_j)P(x_i)}$$

This can be extended to the case where some information is not known. Suppose we are given completely known observations x_i when $i \in K \subseteq \{1, \ldots, n\}$ and have no information about the attributes x_k when $k \in \{1, \ldots, n\} - K$, then for a NBC

$$\pi_j = P(y_j|x_i, i \in K) = P(y_j) \prod_{i \in K} \frac{P(y_j, x_i)}{P(y_j)P(x_i)} \tag{3.2}$$

This can be put in the form of a sum of logarithms

$$\log \pi_j = \log P(y_j) + \sum_{i \in K} \log \left[\frac{P(y_j, x_i)}{P(y_j)P(x_i)} \right] = \log P(y_j) + \sum_{i}^{n} o_i \log \left[\frac{P(y_j, x_i)}{P(y_j)P(x_i)} \right] \tag{3.3}$$

where the indicator variable $o_i = \begin{cases} 1 & i \in K \\ 0 & i \notin K \end{cases} = I_K(i)$ represents whether there is information about a feature i or not.

This can be implemented as a single layer feedforward neural network, with input layer activations o_i, weights $w_{ij} = \log \left[\frac{P(y_j, x_i)}{P(y_j)P(x_i)} \right]$ and biases $\beta_j = \log P(y_j)$. In this way the single layer feed-forward neural network calculates posterior probabilities π_j given the input attributes using an exponential transfer function.

3.3.2 Discrete Valued Attribute Network

If discrete attribute values are represented using indicator variables, a modular structure of the BCPNN follows. Continuous valued attributes can be interval coded, a coding principle abundant in the nervous system. One classical example is the coding of edge orientation in the primary visual cortex (Hubel and Wiesel, 1977), where each orientation minicolumn responds selectively to an interval of orientations. The orientation hypercolumn contains orientation columns covering all angles, and thus represents the local edge orientation pertinent to a given point in visual space. By analogy with this possibly generic cortical structure we have referred to our network model as having a hypercolumnar structure. It should be noted that one BCPNN unit maps naturally to a minicolumn rather than to an individual neuron.

Suppose that each attribute i can take M_i different values, and that we treat the observation of a given value of a given attribute as a new binary attribute marked with double indices, the first indicating the attribute and the second the particular value. Making the necessary labelings in formula 3.2 we get

$$\pi_{jj'} = P(y_{jj'}) \prod_{i \in K} \frac{P(y_{jj'}, x_{ik})}{P(y_{jj'})P(x_{ik})}$$

where for each attribute $i \in \{1, \ldots, n\}$ a unique value x_{ik} is known, where $k \in \{1, \ldots, M_i\}$. Similarly it follows that

$$\pi_{jj'} = P(y_{jj'}) \prod_{i=1}^{n} \sum_{i'=1}^{M_i} \frac{P(y_{jj'}, x_{ii'})}{P(y_{jj'})P(x_{ii'})} o_{ii'}$$

with indicators $o_{ii'} = 1$ if $i' = k$ and zero otherwise.

If we consider the attributes X_i as stochastic variables with values $\{x_{i1}, \ldots, x_{iM_i}\}$, which are explicitly represented in the network, we may view $o_{X_i}(x_{ii'}) := o_{ii'}$ as a degenerate probability $o_{X_i}(x_{ii'}) = \delta_{x_{ik}}(x_{ii'})$ which is zero for all $x_{ii'}$ except for the known value x_{ik}. If we now generalise and replace o_{X_i} with a general probability P_{X_i} we get

$$\hat{\pi}_{jj'} = P(y_{jj'}) \prod_{i=1}^{n} \sum_{i'=1}^{M_i} \frac{P(y_{jj'}, x_{ii'})}{P(y_{jj'})P(x_{ii'})} P_{X_i}(x_{ii'})$$

$$\log(\hat{\pi}_{jj'}) = \log P(y_{jj'}) + \sum_{i=1}^{n} \log \left[\sum_{i'=1}^{M_i} \frac{P(y_{jj'}, x_{ii'})}{P(y_{jj'})P(x_{ii'})} P_{X_i}(x_{ii'}) \right]$$

If the outcomes $x_{ii'}$ of different attributes are independent of each other when conditioned on X_i, $\hat{\pi}_{jj'}$ will be the expectation of $\pi_{jj'}$ given the input X_i.

The corresponding network now has a modular structure. The units ii' in the network, where $i' \in \{1, \ldots, M_i\}$, explicitly representing the values $x_{ii'}$ of X_i may be viewed as a *h*ypercolumn as discussed above. By definition the units of a hypercolumn i have a normalised total activity $\sum_{i'=1}^{M_i} P_{X_i}(x_{ii'}) = 1$.

These procedures estimate the probability of $y_{jj'}$ given uncertain information related to the $x_{ii'}$. In this case the uncertainty of the attribute is reflected in the probability $P_{X_i}(x_{ii'})$ which is the input to the network.

Transforming these equations to the network setting yields

$$h_{jj'} = \beta_{jj'} + \sum_{i}^{N} \log \left(\sum_{i'}^{M_i} w_{ii'jj'} P_{X_i}(x_{ii'}) \right) \tag{3.4}$$

where $h_{jj'}$ is the *s*upport of unit jj'. We make the identifications:

$$\beta_{jj'} = \log(P(y_{jj'})) \tag{3.5}$$

$$w_{ii'jj'} = \frac{P(x_{ii'}, y_{jj'})}{P(x_{ii'})P(y_{jj'})} \tag{3.6}$$

where $\beta_{jj'}$ is the bias term and $w_{ii'jj'}$ is the weight. $\hat{\pi}_{jj'} = f(h_{jj'}) = e^{h_{jj'}}$ can be identified as the output of unit jj', representing the *c*onfidence (heuristic or approximate probability) that attribute j has value j' given the current context.

Since the independence assumption is often only approximately fulfilled and we deal with approximations of probabilities, it is motivated to normalise the output within each hypercolumn:

$$\hat{\pi}_{jj'} = f(h_{jj'}) = \frac{e^{h_{jj'}}}{\sum_{j'} e^{h_{jj'}}} \tag{3.7}$$

3.3.3 Recurrent BCPNN

Now, since both the input and the output of the network represent probabilities this opens up for the possibility of feeding the output back into the network as input creating a fully recurrent network architecture, which can work as an autoassociative memory. The currently observed probability $P_{X_i}(x_{ii'})$ is used as an initial approximation of the true probability of $X_{ii'}$, and used to calculate a posterior probability as a better approximation. This is then fed back and the process is iterated until a stable state is reached. This represents a heuristically estimated probability closest to the observed data and consistent with the already acquired knowledge, that is prior information represented by the learning parameters $\beta_{jj'}$ and $w_{ii'jj'}$. It can be regarded as a decision on the likeliest state of the world.

It should be noted that if the independence assumptions used in the derivation hold exactly the weights will become $w_{ii'jj'} = 1$, corresponding to unconnected units.

In the recurrent setting activations in the network can be updated either discretely or continuously (observe that the $y_{jj'}$ are now incorporated among the $x_{ii'}$). In the discrete case, $\hat{\pi}_{jj'}(t+1)$ is calculated from $\hat{\pi}_{ii'}(t)$, or equivalently, the $h_{jj'}(t+1)$ from $h_{ii'}(t)$ using one iteration of the update rule:

$$h_{jj'}(t+1) = \beta_{jj'} + \sum_{i}^{N} \log \left(\sum_{i'}^{M_i} w_{ii'jj'} f(h_{ii'}(t)) \right)$$

In the continuous case $h_{jj'}(t)$ is updated according to a differential equation, making the approach towards an attractor state continuous.

$$\tau_c \frac{dh_{jj'}(t)}{dt} = \beta_{jj'} + \sum_{i}^{N} \log \left(\sum_{i'}^{M_i} w_{ii'jj'} f(h_{ii'}(t)) \right) - h_{jj'}(t) \tag{3.8}$$

where τ_c is the "membrane time constant" of each unit.

The network is typically used with a training mode where the weights are set and a retrieval mode where inferences are made. Input to the network is introduced by clamping the activation of the relevant units (representing known events or attributes). As the network is updated the activation spreads, creating a posteriori beliefs of other attribute values.

The presence of an energy or Lyapunov function in Hopfield networks guarantees convergence to a fixed point attractor (Hopfield, 1982; Cohen and Grossberg, 1983). It is easy to show that for a recurrent BCPNN without hypercolumns there exists an energy function

$$E = -\frac{1}{2} \sum_{i,j} w_{ij} \hat{\pi}_i \hat{\pi}_j + \sum_i \beta_i \hat{\pi}_i \tag{3.9}$$

that always decreases over time since the weight matrix is symmetric and hence the iteration converges towards a stable attractor state. Unfortunately the form of 3.8

is not amenable to apply the Cohen-Grossberg theorem. Nevertheless the dynamics of the hypercolumn network does not appear to be fundamentally different and does not converge to any limit cycles or strange attractors.

3.3.4 The Prior State

If the activities correspond to the prior probabilities of features, $\hat{\pi}_{ii'} = P(x_{ii'})$ and $h_{ii'} = \log(P(x_{ii'}))$, the network state is in a fixed point:

$$\tau_c \frac{dh_{jj'}(t)}{dt} = \log(P(x_{jj'})) + \sum_i^N \log\left(\sum_{i'}^{M_i} \frac{P(x_{ii'}, x_{jj'})}{P(x_{ii'})P(x_{jj'})} P(x_{ii'})\right) - \log(P(x_{jj'})) \tag{3.10}$$

$$= \sum_i^N \log\left(\sum_{i'}^{M_i} \frac{P(x_{ii'}, x_{jj'})}{P(x_{jj'})}\right) \tag{3.11}$$

$$= \sum_i^N \log\left(\frac{1}{P(x_{jj'})} \sum_{i'}^{M_i} P(x_{ii'}, x_{jj'})\right) \tag{3.12}$$

$$= \sum_i^N \left[-\log(P(x_{jj'})) + \log\left(\sum_{i'}^{M_i} P(x_{ii'}, x_{jj'})\right)\right] \tag{3.13}$$

But the sum of $P(x_{ii'}, x_{jj'})$ across hypercolumn i is $P(x_{jj'})$ since the hypercolumn covers all possible combinations. This reduces the above to

$$\tau_c \frac{dh_{jj'}(t)}{dt} = \sum_i^N [-\log(P(x_{jj'})) + \log(P(x_{jj'}))] = 0 \tag{3.14}$$

and we see that the network is in a fixed point.

The Jacobian matrix of the network update equations in this point is

$$J_{ii'jj'} = -\delta_{ii'jj'} + \frac{P(x_{ii'}, x_{jj'})}{P(x_{ii'})} \tag{3.15}$$

This matrix has at least one eigenvalue with positive real part since not all determinants associated with all upper-left submatrices are negative (the upper 2×2 determinant is either positive or zero since $P(x_{ii'}, x_{jj'}) < P(x_{ii'})$), which is necessary for negative definiteness. Hence this fixed point is unstable, and the network will when perturbed move its state to some other fixed point.

In the above analysis the effect of the normalisation was ignored. However, the only way it could change the stability condition of the point is if eigenvectors of the Jacobian of the normalisation operator with negative eigenvalues exactly coincide with the positive-value eigenvectors of the above Jacobian. But since this Jacobian depends on arbitrary probabilities, this is not the generic case.

3.3.5 Summing Bayesian Learning

To derive the connection weights, estimates of the probabilities $P(x_{ii'})$ and $P(x_{ii'}, x_{jj'})$ are made. If the training data is already present as z observed pattern vectors $\boldsymbol{\xi}^p$ with component events $\xi^p_{ii'}$ the estimates can be easily calculated by counting the number of occurrences of events ii', jj' and $ii'jj'$ in the training data. This constitutes the summing form of the BCPNN learning rule used earlier (Lansner and Ekeberg, 1989; Lansner and Holst, 1996).

$$c_{ii'} = \sum_{p=1}^{z} \xi^p_{jj'} \qquad c_{ii'jj'} = \sum_{p=1}^{z} \xi^p_{ii'} \xi^p_{jj'}$$

giving probability estimates $\hat{p}_{ii'} = c_{ii'}/z$ and $\hat{p}_{ii'jj'} = c_{ii'jj'}/z$. Since logarithms of these values will be used, special care has to be taken with counters that are zero. In practice the logarithm of zero is replaced with a number that is more negative than all the other values of biases or weights in the network (Holst, 1997).

The weights are set to

$$w_{ii'jj'} = \begin{cases} 1 & c_{ii'} = 0 \text{ or } c_{jj'} = 0 \text{ or } i = j \\ 1/z & c_{ii'jj'} = 0 \\ \frac{c_{ii'jj'} z}{c_{ii'} c_{jj'}} & \text{otherwise} \end{cases} \tag{3.16}$$

and the biases to

$$\beta_{ii'} = \begin{cases} \log(1/z^2) & c_{ii'} = 0 \\ \log(c_{ii'}/z) & \text{otherwise} \end{cases} \tag{3.17}$$

On the other hand, if the data is arriving over time, the probability estimates and weights instead have to be estimated from the sequentially available data online. This may be handled by an incremental version of this learning rule.

3.3.6 Incremental Bayesian Learning

A continuously operating network will need to learn incrementally during operation. In order to achieve this, $P(x_{ii'})(t)$ and $P(x_{ii'}, x_{jj'})(t)$ need to be estimated given the information $\{\mathbf{x}(t'),\ t' < t\}$. What we aim for is an estimate with the following properties: *i)* It should converge towards $P(x_{ii'})(t)$ and $P(x_{ii'}, x_{jj'})(t)$ in a stationary environment, *ii)* It should give more weight to recent than remote information and *iii)* It should smooth or filter out noise and adapt to longer trends, in other words lower frequency components of a non-stationary environment.

One such estimator is exponential smoothing (Brown, 1963). It fulfils the above conditions, and can be compared with a moving average with a certain length τ_L and avoids having to store previous values. The incremental Bayesian learning

rule approximates $P(x_{ii'})(t)$ and $P(x_{ii'}, x_{jj'})(t)$ with the exponentially smoothed running averages $\Lambda_{ii'}$ of the activity $\hat{\pi}_{ii'}$ and $\Lambda_{ii'jj'}$ of coincident activity $\hat{\pi}_{ii'}\hat{\pi}_{jj'}$. The units are assumed to be clamped by the input as the learning takes place.

The continuous time version of the update and learning rule takes the following form:

$$\tau_c \frac{dh_{ii'}(t)}{dt} = \beta_{ii'}(t) + \sum_j^N \log\left(\sum_{j'}^{M_i} w_{ii'jj'}(t)\hat{\pi}_{jj'}(t)\right) - h_{ii'}(t) \tag{3.18}$$

$$\hat{\pi}_{ii'}(t) = \frac{e^{h_{ii'}}}{\sum_j e^{h_{ij}}} \tag{3.19}$$

$$\frac{d\Lambda_{ii'}(t)}{dt} = \alpha([(1-\lambda_0)\hat{\pi}_{ii'}(t) + \lambda_0] - \Lambda_{ii'}(t)) \tag{3.20}$$

$$\frac{d\Lambda_{ii'jj'}(t)}{dt} = \alpha([(1-\lambda_0^2)\hat{\pi}_{ii'}(t)\hat{\pi}_{jj'}(t) + \lambda_0^2] - \Lambda_{ii'jj'}(t)) \tag{3.21}$$

$$\beta_{ii'}(t) = \log(\Lambda_{ii'}(t)) \tag{3.22}$$

$$w_{ii'jj'}(t) = \frac{\Lambda_{ii'jj'}(t)}{\Lambda_{ii'}(t)\Lambda_{jj'}(t)} \tag{3.23}$$

The learning rate $\alpha = 1/\tau_L$ is the inverse of the learning time constant; it is a more convenient parameter than τ_L and will be used extensively. By setting α temporarily to zero the network activity can change with no corresponding weight changes, for example during the retrieval mode.

To avoid logarithms of zero in the calculations, a basic low activity $\lambda_0 \ll 1$ is introduced, a kind of noisy background activity that is present regardless of external signals. In the absence of input $\Lambda_{ii'}(t)$ and $\Lambda_{jj'}(t)$ now converge towards λ_0 and $\Lambda_{ii'jj'}(t)$ towards λ_0^2, producing $w_{ii'jj'}(t) = 1$ for large t (corresponding to uncoupled units). The smallest possible weight value if the state variables are initialised to λ_0 and λ_0^2 respectively is $4\lambda_0^2$, and the smallest possible bias $\log(\lambda_0)$. The upper bound on the weights becomes $1/\lambda_0$. In the following, λ_0 is where not otherwise stated set to 0.0001.

Another way of avoiding underflows which is used in chapter 5 is to introduce λ_0 into the weight calculation rather than into the estimates:

$$w_{ii'jj'}(t) = \frac{(1-\lambda_0^2)\Lambda_{ii'jj'}(t) + \lambda_0^2}{((1-\lambda_0)\Lambda_{ii'}(t) + \lambda_0)((1-\lambda_0)\Lambda_{jj'}(t) + \lambda_0)}$$

The difference between these two methods is minor in practice.

Exponential smoothing is a good model if the probabilities are assumed to be constant or change slowly. The estimate will lag for faster changes such as steps, experiencing a transient $\approx \tau_L$. This may cause the model to fail to converge to a good estimate in highly nonstationary environments (e.g. where the probabilities vary in a sinusoidal manner with a higher frequency than α). The variance of the estimate is $\sigma_\Lambda^2 = \alpha\sigma_x^2/(2-\alpha)$ where σ_x^2 is the variance of input noise. Correlated

noise causes the variance of the estimate to rise (Brown, 1963). To counteract this a longer time constant is needed. There is hence a tradeoff between a more correct probability estimate and lesser lag due to transients.

The initial values of the estimates are somewhat arbitrary. One approach is to set $\Lambda_{ii'} = \lambda_0$ and $\Lambda_{ii'jj'} = \lambda_0^2$. This is stable in the absence of activity but does produce transients when units become active (as well as inherently assuming an absence of activity inconsistent with normalisation). Another approach is to set $\Lambda_{ii'} = 1/M_i$ and $\Lambda_{ii'jj} = 1/(M_i M_j)$, corresponding to the assumption that units within a hypercolumn are on average equally active and uncorrelated.

The above probability estimates converge towards the correct values given stationary inputs for sufficiently large time constants. Since the weights of the network depend more on recent than on old data, it appears likely that a Hopfield-like network with the above learning rule would exhibit palimpsest properties.

A further minor complication relates to units that have never participated in any shown pattern. Such units will have a disruptive effect on the network due to extreme valued connections and biases (in a sense the derivation of the BCPNN learning and update rules does not make allowances for events of features that has never happened or have zero probability). By explicitly setting their connections $w_{ij} = 1$ their influence can be removed.

3.4 Network Learning and Dynamics

3.4.1 Behaviour of Single Connection Weights

We first study how the weight $w_{ii'jj'}$ between units ii' and jj' changes with correlation of unit activities. A discretised version of the learning rule was used to derive these examples.

The behaviour of $w_{ii'jj'}$ can be seen in figure 3.1 for two correlated and two anti-correlated units. When both units are active together for a period of time the connection is strengthened significantly, but the strength of the connection decreases during prolonged stimulation. After the stimulation the weight begins to increase again, a result of the fact that the product $\Lambda_{ii'}\Lambda_{jj'}$ decays faster than $\Lambda_{ii'jj'}$. This continues until the estimates level out and the weight goes to zero. It should be noted that the increase in weights is balanced by the behaviour of the bias; the network dynamics remains stable despite the strong weight changes.

Non-stationary Activities

The above situation of two units firing together against a background of very low activity is somewhat extreme given the assumptions of the model. A more canonical example of the changes in weights due to randomly firing units as well as the response to non-stationary activities can be seen in figure 3.2 where the two units are correlated, uncorrelated or anti-correlated with each other at different times. As can be seen, after a brief transient the log weight moves towards a steady state

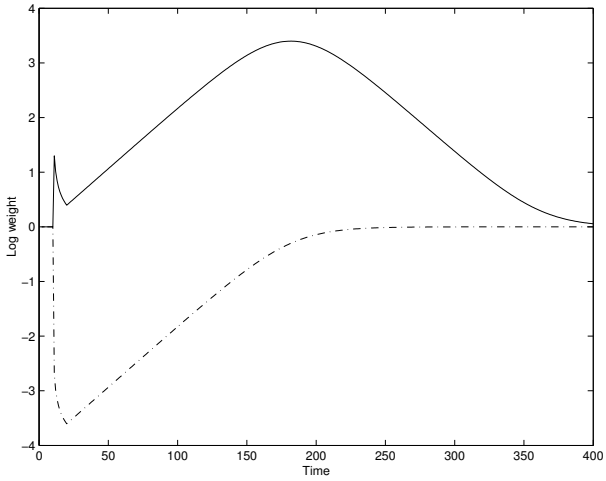


Figure 3.1. Weight between two units that are active together at $10 \leq t \leq 20$ (solid line) and when only one unit is active (dot-dash line) for $\alpha = 0.05$. Note the logarithmic y-scale.

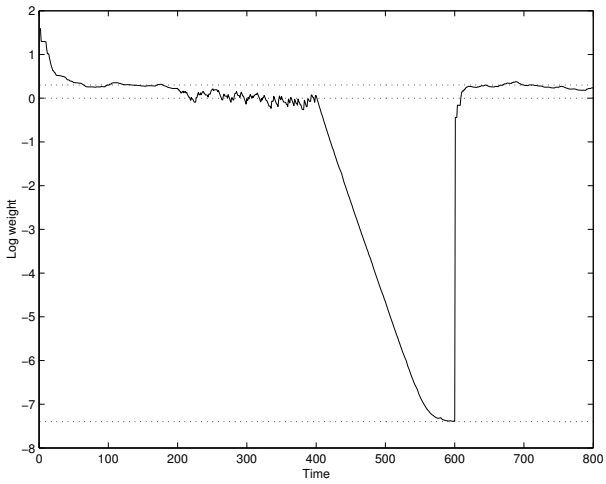


Figure 3.2. Weight between two units that are active 50% of the time, completely correlated for $0 \leq t \leq 200$, uncorrelated for $200 \leq t \leq 400$, completely anti-correlated for $400 \leq t < 600$ and finally correlated again. The dotted lines correspond to the predicted values $\log(2)$, 0 and $\log(4\lambda_0^2)$. In this run $\alpha = 0.05$.

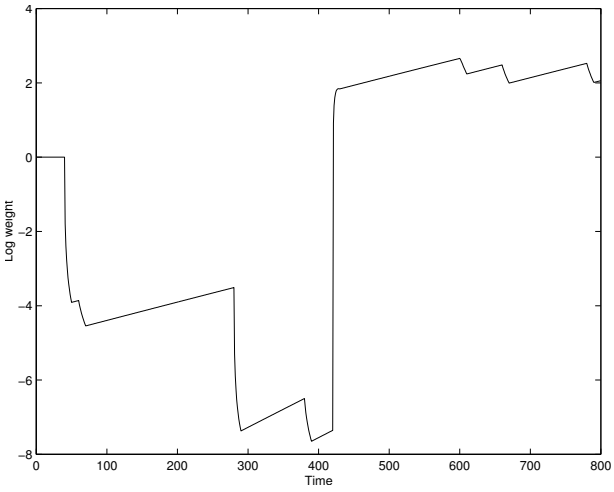


Figure 3.3. Weight between two units during the training of the network with sparse random patterns (10% activity, 10 time steps of presentation of each pattern followed by 10 steps of decay with no activity) with $\alpha = 0.05$. At time 40 one of the units is recruited into a pattern while the other remains silent, causing a strong inhibitory connection to develop. At time 420 both units become recruited by the same pattern, making the connection positive. It then remains positive for the rest of the run, despite occasional activations and coactivations.

value of $\log 2$ as the units remain correlated. The transient is due to initialisation of the estimates to λ_0 and λ_0^2 respectively. As the correlation vanishes the log weight begins to move randomly with a mean close to zero. When the units become anti-correlated the logarithm of the weight decreases practically linearly towards the baseline negative value of $4\lambda_0^2 = 4 \cdot 10^{-8}$. Finally, when the correlations reappear, the weight quickly increases to the steady state value.

When a network is trained with a set of random patterns the logarithms of the weights between two units may change sign depending on whether the units are activated by different patterns (negative logarithm of the weight) or the same pattern (positive log weight) as can be seen in figure 3.3.

Figure 3.4 shows the behavior of a weight plus bias between two units belonging to the same pattern during training. Depending on the learning time constant the weight changes nearly instantly with exposure, or in a slowly accumulative fashion.

3.4.2 Learning and Forgetting

In the following experiments a network with 100 units organised as 10 hypercolumns (where not otherwise stated) was first trained by the repeated presentation of the

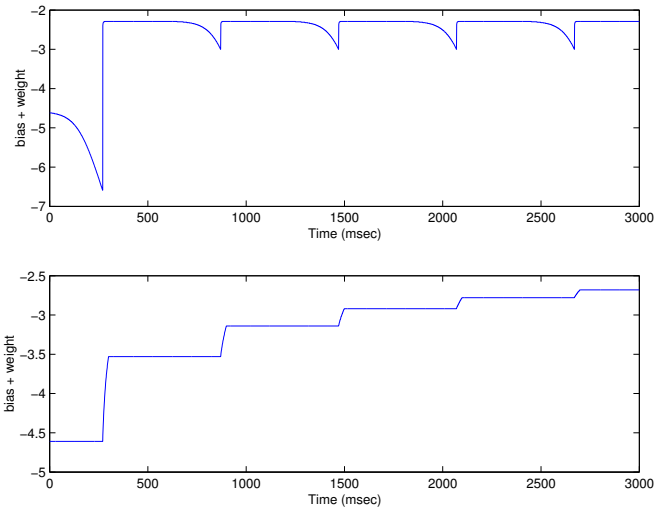


Figure 3.4. Weight plus bias between two units belonging to the same pattern during training, for $\tau_L = 7.2$ s (above) and $\tau_L = 20$ (below). The network was shown the pattern five times for 30 msec.

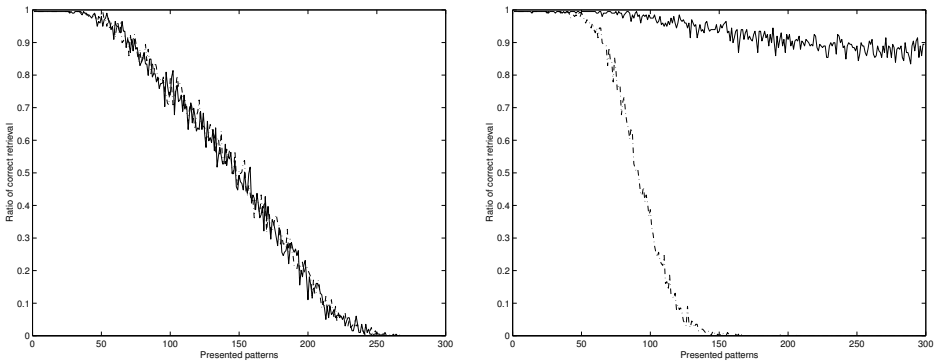
training patterns. Unit activities were clamped to the input patterns for one unit of time. This was followed by a testing period where α was set to zero and no learning took place. The continuous update rule from equations (3.18)–(3.23) was solved using Euler's method with step length $h = 0.1$, $\tau_c = 1$.

The performance for a given pattern $\boldsymbol{\xi}$ was measured as the percentage of perturbed patterns (the activations of two previously active units have randomly been swapped with the activation of other units) which were correctly recalled after relaxation to a tolerance of 0.85 overlap (the overlap was defined to be $\boldsymbol{\xi} \cdot \hat{\boldsymbol{\pi}} / ||\boldsymbol{\xi}|| ||\hat{\boldsymbol{\pi}}||$)

Figure 3.5 shows a comparison between a counter and incremental version of BCPNN. As can be seen the incremental learning rule avoids CF by forgetting the oldest patterns while the recent patterns remain accessible. Figure 3.6 shows this in more detail. The forgetting does not occur immediately: for longer learning time constants the pattern is stored well until a certain number of interfering patterns have been stored, when it starts to gradually fade.

3.4.3 Storage Capacity

Figure 3.7 shows the number of retrievable patterns as a function of α and the size of the training set. For large α the number of retrieved patterns is independent of the size of the training set. For small α a number of patterns up to the maximum capacity of the counter model can be retrieved. For much larger training sets CF



(a) Non-incremental (b) Incremental

Figure 3.5. Comparison of the previous non-incremental BCPNN learning rule and the incremental learning rule for sparse random patterns (one active unit per hypercolumn). The ratio of correctly retrieved patterns (overlap > 0.85) is shown for the first learned pattern $\boldsymbol{\xi}^1$ and the latest $\boldsymbol{\xi}^n$. Solid line: latest learned pattern, dash-dotted line: first learned pattern. $\alpha = 0.01$ in the incremental case.

occurs (as for 400 and 1000 patterns when $\alpha < 10^{-3}$): none of the patterns can be retrieved due to mutual interference.

The same situation but with only one presentation of the training set instead of repetition gives a slightly different picture. Here, even for small training set sizes, the number of patterns retrieved drops for α of about 10^{-4} and below. The reason is that the learning becomes so slow that repeated training is necessary for encoding. For α larger than 10^{-2} the behaviour is identical with or without repetition, i.e. the most recently presented patters are remembered well, regardless of training set size. For large training sets the capacity reaches about 60 % of the maximum capacity. As can be seen in figure 3.7, for the network simulated this occurs for values of α approximately 0.02. In summary, for small α the network operates as a long-term memory of a traditional form, such as the counter BCPNN, thus suffering from CF. For large α it works as a short-term memory with a storage capacity limited by forgetting in the form of weight decay. By modulating α we can tune system performance continuously between these two extremes.

The capacity becomes maximal when the time constant equals the time it takes to run through the entire training set, $\alpha_{opt} = 1/Vz$ where V is the time each pattern is presented and z the number of patterns. For larger α the earliest patterns have faded when the most recent are learned, while for smaller α CF will occur if the training set is larger than can be stored in the network since information in the

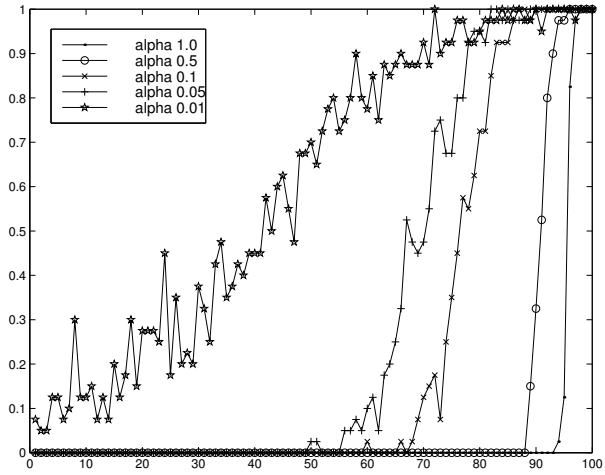


Figure 3.6. Typical forgetting curves after incremental learning. Pattern recall as a function of list position for different values of α (see legend) in a network with 100 units. Recall was estimated as the frequency of retrieval with overlap greater than 0.85 after a presentation of a pattern where two activation of two hypercolumns had been randomly changed followed by 1.0 time units of convergence. Random patterns, 10% activation, 1.0 time units of presentation during training.

estimates will be averaged together. As shown in figure 4.5, there exists a duality between exposure time and learning time constant.

The original BCPNN summing rule without hypercolumns was empirically shown to have a capacity scaling as $N^2/\log^2(N)$ for low-activity patterns (Lansner and Ekeberg, 1989). Johansson et al. (2001) performed similar capacity estimates for the counter and palimpsest versions of the current network with $\sqrt{N}$ equally sized hypercolumns (Figure 3.8. Empirical estimates showed that the counter version had a capacity growing as $\approx N^{1.48}/\log(N)$ and the palimpsest version as $\approx N^{1.52}/\log(N)$, both very close to $N^{3/2}/\log(N)$.

A heuristic estimate of the maximal capacity can be made based on the assumption that synapses can maximally encode a finite and constant amount of information: In a fully connected attractor network of N units where each synapse can encode k bits of information the maximum amount of information that can be stored is $I_{max} = kN^2/2$. This implies that the largest possible number z_{max} of patterns containing I bits that can be retrieved from the network is $z_{max} = kN^2/2I$ For random patterns consisting of $H = N^\rho$ where $0 < \rho < 1$ hypercolumns (which implies an average activity of $N^{\rho-1}$) the information is

$$I_\rho = N^\rho \log_2(N^{1-\rho}) = (1-\rho)N^\rho \log_2(N)$$

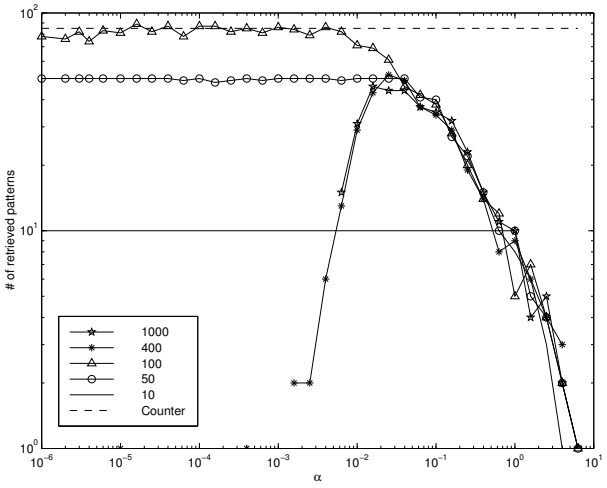


Figure 3.7. Number of correctly retrieved patterns as a function of α for different sizes of the training set. 10, 50, 100, 200, 400 and 1000 patterns were used in the training set, which was repeated 20 times. Successful retrieval is defined as in figure 3.6. As a comparison the maximal capacity of the counter model trained with 100 patterns is drawn as a dashed line.

which gives

$$z_{max} = \frac{kN^2}{2(1-\rho)N^\rho \log_2(N)} = \frac{k}{2(1-\rho)} \frac{N^{2-\rho}}{\log_2(N)}$$

For $\rho = 1/2$ the capacity grows as $N^{3/2}/\log(N)$, which fits well with the experimental results above. This shows that the capacity of the network does grow at the optimal rate, at least within the size range that has been explored. The number of bits per synapse was found to be 0.69 in the counter model and 0.49 in the palimpsest model. The counter model is hence about as efficient as the Willshaw model ($\log 2 \approx 0.693$ bits/synapse) in terms of bits/synapse while the palimpsest memory is about as efficient as the Hopfield network ($1/(\pi \log 2) \approx 0.459$ bits/synapse, (Nadal and Toulouse, 1990)), although still below the limits of what can be achieved with continuous-valued synapses (2 bits/synapse for dense coding and $1/2 \log 2 \approx 0.721$ for sparse coding (Gardner, 1987)).

Note that the theoretical maximal capacity for a fixed N is maximal for ρ close to 0 or 1, i.e. either a few very large hypercolumns or many very small hypercolumns; the smallest feasible hypercolumns are $\rho_{max} = 1 - \log 2/\log N$, corresponding to two units each. The smallest nontrivial network has two hypercolumns, implying $\rho \geq \rho_{min} = \log 2/\log N$.

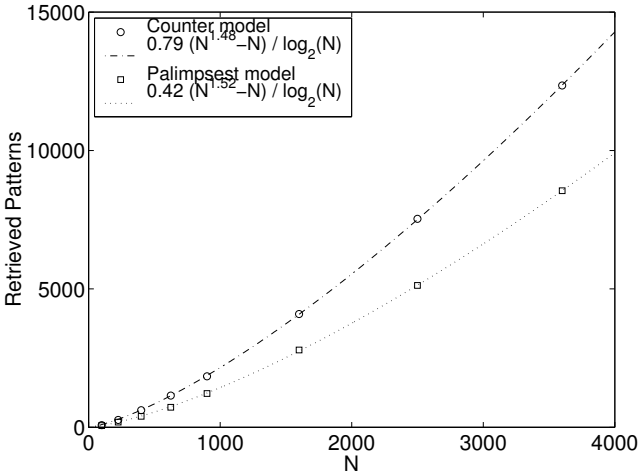


Figure 3.8. Empirical capacity as a function of network size for the counter and incremental version of the learning rule. In the summing case the network was trained with an increasing number of patterns until a maximum number of retrievable patterns was reached, in the incremental case it was trained with a large number of patterns (3–4 times the number that could be stored in the counter model) and the number of retrievable patterns counted. The number of hypercolumns were $N^{1/2}$ and α was shifted optimally in the incremental case. Least mean square fitting was used to fit $kN^e/\log(N)$ to the data. Based on data from (Johansson et al., 2001), used with permission.

However, removing self-connections in hypercolumns complicates things. Instead of $kN^2/2$ storable bits there are $k(N^2 - N^{2-2\rho})/2$, and the maximal number of retrievable patterns becomes

$$z_{max} = \frac{k}{2(1-\rho)} \frac{N^{2-\rho} - N^{2-2\rho}}{\log_2(N)} \tag{3.24}$$

This capacity has one local maximum at ρ_{max} where the capacity grows linearly with N, and a global maximum for a smaller ρ which approaches ρ_{min} as $N \to \infty$ and corresponds to a capacity growing quadratically. An empirical plot is shown in figure 3.9. As can be seen, the empirical capacity follows the theoretical maximal capacity of equation 3.24 fairly well for large ρ and smaller α, and shows a larger deviation for small ρ.

While the capacity grows faster for networks with larger and fewer hypercolumns they are more sensitive to noise in the form of mis-activated columns (since there are fewer other hypercolumns to correct the state), making networks with smaller columns more robust. This can be seen in the response to noise in figure 3.9, where the introduction of disturbed hypercolumns causes the peak to move to higher ρ.

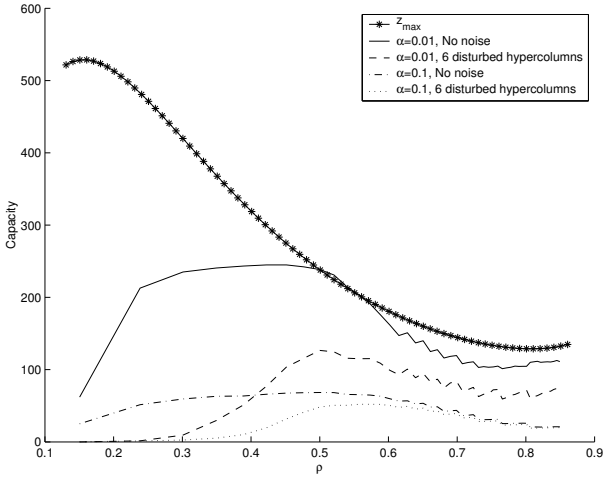


Figure 3.9. Number of retrieved patterns of a 200 unit network ($\alpha = 0.1$) as a function of ρ and α, for 2–50 hypercolumns (left = few large hypercolumns, right = many small hypercolumns). Retrieval from full pattern and patterns where activity in six hypercolumns had been randomly changed are shown, with 100 trials each. Note how increasing noise levels and increasing α move the location of the peak capacity towards the right. The stair pattern on the right is due to the remainder between the network and hypercolumn sizes. Above the simulation plots the maximal capacity of equation 3.24 is plotted, assuming $\log 2$ bits/synapse.

Hence it is reasonable to assume a hypercolumn size with a mid-range ρ if the hypercolumn size is not constrained by the data and the network is assumed to operate on noisy data or with pattern completion. Incremental learning also acts as a noise source, decreasing the optimal hypercolum size for larger α. This can be compared to the similar results for a signal-noise analysis of a Hopfield network with hypercolumns by Johansson et al. (2002).

3.4.4 Convergence Speed

Besides capacity and pattern completion, another characterising property of an attractor neural network is the convergence properties and structure of its state-space. Especially when viewed as a model for working memory retrieval speed becomes relevant. Figure 3.10 and 3.11 shows convergence times for the network as a function of age of patterns and as a function of the number of patterns stored. The stopping criterion used was that the rate of change was below a certain level $||d\hat{\boldsymbol{\pi}}/dt||_1 < 0.05$ (similar to Ruppin and Yeshurun (1991)). Trials where convergence did not occur within 3 time units or where it converged to the wrong attractor were not counted.

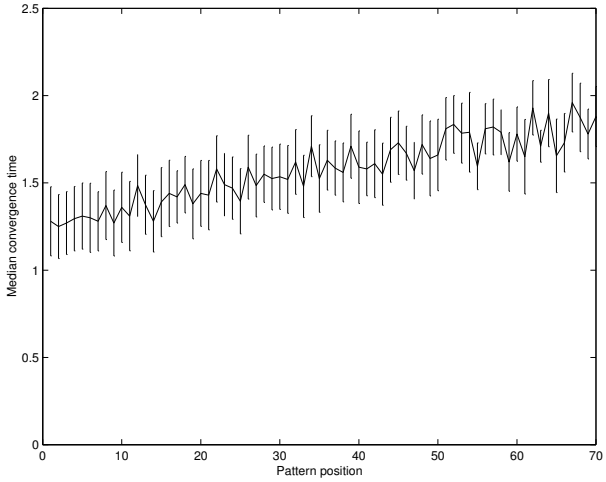


Figure 3.10. Median convergence time and standard deviation for retrieval of stored patterns from a disturbed input, as a function of position in the training set. Pattern 1 is the most recently learned pattern. Only retrieval of correct patterns was counted, and since this was sparser beyond pattern 70 only the first 70 patterns are shown. $\alpha = 0.01$, the network was trained with 100 patterns.

Using this learning rule, as more patterns are learned, the basins of attraction of old patterns become smaller and "shallower" in terms of the energy landscape, eventually disappearing altogether. This also results in a change in convergence speed. There is both a difference between patterns, the latest patterns are completed faster than older patterns, and a roughly linear increase in the median convergence time between networks where few patterns have been learned and networks where the capacity has been reached (figure 3.10, 3.11 and 3.12). Once the network has reached its maximal capacity for a given τ_L, the convergence time remains roughly constant (and retrieval is poor). The distribution of convergence time exhibits a positive skew and the standard deviation of convergence time increases linearly with training set size.

Figure 3.13 shows convergence times when the network is started with a mixture between two patterns (the mixing consists of using 0–10 hypercolumns from one of them and the rest from the other pattern). For the two most recently learned patterns the convergence time is maximal at a 50% mixture (this is practically identical to previous results (Lansner and Ekeberg, 1989) for the non-incremental Bayesian learning rule without hypercolumns). When interpolated between a recent and an old memory the maximum is moved away from the recent memory, a sign that the old memory has a smaller and weaker attractor than the newer memory.

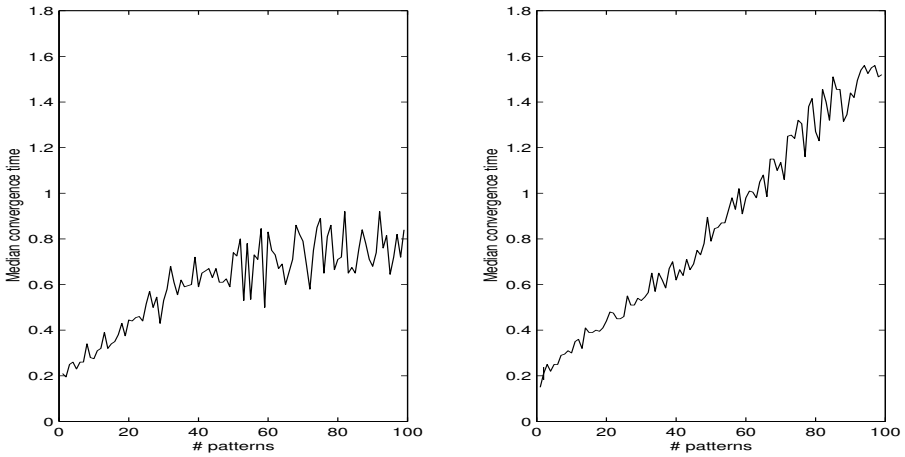


Figure 3.11. Median convergence time and standard deviation for increasing memory load for $\alpha = 0.1$ (left) and $\alpha = 0.01$ (right). The network is trained with up to 100 patterns, but for $\alpha = 0.1$ has capacity for only around 20–30.

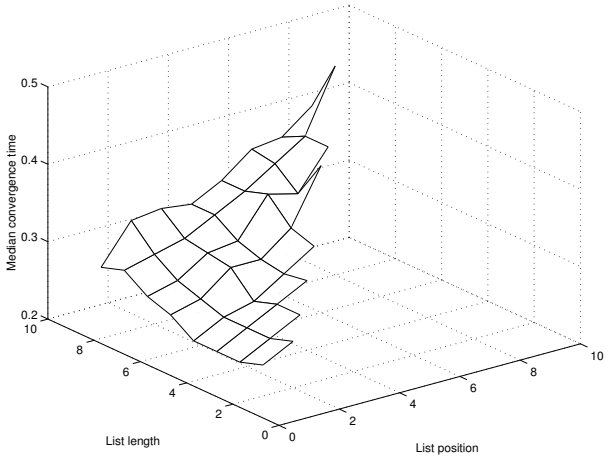


Figure 3.12. Median convergence time for increasing memory load and list position for $\alpha = 0.4$.

It can, however, still be retrieved given a similar enough cue.

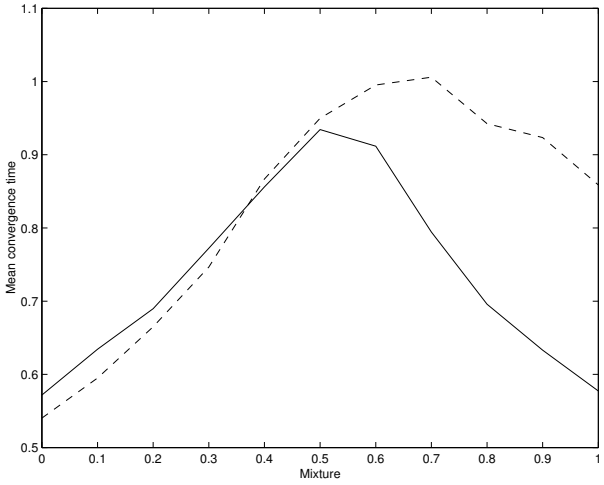


Figure 3.13. Convergence time for mixed input patterns. The input to the network consists of 0–10 hypercolumns of one pattern and the rest from the other pattern. The solid curve represents interpolation between the latest and second latest pattern, the dashed curve interpolation between the latest and an old pattern (#10). 70% of pattern 10 was required to produce convergence to that pattern. The network was trained with $\alpha = 0.01$.

3.4.5 Free Recall with Noise

When stimulated with noise (randomly activated units with the same density as the patterns) the network either converges to one of the learned patterns or to a spurious attractor state. This could be viewed as an abstract model of free recall in memory tests. The probability of convergence towards a certain memory decreases with its age (figure 3.14a). This is in accordance with the results on learning within bounds of Geszti and Pázmándi (1987). They found that as more patterns were stored the basins of attraction of old patterns were more and more flattened energy-wise, and relaxation tended to move the system to a deeper attractor (i.e. a recent pattern).

3.4.6 Comparison with Clipped Weights

We compared our results with the performance of a Hopfield network with clipped weights similar to Parisi (1986). In order to make relevant comparisons we added hypercolumns with sparse activation to the model with clipped weights.

The weights were set based on the sparse Hopfield network (Hertz et al., 1991):

$$w_{ii'jj'}(p+1) = c\left(w_{ii'jj'}(p) + (\xi^p_{ii'} - \sigma)(\xi^p_{jj'} - \sigma)\right)$$

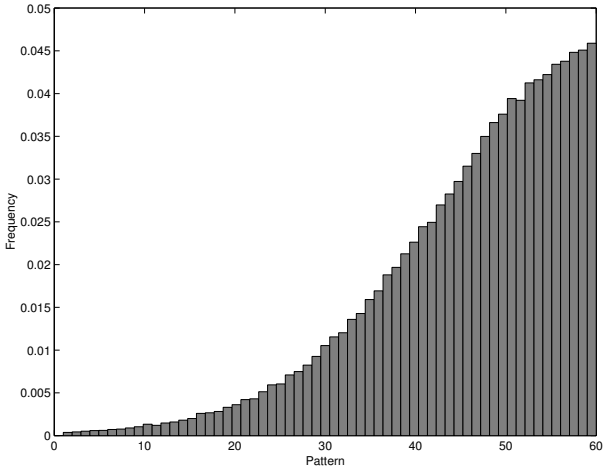


Figure 3.14. Frequency of ending up in the different patterns when activating the network with a random pattern (one randomly selected unit active in each hypercolumn) and allowing it to converge. Each bar corresponds to one pattern. The closest pattern to the resulting end state was used for the diagram. Only patterns with overlap greater than 0.9 are shown; they make up 0.38 of the trials. 7395 trials where the network was trained with 60 patterns and allowed to converge for 100 random patterns; $\alpha = 0.05$.

where c is the same clipping function as in equation 3.1 above and σ the activity level. The update rule used was

$$\tau_c h'_{jj'}(t+1) = \sum_{ii'} w_{ii'jj'} x_{ii'}(t) - h_{jj'}(t)$$

$$x_{ii'}(t+1) = \begin{cases} 0 & h_{ii'}(t+1) < \max_k(h_{ik}(t+1)) \\ 1 & h_{ii'}(t+1) = \max_k(h_{ik}(t+1)) \end{cases}$$

This guarantees a single active unit in each hypercolumn (in cases of ties, one unit is randomly selected).

The forgetting curves (not shown) were similar in character to those presented by Parisi (1986) and also those in figure 3.6, with the clipping range matched to the learning time constant τ_L. The number of correctly retrieved patterns from a disturbed input similar to the one used in the capacity experiments showed a similar behaviour as the incremental BCPNN. The maximal capacity in this model was around 20–30 patterns for 100 units and 400 patterns in the training set. This is significantly lower than the incremental BCPNN (≈ 50 patterns in this case) but still significantly higher than the numbers given by Parisi (1986) for the case with 50% activity and hypercolumns. Our conclusion is that the two models are similar

in character but it appears that the BCPNN has a better performance in terms of storage capacity.

Later comparisions with other rules were done in Johansson et al. (2001), showing similar results.

3.4.7 Different Learning and Forgetting Rates

The basic incremental learning rule has the same learning and forgetting rate; the probability estimates will increase in the presence of activity with the same speed as they decrease in the absence of activity.

A small change of the estimate calculation was tested, were the increase when $\hat{\pi}_{ii'}$ was close to one was α and the decrease when $\hat{\pi}_{ii'}$ was close to zero was β (not to be confused with the bias $\beta_{ii'}$):

$$\frac{d\Lambda_{ii'}(t)}{dt} = (\beta + (\alpha - \beta)\hat{\pi}_{ii'})[(1 - \lambda_0)\hat{\pi}_{ii'}(t) + \lambda_0 - \Lambda_{ii'}(t)] \tag{3.25}$$

$$\frac{d\Lambda_{ii'jj'}(t)}{dt} = (\beta + (\alpha - \beta)\hat{\pi}_{ii'}(t)\hat{\pi}_{jj'}(t))[(1 - \lambda_0)\hat{\pi}_{ii'}(t)\hat{\pi}_{jj'}(t) + \lambda_0 - \Lambda_{ii'jj'}(t)] \tag{3.26}$$

The estimate calculations for a stationary random input still converge to $P(x_{ii'})$ and $P(x_{ii'jj'})$, but now increase and decrease at different rates.

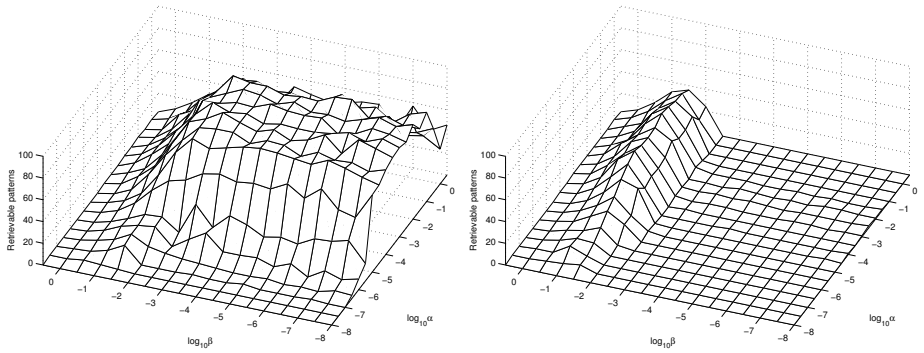
The effect was a network with capacity limited by different factors (figure 3.15): too fast learning tended to overwrite information, too slow learning did not adapt fast enough to patterns shown a single time. Too fast forgetting limited the capacity by making stored information decline, while too slow forgetting made the capacity limited by the size of the network.

It is possible to employ a large α and a low β to achieve a network that learns fast but forgets slowly. However, the number of patterns that can be retrieved are essentially the same as the capacity for $\alpha = \beta$ for an optimal choice of α.

Such a memory might be used as a working memory or episodic memory exhibiting one-shot learning. The stored information would still be vulnerable to being overwritten by new information, including irrelevant "noise". To protect against this some form of gating mechanism would be needed (as has been proposed for the prefrontal cortex (Durstewitz et al., 1999)). The gating mechanism discussed in next chapter based on modulating α could work for this purpose. However, if α is increased to a higher level with the arrival of new information to be stored and kept close to zero the rest of the time, the net effect is similar to using the basic $\alpha = \beta$ network.

3.5 Discussion

We have proposed and characterised an incremental version of a previously described Bayesian learning rule (Lansner and Ekeberg, 1989). The new rule allows



(a) 100 patterns (b) 400 patterns

Figure 3.15. Number of retrievable patterns from a network with fast or slow forgetting. The 100 unit network was trained with 100 (left) or 400 (right) patterns, shown only once. Along the right edge fast forgetting limits the capacity. Along the front edge slow learning limits capacity. Along the back edge overwriting caused by fast learning limits the capacity. The network acts as the counting model and suffers catastrophic forgetting for small values of β.

for continuous real-time learning from a sequence of examples without leading to catastrophic forgetting. Instead, old information is gradually forgotten and only the most recent examples are retained, as in a palimpsest memory. As expected the time for the memory decay scales roughly as the learning time constant τ_L. The memory capacity increases linearly with τ_L up to a limit where it becomes equal to the standard counter BCPNN if repeated enough, and on the same order otherwise. This means that the introduction of palimpsest properties has not reduced the maximal capacity as such. By setting the size of the network and the learning time constant the memory capacity can be regulated from a fast learning and forgetting "working memory" to a slowly learning and forgetting "long-term memory". In the first mode the number of storable patterns is limited by the learning/forgetting rate ("α-limited"), while in the second mode the limit is the maximal capacity of the network and exposure time ("size limited"). The time of convergence to a stored memory state depends both on the age of the memory and the load on the network.

The original BCPNN learning rule is interesting in that it is based on probabilities and statistics rather than being a standard Hebbian outer product rule. Furthermore, palimpsest memories like learning within bounds (Hopfield, 1982; Parisi, 1986) ignore new information supporting an already saturated connection, while non-supporting information will affect it; throwing out old knowledge is favoured

regardless of how much positive evidence has accumulated. The learning rule proposed here does not suffer from this cut-off non-linearity and appears to have a better storage capacity. The learning rule is in some sense similar to marginalist learning (Nadal et al., 1986), where new patterns are exponentially amplified and thus old patterns decay correspondingly relative to these. Unlike the learning rule of Storkey and Valabregue (1999) it does not take local fields into account and does appear to have a somewhat lower amount of information per synapse, but comparision is complicated by the different levels of sparsity employed in the networks (Johansson et al., 2001).

Biological associative synaptic plasticity is generally assumed to be Hebbian and correlation based. This is also the case for the Bayesian-Hebbian learning rule used here. It thus falls in the same category as correlation based learning (Sejnowski, 1989) and the BCM rule (Bienenstock et al., 1982). It exhibits a graded behaviour with multiple synapse activations as well as a more step-wise behaviour for single synapse activation similar to experimental observations in LTP (Petersen et al., 1998). Like the above learning rules the BCPNN rule displays LTP as well as LTD, and with some modifications it provides a phenomenological model for spike-timing dependent plasticity (Wahlgren and Lansner, 2001). The rule presented here contains a bias term which adapts the excitability of the postsynaptic neuron alone. This relates to the phenomenon of EPSP-spike potentiation which has recently received increasing attention (Andersen et al., 1980; Jester et al., 1995).

It is interesting that this formally derived model produces a synaptic learning rule with many similarities to biologically observed phenomena. From a biological point of view it is quite reasonable to implement the learning rule with running averages as we have done here. There are several possibilities how a synapse could realize such computations within the biochemical networks and protein synthesis dependent processes involved in synaptic plasticity (Bhalla and Iyengar, 1999; Frey and Morris, 1997).

It should be noted that the derivation of the BCPNN does not take into account how estimates are calculated. Since they are approximations to the real probabilities of events, there is already a measure of inexactness in the dynamics. Had the network been too sensitive to the correctness or consistency of the estimates it would not have functioned well with incremental learning. From empirical experience the network appears to be robust to changes in estimation method, including having different or changing learning time constants for different estimates. Quantifying this robustness is an important challenge.

While the derivation of the BCPNN induces a particular form of the update equations due to hypercolumns, simpler variants have been found to exhibit essentially similar behaviour. One example is to base a recurrent network on a BCPNN without hypercolumns such as the one in equation 3.3, but with normalisation of groups of units acting as hypercolumns. This has been found to exhibit capacity comparable to the hypercolumn BCPNN (Ström, 2000).

In general the presence of a softmax or winner-takes-all rule appears to improve performance of autoassociative networks. Part of this performance increment is

likely due to constraining the dynamics onto a subspace of state space where the memory states are still reachable. Many components of the noise will be filtered away by the normalisation. This is reinforced by signal-noise analysis of a sparse Hopfield network with a winner-take-all rule (Johansson et al., 2002). It also keeps the activity level fixed, making threshold control in order to maintain a set level of activity unnecessary.

The hypercolumns studied here have many similarities with Potts neurons. K-state Potts neurons (or spins) can take on K possible states, and within the mean field framework the probabilities $P_i(k)$ of neuron i being in state k are normalized $\sum_k P_i(k) = 1$ (Kanter, 1988). Potts neurons have mainly been used in optimization problems (Peterson and Söderberg, 1989, 1998) rather than associative memory.

The input-output function formula 3.4 differs from the standard input-output relation of neural networks $h_i = \beta_i + \sum_j w_{ij} x_j$ with transfer function $x_i = f(h_i)$. It can be viewed as a pi-sigma neural network where $h_i = \beta_i + \log(\prod \sum w_{ij} x_j)$. Such higher-order networks have a higher representational capacity than ordinary perceptron networks while avoiding introducing more free parameters (Shin and Ghosh, 1991). It is also worth noting that the same form of sum of nonlinearly transformed inputs have been observed in simulations of pyramidal cells, although in this case the nonlinearity was found to be sigmoidal rather than logarithmic (Poirazi et al., 2003). Whether this similarity has any relevance remains to be seen; the BCPNN framework assumes input from the same hypercolumn to be summed together, which would imply a degree of synaptic specificity for individual dendrites that appear unlikely if units were assumed to be individual pyramidal cells. The interpretation of BCPNN units as minicolumns rather than individual neurons on the other hand provides a far more complex local circuitry where nonlinear subunit summing could be achieved in a variety of ways.

In our simulations with a fast learning and forgetting memory we have found that the average convergence time increases significantly with memory load. This is similar to the classical finding by Sternberg of a linear reaction time dependence on the number of items held in working memory (Sternberg, 1966) which has been used to support hypotheses of scanning processes underlying working memory (Burle and Bonnet, 2000; Lisman and Idiart, 1995). Our results imply that the psychological phenomena can equally well or better be described by an attractor network with fast synaptic learning-forgetting dynamics. Previous parallel accounts have been based on a limited capacity for processing that has to be shared between all the comparisions (Atkinson and Shiffrin, 1968), but have been criticized on the grounds that the inclusion of another parallel task does not affect the relation between memory set size and reaction time (Sternberg, 1975). This model does not have that drawback, exhibits positively skewed convergence times similar to those observed in experiments (Van Zandt, 2002) and also predicts the observed shorter reaction time to recent items, which is hard to account for in the exhaustive scanning explanation (Forrin and Cunningham, 1973). For mixed patterns the convergence time shows the same slowing behavior at a decision boundary as reported in Ratcliff et al. (1999).

It is interesting to consider the possibility of having a memory system comprised of multiple attractor networks with different learning dynamics and degrees of plasticity, as suggested by Little and Shaw (Little and Shaw, 1975). A quickly adapting network would learn and remember presented objects in working memory, while a more slowly forgetting network might learn from single presentations (as in episodic long-term memory), and even slower learning and forgetting networks would average individual presentation events into a "prototypic" semantic memory (cf. (Brunel et al., 1998) and (Lansner, 1991) for two implementations). Such a memory structure is compatible with what is thought to exist in human memory systems (Squire, 1992).

Furthermore an important aspect of memory is that of relevance information and print-now mechanisms. The learning rule proposed here will result in memory traces that are volatile in the absence of input since the weights are continuously changing to obey the current rate estimates. This leads to a gradual decay of memory over time even when little new information arrives. An alternative possibility is to control the learning rate by some form of relevance or "print-now" signal. In this case, simultaneous pre- and postsynaptic activation is not enough to result in weight changes. It is only when plasticity is enabled by the print-now signal that changes occur, equally for imprinting and decaying (κ in Lansner and Ekeberg (1989)). With regard to our learning rule, we have found that this can easily be implemented as a change in the learning time constant (α is increased with the relevance of the situation) (Sandberg et al., 2001). This will be the subject of chapter 4.

Chapter 4

Memory Modulation

4.1 Relevance Modulation

Long-term memory (LTM) formation in everyday life often occurs incidentally without explicit intention to remember the information processed. It has been suggested that memory formation is the dynamic consequence of information processing and system plasticity (Petersson et al., 1999). Research indicates that specific kinds of information processing contributes to LTM-formation, including meaning-based, context and relational processing and factors like emotional significance and attentional allocation (for a recent review see e.g. (Wagner et al., 1999)). Endogenous processes activated by experience can modulate memory strength in terms of recall probability (McGaugh, 2000). For example, emotionally arousing (Christianson, 1992) or humorous (Schmidt, 1994) experiences are generally better remembered than less affective experiences, and hormones and neuromodulators can affect how strongly experiences are retained (Martinez et al., 1991).

The novelty or uniqueness of a stimulus also plays an important role. The isolation or von Restorff effect consists of improved recall or recognition of an item (the isolate) that is distinct or different from the others in a set, while the other items are less well recalled (retroactive and proactive inhibition) (von Restorff, 1933). While this has mainly been studied in human list recall, a similar effect has been observed in rats (Reed and Richards, 1996) and monkeys (Parker et al., 1998). The effect has been explained in terms of interference among non-isolates (von Restorff, 1933), attention or salience effects, but the interpretations remain controversial (Sikström, 2003).

Some of these factors can be interpreted in the framework of memory consolidation as a relevance modulation of the "print-now" signal by regulating memory encoding and synaptic plasticity. During ordinary events the plasticity is at a low level. When salient, arousing or motivational stimuli arrive one or more relevance detection systems respond to the input and produce a print-now signal increase the

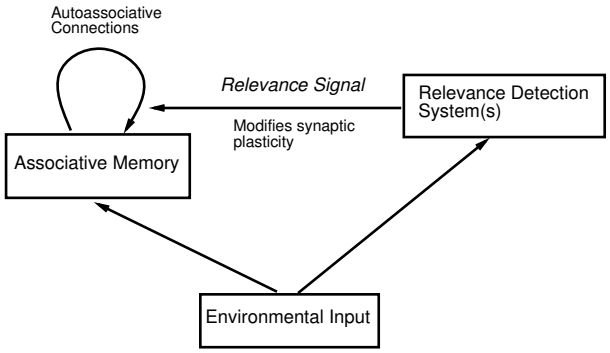


Figure 4.1. Sketch of a possible relevance modulation system. Sensory input arrives both to an associative memory where it can be learned (i.e. the hippocampus) and a relevance estimation system (i.e. the amygdala) which regulates the plasticity of the associative memory.

learning rate, for example through a modulating mechanism. Such a mechanism might relate closely to neuromodulation in the brain, e.g. the effect of dopamine (Wickens and Kötter, 1995) and acetylcholine in synaptic plasticity (Woolf, 1996; Hasselmo et al., 1996).

In an event-related fMRI study of volunteers reading lists of words with semantic, perceptual or emotional isolates activation was observed in the right inferior prefrontal and bilateral posterior fusiform cortices for all kinds of isolates (Strange et al., 2000). Perceptual isolates also caused activation in fusiform cortex, emotional isolates in left amygdala and semantic isolates in the left ventral prefrontal cortex. These results were interpreted as support for the right prefrontal cortex as a monitor of discrepancy between expectation and experience, which in turn would activate an arousing or orienting response. Other results (Parker et al., 1998) implicate perirhinal cortex in this system. In the current model it (together with the attribute-specific regions handling the qualitative character of the relevance) would be a relevance estimation system and activate modulatory pathways affecting short-term memory.

The BCPNN learning rule has a time constant of learning that determines how quickly it will adapt to new information. By modulating this time constant we can model the modulatory regulation of the print-now signal on associative encoding of information into for example long-term memory. This modulation was originally suggested in Lansner and Ekeberg (1989) but not thoroughly investigated.

This paper describes a simple model of an autoassociative network with plasticity modulation for one item, and shows that it can produce the enhanced recall of the isolate, proactive and retroactive inhibition and an inverted U-shape response curve to overall plasticity similar to the one commonly observed in arousal-

performance or dose-response plots. The effect is compared with a mathematical model of how basins of attraction age and the effect of changing the mean overlap between a pattern and the others.

4.2 Effects of Learning Time Constant Modulation

Equation 3.23 was modified to include a time varying relevance/print-now signal $\kappa(t)$, assumed to be sent from a modulator system as a response to the current experience:

$$\frac{d\Lambda_i(t)}{dt} = \kappa(t)\alpha([(1-\lambda_0)\hat{\pi}_i(t) + \lambda_0] - \Lambda_i(t)) \tag{4.1}$$

$$\frac{d\Lambda_{ij}(t)}{dt} = \kappa(t)\alpha([(1-\lambda_0^2)\hat{\pi}_i(t)\hat{\pi}_j(t) + \lambda_0^2] - \Lambda_{ij}(t)) \tag{4.2}$$

In the following experiments it was kept at $\kappa(t) = 1$ except for one pattern in the training set, the isolate, where $\kappa(t)$ was set to κ_i during training. λ_0 was set to 10^{-4} throughout this chapter.

A 100 neuron BCPNN network with 10 hypercolumns of 10 neurons each was trained by clamping unit activity to each training pattern $\boldsymbol{\xi}^p$ for one unit of time, allowing the weights to adapt. Retrieval was tested by activating the neurons with a trained pattern where the activity in three hypercolumns of ten had been randomised, and then allowed to relax for one unit of time (no learning was used during testing). Performance was measured by the overlap $\boldsymbol{\xi}^p \cdot \mathbf{x}/|\boldsymbol{\xi}^p||\mathbf{x}|$.

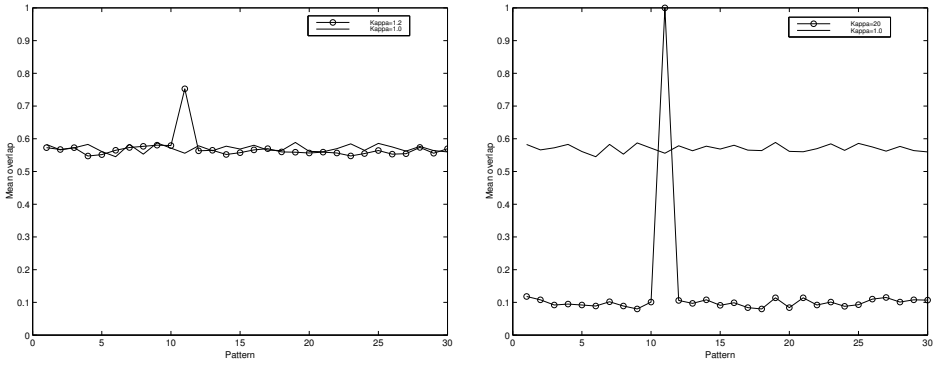
Figure 4.2 shows the selective enhancing effect on recall when an isolate pattern occurs. Note the inhibition of recall of other patterns in the high κ_i condition.

Figure 4.3 shows the mean overlap as a function of the modulation strength for the isolate and for the normal items. As κ_i increases the recall of the isolate becomes better and better, while there is an inhibition effect on the other items. However, it is possible to avoid inhibition in this model for low levels of the relevance signal while still observing a recall enhancement for the isolate. Inhibition is also stronger at higher α, where it mainly impairs retrieval of patterns older than the isolate.

When the network is stimulated by a random pattern it will converge to a given attractor state with a probability depending on the relative volume of the basin of attraction to the total volume of the state space. Again the isolate is more likely to be recalled, suggesting that for these parameters the increased plasticity has enlarged its basin of attraction relative to the other attractors.

If the network has a shorter learning time constant as in figure 4.4, there will be a memory gradient due to fast forgetting. The increase in plasticity caused by a strong modulatory input can prevent encoding of the isolate and retroactively interfere with the early patterns without affecting the subsequent storage of the network, as can be seen in the right subfigure.

As a preliminary for next section, the effect of base time constant and exposure time was checked. Exposing the network to input patterns for a longer time was



(a) $\kappa_i = 1.2$ (b) $\kappa_i = 20$

Figure 4.2. Mean overlap after convergence from noisy cue with or without an isolate pattern. In the dotted run, κ_i was set to 1.2 (left) or 20 (right) for pattern 11. α was set to 10^{-8}.

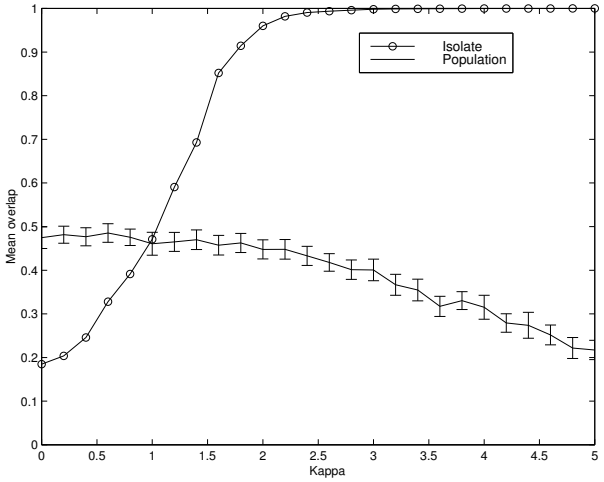
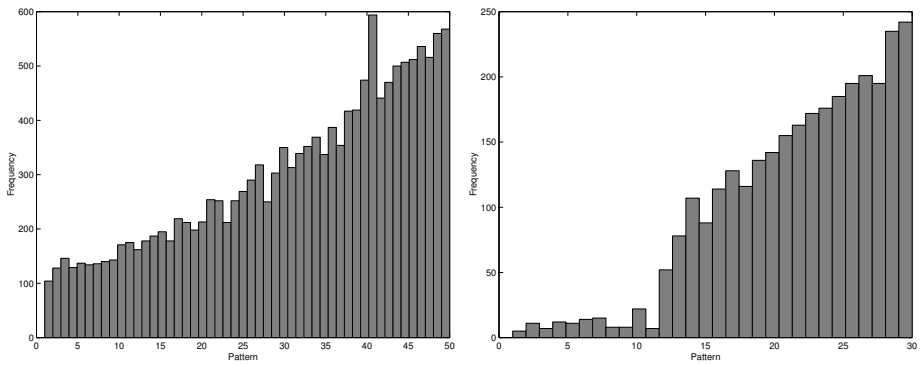


Figure 4.3. Mean overlap after convergence from noisy cue of the isolate pattern and the other patterns as a function of the increase in plasticity κ_i. α was set to 10^{-8}.



(a) $\kappa_i = 2$, $\alpha = 10^{-2}$ (b) $\kappa_i = 20$, $\alpha = 10^{-1.5}$

Figure 4.4. Frequency of ending up in the attractors corresponding to the learned patterns when activating the network with a random pattern (one randomly activated unit per hypercolumn) for a high level of plasticity. The isolate is pattern 41 (left) and 11 (right).

equivalent to a change in time constant: doubling the exposure time corresponds to halving the time constant (Figure 4.5). This shows that merely holding a pattern to be encoded longer can be replaced by a corresponding increase in $\kappa(t)$. Similar to the capacity curves in previous chapter at high α the network learns quickly but forgets the oldest patterns, while at low α the network learns too slowly to learn the patterns.

4.2.1 Performance Model

If the network is placed in an arbitrary state it will converge to an attractor state corresponding to a learned pattern, or possibly a spurious state. The probability $P(\xi^p)$ of reaching a certain state ξ^p is proportional to the volume in state space of the basin of attraction, $V(\xi^p)$. How does the volumes change as new patterns are learned?

As the network learns a large number of patterns it will approach a stationary state where the addition of a new attractor will cause the older attractors to shrink but leave the statistical distributions of attractor sizes time invariant. If we assume there is no specific interference between one pattern and another (e.g. a low degree of overlap, such as in the α-dominated regime of a large network) the volume of the basins will only depend on their age as a decreasing function $V(t)$ where t is their age.

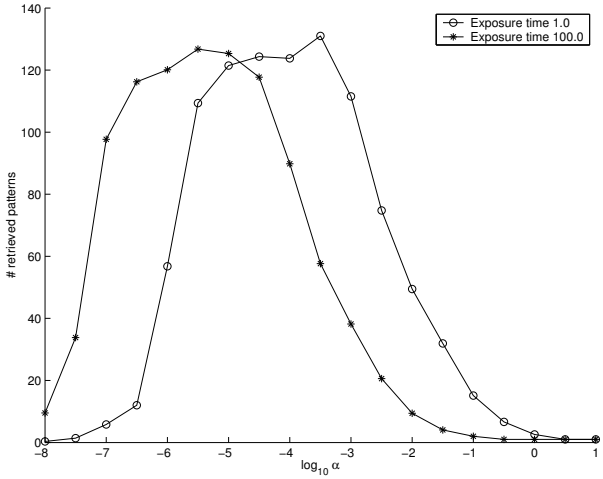


Figure 4.5. Retrievable patterns as a function of α for one time unit exposure and 100 time units exposure (200 unit network, 200 patterns shown once). Scaling up the exposure time is equivalent to scaling down the time constant the same amount.

One of the simplest choices of $V(t)$ and which fits the observed behaviour of the BCPNN for high α (e.g. figure 4.4) is an exponential decay:

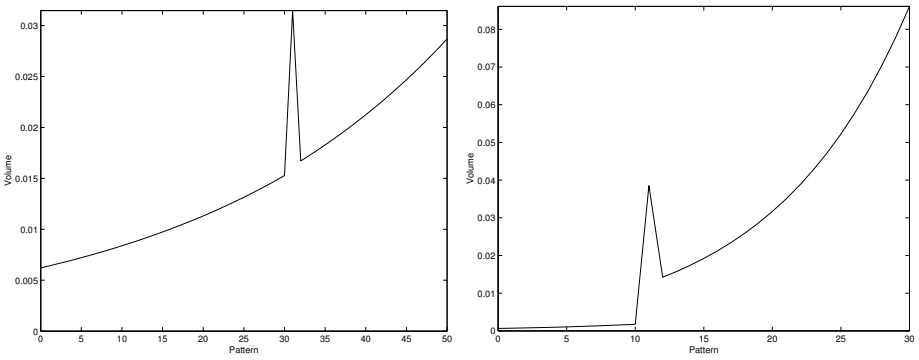
$$V(t) = (1 - e^{-\lambda})e^{-\lambda t} \tag{4.3}$$

(here the volumes are normalised with respect to the total volume of state space; the normalisation constant could be smaller to account for spurious states). λ is assumed proportional to α.

Since a change in the learning time constant is equivalent for this network with a change in exposure time, a halving of the time constant for a pattern is approximately equivalent to exposing the network to the same pattern twice. Hence the time can be reparametrised in terms of $\kappa(t)$ to take print-now modulation into account. Let $\mu(t) = \sum_{s=0}^{t} \kappa(s)$. Then the above equation becomes

$$V(t) = (1 - e^{-\lambda})e^{-\lambda\mu(t)}\kappa(t) \tag{4.4}$$

The above model fits the network behaviour well for small $\kappa(t)$ (figure 4.6). A temporary increase in $\kappa(t)$ increases $V(t)$ but at the expense of earlier memory traces in the simulation. For large print-now signals the effect on $V(t)$ also becomes self-inhibitory as the Λ estimates decline quickly; the isolate pattern becomes less likely to be retrieved compared to other patterns. Here the simulation and model gives different results (compare figure 4.4(b) and 4.6(b)) since the model does not take the self-inhibition into account.



(a) $\kappa = 2$, $\lambda = 0.03$ (b) $\kappa_i = 20$, $\lambda = 0.1$

Figure 4.6. Volume of attractor states estimated using equation 4.4 to match figure 4.4.

The lifetime of a learned pattern depends on the size of its basin of attraction. When the basin of attraction becomes small it is unlikely to be retrievable through free recall, and eventually it disappears entirely making cued recall impossible. In the model this can be estimated by assuming that recall is possible if $V(t) > \epsilon$, where ϵ is a constant dependent on the demands of recall quality. For exponentially decaying patterns (equation 4.3)) the lifetime $L(V_0)$ as a function of the original attractor volume V_0 becomes

$$L(V_0) = \frac{1}{\lambda} \log\left(\frac{V_0}{\epsilon}\right) + \frac{1}{\lambda} \log(1 - e^{-\lambda})$$

(here it is assumed that the learning time constant is not changed after the pattern of interest is encoded; a similar but more complex expression based on equation 4.4 can be derived). If the basin of attraction of the pattern is enlarged by a factor k compared to the basin of attraction of a non-modulated pattern, the relative lifetimes become

$$\frac{L(kV_0)}{L(V_0)} = 1 + \frac{\log(k)}{\lambda L(V_0)}$$

Hence we can expect a pattern to survive proportional to the logarithm of the encoding strength. This fits well with empirical tests of the network (figure 4.7). An implication of the fit is also that $\kappa_i \approx k$, suggesting that at least low levels of time constant modulation regulate the size of the basins of attraction proportionally. The convex lifetime curve makes the benefits of very strong upregulation of κ in terms of pattern lifetime small.

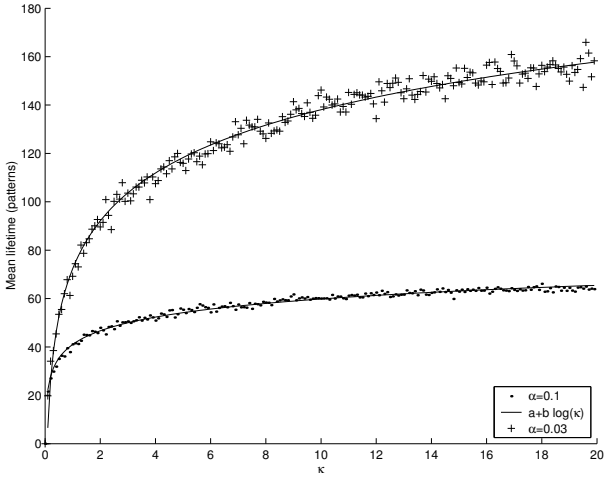


Figure 4.7. Lifetime of isolate patterns as a function of κ_i. 50 patterns were stored in the network before the isolate, and the lifetime estimated as the number of patterns that could be learned before the isolate could not be retrieved with better than 0.85 overlap when cued with itself. The points are the average of 50 trials with $\alpha = 0.1$ and $\alpha = 0.03$. A logarithm function was fitted to the data.

4.3 Correlated Patterns

Another potential cause of the isolation effect is that the isolate has features not found in the other patterns, i.e. it is less correlated with them. As discussed in Sikström (2003) this is not expected to produce an enhancement effect in traditional correlation based learning rules, but by adding a sliding modification threshold to the learning rule decorrelated patterns can become enhanced.

In the BCPNN decreasing the overlap between a pattern and the other patterns increased its lifetime significantly (figure 4.8). Even the minor decrease of average pattern overlap (from 1/81 to 1/90) had a profound effect, and a pattern orthogonal to all subsequent training patterns was not forgotten at all. The reason for this was the difference in the Λ_i estimate of the "rare" units unique to the isolate pattern. Since this estimate was smaller than among the regular units the weights connecting to and from the unit were increased, making the corresponding pattern stronger. In the case of a fully orthogonal pattern all units were such strong units, and further training did not cause any interference. Training with patterns of the same type as the isolate before its presentation abolished most of this enhancement effect (figure 4.8, white bars), since the estimates were now equally high.

The high variance of lifetimes for the decorrelated patterns was due to the different number of unique units randomly generated for each pattern. The average

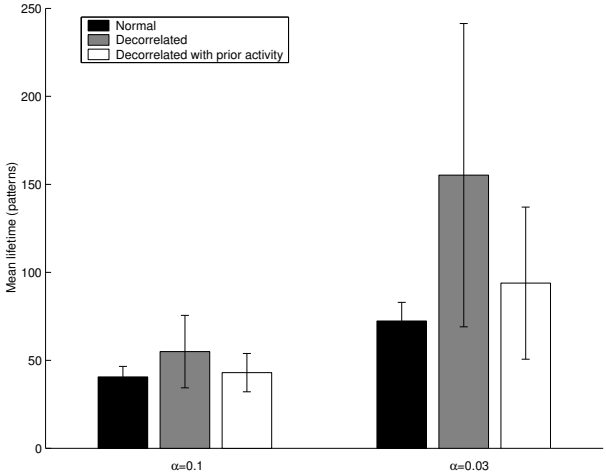


Figure 4.8. Lifetime of stored pattern when normally correlated (black) and decorrelated (gray and white) for $\alpha = 0.1$ and $\alpha = 0.03$. White bars represent decorrelated patterns learned after training with other decorrelated patterns. Lifetimes were estimated as in figure 4.7 for isolate patterns generated by randomly selecting active units among all units of hypercolumns, while the other patterns were generated by selecting from 9 out of the 10 units in each hypercolumn.

lifetime was found to increase linearly with the number of unique units until the majority of the units of the pattern were unique. Larger networks exhibited a less pronounced lifetime enhancement, since the average overlap was smaller due to the sparser high-dimensional patterns.

4.4 Discussion

We have shown that a model of an autoassociative memory with modulated learning rate can exhibit selective encoding enhancement of items associated with a relevance signal. By decreasing the learning time constant the network adapts to a training input faster, producing a stronger attractor state. The increase in trace strength improves cued recall for moderate amounts of signal. The stronger attractor has a competitive effect on other stored patterns and when the signal is too strong inhibition of other items occurs; for high learning rates the inhibition is mostly backwards, while for slowly learning networks inhibition can be both forward and backward. Training the network with a pattern for a longer time was equivalent to a change in the relevance signal.

Such a relevance signal would regulate which patterns would be stored in a flexible way, enabling not just gating of wanted and unwanted memory patterns but also enhancement of more important patterns compared to ordinary patterns. Such a mechanism would enable the network to retrieve isolate patterns more easily during free recall and recall older, but more relevant, patterns long after less relevant patterns have been forgotten.

Emotional modulation of memory could act through such a signal, strengthening encoding of experiences associated with arousing states. This would be especially relevant for an intermediate memory such as the MTL, where later reinstatement into permanent neocortical storage would be influenced by how the primary encoding into the MTL during the experience was modulated by the then current emotional state.

The number of recallable patterns as a function of the basic time constant exhibits an inverted-U curve shape (figure 4.5) reminiscent to the Yerkes-Dodson law (Yerkes and Dodson, 1908) or the observed inverted-U dose-response relationship seen for many memory enhancing drugs (Martinez et al., 1991; Parsons and Gold, 1992). If there is a need to encode relevant experiences at different levels of trace strength the baseline plasticity needs to be in a range where a change in plasticity due to the relevance signal produces a large change in encoding success. Near the maximum of the curve small changes in plasticity have little effect on retrieval performance. If the memory exhibits a U-shaped response to plasticity, this suggests that the baseline plasticity α should not be close to the maximum of the performance curve but rather remain smaller. For small α relevance signals can produce a strong enhancement as they enable the encoding of new patterns in an otherwise static memory. An increased level of baseline arousal would decrease the distinction between relevant and irrelevant memories in such a system. This may relate to the observation that extroverts (with lower baseline arousal level) perform better at working memory tasks than introverts (higher arousal level) (Lieberman, 2000).

There is also the possibility of memories using large α values, where $\kappa(t)$ acts as an inhibitor erasing previous memories when a new pattern arrives. While less plausible as an intermediate memory storing many patterns, it might be useful as a sensory buffer.

The enhancement effect of less correlated patterns appears to be very similar to the isolation effect model analysed by Sikström (2003). There a sliding modification threshold was introduced in the learning rule of a Hopfield-like network with bounded synaptic strengths. The effect of this modification threshold was to make synaptic plasticity decrease for commonly active units, and increase for rare units belonging to less correlated patterns. In the BCPNN learning rule this automatically happens due to the probability estimates. Synapses connecting one or two units belonging to many patterns will require more co-activity Λ_{ij} to balance the high Λ_i than synapses between seldom activated units. While the model of Sikström is deliberately aimed at examining the isolation effect and making predictions about psychological data, the BCPNN implements the same mechanisms due to its statistical derivation.

The two isolation effects can be distinguished by their time-course and effect on other patterns. Plasticity modulation causes retroactive interference and a relatively small increase in memory lifespan, while decorrelation can significantly prolong the survival of a memory even at high levels of plasticity. However, in large networks with sparse activity random patterns will tend to have less overlap with each other, decreasing the strength of this effect.

This suggests that one way of disambiguate encoding modulation through plasticity change and decorrelated encoding is the relative inhibition of previous memories, where plasticity change should cause relatively stronger interference. Adding new modalities or depth of encoding to the stored information would create a more widespread and stable representation but reduce the relative improvement in retrieval of isolate patterns if they are mainly due to decorrelation, while plasticity modulation would not be similarly affected. For example, the observation that memory arts produces improved recall while avoiding proactive and retroactive inhibition (Patten, 1990) suggests that it is due to a more widespread encoding rather than plasticity modulation. Perceptual isolates in the experiment of Strange et al. (2000) exhibited enhancement compared to control after shallow but not deep encoding while emotional isolates were recalled better regardless of encoding depth, which would support a plasticity modulation account.

In general plasticity modulation can be more flexibly applied than decorrelation, since it can be tied to arbitrary memories with no need for a different representation of the information. Decorrelation is a property of the network and learning mechanism, while plasticity modulation can be placed under dynamic control.

While the experiments in this chapter have demonstrated the effect on a neural network memory of a relevance signal, the relevance estimation system remains to be analysed. The exact size of the relevance signal should be a function of the salience of the pattern being stored, which in turn is determined by the expected fitness effect of learning it. This in turn depends on estimates of the likelihood of such patterns and their value. These can be either hardwired through evolution (or design, in an artificial system) or through learning.

Chapter 5

Adaptation

5.1 Synaptic and Cellular Adaptivity

Many kinds of neurons, in particular cortical pyramidal cells (see e.g. (McCormick et al., 1985)), exhibit spike frequency adaptation due to the regulation of the slow afterhyperpolarization phase of action potentials typically by activity-dependent influx of calcium ions that opens Ca^{2+}-dependent potassium channels giving rise to an outward hyperpolarising current. Such adaptation can terminate the activity of self-exciting neuron populations, giving rise to periodic bursts of activity (Lansner, 1982; Lansner and Fransén, 1992; Lansner et al., 1997) or complex aperiodic dynamics (Cartling, 1996, 1997).

In addition, many synapses show depression or facilitation depending on the frequency of presynaptic spikes and neuron classes (Thomson and Deuchars, 1994). Many models of synaptic dynamics have been constructed, mainly models treating the strength of synapses as regulated by the amount of 'resources' available for producing an EPSP (Tsodyks and Markram, 1996; Abbott et al., 1997; Tsodyks et al., 1998). Depletion of transmitter might be one such depression factor, and the rate of replenishment has been shown to regulate the interval between activity bursts in CA3 slices (Staley et al., 1998).

Bibitchkov et al. (2002) studied a sparsely coded associative network with binary neurons with simple resource dynamics, and showed that the addition of adaptation did not change the fixed points of the network, but reduced their basins of attraction significantly, in turn reducing the capacity of the network. Similar results were found by Torres et al. (2002). The behaviour of the binary neuron network was qualitatively the same as for a network of integrate-and-fire neurons, and exhibited a dynamics where the network state moved between different attractor states (Pantic et al., 2002).

Such a dynamics where the network state visits distinct states has many intriguing possibilities, both for free recall, search for one or more matching memory

states to an input and sequence learning. It can be viewed as an implementation of Hebb's phase sequence (Hebb, 1949) or Braitenberg's "pump of thoughts" (Braitenberg, 1984).

5.2 Phenomenological Adaptation and Reinstatement Model

One motivation for exploring adapting networks is to provide a mechanism for free recall and reinstatement dynamics in MTL-NCX interactions. In order to model such reinstatement processes in attractor networks it is necessary to have a method for replaying the attractor states without external retrieval cues. In order to serve as reinstatement the dynamics need to reach all sufficiently large attractors, which is also desirable for free recall. By modulating the learning rate of the network relevant memories can be enhanced and irrelevant memories suppressed (Sandberg et al., 2001). The time spent in each attractor state during reinstatement should ideally reflect the importance of the state.

One possibility is external random stimulation or subcortical disinhibition such as in the model of Bibbig and Wennekers (Bibbig and Wennekers, 1996). A non-specific activation would stimulate the network, followed by convergence to a random attractor state, which could then be reinstated in the cortex. But how does the system get out of the attractor state once it has reached it? If the departure from an attractor state is driven by intrinsic noise, then it seems unlikely the system as a whole could sustain the ordered activity necessary for reinstatement or free recall. If it is externally controlled, for example by the theta rhythm or bursts of noise, then different attractors will be presented for the same time interval and their relative strength will be identical even though they might have very different behavioural relevance.

Another solution is to have an intrinsic mechanism of replay that activates and terminates attractor states. One such possibility would be depression at the synapses involved in the active pattern, alone or in combination with neural adaptation. As the network stays in one attractor the synapses sustaining the activity between the participating cells adapt and weaken and the active neurons accumulate hyperpolarisation. Eventually that group of cells will become unable to sustain their activity. Another attractor will become dominant as local noise is amplified by disinhibited cells. The system can be seen as a bursting intrinsic rhythm generator similar to the model of spontaneous episodic activity in the developing chick spinal cord of Tabak et al. (2000) or the lamprey (Lansner et al., 1997).

This form of replay dynamics does not need any external top-down control, and is regulated by the timescale of synaptic adaptation. The overall dynamics consists of a fast dynamics with timescale τ_c involving convergence to attractors and a slow dynamics with timescale τ_A of adaptation of synapses leading to a shift to a new attractor. The learning dynamics corresponds to a third, even longer timescale τ_L

which changes synaptic connection matrix and thus the attractor states.

Properly speaking the attractor states above are not attractor states, but parts of an attractive limit cycle or strange attractor where the dynamics is slow. They are similar to the quasi-attractor states of Amit (1989) in that they are initially robustly attractive, stable on a timescale longer than τ_c but eventually become destabilized. However, in the following activity states with such slow dynamics in the vicinity of the states that would be fixed point attractors in a non-adapting network will be regarded as "the same" (quasi)attractor states.

In the following we will explore a phenomenological model of adaptation-driven dynamics in an attractor network, where adaptation is modeled using an extra Hebbian associative projection with negative gain. The strength of this projection between two units represents the level of adaptation of their synapses, depression or facilitation. Since the bias of a BCPNN is closely related to the weight dynamics the bias of the adaptation projection is also included, representing firing rate adaptation of the receiving units. Two co-active units will cause their mutual connection in the adaptation projection to increase in weight similar to how learning occurs in normal projections. But since the gain is negative, the net effect will be a decrease of the effective input from the other unit and bias. This is essentially equivalent to the unlearning of van Hemmen (1997), but seen as temporary rather than permanent When units are inactive the projection returns to a low value through "forgetting", restoring the original synaptic efficacy. Since the learning and unlearning processes are equivalent (up to a gain factor and the time constants) they will balance each other.

The strength of the gains of the projections as well as the time constants are plausible targets of neuromodulation. Acetylcholine inhibits associative connections in cortical networks and facilitates the induction of synaptic plasticity while suppressing adaptation (Hasselmo and Cekica, 1996; Hasselmo, 1999; Barkai and Hasselmo, 1994). Similarly noradrenaline can inhibit excitatory intrinsic connectivity producing a more input-controlled network state (Hasselmo et al., 1997) while enabling LTP and blocking LTD and delayed facilitation (Thomas et al., 1996; Katsuki et al., 1997; Cloues et al., 1997). Dopamine also appears to inhibit associative connections in order to produce more localised activation Nunez (1995), which together with the increase in plasticity due to D1/D5 receptors Otmakhova and Lisman (1998) could lead to improved selectivity in sensory learning Bao et al. (2001). In this model these effects could be represented by decreasing τ_L, τ_A, g_A and g_L, while increasing g_I (see below). 5-HT appears to increase the gain of associative fibres and possibly decrease the gain of inhibitory feedback loops Nunez (1995), as well as promoting synaptic facilitation Kozlov et al. (2001). In the model this would correspond to an increase of g_L and g_A. Hence the network could be moved between an input-driven learning state (high catecholamine level, low 5-HT), a point attractor state (intermediate modulation) and an adapting state moving between attractors (high 5-HT, possibly some catecholamine modulation).

5.3 Network

The adaptive synapses were modelled by adding an extra projection between the units of a BCPNN, but with negative gain $-g_A$ and its own learning time constant τ_A. In these simulations τ_A has where not otherwise stated been given a value of 160 msec, corresponding to the decay rate of the action potential related Ca^{2+} pool in the previous biophysically detailed pyramidal cell model (Fransén and Lansner, 1995).

The update equations used are:

$$\begin{aligned}\tau_c \frac{dh_i(t)}{dt} =& g_L \left[\beta_i(t) + \sum_k \log \left(\sum_{j \in H(k)}^{M_k} w_{ij}(t) \hat{\pi}_j(t) \right) \right] \\ & - g_A \left[\gamma_i(t) + \sum_k \log \left(\sum_{j \in H(k)}^{M_k} v_{ij}(t) \hat{\pi}_j(t) \right) \right] \\ & + g_I I_i - h_i(t)\end{aligned} \tag{5.1}$$

$$\hat{\pi}_i(t) = \frac{e^{h_i}}{\sum_{j \in H(i)} e^{h_i}} \tag{5.2}$$

The weights w_{ij} (the normal associative connections) and v_{ij} (the virtual adaptation projection) and their corresponding biases β_i and γ_i are set by the learning rule

$$\kappa(t)\tau_L \frac{d\Lambda_i(t)}{dt} = \hat{\pi}_i(t) - \Lambda_i(t) \tag{5.3}$$

$$\kappa(t)\tau_L \frac{d\Lambda_{ij}(t)}{dt} = \hat{\pi}_i(t)\hat{\pi}_j(t) - \Lambda_{ij}(t) \tag{5.4}$$

$$\beta_i(t) = \log(\Lambda_i(t)) \tag{5.5}$$

$$w_{ij}(t) = \frac{(1-\lambda_0)\Lambda_{ij}(t) + \lambda_0}{((1-\lambda_0)\Lambda_i(t) + \lambda_0)((1-\lambda_0)\Lambda_j(t) + \lambda_0)} \tag{5.6}$$

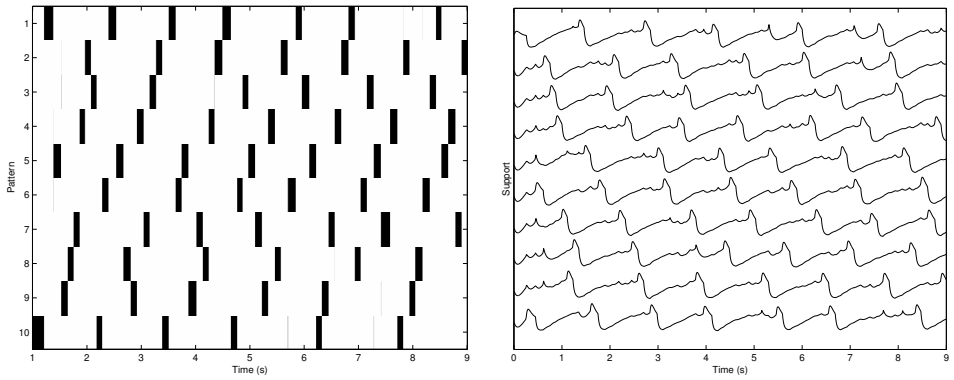
$$\tau_A \frac{d\mu_i(t)}{dt} = \hat{\pi}_i(t) - \mu_i(t) \tag{5.7}$$

$$\tau_A \frac{d\mu_{ij}(t)}{dt} = \hat{\pi}_i(t)\hat{\pi}_j(t) - \mu_{ij}(t) \tag{5.8}$$

$$\gamma_i(t) = \log(\mu_i(t)) \tag{5.9}$$

$$v_{ij}(t) = \frac{(1-\lambda_0)\mu_{ij}(t) + \lambda_0}{((1-\lambda_0)\mu_i(t) + \lambda_0)((1-\lambda_0)\mu_j(t) + \lambda_0)} \tag{5.10}$$

This is a top-down functional model rather than a qualitative attempt to mimic biology. An advantage is that it can be directly combined with the other models developed in this thesis and the unlearning mechanism is easy to balance with the learning dynamics.



(a) Overlap (b) Support

Figure 5.1. Left: overlap between network state and learned patterns over time as the network wanders between quasi-attractors. Right: support (sum of weighted input and bias) for units belonging to the different patterns as a function of time.

5.4 Quasi-Attractor Dynamics

In figure 5.1 a 100 unit network was trained with 10 orthogonal patterns, with a learning time constant of $\tau_L = 7200$ msec (learning was subsequently turned off). It was then allowed to run freely for 9 seconds. The figure shows the overlap of the network activity with different stored patterns as a function of time. As can be seen the network state cycles through the patterns, remaining in each for a brief period of time. The synapses quickly depress and active units adapt, cutting off the activation and producing a refractory period where the pattern cannot be reactivated.

The network converges to a quasi-attractor state on a timescale of τ_c. Each unit belonging to the cell assembly will gain a strong support from the other active units and inhibit competing units. But the activity causes the effective bias $\beta_i - \gamma_i$ to decrease, as well as the support from the other member units via the effective connections (figure 5.2 and 5.3). If $g_A/g_L \geq 1$ the total support becomes too low to sustain the quasi-attractor and the activity decreases quickly. Meanwhile the units belonging to the least inhibited quasi-attractor (due to a low level of adaptation and/or overlaps with the current quasi-attractor that stimulate some of the member units) will increase in activity due to a lessened inhibition and the normalisation, and quickly replace the previous state.

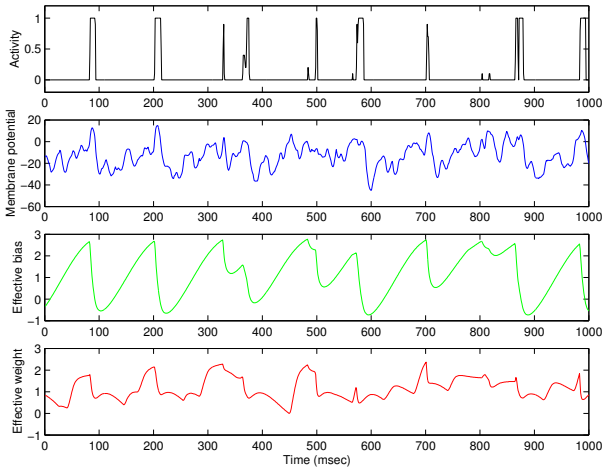


Figure 5.2. Output of a unit, membrane potential (support), effective bias $\beta_i - \gamma_i$ and effective synaptic weight $w_{ij} - v_{ij}$ for a synapse connecting it to another unit associated in the same cell assembly. As the cell assembly is activated, the bias rapidly decreases and synapses quickly depress and cut off the activity. During the inactive periods a gradual recovery occurs, making activation more likely over time.

The mutual information between the current attractor state $S(t)$ (defined as the memory state $\boldsymbol{\xi}^i$ closest to $\hat{\boldsymbol{\pi}}(t)$) and the state at time $t - \Delta$

$$I(\Delta) = \sum_{i,j} P(S(t) = \boldsymbol{\xi}^i, S(t-\Delta) = \boldsymbol{\xi}^j) \log_2 \frac{P(S(t) = \boldsymbol{\xi}^i, S(t-\Delta) = \boldsymbol{\xi}^j)}{P(S(t) = \boldsymbol{\xi}^i) P(S(t-\Delta) = \boldsymbol{\xi}^j)} \quad (5.11)$$

exhibits a non-monotonic decline (figure 5.4). This is caused by the anticorrelation between the current state and the state after adaptation; the dip at ≈ 120 msec implies that the next state will be affected much more by external factors than the current state. The nonmonotonicity also shows that the dynamics is not a true Markov chain, since extra information about the past is "hidden" in the adaptation state. The autocorrelation function averaged over all units shows a similar behavior.

For quickly learning networks (shorter τ_L) only the latest patterns are retrieved, while for more slowly learning networks nearly all patterns are eligible for retrieval (figure 5.5a and 5.6). For networks with fast adaptation (shorter τ_A) more quasi-attractors can be reached in a given time but they are presented more briefly (shorter dwell time), while for slow adaptation they tend to remain in a single quasi-attractor for a long time (figure 5.5b and 5.7a).

If one pattern is learned more strongly than the others the quasi-attractor is visited more often (Figure 5.7b). This can occur both by prolonged or repeated

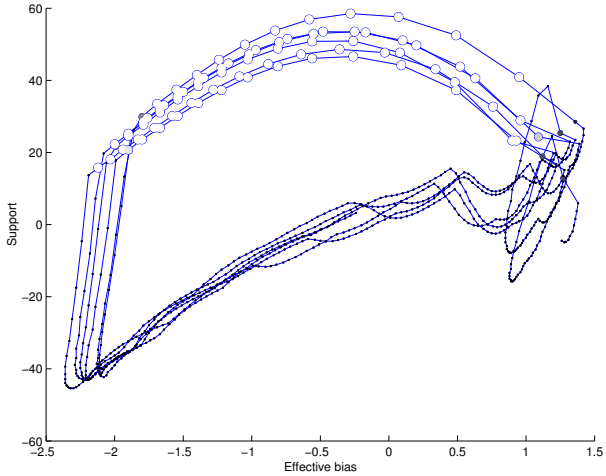


Figure 5.3. Effective bias plotted versus membrane potential for a single unit. The markers signify the activity level π of the neuron; white and large correspond to an active state, small black to an inactive state. The dynamics has a slow component moving it right during inactive periods and left during active periods, and a fast dynamics shifting it up and down between the two modes.

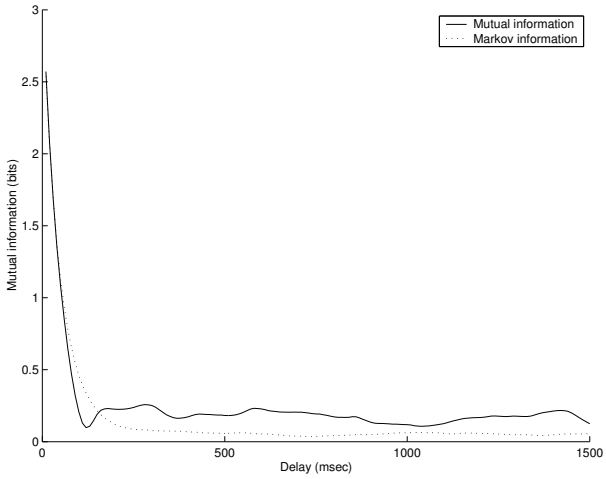
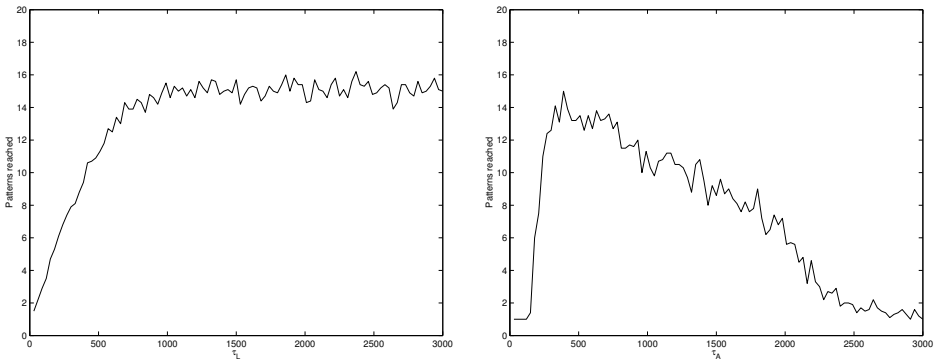
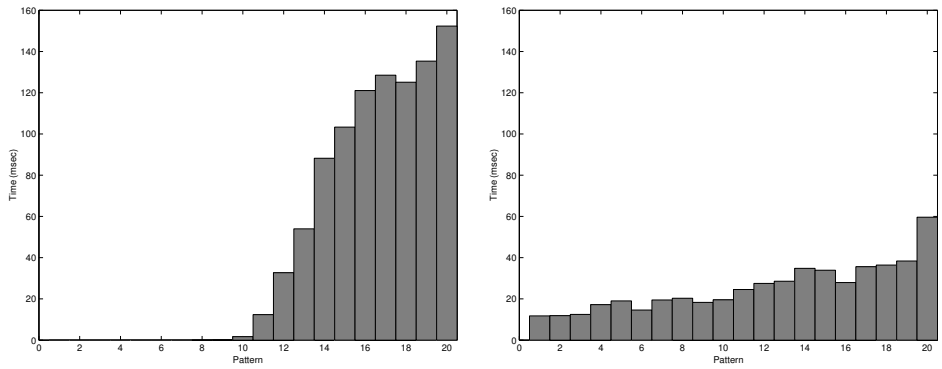


Figure 5.4. Time delayed mutual information $I(\Delta)$ as defined in equation 5.11 for the adaptation dynamics and for data generated from a Markov chain with the same transition probabilities. 9000 msec data from a network trained with 10 orthogonal patterns was used.



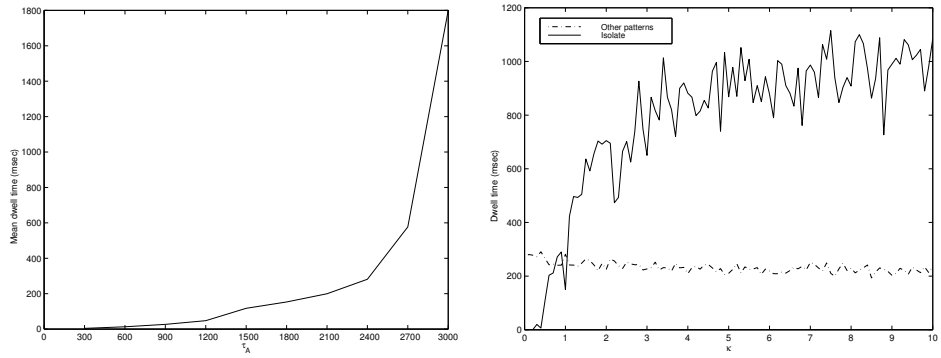
(a) τ_L (b) τ_A

Figure 5.5. Number of patterns visited (overlap > 0.85) during 18 seconds of free recall as a function of learning time constant (left, τ_A=160 msec) and adaptation time constant (right, τ_L=7200 msec). 200 unit network trained with 20 random patterns.



(a) $\tau_L = 360$ msec (b) $\tau_L = 7200$ msec

Figure 5.6. Total time spent near (overlap > 0.85) different quasi-attractors during adaptation, average over 100 runs. To the left for a quickly learning and forgetting network, to the right a slowly learning network.



(a) Dwell time

(b) Modulated dwell time

Figure 5.7. Left: Time spent in each visit to a quasi-attractor state as a function of adaptation time constant (τ_L=360 msec). Right: Dwell time spent in the vicinity of a quasi-attractor that was learned using a modulated learning time constant $\kappa\tau_L$, and the mean time spent in the vicinity of other quasi-attractors. For $\kappa = 1$ all patterns are learned equally, while the isolate pattern is strongly inhibited for $\kappa < 1$ and enhanced for larger values. Compare to figure 4.3b in chapter 4.

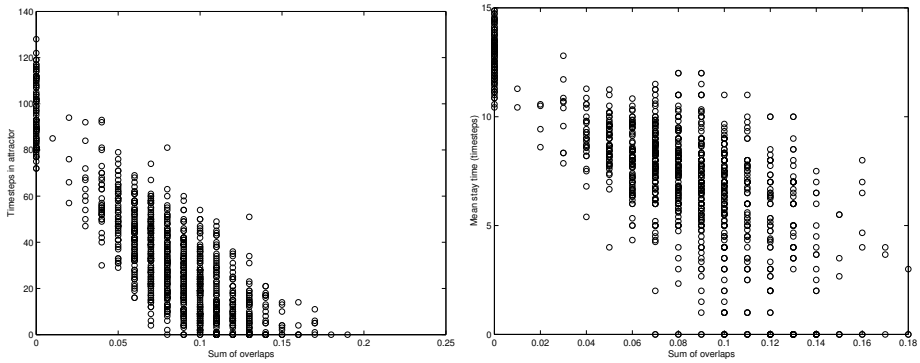
exposure or by a temporary modulation of the learning rate (Sandberg et al., 2001). Since behaviourally important patterns likely have stronger quasi-attractors in the MTL system than irrelevant patters, they will be retrieved more often in the reinstatement dynamics and hence be imprinted more strongly in the cortex.

5.5 Dwell times and Pattern Overlaps

The dynamics of the network depends in a complex way on the relative overlaps of the attractors, but the the distribution of time between the quasi-attractors depends mainly on their relative strength and average overlap with each other. Figure 5.8 shows the distribution of total time spent in different quasi-attractors as a function of their average overlap $s_i = (1/Nz)\sum_p \boldsymbol{\xi}^p \cdot \boldsymbol{\xi}^i$. The total time spent in a certain quasi-attractor is roughly a linear function of $a - bs_i$ plus Gaussian noise. The dwell times (not shown) have a similar distribution.

For small networks the high degree of overlap makes the distribution of dwell times uneven and low, but as the size of the network increases they approach a more even distribution due to increasing orthogonality.

Similar to the experiments in section 4.3 an isolate pattern that is very different from other patterns (low s_i) will be more strongly recalled even without any κ



(a) (b)

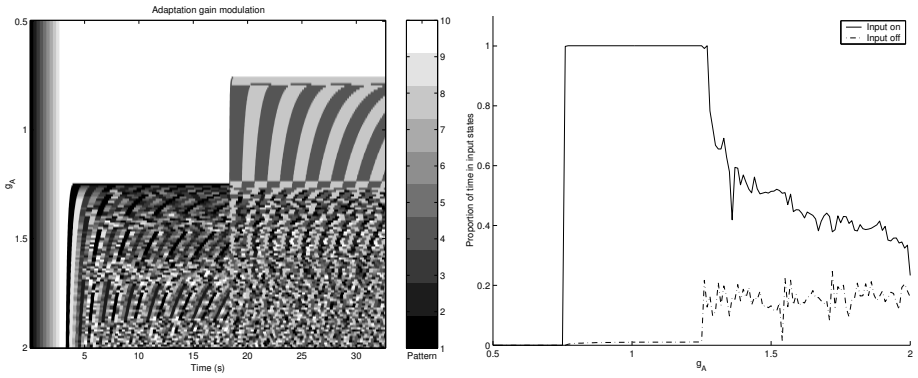
Figure 5.8. (left) Total time spent in quasi-attractor states as a function of sum of overlaps. (right) Mean dwell time in quasi-attractor states as a function of sum of overlaps.Each ring is a quasi-attractor state. 10 random patterns, one of which had no overlap with the others.

modulation. This property remains regardless of whether retrieval is based on convergence from noise input or adaptation, and appears due to a larger sized basin of attraction.

5.6 Second-Best Match

Driving an autoassociative network to a desired state with an external input requires a high input gain or lowering the autoassociative gain enough to overcome the attraction to the current state (which could be a possible role for neuromodulatory influences, e.g. see Hasselmo et al. (1997)). The adaptation dynamics discussed in this chapter suggests an alternative way. As the current attractor state depresses, the network briefly exists in a state where the effective weights (the weights between strongly active units) are close to zero and hence the network is easily influenced by external input. An adapting network will hence tend to approach a "suggested" state, which enables both external control, second-best match and online learning. This is similar to the results of Bibitchkov et al. (2002) for a binary network.

The adaptation dynamics also enables second-best match to input. If a mixture of learned patterns is given as an input during adaptation it will bias the dynamics towards the most closely related quasi-attractors. Figure 5.9 shows the effect of different strengths of adaptation gain on the dynamics as a mixture is shown. For small g_A the network becomes trapped in the last training pattern and does not



(a) (b)

Figure 5.9. (a) Effect on different strengths of adaptation gain g_A on the dynamics. For the first half of the simulation input was turned off, then a 50-50 mixture of pattern 4 and 8 was shown. The network was trained with 10 patterns shown as different grayscales, $\tau_L = 7.2s$. (b) Total dwell time in pattern 4 and 8 without external input (dash-dotted line) and with external input (solid line).

move between the attractors. For higher values it remains in the last quasi-attractor as long as no external input occurs, but begins to wander between the closest quasi-attractors to the input when it is activated. For a narrow range just above $g_A = 1$ it exhibits both free-running dynamics and a lock to the input patterns. For larger gains the dynamics becomes increasingly internally driven; while the input still biases the dynamics towards closely matching patterns all other patterns are also visited. If overlapping patterns are used (not shown) the essential dynamics remains the same, but with the added complexity of different patterns appearing in an order dependent on overlaps.

Similar effects were seen by modulating λ_0 and g_L for moderate strengths of the gain $g_A \approx 1$ but not for stronger $g_A = 4$. The relative dynamical range was wider, suggesting that keeping g_A close to the transition between the modes can enable modulation of other parameters to regulate the network dynamics effectively.

5.7 Online Learning

External input can cause the network state to move during the fast dynamics towards a state which is not a quasi-attractor. If learning is active that state will be reinforced, and may in time become a quasi-attractor state on its own (figure 5.10).

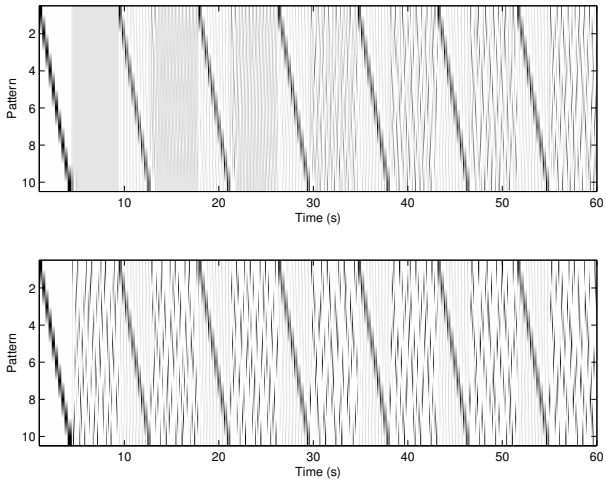


Figure 5.10. Overlap with quasi-attractors (10 orthogonal patterns) as a function of time for $\tau_L = 300$ s (above) and $\tau_S = 30$ s (below). In the first case several repetitions of the patterns are necessary to build up stable quasi-attractors. In the second case just one initial presentation is needed. Once the movement between quasi-attractors has become stable, it interacts strongly with the external input.

If the learning time constant is short the patterns will be learned nearly immediately, and if there is a variation in the input (such as noise or a variable feature) the latest presentation will be recalled, a form of episodic memory. If a more slowly learning network is used more repetitions are needed and the variable features will be averaged together, forming a "semantic" memory.

Since sufficiently strong adaptation dynamics precludes dwelling permanently in a single state, the different quasi-attractors (or a subset of them) will be visited over time and each time slightly reinforced. This form of online memory constantly refreshes itself and retains the capacity of learning from an external input (figure 5.11). Over time the first patterns will be most strongly learned as they have repeated the most, and while they cannot dominate the dynamics completely there will be a notable primacy effect.

5.8 Discussion

We have showed how synaptic depression and unit adaptation in combination in an attractor network can produce a stochastic dynamics, where the network spends much of its time close to stored attractor states but shifts between them on a timescale set by the adaptation time constant. This is in qualitative agreement

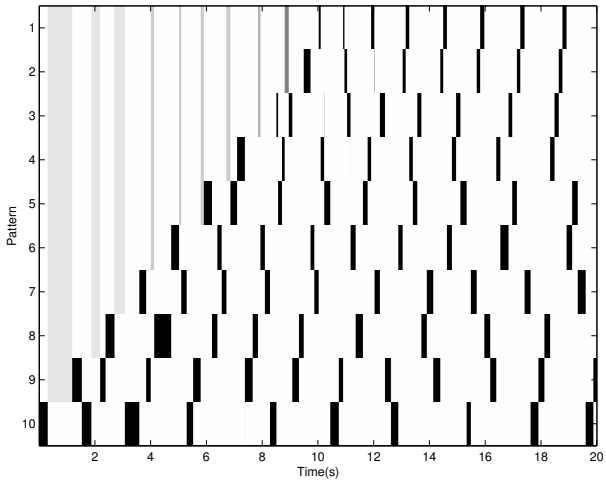


Figure 5.11. Overlap with quasi-attractors (10 orthogonal patterns) as a function of time for online learning. Plasticity and adaptation are active at all times. At the left an input of the training patterns is given with delays between each pattern. The network begins to repeat earlier shown patterns, eventually learning them all.

with biology and could be an intrinsic way of achieving reinstatement of earlier learned information for memory consolidation purposes or free recall.

In general the dynamics appears to exhibit sensitivity to initial conditions without any external source of noise. It is not a random walk between states, although the frequency of visits to quasi-attractors appears proportional to their basins of attraction. Which quasi-attractor is chosen after a given attractor depends on the extent of overlap and adaptation state in a complex way, but the likelihood of ending up in a particular attractor is on average a decreasing function of its uniqueness s_i and a convex and increasing function of the print-now signal during encoding. Member units of a quasi-attractor with high overlap with other quasi-attractors will on average be more depressed than units of an unique quasi-attractor, giving such quasi-attractors an advantage in the competition.

The loss of stability of the quasi-attractor state corresponds to a saddle-node bifurcation with the adaptation acting as a control parameter. As it increases, the radius of the basin of attraction surrounding the current quasi-attractor state decreases. Eventually the basin boundary meets the quasi-attractor state and the stability changes. This bifurcation is affected by parameters that change the effective transfer function such as λ_0 and g_A/g_L. In general the network shifts between quasi-attractors when $g_A/g_L > 1$; the adaptation has to be large enough to overcome the underlying weight matrix. However, as demonstrated in section 5.6 moves

between attractor states can occur for $g_A/g_L < 1$ if an external input is present. The effect of the input is to bias the dynamics of the network, essentially tilting the energy landscape in a certain direction and hence changing the stability of some attractors enough to allow shifts between them.

The dynamics can be interpreted in the BCPNN framework as the deliberate (temporary) exclusion of a certain hypothesis about the world state. After adaptation the effective network weights and biases approximate a situation where the original state has not been learned rather than merely inhibiting the state.

A network such as this repeating its contents has many similarities to connectionist models of serial order such as response competition and competitive queuing models (Houghton, 1990; Houghton and Hartley, 1995). Each quasi-attractor state can be viewed as being queued by its level of adaptation. A strongly adapted state is placed at the end of the queue, a weakly adapted state at the front. Due to competition only one state will become active at a time, and that will likely be the least adapted state.

A limitation of this kind of inhibition or unlearning-based method of accessing multiple memory states is that it will tend to return to earlier states. If the number of stored states is large the first visited states will have recovered from their adaptation before the weaker states have been visited, and the network returns to the early states. While the addition of noise can cause ergodicity, another possibility is to add a second adaptation projection with a time constant on the order of the time needed to pass through all stored patterns, penalising the early patterns to give later patterns a chance. Such an extended adaptation can also counteract the primacy seen in online learning. Biologically it does not appear unlikely that adaptation occurs on a range of timescales, and it does not require the time-varying control signals used in competitive queueing models (Houghton, 1990). A prediction is that pharmacological disruption of slow adaptation effects in biological networks would cause them to more easily fall into short cycles of activity.

As a reinstatement model the adaptation dynamics allows stored patterns to be replayed to other parts of a large learning system with a dwell time based on the strength of encoding. Preliminary results (Liljenkrantz, 2003) show that a model based on adaptation-driven reinstatement can produce results similar to Alvarez and Squire (1994).

This model has many similarities to the models proposed by Lansner and Fransén (1992) and Cartling (1997), which are based on spiking neurons with firing-rate adaptation. As adaptivity is increased the fixed-point attractors become unstable and a limit-cycle or periodic behaviour occurs. Wu and Liljenström (1994) and Cartling (1996) suggested that modulation of the adaptivity could be used as a hierarchical search process, where associative recall starts with a high level of adaptation causing the dynamics to move around widely, gradually narrowing in on the desired category as the level of adaptation was reduced and eventually retrieving the information as a fixed-point attractor. A similar dynamics could be implemented in this network by modulating g_L together with g_A; as shown by Eriksson and Lansner (2003) modulation of g_L can act as clustering, and an additional

modulation of g_A would determine the amount of exploration.

The dynamics is also very similar to the one reported in pure synaptic depression models (Bibitchkov et al., 2002; Torres et al., 2002; Pantic et al., 2002). As remarked in Pantic et al. (2002) cellular adaptation and synaptic depression are to some extent equivalent. This model shows how unlearning also can be seen to belong to this class of similar (but possibly not formally equivalent) models.

Chapter 6

Working Memory

6.1 Introduction

The mechanisms behind the cognitive functions of the human brain remain enigmatic. Relevant experimental data at the microscopic level of ion channels, synapses and neural activity as well as the macroscopic level of psychophysics and cognitive psychology is currently rapidly accumulating. Computational models provide a way to bridge the gap between these levels of description and to integrate information from multiple sources into a coherent picture. A good predictive model allows the experimenter to design maximally informative new experiments.

The memory systems of the brain are key players in cognitive functions. They exist in several different forms and have been characterised along dimensions like episodic-semantic and declarative-procedural. Here we focus specifically on working memory, the retention or maintenance of information for short periods of time, usually linked with an ongoing behavioural task. The paradigmatic experiment has been the delayed response task (see (Goldman-Rakic, 1995, p1)). In the oculomotor version of this task (Funahashi et al., 1989), the location in visual space has to be remembered over a few seconds after which a suitable response should be generated.

For a long time, computational theories and models of memory processes addressing the cellular and network level have focused on long-term memory. More recently, working memory processes have attracted the attention of modellers (Durstewitz et al., 2000; Compte et al., 2000; Wang, 2001; Tegnér et al., 2002). In contrast to network models of long-term memory, current working memory models rely mainly on reverberatory, persistent activity (Goldman-Rakic, 1995; Amit and Brunel, 1997a; Compte et al., 2000) in a network with fixed connectivity. The line attractor version of memory originally proposed as a model for the visual hypercolumn (Ben-Yishai et al., 1995; Hansel and Sompolinsky, 1996) has been the starting point for modelling this kind of working memory, assumed to reside in the prefrontal cortex (PFC).

These models, however, largely disregard important characteristics of the underlying neuronal substrate like synaptic plasticity on the task relevant time-scale 0.1-10 s (Hirsch and Crepel, 1990; Hempel et al., 2000) and neuronal adaptation, which is known to destabilise line attractors (Laing and Longtin, 2001). Their operation is also sensitive to distractors unless additional stabilising factors are included (Compte et al., 2000). In addition, working memory is generally thought to be able to hold several items (7±2) at the same time. A robust mechanism for this has not been demonstrated in the persistent activity working memory models.

Here we investigate the alternative hypothesis that short-term Hebbian plasticity is sufficient to account for the phenomenology in WM tasks. The presentation of a stimulus induces the formation of a corresponding attractor state which can later be read out as reverberatory activity. Persistent activity is still an integral part of this hypothesis but it now acts as an indicator of which stored memory is currently relevant and active. Short-term forms of memory based on fast synaptic plasticity have previously been suggested (von der Malsburg, 1981). A unification of different memory mechanisms acting on a range of time-scales is an attractive consequence of such hypotheses.

Since the Hebbian property is crucial in attractor memory models our hypothesis suggests the existence of fast Hebbian synaptic plasticity in the underlying cortical memory networks, e.g. the PFC. The existence of such forms of synaptic plasticity has not yet been experimentally established, but remains an open possibility.

In the following, we investigate this hypothesis in the context of a recently described attractor network model capable of acting as a long-term as well as short-term palimpsest memory and described in chapter 3.

6.2 The network simulation model

The BCPNN acts as a palimpsest memory where new information overwrites old and memories decay at a rate set by a learning time constant modulated by a print-now signal as in chapter 4. By temporarily up-regulating the print-now signal it is possible to imprint relevant stimuli while partially over-writing information already stored. While the print-now signal is zero no weight changes occur. We have hypothesised that the print-now signal could correspond to dopaminergic modulation, which is well represented in the PFC and facilitates synaptic plasticity (Wickens and Kötter, 1995; Otani et al., 1998; Durstewitz et al., 1999; Cohen et al., 2002).

This model allows us to simulate long-term as well as intermediate and short-term memories. In a short-term memory of this type the effective capacity is set by how strongly new information is imprinted, at the same time forcing old information to decay. With a print-now signal above a certain level, the memory becomes episodic, i.e. no repetition of the stimulus is necessary.

The BCPNN learning rule has previously been used to set the weights in a cortical network model implemented with biologically detailed compartmental model neurons and with cortical minicolumns as its functional units (Fransén and Lansner,

1998). That study suggests that it is reasonable to view the units of the BCPNN network as cortical minicolumns.

To model cellular adaptation and synaptic depression/facilitation we use a simple phenomenological model. The adaptation is modeled as temporary 'unlearning' of the same form as the learning rule. It is implemented as an associative projection with negative gain and a short time constant (see chapter 5). This results in a decrease of the effective synaptic weight between units active together for a long time as well as an increasing negative bias, thus removing valleys in the energy landscape while the network state remains in them. When units are inactive the projection returns to a low value, restoring the original synaptic efficacy. Since the learning and unlearning processes are equivalent (up to a gain factor and time constant) they will balance each other.

6.3 Setup of the delayed oculomotor task

As in previous bump state models of working memory, the canonical experiment was based on the oculomotor delayed response task. In this task a monkey is trained to fixate on a central mark on a screen during a brief presentation of a peripheral cue. The gaze remains on the mark during a subsequent delay period until a signal is given for the animal to make a saccade to the cue position.

In our model the network was first subjected to a period of no input corresponding to the pre-trial period. This was followed by a 300 msec cue stimulus and a simultaneous print-now signal which caused an update of synaptic weights. After the cue the network spent 3 seconds in a delay period with no input and no print-now signals.

After the delay period a reset signal was given in the form of strong stimulation to all neurons in the network together with another print-now signal. While models with persistent activity only need to reset the neural activity itself (Laing and Chow, 2001; Gutkin et al., 2001), in this model the reset signal is also assumed to be associated with a print-now signal erasing the changed synaptic weights. This print-now signal could be the same as that imprints the next stimulus, acting as a gating signal (Durstewitz et al., 1999).

The network consisted of 100 units fully connected to each other with a total normalized activity (this is similar to the setup in (Ben-Yishai et al., 1995); the hypercolumn in this model is, in fact, analogous to several identical hypercolumns of the type used in earlier chapters connected to each other). Each unit received input with different spatial tuning, corresponding to a projection from a population of location sensitive cells in the parietal lobe. For simplicity of display the units were ordered according to their favoured orientation. Each target angle θ corresponded to an input of the form

$$I_i = Ze^{-|i-\frac{\theta N}{2\pi}|^2/\sigma^2} \tag{6.1}$$

Symbol	Parameter	Value
	Cue length	300 msec
	Delay period length	3000 msec
	Reset period length	300 msec
N	Number of neurons	100
σ	Input tuning width	10
g_I	Input gain	1
g_A	Adaptation gain	0 or 0.35
g_N	Noise gain	0.1
τ_L	Learning time constant	7200 msec
τ_c	Membrane time constant	10 msec
τ_A	Adaptation time constant	160 msec
$\kappa(t)$	Print-now signal	0 or 1, 90 for reset

Table 6.1. Default parameters of the model.

where Z is a normalisation constant, N is the number of neurons and $\sigma = 10$. The distance is assumed to wrap around the population, producing a ring-shaped network metric. If not otherwise stated the input gain (g_I) was set to 1, the learning time constant (τ_L) to 7.2 s and the print-now signal to 1 or 0.

As the in the chapter 5 simulations the time constant of the adaptation projection τ_A has (where not otherwise stated) been given a value of 160 msec, corresponding to the decay rate of the action potential related Ca^{2+} pool that contributes to the accumulated after-hyperpolarization in the previous biophysically detailed pyramidal cell model (Fransén and Lansner, 1995). In the initial experiments reported below $g_A = 0$, while in the subsequent experiments with adaptation $g_A = 0.35$.

The noise input to the support was Gaussian with mean 0 and variance 1 and a default gain (g_N) of 0.1. In the inhomogeneity experiments synaptic strengths and time constants (which were set individually for each synapse) were subjected to noise. The noise was uniformly rather than normally distributed in order to avoid negative time constants and synaptic sign reversals.

The model parameters have been collected in Table 6.1.

6.4 Results

The network was tested in the simulated delayed oculomotor task, both with a single and several cues to remember. Different distractors were added during the delay period and the tolerance to noise and parameter inhomogeneity was investigated.

The network exhibited one-shot learning of a bump shaped attractor state. The bump activity profile can be viewed as a tuning curve. When exposed to a cue input consisting of a single bump for 300 msec, the network was able to sustain the bump throughout the delay period due to the change in local excitatory connections

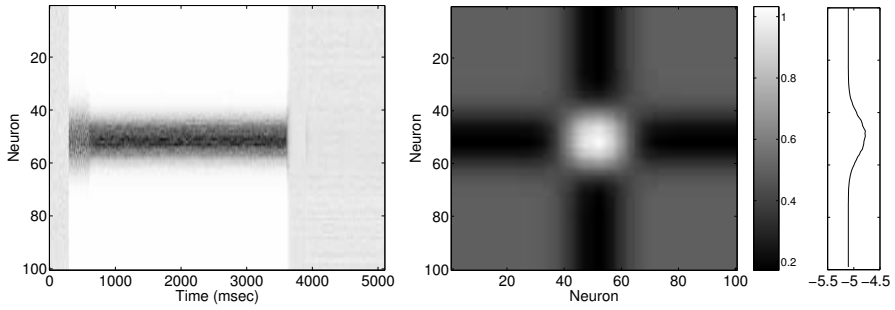


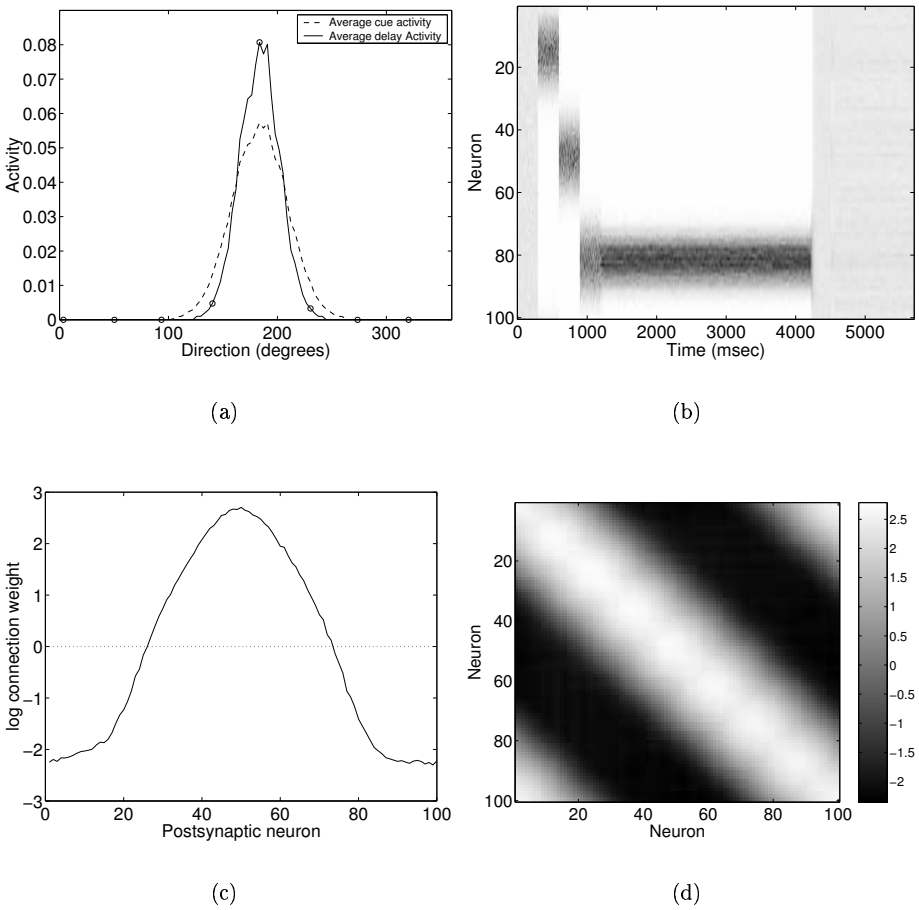
Figure 6.1. (a) Activity of network during a simulated oculomotor delayed response task. Activation of network units is marked by gray-scale. The cue exposure caused a broad activation between 300–600 msec, followed by a persistent activity state until the reset signal at 3600 msec. (b) Weight matrix after the cue period. Bright areas correspond to excitatory interactions and dark areas to inhibition. The units near the target direction have formed a self-exciting group, with lateral inhibition to and from all other units. (c) Bias of the units. Most units have a very low bias, but those units close to the target direction have become more excitable.

and long-range inhibition (Figure 6.1). The increase in lateral inhibition stabilised the location of the bump while the increased local excitation induced a persistent activity state. The reset signal abolished the synaptic changes, thus dissolving the persistent activity state.

The activity of units not corresponding to the cue was reduced during the delay period relative to the spontaneous activity before the cue or after the reset signal.

As a rule, the bump state was resistant even to high levels of noise and did not drift, since its position was determined by the weight matrix and the neuron biases. Since it was the only attractor state of the network the activity profile resumed its shape and location if disturbed or if the activity was reset. Experiments with increasing levels of noise showed a broadening of the bump state, until the noise amplitude was so high that it dominated over the recurrent input. At the same time the population vector remained fixed to the peak of the original cue, i.e. there was no true drift.

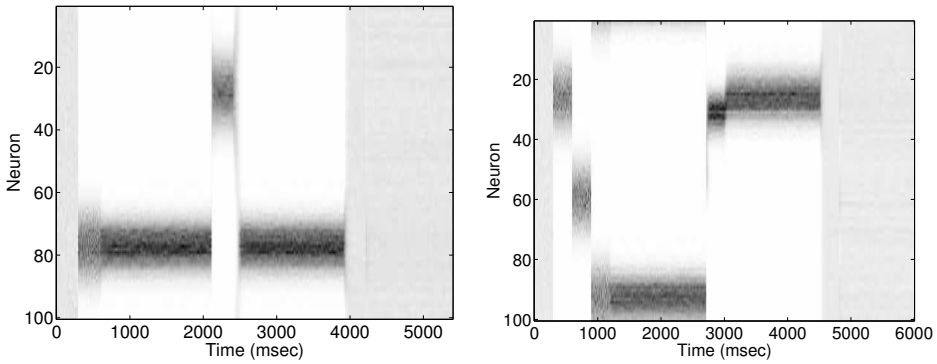
An interesting phenomenon was the sharpening of the bump state relative to the cue signal during the delay period (Figure 6.2(a)). This was due to the creation of a weight matrix with a nonlinear relationship to the original bump size. The learning process of attractor networks does not in general guarantee memory states identical to the input causing them, but it tends to create stable states in the vicinity of the input. The exact relationship between the shape of the original cue signal and the persistent activity pattern will in general depend on details of the learning rule used



(a) (b)

(c) (d)

Figure 6.2. (a): Typical bump activity profile for the network. Mean activity during the cue and delay period for one trial. Eight equidistant points have been marked for comparison with (Funahashi et al., 1989, fig 9). Shape fluctuations due to a noisy cue are imprinted in the delay activity, as can be seen near the top. (b): Activity when the network was exposed to three cues during the cue period. This created a synaptic matrix with three metastable memory states, each individually similar to the state in Figure 6.1. The activity remained in the state corresponding to the last cue until being reset. (c): Plot of weights from a single neuron after learning 8 bump attractors. (d): Weight matrix after learning 8 targets.



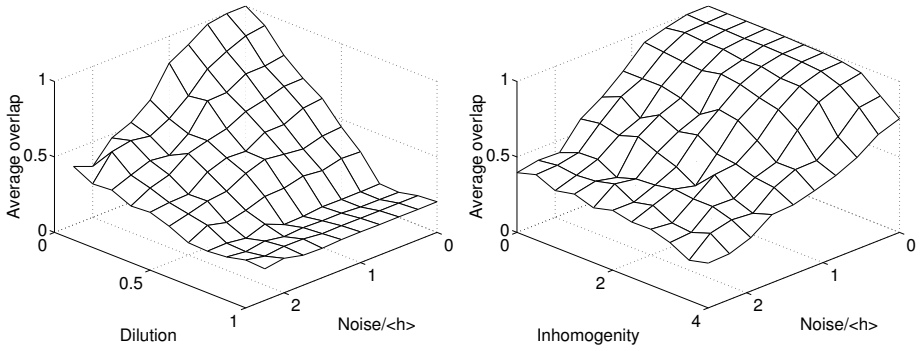
(a) One target (b) Three targets

Figure 6.3. (a) If one pattern was stored, distractor input had no lasting effect: the activity returned to the stored target. (b) If several targets were stored sufficiently strong input allowed the network to shift between them. In the three target simulation the distractor occurred at position 40, but the resulting bump was strongly attracted to the memory state at position 25.

and its time constants; hence biological observations of these relationships can be used to support or rule out proposed network models.

If several patterns were presented in turn during the cue phase (each pattern for 300 msec), the last one remained active (Figure 6.2(b)). If a sequence of targets were shown during the cue period the weight matrix developed into a band matrix with translation symmetry corresponding to a ring attractor family (there is an asymmetry caused by the aging of the memories, but for the current parameters and cue exposure times this is small). Each unit developed excitatory connections to units with similar spatial tuning and inhibitory connections to remote units (Figure 6.2), similar to the rotation invariant pre-wired synaptic matrices used in previous bump state models.

These multiple bump states were metastable attractors: a sufficiently strong stimulus ("distractor") could shift the network state to one of the other stored patterns (Figure 6.3). However, distractors had only temporary effects when a single target had been stored. For increasing input gain it became possible to temporarily shift the bump towards other positions, but when the input ended the network returned to the stored bump state. The shift appeared to be "elastic" and non-linear: for small input gains the distance moved under the influence of external input was proportional to the gain, up to a critical level where the bump instead moved directly to the input direction. When several targets had been stored the distractor would cause a bump intermediate between the distractor location and



(a) Noise and dilution

(b) Noise and inhomogenity

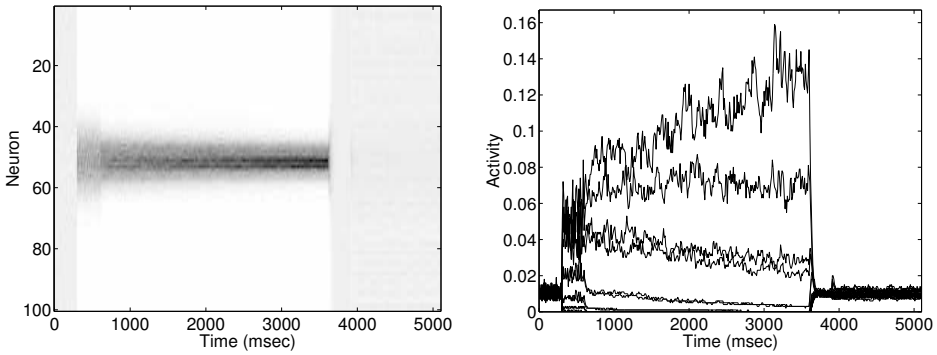
Figure 6.4. Network performance as a function of noise level, synaptic dilution (left) and standard deviation of time constants (right). The network was trained with 8 attractors and placed in one of them. The average overlap between the network state and the original pattern in the period 800-1000 msec afterwards was plotted. Noise was measured relative to the average value of the unit potential h, dilution by the probability of removing a synapse. Time constants were uniformly distributed around 7.2 s. Average over 100 trials in each point.

the closest target; the bumps would be strongly attracted to the stored target. For a large number of targets the behaviour was similar to other line-attractor models, including fast virtual rotation between the original location and the stimulated location.

The network was extremely resistant to network inhomogeneity, synaptic noise and sparse connectivity. Stable, if noisy, bump states persisted both for multiplicative synaptic noise where weights were multiplied by uniform random numbers between 0 and 7 (not shown in figure) and noise injected into the units. Similarly, the learning time constants could vary over a large range without the loss of bump states. Dilution of the connections caused a graceful degradation (Figure 6.4).

The network was also resistant to selective removal of connections between units with distant receptive fields or random removal of connections in specific regions, although the loss of mutual inhibition enabled the co-existence of multiple bumps in different regions at sufficient dilution.

An interesting effect was seen when the print-now signal was not completely turned off during the delay phase, but kept on at 0.1 strength. Strongly activated neurons tended to link to each other, creating an attractor growing smaller and sharper (Figure 6.5). This produced gradual decay of some activities and increase of others, a ramping behaviour similar to what has been observed in unit recordings in working memory experiments (Fuster, 1989; Romo et al., 1999).



(a) (b)

Figure 6.5. (a) Single stored target, with slow learning ($\tau_L = 72$ s) during delay stage. (b) Plot of selected unit activities over time.

6.4.1 Adaptation

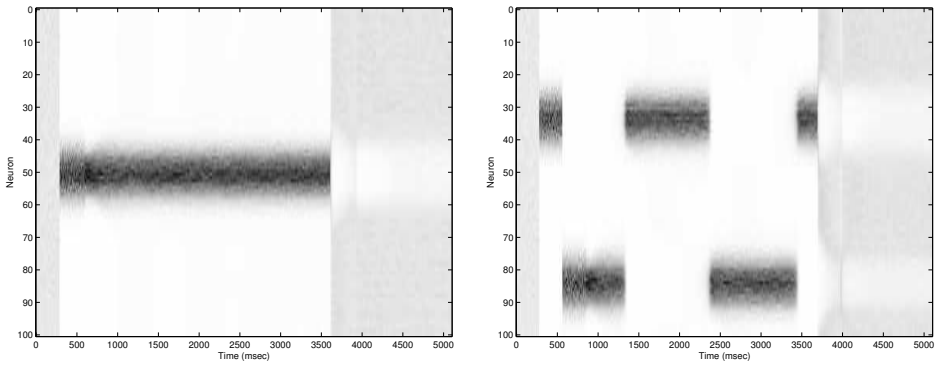
With adaptation turned on ($g_A > 0$) the stability of attractor states became time dependent.

For low values of adaptation the bump state remained stable and the system behaved as before. For strong adaptation the bump state instead became oscillatory, first decaying to a state where most neurons were weakly active and then returning to the bump shape again once the initially adapted neurons and synapses had regained their efficacy. This cycle then repeated itself.

For intermediate values of g_A, the network exhibited a stable bump when trained with one target, while moving between two or more bump states when trained with several targets (Figure 6.6). Hence it was possible to both have a single stable state and time multiplexed multiple attractors without having to change the time constant or strength of adaptation.

The upper limit of the number of targets that could be stored when adaptation occured appeared to be (for these parameters) approximately 10. For more targets increasing overlap between the attractors created a band weight matrix with a single continuous attractor state. Adaptation then caused the delay state to turn into a continous moving wave instead of discrete shifts between targets. This is similar to the adaptation-induced instability of bump states in models with pre-wired band matrices (Laing and Longtin, 2001). For a smaller number of stored targets the regions of strong local excitation in the weight matrix were separated by mutual inhibition, producing a barrier for the translation of the bump state from one location to another.

The network was also tested with several task cycles (cue-delay-reset) in order



(a) One target (b) Two targets

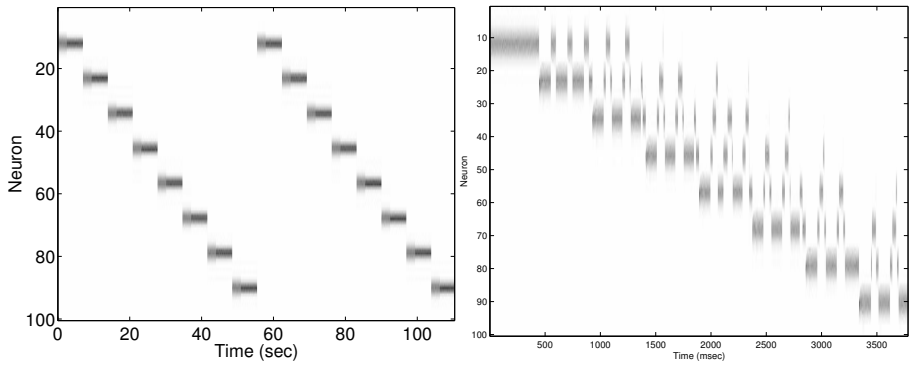
Figure 6.6. Network activity with adaptation, for $g_A = 0.35$. The bump state was stable for a single learned target, while it flipped between different locations if two or more targets were stored.

to check for inter-trial interference. If the imprinting was strong enough previous memory states were abolished (Figure 6.7(a)). For less intense imprinting traces of previous memory states remained and would periodically re-appear when adaptation was used (Figure 6.7(b)). It is worth noting that an explicit reset signal to erase attractor states was not necessary if new cues were imprinted strongly enough; the new cue would erase the old stored information (Figure 6.7).

6.4.2 Non-hebbian plasticity

While the current model has been based on the assumption of the existence of very fast Hebbian synaptic plasticity, there is currently no clear evidence for the presence of this in prefrontal cortex. Hence another related model would assume the existence of a predetermined synaptic matrix (possibly set by long-term synaptic plasticity) and a fast non-Hebbian change in neuron bias such as spike frequency adaptation caused by the cue, or synaptic facilitation. One observed phenomenon that would fit this is the synaptic augmentation observed in prefrontal neurons, which consists of a 40–60% enhancement of synaptic transmission which can be induced after 0.3 s stimulation. This has been suggested as a stabilizing factor for working memory (Hempel et al., 2000).

We used a synaptic matrix based on a continuous attractor and only allowed the bias β_i to change during learning. When exposed to a cue target the network formed a bump attractor which persisted. However, the center of the bump described oscillations around the stored target (Figure 6.8) due to adaptation. When this



(a) $g_A = 0$ (b) $g_A = 0.7$, $g_I = 30$

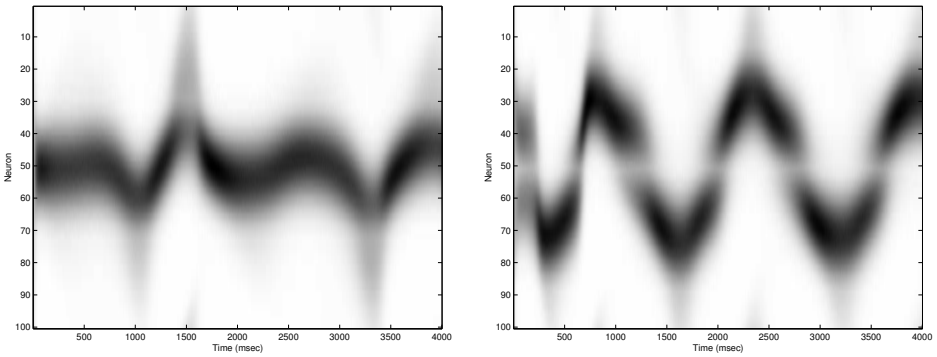
Figure 6.7. Multiple trial cycles. In each trial one cue is learned, remains sustained and is reset by the arrival of the next cue. To the left is a case without adaptation. To the right is a case with adaptation where incomplete erasure occurs, and the network state shifts between the previous four targets. Note the preference for the most recent target.

network was trained with two attractors activity oscillated between them and their centers. The focusing effect of having discrete regions of local excitation separated by long-range inhibition is lacking, enabling the adaptation instability to move the center of the bump state widely.

6.5 Discussion

We have described an attractor neural network with fast Hebbian plasticity that performs as a working memory in the oculomotor delayed response task. It exhibits bump states similar to previous models while being based on synaptic plasticity rather than a hard-wired weight matrix. The modified connectivity maintains the persistent activity and stabilises it against noise, distractors and network inhomogeneity. Unlike line-attractor models this network can store multiple memory states, with external cues activating one memory at a time.

In most models to date the maintenance of line attractors requires fine tuning of network parameters (Wang, 2001; Seung et al., 2000). This network avoids the problem, and the main parameter issue instead becomes setting learning parameters to a range suitable for the task at hand, which is far less sensitive. Learning also stabilises the network against inhomogeneity and noise, making it possible to maintain bump states even when the individual neurons have different parameters and connectivity.



(a) One target (b) Two targets

Figure 6.8. Network activity in model with fixed synaptic matrix and fast change of bias, with adaptation. The attractor states oscillate around the given cue directions.

This is somewhat similar to the models of eye position control and invariant object recognition described by Seung (Seung, 1996, 1998) where line attractors are formed through learning. However, in these models the learning is assumed to occur on a far slower time-scale than the network dynamics. Also, our model is not based on the assumption that the learned attractors form or approximate a line attractor; while multiple overlapping stimuli can produce it, it is no requirement in our model for a functional working memory. It may further be possible that slower forms of plasticity create an underlying line attractor weight matrix, which is transiently modified by faster plasticity (von der Malsburg, 1981).

Persistent activity is an integral part of this model. There is no fundamental contradiction between the hypothesis that persistent activity within cell assemblies is necessary for working memory and the hypothesis that working memory is stored by fast changes in synaptic efficacy. Both possibilities can be combined; the activity is necessary for readout to activate the proper behavioural responses and synaptic plasticity is necessary or helpful in maintaining the activity.

It has been argued that a good reason to assume that working memory is stored as persistent activity is that experimental disruption of such activity causes the animal to make an error (Durstewitz et al., 2000). However, this would also be the case if the memory was stored synaptically and read out as reverberatory activity. Then instead of interfering with the storage as such the read out from memory would be disturbed. Thus, these two hypotheses cannot be distinguished based on such experimental results.

The existence of sufficiently fast Hebbian plasticity is a basic assumption of this model. While non-Hebbian plasticity of suitable speed is known (Hempel et al.,

2000), the existence of fast Hebbian synaptic change in the prefrontal cortex remains conjectural. Long lasting (one day) LTP at hippocampo-prefrontal synapses can be induced by trains of five short bursts of 10 pulses at 200 Hz given at the frequency of the theta rhythm (Doyere et al., 1993), but the time before the LTP is expressed appears to be of the order of seconds to minutes (Gustafsson et al., 1989; Hirsch and Crepel, 1990). On the other hand it is possible that there can exist Hebbian effects within posttetanic potentiation (Bao et al., 1997) and that short-lived Hebbian effects are just not easily demonstrated in the current experimental setups. Non-Hebbian fast adaptation like e.g. augmentation is a less radical hypothesis for the stabilisation of bump states (Hempel et al., 2000), but appears to be insufficient to adequately maintain bump attractor states against destabilising influences such as adaptation.

The extension of the basic model to include adaptation enables attractor dynamics that moves between the different stored states. Due to the existence of a single attractor state in the one cue case, the network avoids the destabilising effect of adaptation reported by Laing and Longtin (Laing and Longtin, 2001). Adaptation acts as an inhibitory current which is preferentially enhanced on the trailing side of a moving bump state, causing it to drive the motion forward. In this model this does not occur due to the non-band structure of the weight matrix, unless enough overlapping memories have been stored to create such a band matrix.

For the same parameter values but with several cues it can also sustain time-multiplexed activity where the network state jumps between the memory states during the delay period. This is similar to the switching in the binocular rivalry model of Laing and Chow (Laing and Chow, 2002) where a "slow" dynamics of spike frequency adaptation and synaptic depression cause bump states to shift between two externally influenced positions.

This appears to partially fit results such as (Constantinidis et al., 2001) where it was found that the presentation of two stimuli (one task relevant, one irrelevant) led to delay activity in neurons tuned to both locations. If the weight matrix was of the band form it would not be able to sustain two simultaneous bumps, unless the connectivity between units tuned to distant orientations were very weak.

In addition to assigning a functional role in working memory for cellular and synaptic phenomena like plasticity and adaptation, the model presented here makes several specific predictions:

- There could be inter-trial interference due to incompletely erased synaptic changes. Mistakes should tend to confuse the current and a similar previous target, as has been observed (Bichot and Schall, 1999). The presentation of several targets at the same place followed by a target nearby should lead to a probability of mistake increasing with the number of times the first target was shown and decreasing with the distance to the second target.
- There should exist a print-now signal resetting the synaptic state of the network concurrent with the reset of neuron activity at the end of the task, regardless of success or failure.

- The shape of the input to the working memory network determines the memory state shape. Hence a modification of the form of the input should result in a corresponding change in the delay activity. If, for example, the previously sharply spatially tuned input was replaced with a broadly tuned input more cells would become less active in the delay period. This would not happen in a fixed synapse model, where the bump shape is set by the pre-existing synaptic structure.

- A prediction and a possible problem with this kind of model is the lack of drift. Although it remains uncertain whether drift of bump states plays a major role in the task (Wang, 2001), the difference between the saccade target and the achieved target in oculomotor tasks appears to increase monotonically with delay time (Funahashi et al., 1989; White et al., 1994; Ploner et al., 1998). In the current model it could be due to the presence of other processes, such as a high intrinsic noise level combined with weak learning during the delay period.

Since the network is not based on pre-set synaptic strengths, it can be generalised to arbitrary attractor states. For example, it can be used to learn from a 2D retinotopic map, enabling the persistence of 2D bumps of activity at locations learned during the cue period. The attractor states could also be distributed representations as in regular attractor networks.

Instead of using a specific kind of network for working memory and another for long-term memory, our model suggests the possibility that the cerebral cortex may be using a canonical network architecture with a spectrum of learning time constants for different functions. Such general networks would become modality specific due to their afferent and efferent connections rather than any particular architectural features, and shift their function due to task demands. This could explain the apparent discrepancy between different studies on the domain specificity of PFC (as discussed in (Undergleider et al., 1998)).

Though we investigated within the context of a specific computational model (BCPNN) we expect the qualitative aspects of our results to be largely independent of the exact simulation model. A possible future extension would be to implement the model using spiking neurons to further examine its generality and properties, such as the effect of spike synchrony in memory reset and the effects of AHP modulation on network dynamics.

In conclusion, we have shown that fast Hebbian learning is sufficient to reproduce many properties of a working memory task which has previously been modelled in terms of persistent activity states in a fixed connectivity attractor network. It remains to be examined to what extent plasticity is necessary to maintain stable activity and how quickly it can be modulated to fit the task.

Chapter 7

Mental Ageing

Plastic processes in the brain occur throughout the lifespan of an animal, but their nature and intensity vary over time. During development and infancy extensive neurogenesis and synaptogenesis occur, linked with experience-dependent synaptic pruning and myelinization. Beyond adolescence, learning plasticity appears to change, possibly due to declines in neuromodulator levels and changes in receptor expression. The overall theme appears to be a steady decline of plasticity from a high level at infancy.

This paper examines a possible evolutionary explanation for age-related episodic memory impairment as antagonistic pleiotropy, the phenomenon that genes with deleterious effects at late ages can be actively selected if they have beneficial effects at young ages. Senescent traits evolve because they have only weak effects on fitness (Medawar, 1952; Rose, 1991; Rose and Mueller, 1998). The effect of a phenotype trait on an individual's total fitness declines monotonically with the age at which it is expressed (Hamilton, 1966). Under a wide variety of conditions a modest fitness benefit early in life can off-set a larger disadvantage later in life that only affects the end of reproductive life (Charlesworth, 1980).

The force of natural selection in humans becomes essentially zero after age 40 (Rose and Mueller, 1998, fig 1, based on data from Charlesworth and Williamson (1975)). Hence changes in brain plasticity that are adaptive for young individuals and that continue beyond maturity to cause impairments in learning would have a net fitness benefit and be selected for.

7.1 Age-Related Memory Impairment

Memory decline in aging humans is associated with both disuse, disease and aging per se. Factoring out disuse and disease, there still appears to exist age-dependent memory impairment although the individual variations are large (Zec, 1995; Small, 2001).

Many age differences in memory can be explained in terms of a decreased mental speed that limits the encoding efficiency and affects decision times (Cerella, 1985; Korsnes and Magnussen, 1994). Aging problems in long-term memory are largely due to problems in encoding and retrieval rather than storage, and rates of forgetting have not been found to be different in old and young adults (Kaszniak, 1986; Zec, 1995).

The variance of memory performance in aging samples increases with age, suggesting that memory decline is not inevitable. However, twin studies show increased correlation between memory measures with age. This suggests genetic factors predisposing towards age-related memory impairments (Rapp and Amaral, 1992; McClearn et al., 1997; Small, 2001; Nilsson et al., 2002).

fMRI data from a feature binding task shows a greater activation in the hippocampus of young but not old adults, suggesting that the decline in such tasks may be due to hippocampal dysfunction (Mitchell et al., 2000).

Age-related volume changes of different brain regions are observed (Jernigan et al., 2001). However, there is considerable controversy over the presence and role of neuron and synapse loss in the hippocampus and cortex in normal aging (Morrison and Hof, 1997; Scheff et al., 2001; Terry and Katzman, 2001); the consensus appears to be emerging that cell loss is not a major cause of memory impairment outside neurodegenerative disorders. On the other hand, cell loss appears to occur in the subcortical modulatory systems (de Lacalle et al., 1991).

During aging human neuromodulation changes. Striatal dopaminergic function averages a 6%-10% decline per decade (Scherman et al., 1989; Rinne et al., 1993; Wang et al., 1998) and age-related dopamine activity decreases appears to impair both motor performance and frontal cognitive functions (Volkow et al., 1998; Kaasinen and Rinne, 2002). Bäckman et al. (2000) found that dopamine D_2 receptor binding was a more important factor than chronological age in predicting variation of perceptual speed and episodic memory performance. Since prefrontal function is important in cognitive function and sensitive to neuromodulatory changes (Arnsten, 1998) this is likely to contribute to age-related impairments.

Rat studies have produced roughly similar results, complementing the human cognitive aging perspective. There is a decline in temporal processing speed for sounds (Mendelson and Ricketts, 2001) similar to human processing slowing. Neurodegeneration occurs both in the hippocampus and among the basal forebrain cholinergic neurons of the aged rat, but hippocampal neuron loss is not inevitable and performance does not correlate with the loss (Rapp and Gallagher, 1996). Aged rats exhibit stable and accurate place fields in the hippocampus during a learning episode, but between episodes they are often rearranged (Barnes et al., 1997). Aged rats with memory impairments show encoding of only part of available context information in the hippocampus, and when cues or tasks change they show decreased plasticity (Tanila et al., 1997; Oler and Markus, 2000). Overall there appears to be a decrease in ability to bind context with internal representations.

The threshold of hippocampal LTP increases with age while the LTD threshold decreases, and LTP decay rate becomes faster (Foster, 1999). Comparisons between

spatial memory performance and LTP decay rate in young and aged rats showed similar relative differences (Barnes and McNaughton, 1985).

Expression of NMDA receptors decline in a subtype-dependent manner in both hippocampus and the cortex (Sonntag et al., 2000). Especially relevant is the decline of expression of the NR2B subunit with age (Clayton and Browning, 2001). Young animals express NR2B almost exclusively compared to NR2A, and the ratio decreases during postnatal development (Sheng et al., 1994). Overexpression of the subunit causes memory improvements (Tang et al., 1999), likely due to the slower kinetics of NR1-NR2B heteromers inducing a more reliable LTP (Monyer et al., 1994).

There is a decrease in cholinergic synaptic transmission over rat lifespan (Shen and Barnes, 1996) and aged rats with memory deficits show a decline in neurotrophin signalling in the basal forebrain (Sugaya et al., 1998). Nucleus basalis lesions caused a more severe impairment in middle-aged and aged rats than in young adult rats, suggesting that the young rats have compensatory responses that are lost with aging (Wellman and Pelleymounter, 1999). ACh is most likely involved in attention and arousal rather than memory per se in these tasks, and the interactions between modulator systems can ameliorate or worsen the impairments (Shen and Barnes, 1996). Lesion experiments show that diffuse cell loss in the basal forebrain cholinergic system most closely mirror aging related memory impairments (Gallagher and Colombo, 1995). Reaction times and response speeds become more variable in rats with dopamine decline (MacRae et al., 1988).

These results imply that cognitive aging may at least partially be due to decline in neuromodulatory tone and receptor expression rather than structural changes. This decline affects brain systems important for memory encoding and management such as the hippocampus in rats and humans and the prefrontal cortex in humans. The different neuromodulatory systems change at different rates both within the same individual (causing differential changes between memory systems) and between individuals (causing increasing performance variance). This decline may be due to genetic factors. In addition, the changes in subunit composition of the NMDA receptor appears to be a regular progression rather than the result of accumulation of errors.

7.2 Episodic and Autobiographical Memory

Episodic memory was originally defined in terms where recall brings with it a sense of time and place of the recalled experience (Tulving, 1972). This is commonly combined with the idea of episodic memories as autobiographical memory, memories and memory functions that relate closely to a persons own narrative. Nevertheless researchers in the autobiographical memory field often distinguish between the two kinds, and may even regard them as different memory systems (Conway, 1990). In the following we will not make any strict distinctions between the two kinds, but

rather focus on the similar properties of being distinct experiences learned through one-shot learning that are recallable across the lifespan.

The main property of autobiographical memory of interest here is the likelihood of retrieval of experiences from different time periods using cued and free recall.

One of the main features of autobiographical memory is the differences in the frequency of reported autobiographical memories depending on their age (Conway, 1990; Rubin et al., 1998). Typically such frequency plots show a low number of accessible memories from the first years of life (infantile or childhood amnesia), an increase in frequency from a particular age range in the second and third decades of life, the "autobiographical bump", and a recency effect for memories from recent years that declines as a power function (Rubin and Schulkind, 1997). The bump is statistically detectable in 50-year-olds and above, but not in 40-year-olds and below (Chu and Downes, 2000).

Normally autobiographical memories are triggered in experiments using verbal cues. If odour cues are used instead, the peak occurs at 6–10 years and then decreases for higher ages. The peak is also higher for the odour cues than the verbal cues (Chu and Downes, 2000).

Childhood amnesia appears to be a universal phenomenon, where autobiographical memories of early childhood become inaccessible. It has been explained in terms of repression, different self-representations, a specific amnestic period or cognitive development (Pillemer, 1998). Social and cultural factors can affect the extent of amnesia, likely due to how much parents discuss early memories with their children (MacDonald et al., 2000). Part of early infantile amnesia is likely due to neurological immaturity. Parts of the hippocampus such as the dentate gyrus are still developing years after birth, and hippocampal-dependent learning is absent before 18-24 months of age (Mangan and Nadel, 1990; Newcombe et al., 1998). At the same time the children are clearly able to do semantic learning such as language acquisition, showing that different memory systems mature at different rates.

Autobiographical memories appear to be linked to prefrontal brain systems. In a fMRI study by Maddock et al. (2001) the caudal part of the left posterior cingulate cortex was most strongly activated during retrieval of autobiographical information cued by names. Maguire et al. (2001) observed increasing activity in the ventrolateral prefrontal region with increasing recency of autobiographic and public event memories (but not the hippocampus).

7.3 Cognitive Aging Models

Most models of cognitive aging are qualitative psychological models rather than quantitative computational models. One strand of models has been based on the concept of diffusely distributed neuron death causing cognitive impairments even in healthy aging (Coleman and Flood, 1987). Birren (1965) proposed a general slowing of processing speed as a hypothesis of the primary cause of cognitive aging. Welford (1965) discussed increased neural noise as another potential cause. The information

loss model of Myerson et al. (1990) explains the slowing due to increased loss of information at each processing step.

Neural network models of synaptic deletion and compensation have been studied as models for Alzheimer's disease, exhibiting gradual degradation of performance and relative sparing of old memories compared to recent (Ruppin and Reggia, 1995). While based on assumptions of heavy neuron/synapse losses, the basic mechanism could be applied as a model for memory decline due to cell death. The compensatory mechanisms could contribute to the inconsistencies found in many studies between memory function and cell loss.

Braver and Barch (2002) present a model based on the PDP framework. Information is held within an active recurrent memory linked to a stimulus-response pathway. A dopamine signal acts both as a learning signal and a gating signal, regulating how much the active memory affects and is affected by the rest of the system. Impairments in these signals cause impairments in context processing, which fit psychological and neuroimaging studies. Proper cognitive control is also likely to be important for long-term memory encoding.

The model of Li et al. (2001) is also based on the assumption of decreasing neuromodulatory tone, but instead treats it as a declining gain of the transfer function in network units (although a change in gain in a backpropagation neural network is equivalent to a scaling of initial synaptic weights and learning rate (Li and Sikström, 2002)). This reduces the distinctiveness of internal representations and increases neuronal noise, causing impairments in many cognitive domains as demonstrated in a series of neural network simulations (Li et al., 2001; Li and Sikström, 2002).

7.4 Evolutionary Neuroscience

Can there be an evolutionary account for cognitive changes in aging? If information acquisition correlates with reproductory fitness at a fertile age, then the learning rate should change so as to maximise the amount of relevant stored information in the reproductory period.

Since this number of experiences is extremely high, and likely far beyond the capacity to store individually, only selected experiences will be stored. This can be achieved by temporary print-now signals (see section 2.1.2 and chapter 4). However, even if only relevant information is stored it will tend to disrupt earlier information. This is especially important for episodic memory, which has to be laid down quickly.

If the learning rate is high early in life when the most new information arrives and then decreases, accidental erasure of old information would be minimised.

The "grandmother/grandfather" hypothesis that long-lived grandparents improve the fitness of their grandchildren might imply that there is an additional benefit to having a few slow-learning, slow-forgetting elderly within the family group. They would maintain knowledge from their own youth that could be beneficial to their kin when rare situations occur (Mergler and Goldstein, 1983; Rubin

et al., 1998). However, this paper does not deal with kin selection effects, so this intriguing hypothesis will not be included in the model.

The current hypothesis is that a network with plasticity decreasing with time at a suitable rate will maximise the total amount of information that can be recalled at a certain age. Evolution would favour organisms that maximise the total amount of available information at reproductive age within the constraints of their nervous system (whose capacity is assumed to be already given; the plasticity traits would be secondary to the anatomical traits that determine the disposition of the nervous system).

At this optimal rate, childhood amnesia will occur due to the plastic loss of the earliest years, and aging-related memory impairments due to lack of plasticity at ages beyond the normal reproductive and child-rearing age.

7.5 Optimal Learning Rate

In the following, we develop a simple toy model of a plastic memory that can be analytically solved in certain cases.

Let $I(t)$ be the strength of the memory traces of information learned at time 0. The learning rate $\alpha(t) \geq 0$ is variable, and affects $I(t)$ through decay proportional to $\alpha(t)$ (corresponding to interference, inhibition effects and trace decay):

$$I'(t) = -\alpha(t)I(t) \tag{7.1}$$

The original strength of the memory trace depends in a nonlinear way on $\alpha(0)$:

$$I(0) = f(\alpha(0)) \tag{7.2}$$

where it is assumed that $f(\alpha) > 0$ and monotonically increasing. The general solution of 7.1 and 7.2 is

$$I(t) = f(\alpha(0))e^{-\int_0^t \alpha(u)du}$$

The total amount of memory trace that has been acquired at time T from traces starting at time 0 and onwards is

$$J(T) = \int_0^T f(\alpha(t))e^{-\int_t^T \alpha(u)du}dt \tag{7.3}$$

$$= \int_0^T f(\alpha(t))e^{\int_0^t \alpha(u)du - \int_0^T \alpha(u)du}dt \tag{7.4}$$

Let

$$\beta(t) = \int_0^t \alpha(u)du$$

$$\beta'(t) = \alpha(t)$$

Equation 7.4 becomes:

$$J(T) = \int_0^T f(\beta'(t))e^{\beta(t)-\beta(T)}dt \tag{7.5}$$

We want to maximise $J(T)$ for a fixed T. This is a variational problem that can be solved with the Beltrami identity: if we seek to extremise $\int f(y, y')dx$ the extremal y satisfies $f - y'\frac{df}{dy'} = C$ for some constant C. Applying this to 7.5 we get:

$$[f(\beta'(t)) - \beta'(t)f'(\beta'(t))]e^{\beta(t)-\beta(T)} = C \tag{7.6}$$

Since $\beta(T)$ can be taken as a parameter it can be included into C, leaving

$$[f(\beta'(t)) - \beta'(t)f'(\beta'(t))]e^{\beta(t)} = C \tag{7.7}$$

7.5.1 Convex f

An important special case occurs if f is convex, $f'' < 0$. $e^{\beta(t)}$ is increasing, which implies that the expression within the brackets must decrease over time in order to make the entire expression constant. Differentiating and using $\alpha(t) = \beta'(t)$ we get

$$f'(\alpha(t))\alpha'(t) - \alpha'(t)f'(\alpha(t)) - \alpha(t)\alpha'(t)f''(\alpha(t)) < 0$$

$$\alpha(t)\alpha'(t)f''(\alpha(t)) > 0$$

Since $\alpha(t) > 0$, $\alpha'(t) < 0$. Hence for a convex learning efficacy f, the optimal learning rate must decrease over time.

The response to a learning stimuli commonly exhibits a convex form, which can be represented by $f(\alpha) = \alpha^k$ for $0 < k < 1$. Putting this into 7.7

$$[(\beta')^k - k(\beta')^k]e^{\beta} = C$$

$$(1-k)(\beta')^k e^{\beta} = C$$

$$(\beta')^k e^{\beta} = C/(1-k)$$

$$e^{\beta/k}d\beta = (C/(1-k))^{1/k}dt$$

$$\int e^{\beta/k}d\beta = \int (C/(1-k))^{1/k}dt$$

$$ke^{\beta/k} = (C/(1-k))^{1/k}t + D$$

$$\beta(t) = k\log[\frac{(C/(1-k))^{1/k}t + D}{k}]$$

Since $\alpha(t) = \beta'(t)$ we get

$$\alpha(t) = \frac{k(C/(1-k))^{1/k}}{(C/(1-k))^{1/k}t + D}$$

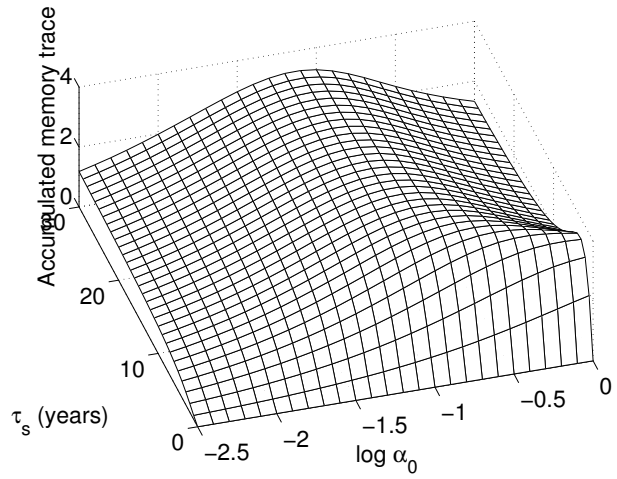


Figure 7.1. Total memory trace $J(T)$ at 25 years of age for $\alpha(t) = \alpha_0 2^{-t/\tau_S}$.

C and D are arbitrary constants, so the general form is

$$\alpha(t) = \frac{k}{t+E} \tag{7.8}$$

where E is a combined constant.

Hence, for learning responses of the form $f(\alpha) = \alpha^k$ $0 < k < 1$ the optimal learning rate has to decrease as $1/t$. Also, for more sharply convex learning efficacies (small k) the learning rate should decrease.

For learning responses of the form $f(\alpha) = \log(\alpha+1)$ such as the one in chapter 4 the corresponding differential equation has no solution in closed form. However, numerical solution shows that $\alpha(t)$ declines fast, and can be approximated with an exponential curve. The effect of using this is shown in figure 7.1. $J(25)$ has a maximum for $\tau_S \approx 6$ years, $\alpha_0 \approx 0.25$, although the values do not change much along the ridge.

7.5.2 Simulation Model

We compared the model of previous section with a neural network simulation of lifespan change of memory plasticity using a BCPNN. The plasticity change was represented by an decrease of the learning rate (the inverse of the time constant τ_L):

$$\alpha(t) = \alpha_0 2^{-t/\tau_S} \tag{7.9}$$

where α_0 is the initial plasticity and τ_S the time constant of plasticity change.

This network model abstracts both the size and time dynamics of memory. The network used 100 units organised into 10 hypercolumns and was exposed to one random pattern to store for each "year" of life. These patterns represented all the experience of the year.

The fast change in time constant ($\tau_S = 10$ years) is very different from the far milder decline in dopamine receptors observed ($\approx 10\%$ per decade). However, mapping from actual synaptic plasticity to this abstract network is not straightforward. The plasticity in the model corresponds to the combined effect of plasticity and interference between a very large number of patterns, while the learning processes in the brain are dependent not just on individual receptor numbers and their plasticity-modulating effect but also their network effects on a long chain of memory systems. The functional relationship translating receptor numbers or modulator concentrations into a general learning rate of the entire brain is likely highly nonlinear, even when leaving out the inverted-u curve effects of high concentrations.

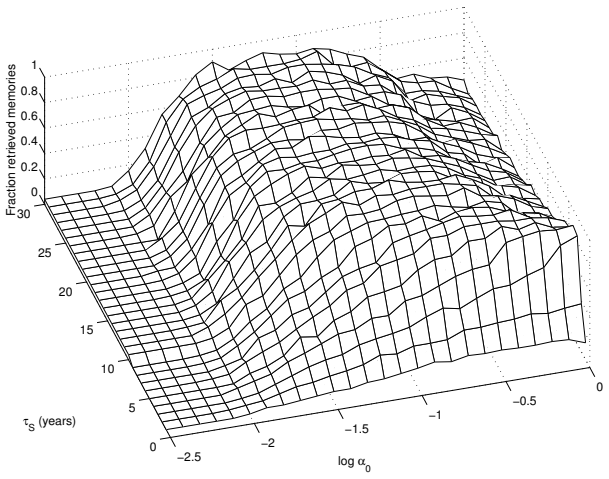
To store m times as many patterns the network needs to be approximately $m^{2/3}$ times larger (Sandberg et al., 2002; Johansson et al., 2001). A network that can store one pattern per day would be at least 51 times larger ($N = 5100$) and for a network storing one pattern per minute 6512 times larger ($N = 651,200$). At the same time the equation 7.9 would be rescaled as $\alpha_m(t) = \alpha_0 2^{-t/m\tau_S}$. However, as long as the network is not operating close to its capacity limits the dynamics would be essentially unchanged. Similarly, a smaller network learning fewer patterns can be used as a model of the larger, more realistic network.

The form of plasticity change in equation 7.9 was selected due to the similarity to neuroanatomical results suggesting a rough exponential decline of receptor levels. The rate of decline should be interpreted as a composite effect of actual memory trace decline, interference and encoding variation. The exact value does not map directly to modulator and receptor levels.

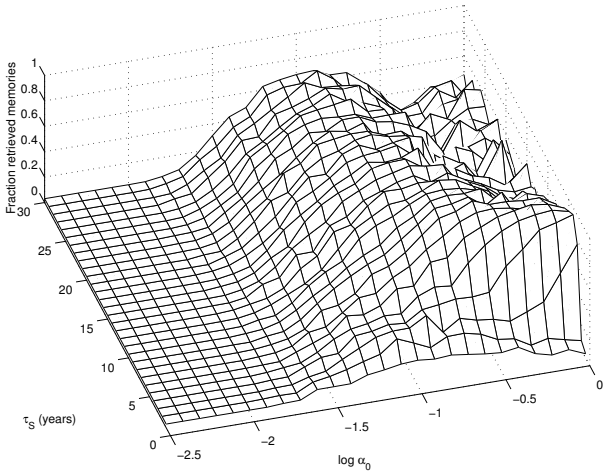
Cued recall was performed from initial states with half of the hypercolumns randomly activated, and the success condition was more than 0.9 overlap after convergence for 2.0 time units. Free recall was performed using convergence from random initial conditions, and if the system state after convergence overlapped more than 0.9 with a stored pattern it was regarded as recalled. Free recall using adaptation as described in chapter 5 was also tested, and performed similarly.

7.5.3 Simulation Results

Figure 7.2 shows the performance at 25 years of age of cued and free recall as a function of α_0 and τ_S. The ridge represents the optimal combination of initial learning rate and decline. The maximum of the cued recall performance is a curve where all patterns are recalled correctly; it is indifferent to the exact choice of α_0 and τ_S. For free recall there is an optimum for very fast learning rates declining very quickly. This is very similar to figure 7.1, showing that the simple mathematical model and the network give nearly the same predictions.



(a) Cued recall



(b) Free recall

Figure 7.2. Learning performance at 25 years of age as a function of α_0 and τ_S for cued recall and free recall. Performance was estimated by the fraction of successful cued retrievals and free recalls. Average of six simulations; for free recall 100 attempts were made.

Recall was tested at age 25, 50 and 80 years for $\alpha_0 = 0.4$, $\tau_S = 10$ (Figure 7.3). In general cued recall was able to retrieve most memories with high probability from a noisy cue, with the exception of early memories and age-related memory decline at old age. Free recall tended to focus on a subset of memories peaking between 20-35 years of age. Cued recall also exhibited a bump, but only when tested at higher ages.

7.6 Discussion

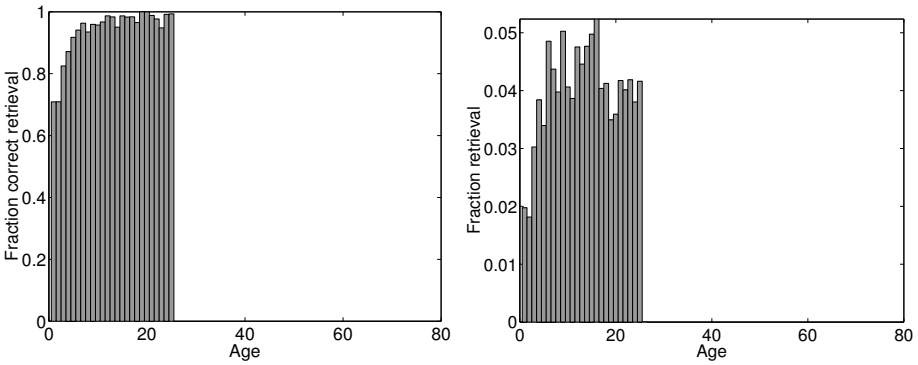
The network model presented here is a simple extension of the one studied in previous chapters of this thesis. It is small and extremely simplistic, but it does account for some features of autobiographical memory as a consequence of an evolutionary account of optimal learning. It leaves out the many effects of disuse and disease, as well as the complex interactions of memory systems, repetition and social feedback. Still, the analytic model shows that under a wide range of conditions a declining learning rate would be favoured by evolution and the neural network simulation manages to produce a plausible autobiographic memory curve exhibiting childhood amnesia and the autobiographical bump.

The recall curves fit well with the early and middle part of empirical autobiographic recall curves, but lack the observed recency. This lack may be due to the lack of a medial temporal lobe in the model. If the MTL is assumed to play a role in "recent" (tens of years) episodic memories (as is suggested by Nadel and Moscovitch (1997)), then it could act as a intermediate-term "add-on" to the long-term cortical memory modelled here similar to the short-term memory in Johansson (2001). Such an extra short-term memory can add its capacity to a long-term memory, contributing to the kind of recency curve observed (Rubin and Schulkind, 1997).

The highly nonlinear decline in learning rate derived from theory and producing autobiographic curves appears to fit observations of relative rapid decline of dopamine transporters during young adulthood followed by less rapid declines during middle age Mozley et al. (1996). This would also fit the observation that noradrenergic innervation in rat frontal cortex decline at an earlier stage of aging (9–13 months) but not at a later stage (13–25) (Ishida et al., 2001).

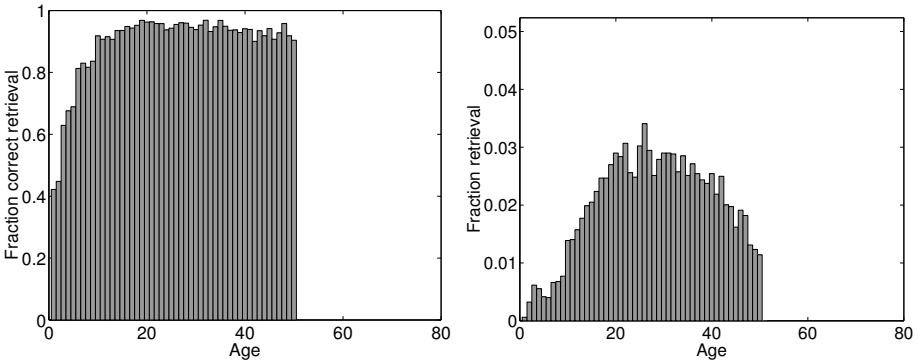
This antagonistic pleiotropic model suggests that the late life decrease in neuromodulators and general plasticity should not just be due to random late-life genes that are not selected away, but a continuation of the genetic programs that underlie the plasticity changes in early life. A "wear-and-tear" model would predict memory impairments due to random cell loss and expression changes with no pattern relating to memory. This model instead predicts a specific decline in memory-related systems.

How can this hypothesis be tested or falsified? One approach would be to use mice overexpressing the NR2B gene (Tang et al., 1999) and study their lifespan learning. The NR2B mice would have a higher overall learning rate leading to a greater degree of interference, reducing their ability to recall remote events.



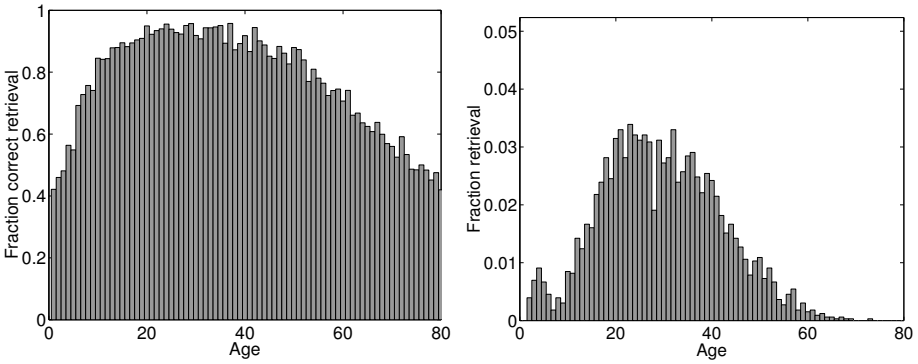
(a) Cued recall, 25 years

(b) Free recall, 25 years



(c) Cued recall, 50 years

(d) Free recall, 50 years



(e) Cued recall, 80 years

(f) Free recall, 80 years

Figure 7.3. Cued and free recall of memories at 25, 50 and 80 years of age.

The genetic regulation of the NR2B:NR2A ratio should remain active across the lifespan, decreasing it at a speed dependent on the age of reproduction. Hence interspecies comparisons of this system should show a dependency on reproduction age rather than maturation age. Similarly the factors determining the neuromodulatory decline should be at least partially under genetic control and dependent on the lifecourse of the species.

The fitness (as measured in 7.2) appears to have a relatively flat optimum, suggesting that the fitness disadvantage of mutants with slightly different rates or shapes of plasticity decrease might be relatively slight. This would predict that the variation of such traits could be relatively large.

Another prediction is that periods with high levels of neural plasticity, for example produced by high neuromodulator levels, would exhibit an increase in frequency of autobiographical memories. This would in a sense be similar to the plasticity modulation modelled in chapter 4 projected on lifelong learning, and would be able to cause similar inhibition effects. A study of Vietnam war veterans hints at such an artificial bump effect in veterans with PTSD (McNally et al., 1994), although interpretation of the data is complicated by social and psychological factors.

Many extensions of the model can be explored. An important extension of the current model would be to include the effect of reinstating earlier information, allowing it to re-imprint itself in memory. The overall effect is a strengthening of already strong or often accessed memories and a weakening of weak or rarely accessed memories. Reinstatement is likely a significant factor both in shaping the general autobiographical memory curve and in childhood amnesia.

This model has treated memory as a learning system black box with no internal structure. More advanced models are possible, and could be used to examine the differential effects of aging on different brain regions and memory systems. A simple parallel model with memory subsystems declining at different rates could easily model the different frequency curves observed for verbal vs. odour cues in Chu and Downes (2000) by assuming that the plasticity of the odour-related memory systems decline faster than for verbal systems.

The current model does not account for early development well, since it does not include the gradual emergence of representations. A more realistic model would consist of a hierarchy of subsystems undergoing maturation at different rates where higher level subsystems require functional low-level systems for their own function. The organism would gain a fitness benefit from a declining learning rate in the lowest subsystem (corresponding, perhaps, to perceptual representations) that quickly learned a stable representation. The second subsystem would be largely useless as long as the first subsystem had not stabilised its representations, and hence it would be adaptive for it to have a slower or delayed maturation rate. But once the lowest subsystem had matured, the learning situation would again be similar to the assumptions in the model, and there would be a fitness advantage in a declining learning rate. This scheme can be continued at increasing hierarchical levels, each with a delay or slowing of learning rate decline relative to the previous level. This would be especially relevant for modelling early development and autobiographical

memory.

The methodology of network models such as that of Li et al. (2001) are complementary to this model. They take the modulatory decrease as the basis and construct a bottom-up model of how the decrease in gain causes memory impairments. This model starts with evolutionary assumptions and derives a declining learning rate over lifespan, which might be expressed in terms of decrease in neuromodulation. There is no fundamental incompatibility between the models in this respect. Rather, the present evolutionary perspective seems to be another possibility of cross-level integration in sensorimotor and cognitive aging that fits with the program (Li and Sikström, 2002).

Chapter 8

Discussion and Conclusions

This thesis started from a heuristic derivation of a neural network and learning rule from statistics, and then gradually extended its functionality by enabling parameters to vary dynamically or adding extra projections. This family of networks were then compared to other models and used to describe different memory phenomena.

A main issue of study was whether there exists necessary architectural differences between neural networks involved in different memory systems such as working memory and long-term memory. A secondary issue was to develop a network that could act as a general robust building block for more complex models. The aim was to be able to construct "networks of networks" of simpler memories that interact in adaptive ways that can be used in more elaborate models of the brain and in "artificial nervous systems" for autonomous learning robots and agents.

8.1 Encoding and Retrieval

We have seen how manipulation of learning time constants within a simple neural network enables a wide spectrum of effects. The main finding was that the same basic network architecture can act as a model of working memory, intermediate memory or a long-term memory depending on its level of plasticity. The plasticity can also be modulated to achieve enhancement of certain memories, both on a short behavioural timescale and over long timescales in order to optimise lifespan learning. These forgetful networks can interpolate between a mode where the number of learnable memories depends on the level of plasticity and a mode where capacity has an upper bound dependent on network size and catastrophic forgetting occurs if they are overloaded.

Is the brain limited by plasticity or size? Catastrophic forgetting is not observed even among the (healthy) very aged, suggesting that either the number of distinct experiences during a human lifetime is smaller than the capacity of cortical memory or that we have some plasticity limitation. Events occurring less than

100 milliseconds apart are usually perceived as simultaneous (VanRullen and Koch, 2003), and introducing new attractors in an online learning network such as the one in chapter 5 requires waiting for the current attractor to adapt, producing roughly the same timescale. Assuming on the order of ten memories per second produces an upper bound of the order of 10^{10} on the number of memories in a human lifetime. This could be fitted into a network of 10^8 minicolumn units (assuming on the order of 10^{10} cortical neurons (Pakkenberg and Gundersen, 1997) with minicolumns containing some 10^2 neurons). If they obey the same capacity scaling as a BCPNN with $\sqrt{N}$ hypercolumns the total capacity would be on the order of 10^{10} patterns. Hence there is likely no strong size limitation, since it is unlikely that all information is retained or even encoded. On the other hand encoding requires plastic change of the cortical networks, and since most learned experience – in order to become useful – requires connecting with or adjusting prior knowledge there is likely a great deal of inter-mnemonic interaction between memory states. Hence the real limiting factor of human memory (beside lack of encoding of many experiences) may be the slow change of synapses subjected to noise and non-specific modulatory input.

The mental aging model in chapter 7 is based on this assumption. Assuming that cortical plasticity rather than space limitations is the main cause of forgetting, it explores the optimal learning rate over the first part of a lifespan and extrapolates memory beyond the reproductive age, finding bump-like effect on memory similar to autobiographical memory. Assuming a size-limited brain would instead lead to a learning rate geared towards preventing catastrophic forgetting before reproductive age but maximising the information content at that age, implying a rapid mental decline at middle age not observed in humans.

While encoding of information into the network can be done in only one way, retrieval can be both by cued recall, free recall or through adaptation. As shown in chapter 3, 4 and 5, all three retrieval methods show roughly the same response to modulation of encoding strength. The size of the basin of attraction determines the amount of noise that can be removed in cued recall as well as the likelihood of ending up in the pattern due to recall from a random initial state and the time spent in the corresponding quasi-attractor during adaptation. Hence control of the size of the basins through plasticity modulation can regulate the strength of the memory traces, a function that can be important in determining the likelihood of retrieval, lifespan of memories and how much they are reinstated during consolidation processes.

The introduction of adaptation and the resulting quasi-attractor dynamics enables a more complex retrieval process which is more sensitive to external input and overlaps between different stored patterns. It also can act as a reinstatement system and allows on-line learning interleaved with recall. The system is still controlled by a small number of parameters: the gain of the different projections, the time constants of learning and adaptation and the auxiliary λ_0. Using these parameters the formation and control of attractors (or rather, their basins of attraction) can be regulated and through them the dynamics of the system.

While the fast plasticity in the adaptation model is purely a phenomenological

model of synaptic and cellular adaptation, it is an integral part of the working memory model of chapter 6. Here fast Hebbian plasticity is used to form a working memory able to reliably store information over behaviourally relevant timescales. While the model still lacks biological verification it shows how a simple assumption can produce a network exhibiting complex dynamics that other ring attractor models have to fine-tune.

8.2 BCPNN as a Memory Model

Attractor neural networks are plausible models of memory since they are naturally associative, instantiate the cell assembly hypothesis and fit the highly recurrent cortical architecture. The convergence to an earlier learned attractor state has many similarities to Gestalt perception and the reconstructive aspects of memory retrieval. Drawbacks for attractor memory models have been catastrophic forgetting, separation of learning and recall phases, identical encoding strength and their lack of dynamics beyond convergence. In this thesis these drawbacks have been ameliorated for the BCPNN family of networks by modifying the basic network or learning rule in various ways. Each step has been motivated by a perceived drawback as a memory model or the need for a particular feature, and have each been constrained to or inspired by biology. Incremental learning fits the reversible nature of LTP/LTD, the phenomenological adaptation model used mimics certain forms of cellular adaptation and synaptic depression and plasticity modulation fits theories of memory modulation. While this does not guarantee that the resulting models are correct models of biology, it does give them a link to biological plausibility without sacrificing the original level of abstraction.

The BCPNN family of memory models is attractive due to their direct interpretation in terms of probability estimation. Learning consists of updating probability estimates of external or internal events, while the flow of activity approximates inference. This can be viewed from a top-down perspective as a model of memory and plausible reasoning (Jaynes, 1996).

The requirement of synaptic plasticity or adaptation at the timescale of network dynamics used in the adaptation and working memory models removes us from much previous theoretical work in which neural dynamics and synaptic plasticity has been assumed to be uncoupled since they operate on very different time-scales. This was formulated by Caianiello as the "adiabatic learning hypothesis" (Caianiello, 1961, 1989) and removing this assumption makes theoretical analysis considerably harder. However, we know today that neurobiological processes do occur on a continuum of timescales from milliseconds to years with no gaps suitable for strict uncoupling.

One of the main drawbacks of the BCPNN approach is that on the way from Bayes' theorem to a functioning network assumptions and additional mechanisms are introduced making mathematical analysis intractable. The nonlinear summation of dendritic contributions, the deviations from the independence assumptions

as well as the normalisation deflect the normal tools of analysis and force a more empirical approach to exploring the properties of the network. This makes much of this thesis a prelude to experimental mathematics rather than a theorem-proof exposition. The results demonstrated show many intriguing possibilities that remain to be formalised and throughly explored for their own sake, regardless of applying the network to memory models. For example, a deeper understanding of the statistical mechanics of BCPNN and its representational capabilities still remains to be elucidated.

However, as demonstrated by the simplified models of chapter 4 and 7, the properties of the network can be described phenomenologically in useful ways that enables predictions about the network behaviour. Again this points at the benefit of having models at different levels of abstraction.

Overall the BCPNN has with relatively minor modifications been able to function as different kinds of memory systems, reproducing not just the expected timescale of storage but also other properties. This suggests that there is no necessary architectural difference between different memory systems, at least not the systems that have been studied here (working memory, intermediate memory, long-term declarative memory). This answers one of the main questions posed in the thesis and also opens up for building more complex structures from standard building blocks.

A number of simple networks are found in many brain areas and circuits, especially pattern associators, autoassociative networks and competitive networks (Rolls and Treves, 1998). Although these have very different functions they can be derived as phenotypes from the same model of genetic evolution (Rolls and Stringer, 2000). It is interesting to speculate that their functions may partially be derivable from the same architecture through different modulatory inputs, explaining how the cortex achieves its multifunctionality. Even if the number of synapses onto different neurons from different projections might not be the theoretically optimal values for a particular function in such a multifunctional network, there would likely be a fitness benefit to an organism living in a changing environment in having a "standard cortex" since the organism could adapt it to the environment and changing demands without the need to have a rigid specialisation.

8.3 Current Work

Analysis, extensions and application of the BCPNN model has been ongoing, and the results described in this thesis cover only part of this field.

One current area of research is applying the BCPNN to reinforcement learning and other learning paradigms. The BCPNNRL system of Johansson and Lansner (2002b) and Johansson et al. (2003) consists of several populations of BCPNN units representing world state, rewards and actions. Despite being derived in terms of unsupervised learning, preliminary results show that the BCPNN components can be combined into a reinforcement learning system with a performance comparable

to Monte Carlo learning and the Associative Reward-Penalty algorithm. BCPNN has also been applied to classical conditioning, where it reproduces a wide range of typical experiments such as extinction, blocking, inhibition and secondary conditioning (Johansson and Lansner, 2002a).

An important part of models of semantic memory is category learning and clustering. Originally explored in Levin (1995), it was extended to clustering based on varying the learning time constant and λ_0 (Gars and Tamsen, 1999) and gain (Eriksson and Lansner, 2003). These investigations have shown a rich clustering dynamics that can be regulated in several ways, including biologically plausible mechanisms such as gain and plasticity level during encoding.

Detailed cortical network models with spiking neurons and column structure are also being investigated, connecting the BCPNN architecture more closely with biological reality. In these models more biologically inspired cell models are used to achieve the properties of BCPNN units and hypercolumns.

The heuristic derivation of chapter 3 works well for spiking units, although the learning rule needs to define an additional timescale for spike coincidence. This can easily be done in the incremental framework by using higher-order exponential smoothing: one smoothed estimate of unit activity is used as the input to another estimate, which in turn may be input to a third and so on. This approach was the basis for the phenomenological model for spike-timing dependent plasticity of Wahlgren and Lansner (2001).

In a network unit activities can be used to drive Poisson-spiking output rather than rate codes, and preliminary experiments show that spiking BCPNN networks function well. Another benefit of a spiking rather than rate coded BCPNN is that the update equation (and efficient parallel implementation) becomes simpler. If one unit at a time is assumed to be active within a hypercolumn, only one term in the inner sum of equation 3.8 will be nonzero and the update can be viewed as summing log-weights to the support (Lansner and Holst, 1996).

8.3.1 Networks of Networks

Rather than being a homogeneous network the cerebral cortex consists of interconnected subsystems organised in more or less hierarchical patterns Scannell et al. (1995). Different memory systems also appear to form relatively dissociable groups which act together in cognition. Hence it is natural to move beyond single network models and study models where different networks are linked to each other.

In networks of buffers and memories such as the ones discussed by Baddeley (2000) it is necessary to have stores that do not distort incoming information according to old information but still take advantage of existing information. Simple BCPNN models of connected short-term and long-term memories show the ability to both learn quickly and retain information for longer, as well as perform simple binding between items in STM and LTM (Johansson, 2001).

Information transfer between network modules is another area worth exploring. When a set of items are learned by a memory store their encoding strength will be

dependent on the number of times they have been shown, which affects their probability of being reinstated during free recall or adaptation. This makes encoding of items in a secondary memory store through reinstatement introduce a further bias of encoding strength. While a faithful representation of experience would benefit from a linear exposure-reinstatement relation, nonlinearities in this relation could also have adaptive effects: if strong memories are significantly enhanced the secondary store would learn only the most relevant items (while the less relevant might be accessible in the primary store before being dislodged by new learning). If weaker memories are more likely to be retrieved the reinstatement would be noisier, but also less likely to lose information.

The second-best match behaviour suggests that the network can be used as a building block for making perceptual hypotheses. A single network will shift between internally consistent but alternative matches to a given input, providing hypotheses for downstream networks. Groups of networks receiving different input modalities could also influence each other through mutual connections, allowing different hypotheses to reinforce each other when the external evidence is not strong enough to force the networks into specific attractor states. This may also suggest a way for information from higher order association cortex to influence more primary sensory processing in a top-down way.

The modulatory system of chapter 4 has not been explicitly simulated in this thesis. Extending the model with an explicit neural network to detect significant input appears relatively simple and has partially been done in the reward system of Johansson et al. (2003). By adding several such modulatory systems affecting the print-now signal $\kappa(t)$ as a common output channel learning can be made dependent on different motivational factors as well as specific systems such as familiarity detection (which could plausibly be implemented within the BCPNN framework through a mechanism similar to Bogacz and Brown (2003)). An interesting issue is the learning of secondary reinforcers by prediction of primary (hardwired) reinforcers and how to maintain the overall stability of a system with self-regulated plasticity.

8.4 Open Issues and Further Research

One thing that is clearly lacking in this thesis is recognition and storage of temporal sequences rather than point attractors. In order to model more than just retrieval of contexts the network needs to handle temporal interrelations. The basic BCPNN framework does not make any statement on whether the features are features observed at the same time or at different times, but an implementation needs to represent time-dependent information in a suitable manner. For example, recognition of time series have been attempted using banks of delayed averages. However, retrieval of time series also requires that the network dynamics mirrors learned (arbitrary) patterns, which suggests the need for a different update dynamics than the current. One way of allowing temporal association is to have different

learning time constants for the probability estimates on the pre- and postsynaptic sides of the BCPNN learning rule (this is especially natural in the context of higher order estimates used in a spiking network). This would create an asymmetric weight matrix that could associate one state to the next. Other extensions involve variants of the adaptation dynamics of chapter 5. However, it remains to be seen how flexibly and optimally such approaches can produce sequence retrieval.

Another key area that remains to be developed is self-organising internal representations. At present the hypercolumn structure is given, either due to a particular form of data or just as an arbitrary architecture. In biological reality the division of labour among minicolumns likely occur through a complex process of competition and cooperativity; in the BCPNN this would correspond to a partitioning of units depending on e.g. correlation or mutual information measures (Holst, 1997). There is also a need for feature detectors and map formation in the BCPNN. Implementing anti-Hebbian learning is relatively easy using the same unlearning mechanism as in the adaptation model; hence some competitive map formation can likely be reproduced within the BCPNN framework.

One of the deepest issues with the BCPNN is whether the network has a privileged position in the space of neural network rules. Practical experience has shown that it is extraordinarily resilient to variations in implementation, estimate calculation and changes to the update or learning equations (both deliberate changes and accidental bugs) compared to other attractor type ANNs. While this is merely anecdotal evidence, it seems likely that there should exist a particular relationship between the attractor states of the network and the probability estimates in weights and biases that could have an information theoretic interpretation. The energy function for a recurrent BCPNN without hypercolumns (equation 3.9) appears to maximise the weighted sum of information components (Orre, 1998) $\sum_{ij} \log(P(x_{ij})/P(x_i)P(x_j))\hat{\pi}_i\hat{\pi}_j$ with an additional soft constraint of minimising $\sum_i \log(P(x_i))\hat{\pi}_i$. The first term favours $\hat{\pi}$ that fits the correlation structure of the world, while the second term can be seen as emphasising the most informative features. It appears likely that convergence corresponds to extremising some information measure such as the Kullback-Leibler distance or entropy. If this is so, then the different BCPNN networks discussed in this thesis are performing an information maximisation operation or an approximation to it, which could explain the robustness of its operation.

The interpretation of cell assemblies as self-consistent estimates of the world state provides an interesting starting point for deeper exploration of cortical memory. Beyond memory as mere information storage lies memory as formation of knowledge. The emergence of representations binding together different concepts, modalities and levels of abstraction into consistent wholes is the growth of knowledge, and associative retrieval/reconstruction is an important part of reasoning. Knowledge consists of prior information that help shape and direct newly acquired information, both in order to cause behaviour and to update what is known.

This process cannot fully be studied in isolation in single networks as in this thesis, but must be studied in terms of adaptive systems interacting with their

environment. Memory and knowledge are not just about retaining information but seeking it out and putting it into practical use, be it enhancing evolutionary fitness or achieving individual goals. Hence an exploration of memory needs to involve motivational systems orienting us towards relevant or rich sources of information:

> All men by nature desire to know. An indication of this is the delight we take in our senses; for even apart from their usefulness they are loved for themselves; and above all others the sense of sight. For not only with a view to action, but even when we are not going to do anything, we prefer seeing (one might say) to everything else. The reason is that this, most of all the senses, makes us know and brings to light many differences between things.
>
> By nature animals are born with the faculty of sensation, and from sensation memory is produced in some of them, though not in others. And therefore the former are more intelligent and apt at learning than those which cannot remember.
>
> – *Aristotle, Metaphysics*

Bibliography

Abbott, L., Varela, J., Sen, K., and Nelson, S. (1997). Synaptic depression and cortical gain control. *Science*, 275:220–224.

Abel, T. and Lattal, K. M. (2001). Molecular mechanisms of memory acquistion, consolidation and retrieval. *Current Opinion in Neurobiology*, 11:180–187.

Abeles, M. and Gerstein, G. L. (1988). Detecting spatiotemporal firing patterns among simultaneously recorded single neurons. *J. Neurophysiol.*, 60:909–924.

Abeles, M., Vaadia, E., and Bergman, H. (1990). Firing patterns of single units in the prefrontal cortex and neural network models. *Network*, 1:13–35.

Adams, D. L. and Horton, J. C. (2003). Capricious expression of cortical columns in the primate brain. *Nature Neuroscience*, 6(2):113–114.

Alberini, C. M. (1999). Genes to remember. *Journal of Experimental Biology*, 202:2887–2891.

Alvarez, P. and Squire, L. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proc. Natl. Acad. Sci USA*, 91:7041–7045.

Amari, S.-I. (1977a). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.*, 27:77–87.

Amari, S.-I. (1977b). Neural theory of association and concept-formation. *Biological Cybernetics*, 26:175–185.

Amit, D. (1989). *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press, Cambridge.

Amit, D. and Brunel, N. (1995). Learning internal representations in an attractor neural network. *Network*, 6:359.

Amit, D. and Brunel, N. (1997a). Dynamics of a recurrent network of spiking neurons before and following learning. *Network: Computation in Neural Systems*, 8:373–404.

Amit, D. and Brunel, N. (1997b). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex*, 7:237–252.

Amit, D., Gutfreund, H., and Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. lett.*, 55:1530–1533.

Amit, D. and Tsodyks, M. (1991a). Quantitative study of attractor neural networks retrieving at low spike rates I: Substrate - spikes, rates and neuronal gain. *Network*, 2:259–273.

Amit, D. and Tsodyks, M. (1991b). Quantitative study of attractor neural networks retrieving at low spike rates II: Low rate retrieval in symmetric networks. *Network*, 2:275–294.

Amit, D. J. (1994). The hebbian paradigm reintegrated: local reverberations as internal representations. *Behavioral and Brain Sciences*, 18(4):617–626.

Anagnostaras, S., Maren, S., and Franselow, M. (1999). Temporally graded retrograde amnesia of contextual fear after hippocampal damage in rats: within-subjects examination. *J. Neurosci.*, 19(3):1106–14.

Andersen, P., Sundberg, S., Sveen, O., Swann, J., and Wigström, H. (1980). Possible mechanisms for long-lasting potentiation of synaptic transmission in hippocampal slices from guinea-pigs. *J. Physiology*, 302:463–482.

Anderson, J., Silverstein, J., Ritz, S., and Jones, R. (1977). Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychological Review*, 84:413–451.

Arbib, M. A., Érdi, P., and Szentágothai, J. (1998). *Neural Organization: Structure, Function and Dynamics*, chapter 8, Cerebral Cortex. MIT Press, Cambridge, MA.

Aristotle (350a). *On Memory and Reminiscence.*

Aristotle (350b). *On the Soul (de Anima).*

Arnsten, A. F. T. (1998). Catecholamine modulation of prefrontal cortical cognitive function. *Trends in Cognitive Sciences*, 2(11):436–447.

Atkinson, R. and Shiffrin, R. (1968). Human memory: a proposed system and its control processes. In Spence, K. and Spence, J., editors, *The psychology of learning and motivation: advances in research and theory.* Academic Press, New York.

Atkinson, R. and Shiffrin, R. (1971). The control of short-term memory. *Scientific American*, 225:82–90.

Baddeley, A. (1966a). The influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology*, 18:302–309.

Baddeley, A. (1966b). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, 18:362–365.

Baddeley, A. (1999). Memory. In Wilson, R. and Keil, F., editors, *Encyclopedia of the Cognitive Sciences.* MIT Press, Cambridge, MA.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423.

Baddeley, A. and Hitch, G. (1974). Working memory. In Bower, G., editor, *The Psychology of Learning and Motivation*, pages 47–89. Academic Press, New York.

Bailey, C., Bartsch, D., and Kandel, E. (1996). Toward a molecular definition of long-term memory storage. *PNAS*, 93(24):13445–13452.

Bao, J.-X., Kandel, E. R., and Hawkins, R. D. (1997). Involvement of pre- and postsynaptic mechanisms in posttetaic potentiation at aplysia synapses. *Science*, 275:969–970.

Bao, S., Chan, V., and Merzenich, M. (2001). Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature*, 412:79–83.

Barkai, E. and Hasselmo, M. (1994). Modulation of the input/output function of rat piriform cortex pyramidal cells. *Journal of Neurophysiology*, pages 644–658.

Barnes, C. and McNaughton, B. (1985). An age comparison of the rates of acquistion and forgetting of spatial information in relation to long-term enhancement of hippocampal synapses. *Behavioral Neuroscience*, 99(6):1040–1048.

Barnes, C., Suster, M., Shen, J., and McNaughton, B. L. (1997). Multistability of cognitive maps in the hippocampus of old rats. *Nature*, 388:272–275.

Baudry, M. and Davis, J. L., editors (1996). *Long-Term Potentiation*, volume 3. MIT Press, Cambridge, MA.

Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances. *Biometrica (reprint of original article in Philos. Trans. R. Soc. London 53, pp. 370–418, 1763)*, 45:296–315.

Bäckman, L., Ginovart, N., Dixon, R. A., Wahlin, T.-B. R., Åke Wahlin, Halldin, C., and Farde, L. (2000). Age-related cognitive deficits mediated by changes in striatal dopamine system. *Am. J. Psychiatry*, 157:635–637.

Bear, M. (1996). A synaptic basis for memory storage in the cerebral cortex. *PNAS*, 93(24):13453–13459.

Bell, C., Han, V., Sugawara, Y., and Grant, K. (1997). Synaptic plasticity in a cerebellum-like structure depends on temporal order. *Nature*, 387:278–281.

Ben-Yishai, R., Bar-Or, R. L., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. USA*, 92:3844–3848.

Bhalla, U. and Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science*, 283:381–387.

Bi, G.-Q. (2002). Spatiotemporal specificity of synaptic plasticity: cellular rules and mechanisms. *Biological Cybernetics*, 87:319–332.

Bibbig, A. and Wennekers, T. (1996). Hippocampal two-stage learning and memory consolidation. In *Proceedings of the 13th European Meeting on Cybernetics and Systems Research*, Vienna.

Bibbig, A., Wennekers, T., and Palm, G. (1995). A neural network model of the cortico-hippocampal interplay and the representation of contexts. *Behav Brain Res*, 66(1-2):169–75.

Bibitchkov, D., Herrmann, J., and Geisel, T. (2002). Pattern storage and processing in attractor networks with short-time synaptic dynamics. *Network: Comput. Neural. Syst.*, 13:115–129.

Bichot, N. P. and Schall, J. D. (1999). Effects of similarity and history on neural mechanisms of visual selection. *Nature Neuroscience*, 2(6):549–554.

Bienenstock, E., Cooper, L., and Munro, P. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2:32–48.

Birren, J. (1965). Age changes in speed of behavior: its central nature and physiological correlates. In Welford, A. and Birren, J., editors, *Behavior, aging, and the nervous System*, pages 191–216. Thomas, Springfield, IL.

Bliss, T. and Collingridge, G. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361:31–39.

Bliss, T. and Gardner-Medwin, A. (1973). Long-lasting potentation of synaptic transmission in the dendate area of unanaesthetized rabbit following stimulation of the perforant path. *J. Physiol.*, 232:357–374.

Bliss, T. and Lømo, T. (1973). Long-lasting potentation of synaptic transmission in the dendate area of anaesthetized rabbit following stimulation of the perforant path. *J. Physiol*, 232:331–356.

Boccia, M. M., Kopf, S. R., and Baratti, C. M. (1999). Phlorizin, a competitive inhibitor of glucose transport, facilitates memory storage in mice. *Neurobiol Learn Mem*, 71(1):104–12.

Bogacz, R. and Brown, M. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, 13:494–524.

Bonnaz, D. (1997). Storage capacity of generalized palimpsests. *J. Phys. I France*, 7:1709–1721. December.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA.

Braitenberg, V. (2001). Brain size and number of neurons: an exercose in synthetic neuroanatomy. *Journal of Computational Neuroscience*, 10:71–77.

Braitenberg, V. and Schüz, A. (1991). *The anatomy of the cortex. Statistics and Geometry*. Springer, Berlin.

Braver, T. S. and Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews*, 26(7):809–817.

Bressler, S. L. (1990). The gamma wave: a cortical information carrier? *Trends in Neurosciences*, 13(5):161–162.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 10:12–21.

Brown, R. G. (1963). *Smoothing, Forecasting and Prediction of Discrete Time Series*. Prentice-Hall.

Brunel, N., Carusi, F., and Fusi, S. (1998). Slow stochastic Hebbian learning of classes of stimuli in a recurrent neural network. *Network*, 9:123–152.

Brunel, N. and Wang, X.-J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of Computational Neuroscience*, 11:63–85.

Buckner, R., Kelley, W., and Petersen, S. (1999). Frontal cortex contributes to human memory formation. *Nature Neuroscience*, 2:311–314.

Bugbee, N. and Goldman-Rakic, P. (1983). Columnar organization of corticocortical projections in squirrel and rhesus monkeys: similarity of column width in species differing in corticaul volume. *J. Comp. Neurol.*, 220:355–364.

Burle, B. and Bonnet, M. (2000). High-speed memory scanning: a behavioral argument for a serial oscillatory model. *Cognitive Brain Research*, 9:327–337.

Buzsáki, G. and Solt, V. (1995). Slow wave sleep contribution to memory consolidation. *Sleep Research Society Bulletin*, 1(2).

Cahill, L. and McGaugh, J. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends Neurosci*, 21(7):294–299.

Caianiello, E. (1961). Outline of a theory of thought processes and thinking machines. *Journal of Theor. Biology*, 2:204–235.

Caianiello, E. (1989). A theory of neural networks. In Aleksander, I., editor, *Neural Computing Architectures*. MIT Press.

Calvin, W. (1995). Cortical columns, modules, and hebbian cell assemblies. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 269–272. MIT Press, Cambridge, MA.

Camperi, M. and Wang, X.-J. (1998). A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *Journal of Computational Neuroscience*, 5:383–405.

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

Carpenter, G. and Grossberg, S. (1987). ART 2: Self-organization ofstable category recognition codes for analog input patterns. *Applied Optics*, 26:4919–4930.

Cartling, B. (1996). Dynamics control of semantic processes in a hierarchical associative memory. *Biol Cybern*, 74(1):63–71. published erratum appears in Biol Cybern 1996 Apr;74(4):385.

Cartling, B. (1997). Control of computational dynamics of coupled integrate-and-fire neurons. *Biol Cybern*, 76(5):383–95.

Cerella, J. (1985). Information processing rates in the elderly. *Psychological bulletin*, 98:67–83.

Charlesworth, B., editor (1980). *Evolution in age-structured populations*. Cambridge University Press, Cambridge.

Charlesworth, B. and Williamson, J. (1975). The probability of survival of a mutant gene in agestructured population and implications for the evolution of life-histories. *Genet. Res*, 26:1–10.

Charniak, E. and McDermott, D. (1985). *Introduction to Artificial Intelligence*. Addison-Wesley Publishing, Reading, Massachusetts. Chapter 8.

Christianson, S.-A., editor (1992). *Handbook of Emotion and Memory: Current Research and Theory*. Erlbaum, Hillsdale, NJ.

Christie, B., Kerr, D., and Abraham, W. (1994). Flip side of synaptic plasticity: long-term depression mechanisms in the hippocampus. *Hippocampus*, 4:127–135.

Christos, G. A. (1996). Investigation of the crick-mitchison reverse-learning dream sleep hypothesis in a dynamical setting. *Neural Networks*, 9(3):427–434.

Chu, S. and Downes, J. J. (2000). Long live proust: the odour-cued autobiographical memory bump. *Cognition*, pages B41–B50.

Clark, K. B., Naritoku, D. K., Smith, D. C., Browning, R. A., and Jensen, R. A. (1999). Enhanced recognition memory following vagus nerve stimulation in human subjects. *Nat Neurosci*, 2(1):94–8.

Clayton, D. and Browning, M. (2001). Deficits in the expression of the NR2B subunit in the hippocampus of aged fisher 344 rats. *Neurobiology of Ageing*, 22:165–168.

Cloues, R., Tavalin, S., and Marrion, N. (1997). β-adrenergic stimulation selectively inhibits long-lasting L-type calcium channel facilitation in hippocampal pyramidal neurons. *Journal of Neuroscience*, 17(17):6493–6503.

Cohen, J., Braver, T., and Brown, J. (2002). Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology*, 12:223–229.

Cohen, M. and Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Trans on Sys. Man and Cyber.*, 13(5):815–826.

Cohen, N. and Squire, L. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science*, 210:207–210.

Coleman, P. and Flood, D. (1987). Neuron numbers and dendritic extent in normal aging and alzheimer's disease. *Neurobiology of Aging*, 8(6):521–545.

Compte, A., Brunel, N., Goldman-Rakic, P. S., and Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10:910–923.

Constantinidis, C., Franowicz, M. N., and Goldman-Rakic, P. S. (2001). The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nature Neuroscience*, 4(3):311–316.

Conway, M. A. (1990). *Autobiographical Memory: An Introduction*. Open University Press, Milton Keynes, Philadelphia.

Craik, F. and Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11:671–684.

Craik, F. and Watkins, M. (1973). The role of rehearsal in short-term memory. *J Verb Learn Behav*, 12:599–607.

Crick, F. and Mitchison, G. (1983). The function of dream sleep. *Nature*, 304:111–114.

de Lacalle, S., Irazoz, I., and Gonzalo, L. M. (1991). Differential changes in cell size and number in topographic subdivisions of human basal nucleus in normal aging. *Neuroscience*, 43(2/3):445–456.

de Nó, R. L. (1938). Analysis of the activity of the chains of internuncial neurons. *J. Neurophysiol.*, 1:207–244.

de Nó, R. L. (1938). The cerebral cortex: Architexcture, intracortical connections and motor projections. In Fulton, J., editor, *Physiology of the Nervous System*, pages 291–325. Oxford University Press.

De Schutter, E. (1994). Modelling the cerebellar purkinje cell: Experiments in computo. In van Pelt, J., Corner, M., Uylings, H., van Veen, M., and van Ooyen, A., editors, *The Self-Organizing Brain: From Growth Cones to Functional Networks*, volume 102 of *Progress in Brain Research*, pages 427–441. Elsevier Science, Amsterdam.

Doyere, V., Burette, F., Negro, C. R.-D., and Laroche, S. (1993). Long-term potentiation of hippocampal afferents and efferents to prefrontal cortex: implications for associative learning. *Neuropsychologia*, 31(10):1031–1053.

Durstewitz, D., Kelc, M., and Güntürkün, O. (1999). A neurocomputational theory of the dopaminergic modulation of working memory functions. *Journal of Neuroscience*, 19(7):2807–2822.

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

Eckhorn, R., Bauer, R., Jordan, W., Brosch, S., Kruse, W., Munk, M., and Reitboeck, H. J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological Cybernetics*, 60:121–130.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review*, 89:627–661.

Eichenbaum, H. (1993). Thinking about brain cell assemblies. *Science*, 261:993–994.

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience*, 1:41–50.

Eichenbaum, H. and Cohen, N. J. (2001). *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. Oxford University Press.

Eldridge, L. L., Knowlton, B. J., Furmanski, C. S., Bookheimer, S. Y., and Engel, S. A. (2000). Remembering episodes: a selective role for the hippocampus during retrieval. *Nature Neuroscience*, 3(11):1149–1152.

Ennaceur, A. and Delacour, J. (1987). Effect of combined or separate administration of piracetam and choline on learning and memory in the rat. *Psychopharmacology*, 92:58–67.

Eriksson, D. and Lansner, A. (2003). Dynamic hierarchical attractor-clustering. Technical Report TRITA-NA-P0304, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden.

Eriksson, P. S., Perfilieva, E., Björk-Eriksson, T., Alborn, A.-M., Nordborg, C., Peterson, D. A., and Gage, F. H. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine*, 4(11):1313 – 1317.

Fanselow, M. S. and LeDoux, J. E. (1999). Why we think plasticity underlying Pavlovian fear conditioning occurs in the basolateral amygdala. *Neuron*, 23:229–232.

Felleman, D. and van Essen, D. (1991). Distributed hierarchival processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47.

Forrin, B. and Cunningham, K. (1973). Recognition time and serial position of probed item in short-term memory. *Journal of Experimental Psychology*, 99(2):272–279.

Foster, T. C. (1999). Involvement of hippocampal synaptic plasticity in age-related memory decline. *Brain Research Reviews*, 30:236–249.

Fransén, E. (1996). *Biophysical Simulation of Cortical Associative Memory*. PhD thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden. TRITA-NA-P96/28.

Fransén, E. and Lansner, A. (1990). Modelling Hebbian cell assemblies comprised of cortical neurons. In *Proc. Open Network Conference on Neural Mechanisms of Learning and Memory*, page 4:9, London.

Fransén, E. and Lansner, A. (1995). Low spiking rates in a population of mutually exciting pyramidal cells. *Network*, 6(2):271–288.

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

Freedman, M. and Oscar-Berman, M. (1986). Bilateral frontal lobe disease and selective delayed response deficits in humans. *Behav Neurosci*, 100:337–42.

Freeman, W. (1991). The physiology of perception. *Scientific American*, 264:78–85.

French, R. (1999). Catastrophic forgetting in connectionist networks: Causes, consequences and solutions. *Trends in Cognitive Science*, 3(4):128–135.

Frey, U. and Morris, R. (1997). Synaptic tagging – synapse specificity during protein synthesis-dependent long-term potentiation. *Nature*, 385:533–536.

Fried, I., MacDonald, K. A., and Wilson, C. L. (1997). Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron*, 18(5):753–65.

Funahashi, S., Bruce, C. J., and Goldman-Rakic, P. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, 61(2):331–349.

Funahashi, S., Chafee, M., and Goldman-Rakic, P. (1993). Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature*, 365:753–756.

Fuster, J. (1989). *The Prefrontal Cortex*. Raven, New York, 2nd edition.

Fuster, J. (1995). *Memory in the Cerebral Cortex*. MIT Press, Cambridge, Massachusetts.

Fuster, J. and Alexander, G. (1971). Neuron activity related to short-term memory. *Science*, 73:652–654.

Fuster, J., Bauer, R., and J.P, J. (1982). Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks. *Exp. Neurol*, 77:679–694.

Gage, F. H. (2002). Neurogenesis in the adult brain. *Journal of Neuroscience*, 22(3):612–613.

Gallagher, M. and Colombo, P. J. (1995). Ageing: the cholinergic hypothesis of cognitive decline. *Current Opinion in Neurobiology*, 5(2):161–168.

Gardner, E. (1987). Maximum storage capacity in neural networks. *Europhysics Letters*, 4:481–485.

Gars, J. and Tamsen, F. (1999). Clustering in bayesian neural networks. Technical Report TRITA-NA-P9910, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

Geszti, T. and Pázmándi, F. (1987). Learning within bounds and dream sleep. *J Phys. A: Math Gen*, 20:L1299–L1303.

Gilbert, C. (1993). Circuitry, architecture and functional dynamics of visual cortex. *Cerebral Cortex*, 3:373–386.

Gilbert, C., Hirsch, J., and Wiesel, T. (1990). *Lateral interactions in the visual cortex*, volume LV of *Cold Spring Harbor Symposia on Quantiative Biology*. Cold Spring Harbor Laboratory Press.

Gilbert, C. and Wiesel, T. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.*, 9(7):2432–2442.

Gleissner, U., Helmstaedter, C., Kurthen, M., and Elger, C. E. (1997). Evidence of very fast memory consolidation: an intracarotid amytal study. *NeuroReport*, 8:2893–2896.

Goldman-Rakic, P. (1995). Cellular basis of working memory. *Neuron*, 14:477–485.

Goldman-Rakic, P. and Schwartz, M. (1982). Interdigitation of contralateral and ipsilateral columnar projections to frontal association cortex in primates. *Scoemce*, 216:755–757.

Goldman-Rakic, P. S., Ó Scalaidhe, S., and Chafee, M. (1999). Domain specificity in cognitive systems. In Gazzaniga, M. S., editor, *The New Cognitive Neurosciences*, pages 733–742. MIT Press, Cambridge, MA, 2nd edition.

Good, I. (1950). *Probability and the weighing of evidence*. Charles Griffin, London.

Graham, B. and Willshaw, D. (1997). Capacity and information efficiency of the associative net. *Network*, 8:35–54.

Graham, K., Patterson, K., and Hodges, J. (1999). Episodic memory: new insights from the study of semantic demential. *Current Opinion in Neurobiology*, 9:245–250.

Graham, K. S., Simons, J. S., Pratt, K. H., Patterson, K., and Hodges, J. R. (2000). Insights from semantic dementia on the relationship between episodic and semantic memory. *Neuropsychologia*, 38:313–324.

Gray, C., König, P., Engel, A., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338:334–337.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23:121–134.

Grossberg, S. (1982). *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control*. Reidel, Boston.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63.

Gupta, A., Wang, Y., and Markram, H. (2000). Organizing principles for a diversity of GABAergic interneurons and synapses in the neocortex. *Science*, 287:273–278.

Gustafsson, B., Asztely, F., Hanse, E., and Wigström, H. (1989). Onset characteristics of long-term potentiation in the guinea-pig hippocampal CA1 region in vitro. *Eur J Neurosci*, 1(4):382–394.

Gustafsson, B. and Wigström, H. (1988). Physiological mechanisms underlying long-term potentiation. *Trends in Neurosciences*, 11:156–162.

Gutkin, B. S., Laing, C. R., Colby, C. L., Chow, C. C., and Ermentrout, G. B. (2001). Turning on and off with excitation: the role of spike-timing asynchrony and synchrony in sustained neural activity. *Journal of computational neuroscience*, 11:121–134.

Haberly, L. and Bower, J. (1989). Olfactory cortex: model circuit for study of associative memory? *Trends Neurosci*, 12(7):258–64.

Haist, F., Gore, J. B., and Mao, H. (2001). Consolidation of human memory over decades revealed by functional magnetic resonance imaging. *Nature Neuroscience*, 4(11):1139–1145.

Hamilton, W. (1966). The moulding of senescence by natural selection. *J. Theoretical Biology*, 12:12–45.

Hamker, F. H. (2001). Life-long learning cell structures – continuously learning without carastrophic interference. *Neural Networks*, 14:551–573.

Hammarlund, P. and Ekeberg, Ö. (1998). Large neural network simulations on multiple hardware platforms. *Journal of Computational Neuroscience*, 5:443–459.

Hansel, D. and Sompolinsky, H. (1996). Chaos and synchrony in a model of a hypercolumn in visual cortex. *Journal of Computational Neuroscience*, 3(1):7–34.

Hasselmo, M. (1999). Neuromodulation: acetylcholine and memory consolidation. *Trends in Cognitive Sciences*, 3(9):351–359.

Hasselmo., M., Anderson, B., and Bower, J. (1992). Cholinergic modulation of cortical associative memory function. *J. Neurophysiol.*, 67:1230–1246.

Hasselmo, M. and Cekica, M. (1996). Suppression of synaptic transmission may allow combination of associative feedback and self-organizing feedforward connections in the neocortex. *Behavioural Brain Research*, 79(1–2):153–161.

Hasselmo, M., Linster, C., Patil, M., Ma, D., and Cekic, M. (1997). Noradrenergic suppression of synaptic transmission may influence cortical signal-to-noise ratio. *Journal of Neurophysiology*, 77(6):3326–3339.

Hasselmo, M. E., Wyble, B. P., and Wallenstein, G. V. (1996). Encoding and retrieval of episodic memories: role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus*, 6(6):693–708.

Hebb, D. (1949). *The Organization of Behavior*. John Wiley Inc., New York.

Hebb, D. (1959). A neuropsychological theory. In Koch, S., editor, *Psychology: A Study of a Science*, volume 1. McGraw-Hill, New York.

Hempel, C. M., Hartman, K. H., Wang, X.-J., Turrigiano, G. G., and Nelson, S. B. (2000). Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J Neurophysiol*, 83(5):3031–3041.

Hertz, J., Krogh, A., and Palmer, R. (1991). *Introduction to the Theory of Neural Computation*, volume 1 of *Lecture Notes*. Addison-Wesley, Santa Fe Institute for studies in the sciences of complexity.

Hevner, R. and Wong-Riley, M. (1992). Entorhinal cortex of the human, monkey, and rat: metabolic map as revealed by cytochrome oxidase. *J. Comp. Neurol.*, 326:451–469.

Hinton, G. and Anderson, J., editors (1981). *Parallel models of associative memory*. Erlbaum, New York, Hillsdale.

Hirsch, J. and Crepel, F. (1990). Use-dependent changes in synaptic efficacy in rat prefrontal neurons *in vitro*. *J. Physiol.*, 427:31–49.

Hirsch, J. and Gilbert, C. (1991). Synaptic physiology of horizontal connections in the cats visual cortex. *J. Neurosci*, 11(6):1800–1809.

Hoffman, K. L. and McNaughton, B. L. (2002). Coordinated Reactivation of Distributed Memory Traces in Primate Neocortex. *Science*, 297(5589):2070–2073.

Holland, P. and Bouton, M. (1999). Hippocampus and context in classical conditioning. *Curr Opin Neurobiol*, 9(2):195–202.

Holst, A. (1997). *The Use of a Bayesian Neural Network Model for Classification Tasks*. PhD thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden. TRITA-NA-P9708.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A*, 79(8):2554–8.

Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *PNAS*, 81:3088–3092.

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In Dale, R., Mellish, C., and Zock, M., editors, *Current research in natural language generation*. Academic Press, London.

Houghton, G. and Hartley, T. (1995). Parallel models of serial behaviour: Lashley revisited. *Psyche*, 2(25).

Hubel, D. and Wiesel, T. (1977). The functional architecture of the macaque visual cortex. The Ferrier lecture. *Proc. Royal. Soc. B*, 198:1–59.

Humphreys, M., Bain, J., and Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic and procedural tasks. *Psychological Review*, 96:208–233.

Imig, T. and Adrian, H. (1977). Binaural columns in the primary field (A1) of cat auditory cortex. *Brain Research*, 138:241–257.

Ishida, Y., Shirokawa, T., Miyausgu, O., Komatsu, Y., and Isobe, K. (2001). Age-dependent changes in noradrenergic innervations of the frontal cortex of F344 rats. *Neurobiology of aging*, 22:283–286.

Issa, N. P., Trepel, C., and Stryker, M. P. (2000). Spatial frequency maps in cat visual cortex. *Journal of Neuroscience*, 20(22):8504–8514.

Izquierdo, I. and Medina, J. H. (1997). Memory formation: the sequence of biochemical events in the hippocampus and its connection to activity in other brain structures. *Neurobiol Learn Mem*, 68(3):285–316.

Jacobsen, C. (1935). Functions of ftrontal association area in primates. *Arch. Neurol. Psychiat.*, 33:558–569.

James, W. (1890). *The Principles of Psychology*. Holt, Rinehart and Winston, New York.

James, W. (1892). *Psychology (Briefer Course)*. Collier, New York. chapter 16.

Jaynes, E. T. (1996). Probability theory: The logic of science. http://omega.math.albany.edu:8008/JaynesBook.html.

Jernigan, T., Archibald, S., Fennema-Notestine, C., Gamst, A., Stout, J., Bonner, J., and Hesselink, J. (2001). Effects of age on tissues and regions of the cerebrum and cerebellum. *Neurobiology of Aging*, 22:581–594.

Jester, J., Campbell, L., and Sejnowski, T. (1995). Associative EPSP-spike potentiation induced by pairing orthodromic and antidromic stimulation in rat hippocampal slices. *J. Physiology*, 484:689–705.

Johansson, C. (2001). A study of interacting bayesian recurrent neural networks with incremental learning. Master's thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology. TRITA-NA-E0110, http://www.nada.kth.se/~cjo/documents/exjobb.pdf.

Johansson, C. and Lansner, A. (2002a). An associative neural network model of classical conditioning. Technical Report TRITA-NA-P0217, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden.

Johansson, C. and Lansner, A. (2002b). A neural reinforcement learning system. Technical Report TRITA-NA-P0215, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden.

Johansson, C., Raicevic, P., and Lansner, A. (2003). Reinforcement learning based on a bayesian confidence propagating neural network. April 10-11, SAIS-SSLS Joint Workshop, Center for Applied Autonomous Sensor Systems, Örebro, Sweden.

Johansson, C., Sandberg, A., and Lansner, A. (2001). A capacity study of a bayesian neural network with hypercolumns. Technical Report TRITA-NA-P0120, Department of Numerical Analysis and Computer Science, Royal Institute of Technology. http://www.nada.kth.se/~cjo/documents/bcpnn_capacity.pdf.

Johansson, C., Sandberg, A., and Lansner, A. (2002). A neural network with hypercolumns. In *ICANN 2002*, pages 192–197, Berlin. Springer-Verlag. LNCS 2415.

Kaasinen, V. and Rinne, J. O. (2002). Functional imaging studies of dopamine system and cognition in normal aging and parkinson's disease. *Neuroscience & Biobehavioral Reviews*, 26(7):785–793.

Kanerva, P. (1988). *Sparse Distributed Memory*. MIT Press, Cambridge, MA.

Kanter, I. (1988). Potts glass models of neural networks. *Phys. Rev. A*, 37(7).

Kaszniak, A. (1986). The neuropsychology of dementia. In Grant, I. and Adams, K., editors, *Neuropsychology Assessment of Neuropsychiatric Disorders*, pages 172–220. Oxford, New York.

Katsuki, H., Izumi, Y., and Zorumski, C. (1997). Noradrenergic regulation of synaptic plasticity in the hippocampal CA1 region. *Journal of Neurophysiology*, 77(6):3013–3020.

Klingberg, T. (1997). *The Neurophysology of Working Memory – functional mapping of the human brain with positron emission tomography*. PhD thesis, Karolinska Institute.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers*, 21:353–359.

Kononenko, I. (1989). Bayesian neural networks. *Biological Cybernetics*, 61:361–370.

Korsnes, M. and Magnussen, S. (1994). Age comparisions of serial position effects in short-term memory. *Acta Psychologica*, 94:133–143.

Kozlov, A., Hellgren-Kotaleski, J., Aurell, E., Grillner, S., and Lansner, A. (2001). Modeling of substance P and 5-HT induced synaptic plasticity in the lamprey spinal CPG – consequences for network pattern generation. *J. Computational Neuroscience*, 11:183–200.

Kozlov, A., Lansner, A., and Grillner, S. (2003). Burst dynamics under mixed NMDA and AMPA dirve in the models of the lamprey spinal CPG. *Neurocomputing*. In press.

Laing, C. R. and Chow, C. C. (2001). Stationary bumps in networks of spiking neurons. *Neural Computation*, 13:1473–1494.

Laing, C. R. and Chow, C. C. (2002). A spiking neuron model for binocular rivalry. *Journal of Computational Neuroscience*, 12:39–53.

Laing, C. R. and Longtin, A. (2001). Noise-induced stabilization of bumps in systems with long-range spatial coupling. *Physica D*, 160:149–172.

Lansner, A. (1982). Information processing in a network of model neurons. A computer simulation study. Tech. Rep. TRITA-NA-8211, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden.

Lansner, A. (1991). A recurrent bayesian ANN capable of extracting prototypes from unlabeled and noisy examples. In Kohonen, T., Mäkisara, K., Simula, O., and Kangas, J., editors, *Artificial Neural Networks*, pages 247–254, Espoo, Finland. Elsevier, Amsterdam. Proc. ICANN-91.

Lansner, A. and Ekeberg, Ö. (1987). An associative network solving the "4-Bit ADDER problem". In Caudill, M. and Butler, C., editors, *IEEE First International Conference on Neural Networks*, pages II–549, San Diego, CA.

Lansner, A. and Ekeberg, Ö. (1989). A one-layer feedback artificial neural network with a bayesian learning rule. *Int. J. Neural Systems*, 1(1):77–87.

Lansner, A. and Fransén, E. (1992). Modeling Hebbian cell assemblies comprised of cortical neurons. *Network*, 3(2):105–119.

Lansner, A., Fransen, E., and Sandberg, A. (2003). Cell assembly dynamics in detailed and abstract attractor models of cortical associative memory. *Theory in Biosciences*. in press.

Lansner, A., Hellgren-Kotaleski, J., Ullström, M., and Grillner, S. (1997). Local spinal modulation of the calcium dependent potassium channel underlying slow adaptation in a model of the lamprey cpg. In Bower, J. M., editor, *Computational Neuroscience: Trends in Research, 1998*, pages 429–434, Big Sky, Montana. Plenum Press.

Lansner, A. and Holst, A. (1996). A higher order Bayesian neural network with spiking units. *Int. J. Neural Systems*, 7(2):115–128.

LeVay, S. and Gilbert, C. (1976). Laminar patterns of geniculocortical projections in the cat. *Brain Res.*, 113:1–19.

Levin, B. (1995). *On Extensions, Parallel Implementation and Applications of a Bayesian Neural Network*. PhD thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden. TRITA-NA-P9515.

Levin, E. D. and Simon, B. B. (1998). Nicotinic acetylcholine involvement in cognitive function in animals. *Psychopharmacology (Berl)*, 138(3-4):217–30.

Levy, W. and Desmond, N. (1985). The rules of elemental synaptic plasticity. In Levy, W., Anderson, J., and Lehmkule, S., editors, *Synaptic Modification, Neuron Selectivity, and Nervous System Organization*, Hillsdale. Lawrence Erlbaum Associates.

Levy, W. and Steward, O. (1979). Synapses as associative memory elements in the hippocampal formation. *Brain Research*, 179:233–245.

Levy, W. and Steward, O. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, 8:791–797.

Li, S.-C., Lindenberger, U., and Sikström, S. (2001). Aging cognition: from neuromodulation to representation. *Trends in cognitive sciences*, 5(11):479–486.

Li, S.-C. and Sikström, S. (2002). Integrative neurocomputational perspectives on cognitive agingm neuromodulation, and representation. *Neuroscience and Biobehavioral Reviews*, 26:795–808.

Lieberman, M. D. (2000). Introversion and working memory: central executive differences. *Personality and Individual Differences*, 28:479–486.

Liljenkrantz, A. (2003). Memory consolidation in artificial neural networks. Master's thesis, Dept. of Numerical Analysis and Computing Science. In press.

Lisman, J. E. and Idiart, M. A. (1995). Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science*, 267(5203):1512–5.

Little, W. (1974). The existence of persistent states in the brain. *Mathematical Biosciences*, 19:101–120.

Little, W. and Shaw, G. (1975). A statistical theory of short and long term memory. *Behavioral Biology*, 14:115–133.

Lund, J., Yoshioka, T., and Levitt, J. (1993). Comparision of the intrinsic connectivity in different areas of the macaque monkey cerebral cortex. *Cerebral Cortex*, 3:148–162.

Lynch, G. (1998). Memory and the brain: Unexpected chemistries and a new pharmacology. *Neurobiol Learn Mem*, 70(1/2):82–100.

MacDonald, S., Uesiliana, K., and Hayne, H. (2000). Cross-cultural and gender differences in childhood amnesia. *Memory*, 8(6):365–376.

MacGregor, R. and McMullen, T. (1978). Computer simulation of diffusely-connected neuronal populations. *Biological Cybernetics*, 28:121–127.

MacGregor, R. and Palasek, R. (1974). Computer simulation of rhytmic oscillations in neuron pools. *Kybernetik*, 16:79–86.

MacKay, D. (1995). Probable networks and plausible predictions — a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6:469–505.

MacRae, P., Spirduso, W., and Wilcox, R. (1988). Reaction time and nigrostriatal dopamine function: the effect of age and practice. *Brain Res*, 451:139–146.

Maddock, R., Garret, A., and Buonocore, M. (2001). Remembering familiar people: the posterior cingulate cortex and autobiographical memory retrieval. *Neuroscience*, 104(3):667–676.

Maguire, E., Henson, R., Mummery, C., and Frith, C. (2001). Activity in prefrontal cortex, not hippocampus, varies parametrically with the increasing remoteness of memories. *NeuroReport*, 12(3):441–444.

Maguire, E. A., Burgess, N., Donnett, J. G., Frackowiak, R. S., Frith, C. D., and O'Keefe, J. (1998). Knowing where and getting there: a human navigation network. *Science*, 280(5365):921–4.

Mangan, P. and Nadel, L. (1990). Development of spatial memory in the human infant. *Psychnomic Soc. Abstr*, 31:35–36.

Maren, S. and Fanselow, M. S. (1996). The amygdala and fear conditioning: Has the nut been cracked? *Neuron*, 16:237–240.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 245:213–215.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology*, 202:437–470.

Marr, D. (1971). Simple memory: a theory for the archicortex. *Philosophical Transactions of the Royal Society of London, B*, 262::23–81.

Martinez, J. L., Schulteis, G., and Weinberger, S. B. (1991). How to increase and decrease the strength of memory traces: the effects of drugs and hormones. In Martinez, J. L. and Kesner, R. P., editors, *Learning and Memory: A Biological View*, chapter 4, pages 149–198. Academic Press, second edition.

McClearn, G. E., Johansson, B., Berg, S., Pedersen, N. L., Ahern, F., Petrill, S. A., and Plomin, R. (1997). Substantial Genetic Influence on Cognitive Abilities in Twins 80 or More Years Old. *Science*, 276(5318):1560–1563.

McClelland, J., McNaughton, B., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the success and failures of connectionist models of learning and memory. *Psychological Review*, 102:419–457.

McClelland, J. L. (1994). The organization of memory. a parallel distributed processing perspective. *Rev Neurol (Paris)*, 150(8-9):570–9.

McClelland, J. L. and Goddard, N. H. (1997). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, 6:654–665.

McCormick, D., Connors, B., Lighthall, J., and OPrince, D. (1985). Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *J. Neurophysiol*, 54:782–805.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

McGaugh, J. L. (2000). Memory – a century of consolidation. *Science*, 287:248–251.

McNally, R., Lasko, N., Macklin, M., and Pitman, R. (1994). Autobiographical memory disturbance in combat-related posttrumatic stress disorder. *Behav. Res. Ther.*, 33(6):619–630.

Medawar, P., editor (1952). *An Unsolved Problem of Biology*. Lewis, London.

Medina, J., Repa, J., M, D. M., and LeDoux, J. (2002). Parallels between cerebellum- and amygdala-dependent conditioning. *Nature Reviews Neuroscience*, 3(2):122–131.

Mendelson, J. R. and Ricketts, C. (2001). Age-related temporal processing speed deterioration in auditory cortex. *Hearing Research*, 158(1–2):84–94.

Mergler, N. and Goldstein, M. (1983). Why are there old people: senescence as biological and cultural preparedness for the transmission of information. *Human Development*, 26:72–90.

Mezard, M., Nadal, J., and Toulouse, G. (1986). Solvable models of working memories. *Journal de Physique*, 47:1457–1462.

Middlebrooks, J., Dykes, J., and Merzenich, M. (1980). Binaural response-specific bands in primary auditory cortex (A1) of the cat: topographical organization orthogonal to isofrequency contours. *Brain Research*, 181:31–48.

Milner, P. (1957). The cell assembly: Mark II. *Psychol. Rev.*, 64:242–252.

Mishkin, M., Malamut, B., and Bachevalier, J. (1984). Memories and habits: two neural systems. In Lynch, G., McGaugh, J., and Weinberger, N., editors, *Neurobiology of Learning and Memory*, pages 65–77. Guilford Press, New York.

Mitchell, K. J., Johnson, M. K., Raye, C. L., and D'Esposito, M. (2000). fMRI evidence of age-related hippocampal dysnfunction in feature binding in working memory. *Cognitive Brain Research*, 10:197–206.

Moita, M. A., Rosis, S., Zhou, Y., LeDoux, J. E., and Blair, H. T. (2003). Hippocampal place cells acquire location-specific responses to the conditioned stimulus during auditory fear conditioning. *Neuron*, 37:485–497.

Monyer, H., Burnashev, N., Laurie, D., Sakmann, B., and Seeburg, P. (1994). Developmental and regional expression in the rat brain and functional properties of four NMDA receptors. *Neuron*, 12:529–540.

Morris, R. (1996). Learning, memory and synaptic plasticity: cellular mechanisms, network architecture and the recording of attended experience. In Magnusson, D., editor, *The lifespan development of individuals: behavioral, neurobiological, and psychosocial perspectives*, chapter 7, pages 139–161. Cambridge University press.

Morrison, J. H. and Hof, P. R. (1997). Life and Death of Neurons in the Aging Brain. *Science*, 278(5337):412–419.

Mountcastle, V. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiol.*, 20:408–434.

Mountcastle, V. (1978). An organizing principle for cerebral function: the unit module and distributed function. In Edelman, G. and Mountcastle, V. B., editors, *The Mindful Brain*. Cambridge, MIT Press.

Mountcastle, V. (1998). *Perceptual neuroscience: The Cerebral Cortex*. Harvard University Press, Cambridge, Massachisetts.

Mozley, P., Kim, H., Gur, R., Tatsch, K., Muenz, L., McElgin, W., Kung, M., Mu, M., Myers, A., and Kung, H. (1996). Iodine-123-IPT SPECT imaging of CNS dopamine transporters: nonlinear effects of normal aging on striatal uptake values. *J Nucl Med*, 37(12):1965–1970.

Murdock, B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89(6):609–626.

Murre, J. M. (1996). Tracelink: a model of amnesia and consolidation of memory. *Hippocampus*, 6(6):675–84.

Myerson, J., Hale, S., Wagstaff, D., Poon, L., and Smith, G. (1990). The information-loss model: A mathematical theory of age-related cognitive slowing. *Psychological review*, 4:475–487.

Nadal, J. and Toulouse, G. (1990). Information storage in sparsely coded memory nets. *Network*, 1:61–74.

Nadal, J., Toulouse, G., Changeux, J., and Dehaene, S. (1986). Networks of formal neurons and memory palimpsests. *Europhysics Letters*, 1(10):535–542.

Nadel, L. and Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr Opin Neurobiol*, 7(2):217–27.

Nadel, L. and Moscovitch, M. (1998). Hippocampal contributions to cortical plasticity. *Neuropharmacology*, 37:431–439.

Newcombe, N., Huttenlocher, J., Drummey, A., and Wiley, J. (1998). The development of spatial location coding: place learning and dead reckoning in the second and third years. *Cognitive Dev.*, 13:185–201.

Nilsson, L.-G., Nyberg, L., and Bäckman, L. (2002). Genetic variation in memory functioning. *Neuroscience and Biobehavioral Reviews*, 26(7):841–848.

Nunez, P. (1995). *Neocortical Dynamics and Human EEG Rhythms*. Oxford University Press.

Nyberg, L., Habib, R., McIntosh, A. R., and Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *PNAS*, 97(20):11120–11124.

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map. preliminary evidence from unit activity in the the freely-moving rat. *Brain Res.*, 34:171–175.

Oler, J. and Markus, E. (2000). Age-related deficits in the ability to encode contextual change: A place cell analysis. *Hippocampus*, 10:338–350.

Orre, R. (1998). Data mining and process modelling using a bayesian confidence propagation neural network. Licentiate degree thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden. TRITA-NA-P9810, ISSN 1101-2250, ISRN KTH/NA/P-98/10-SE, ISBN 91-7170-273-3.

Otani, S., Blond, O., Desche, J.-M., and Crépel, F. (1998). Dopamine facilitates long-term depression of glutamatergic transmission in rat prefrontal cortex. *Neuroscience*, 85(3):669–676.

Otmakhova, N. and Lisman, J. E. (1998). D1/D5 dopamine receptors inhibit depotentiation at CA1 synapses via cAMP-dependent mechanism. *Journal of Neuroscience*, 18(4):1270–1279.

Packard, M. G. and McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol Learn Mem*, 65(1):65–72.

Packard, M. G. and Teather, L. A. (1998). Amygdala modulation of multiple memory systems: hippocampus and caudate-putamen. *Neurobiol Learn Mem*, 69(2):163–203.

Pakkenberg, B. and Gundersen, H. J. G. (1997). Neocortical neuron number in humans: effect of sex and age. *Journal of Comparative Neurology*, 384:312–320.

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

Palm, G. (1981). On the storage capacity of an associative memory with randomly distributed storage elements. *Biological Cybernetics*, 39:125–127.

Palm, G. (1982). *Neural Assemblies. An Alternative Approach to Artificial Intelligence*. Springer, Berlin.

Palm, G. (1990). Cell assemblies as a guideline for brain research. *Concepts in Neuroscience*, 1:133–147.

Pantic, L., Torres, J., Kappen, H., and Gielen, S. C. (2002). Associative memory with dynamic synapses. *Neural Computation*, 14(12):2903–2923.

Parisi, G. (1986). A memory which forgets. *J. Phys. A: Math. Gen*, 19:L617–620.

Parker, A., Wilding, E., and Akerman, C. (1998). The von restorff effect in visual object recognition memory in humans and monkeys: The role of frontal/perirhinal interaction. *Journal of Cognitive Neuroscience*, 10(6):691–703.

Parsons, M. and Gold, P. (1992). Glucose enhancement of memory in elderly humans: an inverted-u dose-response curve. *Neurobiol Aging*, 13(3):401–4.

Passingham, R. (1975). Delayed matching after selective prefrontal lesions in monkeys. *Brain Research*, 92:89–102.

Patten, B. M. (1990). The history of memory arts. *Neurology*, 40:346–352.

Petersen, C., Malenka, R., Nicoll, R., and Hopfield, J. (1998). All-or-none potentiation at CA3-CA1 synapses. *Proc Natl Acad Sci U S A*, 95(8):4732–7.

Peterson, C. and Söderberg, B. (1989). A new method for mapping optimization problems onto neural networks. *International Journal of Neural Systems*, 1(1):3–22.

Peterson, C. and Söderberg, B. (1998). Neural optimization. In Arbib, M., editor, *The Handbook of Brain Research and Neural Networks*, pages 617–622. MIT Press, Cambdrige, MA, 2nd edition edition.

Peterson, L. and Peterson, M. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58:193–198.

Petersson, K., Elfgren, C., and Ingvar, M. (1997). A dynamic role of the medial temporal lobe during retrieval of declarative memory in man. *Neuroimage*, 6:1–11.

Petersson, K., Elfgren, C., and Ingvar, M. (1999). Dynamic changes in the functional anatomy of the human brain during recall of abstract designs related to practice. *Neuropsychologia*, 37:567–587.

Pike, R. (1984). A comparison of convolution and matrix distributed memory systems. *Psychological Review*, 91:281–294.

Pillemer, D. B. (1998). What is remembered about early childhood events? *Clinical psychology review*, 18(8):895–913.

Plato (360). *Theaetetus*.

Ploner, C. J., Gaymard, B., Rivaud, S., Agid, Y., and Pierrot-Deseilligny, C. (1998). Temporal limits of spatial working memory in humans. *European Journal of Neuroscience*, 10:794–797.

Poirazi, P., Brannon, T., and Mel, B. W. (2003). Pyramidal neuron as two-layer neural network. *Neuron*, 37:989–999.

Pribram, K., Mishkin, M., Rosvold, H., and Kaplan, S. (1952). Effects on delayed-response preformance of lesions of dorsolateral and ventromedial frontal cortex of baboons. *J Comp Physiol Psychol*, 45:565–575.

Purves, D., Riddle, D., and LaMantia, A.-S. (1992). Iterated patterns of brain circuitry (or how the cortex gets its spots). *TINS*, 15:362–368.

Quillfeldt, J. A., Zanatta, M. S., Schmitz, P. K., Quevedo, J., Schaeffer, E., Lima, J. B., Medina, J. H., and Izquierdo, I. (1996). Different brain areas are involved in memory expression at different times from training. *Neurobiol Learn Mem*, 66(2):97–101.

Quinlan, P. (1991). *Connectionism and Psychology. A Psychological Perspective on Connectionist Research*. Harvester, Wheatsheaf, New York.

Ragozzino, M. E., Unick, K. E., and Gold, P. E. (1996). Hippocampal acetylcholine release during memory testing in rats: Augmentation by glucose. *Proc. Natl. Acad. Sci. USA*, 93:4693–4698.

Ramirez-Amaya, V., Balderas, I., Sandoval, J., Escobar, M. L., and Bermudez-Rattoni, F. (2001). Spatial Long-Term Memory Is Related to Mossy Fiber Synaptogenesis. *J. Neurosci.*, 21(18):7340–7348.

Rapp, P. and Amaral, D. (1992). Individual differences in the cognitive and neurobiological consequences of normal aging. *Trends Neurosci*, 15:340–345.

Rapp, P. R. and Gallagher, M. (1996). Preserved neuron number in the hippocampus of aged rats with spatial learning deficits. *PNAS*, 93(18):9926–9930.

Ratcliff, R., Van Zandt, T., and McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, 106:261–300.

Redish, A., Elga, A., and Touretzky, D. (1996). A coupled attractor model of the rodent head direction system. *Network*, 7:671–685.

Reed, P. and Richards, A. (1996). The von restorff effect in rats (rattus norvegius). *Journal och Comparative Psychology*, 110(2):193–198.

Ribot, T. (1882). *Diseases of Memory*. Appleton, New York.

Rinne, J., Hietala, J., Ruotsalainen, U., Sako, E., Laihinen, A., Nagren, K., Lehikinen, P., Oikonen, V., and Syvalahti, E. (1993). Decrease in human striatal dopamine D2 receptor density with age: a PET study with [11C]raclopride. *J. Cereb Blood Flow Metab*, 13:310–314.

Robins, A. (1996). Consolidation in neural networks and in the sleeping brain. *Connection Science*, 8(2):259–275.

Rochester, N., Holland, J., Haibt, L., and Duda, W. (1956). Tests on a cell assembly theory of the action of the brain, using a large scale digital computer. *IRE Transactions of Information Theory*, IT-2:80–93.

Rolls, E. and Stringer, S. (2000). On the design of neural networks in the brain by genetic evolution. *Progress in Neurobiology*, 61:557–579.

Rolls, E. and Treves, A. (1998). *Neural Networks and Brain Function*. Oxford University Press.

Romo, R., Brody, C. D., Hernandez, A., and Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399:470–473.

Rose, M., editor (1991). *Evolutionary Biology of Aging*. Oxford University Press, Oxford.

Rose, M. and Mueller, L. (1998). Evolution of human lifespan: Past, future, and present. *American journal of human biology*, 10:409–420.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan books, Washington, D.C.

Rubin, D., Rahhal, T., and Poon, L. (1998). Things learned in early adulthood are remembered best. *Memory and Cognition*, 26:3–19.

Rubin, D. and Schulkind, M. (1997). Distribution of autobiographical memories across the lifespan. *Memory and Cognition*, 25:859–866.

Rumelhart, D. E., McClelland, J. L., and PDP Research Group, editors (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1–2. MIT Press, Cambridge.

Ruppin, E. (1995). Neural modeling of psychiatric disorders. *Network : Computation in Neural Systems*, 6:635–656.

Ruppin, E. and Reggia, J. (1995). A neural model of memory impairment in diffuse cerebral atrophy. *Br. J. of Psychiatry*, 166(1):19–28.

Ruppin, E. and Yeshurun, Y. (1991). Recall and recognition in an attractor neural network of memory retrieval. *Connection Science*, 3(4):381–400.

Sakurai, Y. (1998). Cell-assembly coding in several memory processes. *Neurobiology of learning and memory*, 70:212–225.

Sandberg, A., Lansner, A., and Petersson, K. (2001). Selective enhancement of recall through plasticity modulation in an autoassociative memory. *Neurocomputing*, 38–40:867–873.

Sandberg, A., Lansner, A., Petersson, K.-M., and Ekeberg, Ö. (2002). A bayesian attractor network with incremental learning. *Network*, 13(2):179–194.

Sarle, W. S. (2002). comp.ai.neural-nets FAQ, part 2: Learning. http://www.faqs.org/faqs/ai-faq/neural-nets/part2/section-2.html.

Sawaguchi, T. and Goldman-Rakic, P. (1991). D1 dopamine receptors in prefrontal cortex: involvement in working memory. *Science*, 251:947–950.

Scannell, J., Blakemore, C., and Young, M. (1995). Analysis of connectivity in the cat cerebral cortex. *J. Neurosci.*, 15:1463–1483.

Schachter, D. and Tulving, E. (1994). What are the memory systems of 1994? In Schachter, D. and Tulving, E., editors, *Memory Systems*, pages 1–38. MIT Press, Cambridge, MA.

Scheff, S. W., Price, D. A., and Sparks, D. L. (2001). Quantitative assessment of possible age-related change in synaptic numbers in the human frontal cortex. *Neurobiology of Aging*, 22:355–265.

Scherman, D., Desnos, C., Darchen, F., Pollack, P., Javoy-Agid, F., and Agid, Y. (1989). Striatal dopamine deficiency in parkinson's disease: role of aging. *Ann. Neurol.*, 26:551–557.

Schmidt, S. (1994). The effects of humor on sentence memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20:953–967.

Schüz, A. (1994). Patchiness as a means to get a message across. *Trends in Neuroscience*, 17(9):265.

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

Scoville, W. and Milner, B. (1957). Loss of recent memories after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry*, 20:11–21.

Seitz, A. R. and Watanabe, T. (2003). Is subliminal learning really passive? *Nature*, 422:36.

Sejnowski, T. (1989). Induction of synaptic plasticity by Hebbian covariance in the hippocampus. In Durbin, R., Miall, C., and Mitchison, G., editors, *The Computing Neuron*, Wokingham, U.K. Addison-Wesley.

Seung, H. (1996). How the brain keeps the eyes still. *Proc. Natl. Acad. Sci. USA*, 93:13339–13344.

Seung, H. (1998). Learning continuous attractors in recurrent networks. *Advances in Neural Information Processing Systems*, 10:654–660.

Seung, H., Lee, D., Reis, B., and Tank, D. (2000). Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron*, 26:259–271.

Shallice, T. and Warrington, E. (1970). Independent functioning of verbal memory stores: a neuropsychological study. *Quarterly Journal of Experimental Psychology*, 22:261–273.

Shen, J. and Barnes, C. A. (1996). Age-related decrease in cholinergic synaptic transmission in three hippocampal subfields. *Neurobiology of Ageing*, 17(3):439–451.

Sheng, M., Cummings, J., Roldan, L., Jan, Y., and Jan, L. (1994). Changing subunit composition of heteromeric NMDA receptors during development of rat cortex. *Nature*, 368:144–147.

Shepherd, G., editor (1998). *The Synaptic Organization of the Brain*. Oxford University press, New York, 4th edition.

Shin, Y. and Ghosh, J. (1991). The pi-sigma network: An efficient higher-order neural network for pattern classification and function approximation. In *Proc. IJCNN*, Seattle.

Shors, T. J., Miesegaes, G., Beylin, A., Zhao, M., Rydel, T., and Gould, E. (2001). Neurogenesis in the adult is involved in the formation of trace memories. *Nature*, 410:372–376.

Sikström, S. (2003). The isolation effect and serial position curves predicted by mechanisms in synaptic plasticity. Submitted.

Singer, W., Engel, A. K., Kreiter, A. K., Munk, M. H. J., Neuenschwander, S., and Roelfsema, P. R. (1997). Neuronal assemblies: necessity, signature and detectability. *Trends Cognitive Science*, 1(7):252–261.

Skaggs, W., Knierim, J., Kudrimoti, H., and McNaughton, B. (1995). A model of the neural basis of the rat's sense of direction. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in neural information processing systems*, volume 7, pages 173–180. MIT Press, Cambdrige, MA.

Small, S. A. (2001). Age-related memory decline. *Arch. Neurol.*, 58:360–364.

Sommer, F. and Dayan, P. (1998). Bayesian retrieval in associative memories with storage errors. *IEEE Transations on Neural Networks*, 9(4):705–711.

Sonntag, W., Bennett, S., Khan, A., Thornton, P., Xu, X., Ingram, R., and Brunso-Bechtold, J. (2000). Age and insulin-like growth factor-1 modulate n-methyl-d-aspartate receptor subtype expression in rats. *Brain Research Bulletin*, 51(4):331–338.

Sperling, G. (1960). The information available in brief presentation. *Psychological Monograph*, 74(498):1–29.

Squire, L. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99:195–231.

Squire, L., Ojemann, J., Miezin, F., Petersen, S., Videen, T., and Raichle, M. (1992). Activation of the hippocampus in normal humans: a functional anatomical study of memory. *PNAS*, 89:1837–1841.

Squire, L. R. and Alvarez, P. (1995). Retogerade amnesia and memory consolidation: a neurobiologival perspective. *Current Opinion in Neurobiology*, 5:169–177.

Squire, L. R., Knowlton, B., and Musen, G. (1993). The structure and organization of memory. *Annu Rev Psychol*, 44:453–95.

Squire, L. R. and Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, 253:1380–1386.

Staley, K., Longacher, M., Bains, J., and Yee, A. (1998). Presynaptic modulation of CA3 network activity. *Nature Neuroscience*, 1(3):201–209.

Steinbuch, K. (1961). Die lernmatrix. *Kybernetik*, 1:36–45.

Steinbuch, K. and Piske, U. (1963). Learning matrices and their applications. *IEEE Transactions on Electronic Computers*, pages 846–862.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, 153(736):652–4.

Sternberg, S. (1975). Memory scanning: new findings and current controversies. *Q. J. Exp. Psychol*, 27:1–32.

Storkey, A. (1999). *Efficient Covariance Matrix Methods for Bayesian Gaussian Processes and Hopfield Neural Networks*. PhD thesis, University of London.

Storkey, A. and Valabregue, R. (1997). A new learning rule with high capacity storage of time-correlated patterns. *Electronic Letters*, 33:1803–1804.

Storkey, A. and Valabregue, R. (1999). The basins of attraction of a new hopfield learning rule. *Neural Networks*, 12:869–876.

Strange, B., Henson, R., Friston, K., and Dolan, R. (2000). Brain mechanisms for detecting perceptual, semantic, and emotional deviance. *NeuroImage*, 12:425–433.

Strangman, G. (1996). Searching for cell assemblies: how many electrodes do I need? *J. Comput. Neurosci.*, 3:111–124.

Ström, B. (2000). A model of neocortical memory, using hypercolumns and a bayesian learning rule. Master's thesis, Dept. of Numerical Analysis and Computing Science. TRITA-NA-E0020.

Sugaya, K., Greene, R., Personett, D., Robbins, M., Kent, C., Bryan, D., Skiba, E., Gallagher, M., and Mckinney, M. (1998). Septo-hippocampal cholinergic and neurotrophin markers in age-induced cognitive decline. *Neurobiology of aging*, 19(4):351–361.

Tabak, J., Senn, W., O'Donovan, M., and Rinzel, J. (2000). Modeling of spontaneous activity in developing spinal cord using activity-dependent depression in an excitatory network. *Journal of Neuroscience*, 20(8):3041–3056.

Tanaka, T. and Yamada, M. (1993). The characteristics of the convergence time of associative neural networks. *Neural Computation*, 5(3):463–472.

Tang, Y.-P., Shimizu, E., Dube, G. R., Rampon, C., Kerchner, G. A., Zhuo, M., Liu, G., and Tsien, J. Z. (1999). Genetic enhancement of learning and memory in mice. *Nature*, 401:63–69.

Tanila, H., Shapiro, M., Gallagher, M., and Eichenbaum, H. (1997). Brain aging: Changes in the nature of information coding by the hippocampus. *The Journal of Neuroscience*, 17(13):5155–5166.

Tegnér, J., Compte, A., and Wang, X. (2002). The dynamical stability of reverberatory dynamics. *Biological Cybernetics*, 87(5–6):471–481.

Terry, R. and Katzman, R. (2001). Life span and synapses: will there be a primary senile dementia? *Neurobiology of Aging*, 22:347–348.

Thach, W. (1996). On the specific role of the cerebellum in motor learning and cognition: Clues from PET activation and lesion studies in man. *Behavioral and Brain Sciences*, 19(3):411–431.

Thomas, M., Moody, T., Makhinson, M., and O'Dell, T. (1996). Activity dependent β-adrenergic modulation of low frequency stimulation induced LTP in the hippocampal CA1 region. *Neuron*, 17:475–482.

Thomson, A. and Deuchars, J. (1994). Temporal and spatial properties of local circuits in neocortex. *TINS*, 17:119–126.

Thomson, A. M. and Bannister, A. P. (2003). Interlaminar connections in the neocortex. *Cerebral Cortex*, 13:5–14.

Torres, J. J., Pantic, L., and Kappen, H. J. (2002). Storage capacity of attractor neural networks with depressing synapses. *Physical Review E*, 66:061910.

Tsodyks, M. and Markram, H. (1996). Plasticity of neocortical synapses enables transitions between rate and temporal coding. *Lect. Nodes Comput. Sci.*, 1112:445–450.

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. *Neural Computation*, 10:821–835.

Tulving, E. (1972). Episodic and semantic memory. In Tulving, E. and Donaldson, W., editors, *Organization of memory*, pages 382–403. Academic Press, New York.

Tulving, E. (1999). Introduction. In Gazzaniga, M. S., editor, *The New Cognitive Neurosciences*, page 728. MIT Press, Cambridge, MA, 2nd edition.

Undergleider, L. G., Courtney, S. M., and Haxby, J. V. (1998). A neural system for human visual working memory. *Proc. Natl. Acad. Sci. USA*, 95:883–890.

Vakil, E. and Herishanu-Naaman, S. (1998). Declarative and procedural learning in parkinson's disease patients having tremor or bradykinesia as the predominant symptom. *Cortex*, 34:611–620.

van Hemmen, J. (1997). Hebbian learning, its correlation catastrophe, and unlearning. *Network: Comput. Neural Syste.*, 8:V1–V17.

Van Zandt, T. (2002). Analysis of response time distributions. In Wixted, J. T. and Pashler, H., editors, *Stevens' Handbook of Experimental Psychology*, volume 4, pages 461–516. Wiley Press, New York, 3rd edition edition.

VanRullen, R. and Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences*, 7(5).

Volkow, N. D., Gur, R. C., Wang, G.-J., Fowler, J. S., Moberg, P. J., Ding, Y.-S., Hitzemann, R., Smith, G., and Logan, J. (1998). Association Between Decline in Brain Dopamine Activity With Age and Cognitive and Motor Impairment in Healthy Individuals. *Am J Psychiatry*, 155(3):344–349.

von der Malsburg, C. (1981). The correlation theory of brain function. Technical report, Dept. of Neurobiology, Max-Planck-Institute for Biophysical Chemistry, Göttingen. Reprinted in: Models of Neural networks II, edited by E. Domany, J.L. van Hemmen, and K. Schulten (Springer, Berlin, 1994) Chapter 2, pp. 95-119.

von der Malsburg, C. (1986). Am I thinking assemblies? In Palm, G. and Aertsen, A., editors, *Brain Theory*. Springer, Berlin.

von Hayek, F. (1952). *The Sensory Order*. University of Chicago Press, Chicago.

von Restorff, H. (1933). Analyse von vorgängen in spurenfeld. *Psychologice Forschung*, 18:299–342.

Wagner, A., Koutstaal, W., and Schacter, D. (1999). When encoding yields remembering: insights from event-related neuroimaging. *Phil. Trans. R. Soc. Lond. B*, 121:1307–1324.

Wahlgren, N. and Lansner, A. (2001). Biological evaluation of a Hebbian-Bayesian learning rule. *Neurocomputing*, 38–40:433–438.

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent memory. *Trends in Neurosciences*, 24(8):455–463.

Wang, Y., Chan, G. L., Holden, J. E., Dobko, T., Mak, E., Schulzer, M., Huser, J. M., Snow, B. J., Ruth, T. J., Calne, D. B., and Stoessl, A. J. (1998). Age-dependent decline of dopamine D1 receptors in human brain: A PET study. *Synapse*, 30(1):56–61.

Warburton, D., Rusted, J., and Fowler, J. (1992). A comparision of the attentional and consolidation hypotheses for the facilitation of memory by nicotine. *Psychopharmacology*, 108:443–447.

Watts, D. and Strogatz, D. (1998). Collective dynamics of small-world networks. *Nature*, 393:440–442.

Waughn, N. and Norman, D. (1965). Primary memory. *Psychological Review*, 72:89–104.

Welford, A. (1965). Performance, biological mechanisms and age: a theoretical sketch. In Welford, A. and Birren, J., editors, *Behavior, aging, and the nervous System*, pages 3–20. Thomas, Springfield, IL.

Wellman, C. L. and Pelleymounter, M. A. (1999). Differential effects of nucleus basalis lesions in young adult and aging rats. *Neurobiology of Aging*, 20:381–393.

Wheeler, M. E., Petersen, S. E., and Buckner, R. L. (2000). Memory's echo: vivid remembering reactivates sensory-specific cortex. *PNAS*, 97(20):11125–11129.

White, J. M., Sparks, D. L., and Stanford, T. R. (1994). Saccades to remembered target locations: an analysis of systematic and variable errors. *Vision Research*, 34(1):79–92.

Wickelgren, W. A. (1992). Webs, cell assemblies, and chunking in neural nets. *Concepts in Neuroscience*, 3:1–53.

Wickens, J. and Kötter, R. (1995). Cellular models of reinforcement. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 187–214. MIT Press.

Willingham, D. B. and Preuss, L. (1995). The death of implicit memory. *Psyche*, 2(15).

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

Wilson, H. and Cowan, J. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13:55–80.

Wilson, M. A. and McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, 261(5124):1055–8.

Wilson, M. A. and McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265:676–679.

Winder, R. and Borril, J. (1998). Fuels for memory: the role of oxygen and glucose in memory enhancement. *Psychopharmacology*, 136:349–356.

Wolpert, D. H. and Macready, W. G. (1995). No free lunch theorems for search. Technical Report SFI-TR-95-02-010, Santa Fe, NM.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Wood, E., Dudchenko, P., and Eichenbaum, H. (1999). The global record of memory in hippocampal neuronal activity. *Nature*, 397:613–616.

Woolf, N. J. (1996). The critical role of cholinergic basal forebrain neurons in morphological change and memory encoding: a hypothesis. *Neurobiol Learn Mem*, 66(3):258–66.

Wu, X. and Liljenström, H. (1994). Regulating the nonlinear dynamics of olfactory cortex. *Network*, 5:47–60.

Xu, L., Anwyl, R., and Rowan, M. J. (1998). Spatial exploration induces a persistent reversal of long-term potentiation in rat hippocampus. *Nature*, 394(6696):891–4.

Yasuno, F., Nishikawa, T., Tokunaga, H., and *et al.* (2000). The neural basis of perceptual and conceptual word priming - a PET study. *Cortex*, 36:59–69.

Yerkes, R. and Dodson, J. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurological Psychology*, 18:459–482.

Zec, R. F. (1995). The neuropsychology of aging. *Exp Gerontol*, 30(3-4):431–42.

Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensembles: a theory. *J. Neuroscience*, 16:2112–2126.

Zola-Morgan, S. and Squire, L. R. (1993). Neuroanatomy of memory. *Annu. Rev. Neurosci.*, 16:547–63.