Memory Stability and Synaptic Plasticity

Guy Billings

Doctor of Philosophy

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh

2008

Abstract

Numerous experiments have demonstrated that the activity of neurons can alter the strength of excitatory synapses. This synaptic plasticity is bidirectional and synapses can be strengthened (potentiation) or weakened (depression). Synaptic plasticity offers a mechanism that links the ongoing activity of the brain with persistent physical changes to its structure. For this reason it is widely believed that synaptic plasticity mediates learning and memory.

The hypothesis that synapses store memories by modifying their strengths raises an important issue. There should be a balance between the necessity that synapses change frequently, allowing new memories to be stored with high fidelity, and the necessity that synapses retain previously stored information. This is the plasticity stability dilemma. In this thesis the plasticity stability dilemma is studied in the context of the two dominant paradigms of activity dependent synaptic plasticity: Spike timing dependent plasticity (STDP) and long term potentiation and depression (LTP/D). Models of biological synapses are analysed and processes that might ameliorate the plasticity stability dilemma are identified.

Two popular existing models of STDP are compared. Through this comparison it is demonstrated that the synaptic weight dynamics of STDP has a large impact upon the retention time of correlation between the weights of a single neuron and a memory. In networks it is shown that lateral inhibition stabilises the synaptic weights and receptive fields.

To analyse LTP a novel model of LTP/D is proposed. The model centres on the distinction between early LTP/D, when synaptic modifications are persistent on a short timescale, and late LTP/D when synaptic modifications are persistent on a long timescale. In the context of the hippocampus it is proposed that early LTP/D allows the rapid and continuous storage of short lasting memory traces over a long lasting trace established with late LTP/D. It is shown that this might confer a longer memory retention time than in a system with only one phase of LTP/D. Experimental predictions about the dynamics of amnesia based upon this model are proposed.

Synaptic tagging is a phenomenon whereby early LTP can be converted into late LTP, by subsequent induction of late LTP in a separate but nearby input. Synaptic tagging is incorporated into the LTP/D framework. Using this model it is demonstrated that synaptic tagging could lead to the conversion of a short lasting memory trace into a longer lasting trace. It is proposed that this allows the rescue of memory traces that were initially destined for complete decay. When combined with early and late LTP/D

synaptic tagging might allow the management of hippocampal memory traces, such that not all memories must be stored on the longest, most stable late phase timescale. This lessens the plasticity stability dilemma in the hippocampus, where it has been hypothesised that memory traces must be frequently and vividly formed, but that not all traces demand eventual consolidation at the systems level.

Acknowledgements

Mark van Rossum has been an outstanding mentor. I thank Mark for his consistent encouragement, support and advice throughout my time in Edinburgh. My interest in applying statistical physics to synaptic plasticity was inspired by Mark. I shall always have fond memories of my time as his student thanks to his intelligence, dry wit and professionalism. Coming to Edinburgh to study with him was one of the best decisions I have ever made.

I also thank my second supervisor Richard Morris for many exciting discussions about synaptic plasticity. My interpretations of experimental data were greatly aided by his clarity of thought and incisiveness. The philosophical underpinnings of this work are due in large part to Richard's ideas about synaptic plasticity and memory. I thank Steve Martin and Roger Redondo for helpful discussions and for letting me look at raw data.

During my PhD I was lucky enough to spend several months visiting the computational neurobiology lab at the Salk Institute. I thank Mark van Rossum and Terrence Sejnowski for allowing this to happen. During this visit I wrote most of the code that was used to simulate the state based models of LTP and I received much friendly guidance and help from the members of CNL. Thanks for giving me such a warm Californian welcome!

My thanks also go out to the members of the Doctoral Training Centre and the Institute for Adaptive and Neural Computation. These institutions have provided a unique learning environment that successfully blends exploration and rigor. In particular I would like to thank Adam Barrett, Jesus Cortes, David Sterratt, Matthias Hennig, Rowland Sillito, Mark Longair and John Clayden for many great coffee (or alcohol) fueled discussions. Finally I thank Angus Silver, the head of my new lab, for his patience with me during the final stages of my PhD.

While undertaking this thesis I have been given considerable personal support. For this I thank my wonderful girlfriend, Hannah (this time, the thesis really is finished) and my family for helping me to find and to follow the path.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Guy Billings*)

Table of Contents

Chapter 1

Synaptic plasticity and memory

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

Neuroscience suggests that experience of the world evokes patterns of activity in the brain. This causes changes to synaptic efficacy by means of activity dependent synaptic plasticity. Once altered, synaptic efficacies lay dormant until a subsequent experience reactivates the synapses. The idea that this process embodies the encoding, storage and recall of memory, has been referred to as the synaptic plasticity and memory hypothesis (SPM) (Morris et al., 2003). This hypothesis is often introduced by a statement of Hebb's famous postulate, quoted above.

There have been major advances in support of our understanding of Hebb's prescient postulate and the SPM; such as the discovery that synapses can be modified with long term potentiation (LTP) (Bliss and Gardner-Medwin, 1973; Bliss and Lomo, 1973; Malenka and Bear, 2004) and spike timing dependent plasticity (STDP) (Levy and Steward, 1983; Markram et al., 1997; Bi and Poo, 1998), the discovery of evidence that links synaptic modification to memory in the intact animal (Steele and Morris, 1999; Moser and Morris, 1998; Pastalkova et al., 2006; Whitlock et al., 2006) and the observation of replayed activity in neural ensembles (Lin et al., 2007).

LTP and STDP are the current dominant paradigms for activity dependent long term synaptic modification. It is widely believed that if the SPM holds then the encoding and storage of memory is mediated by processes such as LTP and STDP. In this chapter an overview of the biology underlying the SPM is reviewed.

The hippocampus occupies a pivotal position in memory research. Most theories

of hippocampal function, while disagreeing upon the details of processing, do agree that somewhere in the hippocampus patterns of synaptic efficacy are laid down and stored in a memory trace. Both LTP and STDP have been observed in hippocampal neurons, strengthening the implied role of these process in memory storage via synaptic modification. The memory trace can thus be defined as a set of synaptic efficacies, laid down by processes such as STDP and LTP, whose precise values are relevant to some cognitive process such as episodic memory. In this chapter major theories of hippocampal function are briefly reviewed with the aim of demonstrating that the notion of a memory trace is implicit in all of them.

The rest of the thesis is devoted to studying the dynamics of the memory trace in various contexts, assuming that it is mediated by STDP or LTP, while making minimal assumptions about the neural code. If the elements of memory are encoding, storage and retrieval, the work in this thesis aims to contribute to our understanding of storage: How memory traces can be retained within a population of plastic synapses when those synapses must accommodate ongoing learning.

1.1 Long term potentiation

Long Term Potentiation is the phenomenon whereby synapses can be strengthened for an extended period when the neurons they connect are made to fire at particular frequencies (Bliss and Lomo, 1973; Bliss and Gardner-Medwin, 1973). Long term potentiation has been extensively studied in the hippocampus. Typically, stimulation protocols are applied to either the synapses between the Schaffer collateral-commissural axons and the apical dendrites of CA1 pyramidal cells, or synapses between the mossy fibre terminals of dentate granule cells and the dendrites of CA3 granule cells. Experiments are performed on slice preparations in vitro or using chronic implants in behaving animals.

Classically, repetitive high frequency stimulation (HFS) is required for LTP induction. This protocol hits the 'sweet spot' of NMDA receptor activation by ensuring that there is sufficient coincidence of post-synaptic depolarization and pre-synaptic glutamate input for potentiation to occur. Broadly speaking, potentiation is induced by elevated calcium levels in the cell, resulting from NMDA receptor activation (Kauer, Malenka, and Nicoll, 1988). Analogously, LTD can be induced by producing low but sustained intracellular calcium levels which can result from low frequency stimulation (LFS).

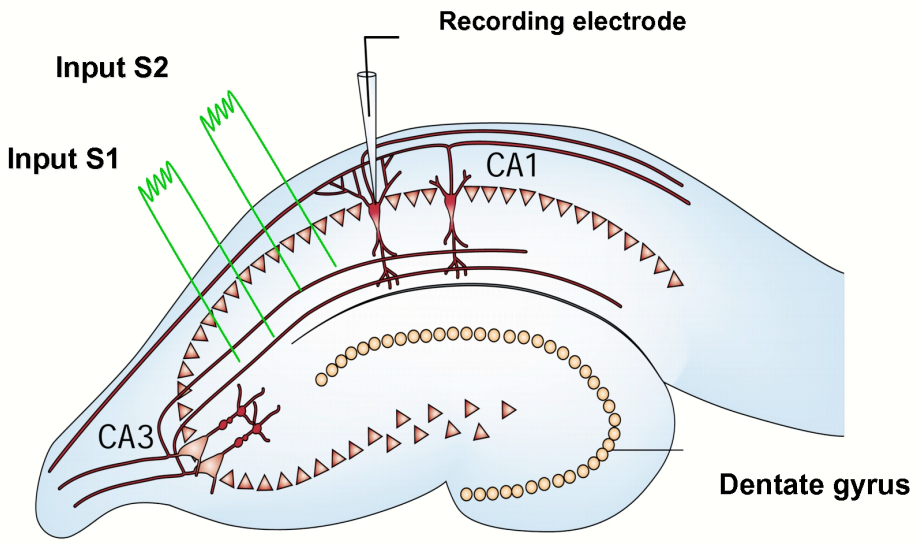

Figure 1.1: Schematic of a transverse slice through the Hippocampus. This preparation is typically used in Hippocampal LTP experiments. LTP or LTD can be induced in the Schaffer colateral pathway running from the CA3 region to the CA1 region. Inputs S1 and S2 show possible locations for stimulating electrodes used to apply the protocols in Tables 1.1+1.2. The recording electrode is placed near to cell bodies in the CA1 region. In tagging protocols both inputs S1 and S2 are used to apply plasticity protocols to separate synaptic inputs. Figure obtained from Roger Redondo.

Right from the early days of LTP research it seemed clear that the persistence of LTP was variable, lasting from hours to days up to weeks in chronically implanted freely behaving animals (Bliss and Gardner-Medwin, 1973; Abraham, 2003). It has since become apparent that there are at least two phases of LTP, late-LTP (lLTP) and early-LTP (eLTP) where lLTP can be induced by applying more cycles of HFS to a preparation than are required to induce eLTP. An important distinction between these phases, which are phenomenologically differentiated by the extreme persistence of late lLTP up to many weeks, is that lLTP requires protein synthesis whereas eLTP does not (Malenka and Bear, 2004). Application of an mRNA translation inhibitor, anisomycin, completely blocks late LTP (Sajikumar and Frey, 2003).

One of the questions that this thesis addresses is: Why are there different decay timescales of LTP/D? It shall be argued later that having more than one decay timescale of LTP/D might confer benefits upon the storage of the memory trace.

1.1.1 LTP induction protocols and decay timescales

Both the magnitude and the timescale of decay of LTP/D is dependent upon the stimulation protocol. Furthermore, when looking at the data one notices that the magnitude of synaptic modification and the decay timescale are to some extent independent, depending upon the protocol. Similar magnitudes of potentiation can be elicited but with differing decay timescales. This is seen again and again in the literature (Tables 1.1+1.2). There are vast numbers of experimental studies dealing with LTP. Studies included here were selected with the following criteria:

1. Produced electrophysiological data measured in terms EPSP slope
2. Performed in hippocampal slice preparations from rats
3. Concerned synapses between CA3 and CA1, Fig. 1.1
4. Produced data where the timecourse of decay could be estimated (i.e. where measurements of the EPSP slope after induction were taken on a timescale of at least the order of an hour).

Early LTP is typically elicited with one burst of HFS (60-100Hz) and lasts on the order of hours. Of the studies surveyed in Table 1.1 the typical induction time[1] for early LTP with HFS was 30s-120s. Early LTD is obtained with LFS (1-10Hz). Early LTD appears to have a similar decay timecourse as early LTP although the induction times are generally much larger (partly as a result of the time required to deliver many pulses at 1Hz) and range from 450s to 21min, Table 1.2.

Late LTP can be elicited with HFS by repeating stimulation bursts with some interval between them, often 3-5 bursts with a 10min interburst interval. Due to the interburst interval, late LTP administered in this fashion takes 30-50 mins to induce. Late LTD can be elicited with a lengthy duration of very low frequency stimulation such as 1Hz for 15 mins. Late LTD seems to exhibit similar persistence as does late LTP. The induction time for late LTD with these protocols is 15-20mins.

[1] The induction time was taken as the minimum time of LTP/D induction as extracted from the graphs in the respective studies. Often the minimum time that could be measured was simply the time interval between one data point prior to the application of a protocol and another data point after application of the protocol. Typically the distance between data points was several minutes. This time is of relevance because it is used to match the LTP models in chapter 6 to the data, such that the same rate of LTP is achieved as in experiments. However the conclusions are not dependent upon a precise knowledge of the induction timescale.

The LTP protocols mentioned above take a relatively long period to apply, often around 10 mins. Furthermore, the volleys of activity in standard protocols are biologically atypical. Another type of protocol named Θ burst stimulation (TBS) has been developed that succeeds in creating more rapid potentiation with a biologically plausible pattern of bursts (Larson and Lynch, 1986; Staubli and Lynch, 1987). TBS consists of many volleys of short high frequency stimulation with interburst intervals of hundreds of milliseconds. TBS is highly effective at eliciting LTP with induction times of the order of seconds rather than minutes.

1.1.1.1 Depotentiation

Depotentiation is the removal of recently induced potentiation by application of low frequency stimulation (LFS). There is evidence that both early and late LTP can be depotentiated (Bashir and Collingridge, 1994; Sajaykumar and Frey, 2004). The distinction between depotentiation and early LTD is that depotentiation can be elicited with LFS that has no effect on the synaptic weight when applied to synapses that have not recently undergone potentiation. However if the synapses have recently been potentiated, the LFS reverses that potentiation, returning the measured EPSP slope back to its original baseline value. LTP induced with TBS can be reversed by applying a low frequency 2Hz stimulation protocol with a duration of around 35s (Staubli and Chun, 1996), a shorter duration than required to achieve LTD.

1.1.1.2 LTP decay timescales

In confirmation of previous reviews, the papers summarised in Table 1.1 imply that it is reasonable to take the timescale of decay of early LTP to be of the order of one hour. The timescale of late LTP cannot be definitively identified because it is often larger than the duration of the experiments. However the experiments in which slices have been maintained over a long period indicate that it is in excess of 10 hours. In vivo studies demonstrate the maintenance of LTP for longer than this, for time periods of weeks or months (Abraham, 2003; Bliss and Gardner-Medwin, 1973). Thus we can conservatively consider late LTP as having a decay timescale that is at least an order of magnitude greater than early LTP. This is assumed in chapters 6 and 7.

Study	LTP induction protocol	LTP induction timescale	LTP decay timescale	LTP magnitude
Dudek 1992	900 pulses @ 50 Hz	150s	>45min	135%
Dudek 1993	TBS : 10 trains @ 5Hz, 4 pulses @ 100Hz per train	6min	>40min	175%
Bashir 1994	1s @ 100Hz	30s	>60min	200%
Alpermann 2006	1s @ 100Hz	5min	>360min	180%
Frey 1999	3 x 100s @ 100Hz (10 min intertrain)	1min	>480min	175%
Kelleher 1994	2 x 1s @ 100Hz (30s intertrain)	1min	>60min	250%
	4 x 1s @ 100Hz (5min intertrain)	20min	>200min	295%
Frey 1997	3 x 100s @ 100Hz (10min intertrain)	30min	>450min	180%
	21s @ 100Hz	2min	-	-
Woo 2007	1s @ 100Hz	60s	18min	250%
	4 x 1s @ 100Hz (5mins intertrain)	20min	>130min	500%
Nguyen 1997	TBS: 15 trains, 4 pulses @ 100Hz (200ms intertrain)	2min	>180min	250%
	1s @ 60Hz	1min	50min	175%
Sajikumar 2005	3 trains 100 pulses @ 100Hz, (10min intertrain)	35min	>540min	160%
	eLTP: 1 x 100 Hz train	1min	90min	160%
Sanna 2002	2 x 500ms @ 100Hz	3min	-	200%
Staubli 1996	TBS: 10 trains, 30ms pulses @ 100Hz (200ms intertrain)	30s	>165min	180%

Table 1.1: A summary of LTP induction protocols taken from studies investigating CA1 of hippocampal slices in rats. The results of LTP experiments are remarkably reproducible. The key points demonstrated by the experiments in this table are that: 1) there are 2 clear timescales of synaptic modification, early phase LTP and late phase LTP. 2) The same degree of potentiation can be achieved regardless of whether late or early LTP are induced. Therefore, the timescale of persistence of LTP can be varied independently of the magnitude of that change in synaptic efficacy. In this thesis, the implications of this for memory storage are explored.

Study	LTD induction protocol	LTD induction time	LTD decay time	LTD magnitude
Dudek 1992	900 pulses @ 10 Hz	5min	10min	70%
	900 pulses @ 3 Hz	5min	>45min	65%
	900 pulses @ 1Hz	15min	>45min	60%
Dudek 1993	900 pulses @ 1Hz	15min	>45min	50%
Bashir 1994	depotentiation: 450s @ 2Hz	450s	-	-
Woo 2007	depotentiation: 480s @ 3Hz	480s	-	-
Sajikumar 2005	900 trains of 3 bursts @ 20Hz	21min	>540min	50%
	900 pulses @ 1Hz	21min	180min	70%
	depotentiation: 250 pulses @ 1Hz			
Staubli 1996	150x 30ms trains @ 100Hz (200ms intertrain)	-	-	-

Table 1.2: A summary of early and late LTD and depotentiation induction protocols taken from studies investigating synaptic connections within CA1 of hippocampal slices taken from rats. This table demonstrates that LTD has similar properties to LTP: Its magnitude and persistence can be varied independently.

1.1.2 Mechanisms of LTP induction and maintenance

The hypothesis that the efficacy of synapses determine memory storage and that changes to those efficacies are mediated by plasticity, demands that we consider how synaptic efficacy is regulated. The arrival of the action potential triggers an increase in Ca^{2+} in the presynaptic terminal and results in the release of neurotransmitter into the synaptic cleft. Glutamaturgic synapses are responsible for the behavior seen in the plasticity experiments reviewed in Tables 1.1+1.2. In these synapses AMPA receptors (AMPAR) are bound by glutamate that has crossed the synaptic cleft. This leads to conformational changes to the AMPAR protein subunits such that they flux positively charged ions, predominantly Na^{+}, which depolarises the postsynaptic cell. When the cell is sufficiently depolarised such that the membrane potential crosses a critical threshold (typically around $-50mV$), the post synaptic cell fires an action potential (spike).

When the strength of a synapse is increased by LTP, it is more effective at depolarizing the post-synaptic cell and allows more positive charge in to the post synaptic cell upon activation of the synapse. There are many possible ways of increasing the AMPA current. One possibility is the modification of AMPARs such that more positive charge is permitted to enter. It is known that the GluR1 and GluR2 subunits of the AMPA receptor can be phosphorylated by CaMKII and PKA. Furthermore, phosphorylation of GluR1 alters the channel properties (Roche et al., 1996; Lee et al., 2000; Chung et al., 2000; Malenka and Siegelbaum, 2001).

Phosphorylation may also play a role in the maintenance of late LTP. An important protein in the stabilisation of AMPARs in the synapse is PSD 95. It has been found that PSD 95 can be phosphorylated and that this leads to the accumulation of PSD 95 resulting in more persistent retention of AMPARs (Kim et al., 2007).

In this thesis it is assumed that LTP/D on long timescales is the result of postsynaptic modifications. Postsynaptic AMPAR regulation processes inspire the models LTP/D in chapters 6-8. Next, this literature is briefly reviewed.

1.1.2.1 AMPA receptor regulation

There is evidence that potentiation of synapses in the cerebellum leads to the insertion of AMPAR (Shigemoto, 2006). Furthermore, synaptic boutons enlarge subsequent to protein synthesis dependent lLTP (Malenka and Bear, 2004; Harris, Fiala, and Ostroff, 2003) suggesting that more proteins are incorporated into the spine after potentiation. There are three putative processes that enable the number of post-synaptic AMPARs

to be regulated. These are; endo/exo cytosis of receptors, association of AMPARs with proteins in the post-synaptic density and lateral diffusion of AMPAR receptors.

Endo/Exocytosis: Experiments exploiting immunoflourescence tagging of AMPA receptors have shown that they are internalized from the plasma membranes of hippocampal cells in vitro via endocytosis and that AMPAR inside endosomes within the cell can be inserted in to the membrane via exocytosis (Park et al., 2004). AMPA receptor internalisation is dynamin dependent suggesting endocytosis as the mechanism of internalisation. Blocking AMPAR exocytosis leads to rapid synaptic rundown suggesting that it is necessary to maintain synaptic potentiation (Lin et al., 2000; Man et al., 2000; Carroll et al., 2001).

The half life of decay of the ratio of internal florescence to total fluorescence for labeled AMPA receptors in cultured hippocampal cells was found to be of the order of 10 minutes in basal conditions (Lin et al., 2000). This figure was obtained by labeling AMPAR receptors with a PH sensitive dye. The dye is inactivated when AMPAR are endocytosed allowing the fraction of AMPAR that remain in the whole cell membrane to be measured as a function of time. The observed fluorescence decay timeconstant is very rapid in comparison to the lifetime of early and late LTP. Although these results apply to the whole cell membrane as opposed to the post synaptic zone, the high AMPAR turnover rate suggests that the maintenance of LTP is dynamic: That is to say that rather than merely 'sticking' in synapses, receptors are perpetually inserted and removed and that the balance between these processes is what gives rise to the long term synaptic efficacy.

Interestingly the AMPAR turnover rate was reduced to around 3 minutes in the case that AMPA was applied, suggesting that activation of the synapse increases AMPAR turnover rate. Some evidence implies that NMDA dependent calcium influx can be responsible for AMPA receptor internalisation through the activation of the calcium dependent enzyme calcineurin (Beattie et al., 2000).

The studies cited above show that endo/exocytosis directly controls the number of AMPA receptors in the plasma membrane. Thus it is likely that these processes would be capable of exerting an influence on the number of AMPA receptors present in the post-synaptic apparatus. Thus LTP might result from the exocytosis of AMPA receptors directly into the synapse and LTD might result from their internalisation. At any given time the weight of a synapse would be determined by the balance of addition and removal of AMPA receptors in to the plasma membrane at that synapse.

The machinery of endocytosis is the endocytotic pit in which clatharin pinches off

sections of membrane to create vesicles containing AMPARs. Such pits are observed primarily outside of the synaptic zone (Triller and Choquet, 2005). Furthermore vesicles containing AMPAR, presumably the precursor to exocytosis or the consequence of endocytosis, have not been observed directly beneath the PSD. In fact they have only been observed in the periphery of synapses (Choquet and Triller, 2003)[2]. These observations point to the need for more than endo/exocytosis alone in regulation of AMPARs at the synapse.

Endo/Exocytosis regulates the number of AMPA receptors in the whole plasma membrane, but the observations mentioned above do not prove that these processes alone account for changes in the number of receptors within the postsynaptic membrane of a single spine. If the number of AMPARs is globally regulated by the cell, are there other mechanisms that could account for local regulation of the number of AMPAR in each spine?

'Slot' proteins: The cell membrane contains freely and semi confined diffusing AMPAR (Borgdoff and Choquet, 2002; Choquet and Triller, 2003; Tardin et al., 2003; Triller and Choquet, 2003; Groc et al., 2004; Triller and Choquet, 2005) that are added and removed at sites away from the synapse. There is evidence to suggest that the PSD performs the role of providing a substrate for 'slot' proteins that stabilise AMPARs by binding to their cytoplasmic elements (Turrigiano, 2000). This might allow the PSD to regulate the rate of loss and capture of AMPA receptors in the synapse from the surrounding plasma membrane, removing the need for endo and exocytosis within the spine.

It is known that each GluR AMPAR subunit has specific interactions with various proteins (Bredt et al., 2004; Duprat et al., 2003; Sheng and Lee, 2003). These interactions play a role in both dynamic regulation of receptor cycling (for example PICK/GRIP1 and Stargazin) but may also allow the retention of AMPARs at synaptic sites depending upon their subunit composition. The PSD might provide this functionality with PSD95 as the protein that seems a likely candidate for the role of being a slot due to its interaction with Stargazin and hence indirectly, AMPARs (Bredt et al., 2004).

AMPAR lateral diffusion: If AMPA receptors are regulated by capture and release

[2]In order to extract specific proteins, cells are often centrifuged, destroying nearly all structure. However, one element that does survive is the post synaptic density, resembling a tiny coin (Dimitri Kullman, personal communication). Any exocytosis of AMPAR directly into the synapse would necessitate that large proteins (the AMPAR) should penetrate the PSD. But under the conditions of centrifuge the PSD is clearly a very solid structure in cellular terms.

at the PSD, then there must a transport process to and from the synaptic site. AMPA receptor diffusion has been directly observed on the surface of hippocampal cells using both latex beads and single molecule fluorescence microscopy. (Borgdoff and Choquet, 2002; Tardin et al., 2003; Groc et al., 2004). These studies have discovered that within sub μm^2 domains of extra-synaptic space the mean squared displacement (MSD) of the random AMPAR movements observed scales linearly with time, as is characteristic of Brownian motion.

Experiments suggest that the cell membrane incorporates a patch work of 'picket fences' composed from proteins anchored to the cytoskeleton. These fences are not fixed but can provide reconfigurable confinement of freely diffusing membrane proteins. We can regard the AMPARs within sub-domains of these corrals to be freely diffusing, accounting for the observed MSD of extra-synaptic AMPARs (Triller and Choquet, 2003). Confinement of AMPARs to synaptic domains is most likely accounted for by interactions of AMPARs with the PSD and trans-membrane proteins (Triller and Choquet, 2005).

Finally, a particularly intriguing aspect of diffusion of AMPARs is that the mobility of AMPARs in the cell membrane may be altered by calcium concentration (Borgdoff and Choquet, 2002). Chelation of Ca^{2+} with BAPTA leads to more than a two-fold increase in AMPAR mobility in cultured hippocampal neurons while UV uncaging of Ca^{2+} leads to an almost complete obliteration of AMPAR diffusion that subsides around $100s$ later. This points to the possibility that AMPAR diffusion is altered during the induction of LTP/LTD, thus allowing the number of AMPA receptors to be changed and causing modification of the synaptic efficacy.

1.2 Synaptic tagging and capture

As stated above, experiments have suggested that late LTP is protein synthesis dependent and that it persists for a long period of time, while early LTP is not protein synthesis dependent and decays more rapidly. This raises the question of how the proteins required for synaptic weight stabilisation are targeted to the correct synapse when that synapse undergoes late LTP.

One explanation would be that the plasticity related proteins (PRPs) required for eLTP to become lLTP are manufactured locally to the synapse, thus negating the targeting problem. This solution seems unlikely since it would require protein synthesis within dendritic spines. Alternatively, the cell might operate an indexing system of

some sort, allowing proteins to be precisely targeted by means of a sophisticated transport system.

While investigating the targeting of PRPs, Frey and Morris (Frey and Morris, 1997) found an interesting result that cannot be explained by either of the mechanisms mentioned above. Late LTP of one pathway (eg. input S1, Fig. 1.1) in hippocampal slice preparations could be induced even after blockade of protein synthesis with anisomycin provided that late LTP had been previously induced on a second, anatomically nearby pathway (eg. input S2, Fig. 1.1). This implies that the induction of late LTP triggers a release of PRPs that is widespread enough for other synapses to capture them. This finding is not consistent with either of the above explanations for the specificity of PRP uptake, since neither of these explanations can explain the paradoxical transformation of early LTP into late LTP on the second pathway.

The discovery of paradoxical late LTP, induced under protein synthesis blockade, gave way to the possibility that early LTP might be converted to late LTP under other circumstances. Another experimental result showed that indeed this is the case. After inducing lLTP on one input (S1) using strong HFS, subsequent weak HFS applied to a second nearby input (S2) leads to the induction of lLTP at that input (S2) also (Frey and Morris, 1997; Frey and Morris, 1998). Weak stimulation that would have otherwise only induced eLTP in S2 instead induces lLTP due to the strong stimulation of another input (S1).

Paradoxical late LTP can be explained with the hypothesis that strong HFS does two things: Firstly it triggers some global process of transcription and translation (involving the polyribosomes of the particular dendritic branch or throughout the whole cell), secondly, the strong HFS sets a marker or 'tag' on the synapse to indicate that new proteins should be used to stabilise the synaptic weight. In addition it is hypothesised that weak HFS sets a tag but does not activate protein synthesis. Paradoxical late LTP can now be explained because the second input (S2) is able to sequester PRPs that were manufactured due to the strong stimulation of the first input (S1), leading to stabilisation. This is the synaptic tagging hypothesis, Fig. 1.2.

It should be emphasised that tagging can allow the consolidation of early LTP after a considerable time elapse, as shown in experiments where first strong HFS is administered to input S1 followed 35mins later by a dose of Anisomycin to block protein synthesis. Performing strong HFS at another input S2 after a further 25mins provokes lLTP despite ablation of protein synthesis. Thus PRPs are capable of being retained for some time.

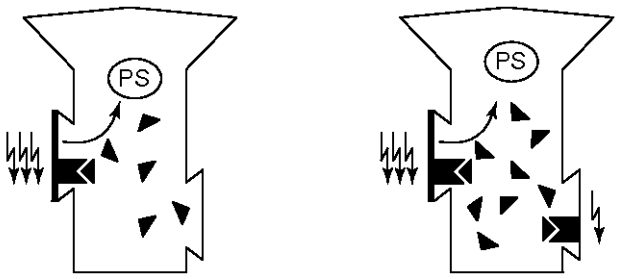

Figure 1.2: The Synaptic tagging hypothesis. Each diagram is a schematic of a cell body attached to a dendrite having two synapses. In experiments, these cells are located in the CA3 region of the hippocampus. When late LTP is induced in one of the synapses (left), then the synapse is potentiated, a tag is set and protein synthesis (PS) is activated. The protein synthesis produces plasticity related proteins (triangles) that stabilise the synaptic modification. The tag captures plasticity related proteins. Subsequently, early LTP is induced in a second synapse (right). Early LTP does not activate protein synthesis, but does set a tag. The tag captures plasticity related proteins that are still present due to prior protein synthesis from late LTP. Thus the synaptic modification induced by early LTP is stabilised, converting early LTP in to late LTP. This figure was taken from Frey Morris 1998.

Further interesting results suggest that the above phenomenon also holds for LTD. Again lLTD requires protein synthesis and is blocked by Anisomycin. It is possible to convert eLTD into lLTD with a strong low frequency stimulus (LFS) applied to another input within a time window of the order of 80mins. More exotically, 'cross-tagging' can be induced, whereby strong LFS in one input can allow another input to display lLTP after application of weak high frequency stimulation (HFS) and visa-versa (Sajaykumar and Frey, 2004).

As previously mentioned, LTP experiments that use the hippocampus typically stimulate many afferent fibres. In doing so it is likely that in addition to the fibres being targeted, dopamine inputs are also stimulated. There is evidence that this dopamine release and subsequent activation of D1/D5 receptors is necessary for lLTP and late associativity (Li et al., 2003; Sajaykumar and Frey, 2004).

The molecular identities of the tag and the PRPs are as yet unknown. It is plausible that the tag is the result of phosphorylation of some synaptic molecule that would enable the synapse to incorporate PRPs. This is a difficult idea to test experimentally because phosphorylation of molecules is a ubiquitous process in synaptic plasticity. Non specific pharmacological blockade of phosphorylation simply blocks plasticity it-

self (Malenka, 1994). In theory it might be possible to find a drug that can specifically block phosphorylation of the substrate molecule (this would isolate the tag) but this is at least as hard as finding a needle in a haystack. The identity of the PRPs is also uncertain. It has been suggested that they are not incorporated into the synapses at all, but that rather they are enzymes allowing alteration of the composition of the postsynaptic density. (Sajikumar and Frey, 2004).

1.3 The Hippocampus and declarative memory

It has long been known that loss of the hippocampus leads to terrible memory deficits in human beings. Most famously in 1953 the patient H.M. underwent bilateral removal of his hippocampus in order to cure intractable epilepsy. Although his epilepsy improved, the procedure rendered H.M. unable to retain new memories (anterograde amnesia) (Schoville and Milner, 1957; Corkin, 2002). H.M. can no longer learn new episodic or semantic memories (declarative memories), although his recall of earlier events, language and facts is undisrupted (Corkin, 2002). This evidence strongly supports the idea that the hippocampus is required in order to retain incoming declarative memories, but that it is not the repository of long term declarative memory. Subsequent lesioning studies in animals (Squire and Zola-Morgan, 1991; Alvarez, Zola-Morgan, and Squire, 1995), pharmacological investigation by blockade of NMDA receptors with AP5 (Steele and Morris, 1999) and electrophysiological intervention in the hippocampus (Moser and Morris, 1998) all lend credence to this hypothesis.

Each individual experiment always suffers from confounding factors. For example in vitro studies of hippocampal tissue can be more tightly controlled but are atypical of conditions in vivo, while in vivo studies are more realistic, but any intervention with the hippocampus, be it surgical or pharmacological must have effects on other parts of the brain. Despite these caveats, the sheer weight and diversity of convergent evidence that the hippocampus is a short - or perhaps more accurately *variable* - term memory store that is important for formation of declarative memories is overwhelmingly persuasive.

The first detailed computational model of the hippocampus was developed by David Marr (Marr, 1971). In Marr's scheme the hippocampus is regarded as a temporary content addressable memory store. In this model, the hippocampus stores incoming information by modifying the synaptic weight between two or three layers of simple binary units (both configurations were investigated) such that upon subsequent presentation of an incomplete version of the pattern, the original pattern is reproduced. The argument

is made that it would be inefficient for the neocortex to store all of the new information that the animal encounters. For this reason Marr postulates that the hippocampus is a non specific memory store, retaining new information until its usefulness can be assessed. Marr's model has been hugely influential, however the validity of assumptions that were made about the anatomy and hence the solidity of Marr's conclusions have since been questioned (Willshaw and Buckingham, 1990).

Later theories built upon Marr's model by proposing specific functions for subhippocampal regions within a more developed theoretical framework. The importance of correlations stored within an associative matrix mediated by the numerous recurrent connections in CA3 was emphasized by McNaughton and Morris (McNaughton and Morris, 1987) and built upon the previous development of the theory of associative networks (Willshaw, Buneman, and Longuet-Higgins, 1969). Also centering upon the possible importance of CA3 as an associative network, a large body of theoretical work was undertaken by Treves and Rolls (Rolls, 1996; Rolls and Treves, 1998). In this model of the hippocampus, CA1 is taken to be a relay centre from CA3 to the wider brain. As part of this relay process CA1 is hypothesized to recode neural activity in CA3. Representing conjunctions of firing cell assemblies in CA3 (episodic memories) as the firing of a single cell assembly in CA1, where the number of neurons used to represent the recoding is larger in CA1 and leads to reduction in 'density' of the original code (sparsification). Calculations showed that this process requires associatively modifiable synapses between CA3 and CA1 in the Shaffer collateral pathway.

The notion that the hippocampus functions to associate information across modalities is pervasive, but there are still many undecided details as to how structure relates to function and how function relates to behavior. One important idea incorporated into the hippocampal model of O'Reilly and McClelland (O'Reilly and McClelland, 1994) is that incoming patterns, from the ethorinal cortex to CA3 should be orthogonalised by the dentate gyrus in order to minimise the overlap of synaptic activation. This maximises representational efficiency. In this case the hippocampus can be regarded as not only having the ability for pattern completion (as an autoassociative network) but also the ability for pattern separation (by orthogonalisation of inputs). There is evidence that both CA1 and CA3 are capable of pattern completion and separation (Guzowski and Knierim, 2004), although recent functional imaging data suggests that CA1 is concerned with pattern completion, while CA3 and the dentate gyrus are concerned with pattern separation (Bakker et al., 2008). This dual function of the hippocampus could allow orthogonalised CA3 patterns to be re-associated with the ethorinal inputs that

gave rise to them such they can be indexed by the cortex.

The above discussion has made no mention of the other functions of the hippocampus. The models above all cast the hippocampus as primarily a device for the acquisition of declarative memory. However there are numerous observations that spatial location is coded by activity of cells in the hippocampus (O'Keefe and Dostrovsky, 1971; O'Keefe and Conway, 1978; O'Keefe and Burgess, 1996). In place field research, the role of the hippocampus in spatial learning and memory is investigated. Place coding in hippocampal function, while important in its own right, is also probably relevant to the acquisition of declarative memories in that it provides a method to bind multimodal memories together with an appropriate que [3]. However, the influence of place cells is not the domain of the models of LTP to be discussed in this thesis. For the purposes of what is to come it is only necessary for us to acknowledge that in making memories the hippocampus utilises a memory trace, and that the modification of synaptic efficacies acts in a manner that has some functional relationship to input and output activity. Therefore it is the arrangement of synaptic efficacies that provides the substrate for recall of information. To this end we note that all major theories of hippocampal function agree on this point although they might disagree upon the mechanisms at play and the precise nature of the memories mediated. Furthermore, the synapses between CA3 and CA1 are good candidates for the location of memory traces, although the existence of memory traces is by no means necessarily limited to this synaptic population.

1.4 Spike Timing Dependent Plasticity

The LTP/D induction protocols mentioned previously activate large numbers of synapses, creating a large synaptic drive in a non specific manner. In these protocols there is no specific structure in the time intervals between individual presynaptic and postsynaptic spikes. Alternatively, it is reasonable to wonder how synapses behave when there is a particular sequence of timings between presynaptic and postsynaptic spikes. In several systems it has been demonstrated that the precise spike timing has a large effect on the outcome of synaptic plasticity (Levy and Steward, 1983; Bell et al., 1997; Markram et al., 1997; Bi and Poo, 1998; Sjöström, Turrigiano, and Nelson, 2001; Froemke and Dan, 2002; Dan and Poo, 2006). Importantly, at a fixed low frequency,

[3] This is perhaps of most importance in episodic memory for which 'where' is one of the tripartite elements of its definition: Where, what, when. However it can also be relevant for semantic memory, as attested to by memory techniques that boost ones acquisition of facts by association with previously learned spaces, such as the interior of a familiar house.

the exact ordering of pre and postsynaptic events can determine whether the synapse undergoes potentiation or depression. This phenomenon is known as Spike Timing Dependent Plasticity (STDP). The relationship of STDP to classical LTP is still not understood, although it is often presumed that STDP 'learning rules' give rise to classical LTP behavior in the limit of a large number of spikes.

Spike timing dependent plasticity can be induced between cortical layer 5 pyramidal neurons in slice preparations. Individual pre-synaptic spikes (leading to post-synaptic EPSPs) cause synaptic potentiation when paired with post-synaptic spikes, provided that the post synaptic spikes follow the EPSPs in time (Markram et al., 1997). Later experiments in cultures of hippocampal pyramidal neurons, established that the weight modification obtained by the spike-pairing is dependent upon the timing between those spikes in such a fashion so as to mark out a 'plasticity window'. Should the post-synaptic spike follow the EPSP then synaptic modification is large and positive. Conversely if the order of EPSP and spike is reversed, the synaptic modification becomes negative, Fig. 1.3, and in between these limits the magnitude of synaptic modification is continuously graded (Bi and Poo, 1998; Bi and Poo, 2001).

Upon first sight, the plasticity window implies that when some pre synaptic spike train interacts with a post synaptic spike train, then the total resulting synaptic modification is some combination of the modifications from individual spike interactions as determined by the single pairing STDP window. In the simplest case this is the linear sum of all of the individual possible contributions. It is now known that this simplest scenario of linear interaction of the modifications due to each spike paring, does not occur (Bi and Poo, 1998; Froemke and Dan, 2002; Wang et al., 2005). This can be tested by performing 'higher order' probes of spike pairing, where instead of inducing a pre spike and then a post spike, other spikes are introduced. For example, one might probe with spike triplets, in which case a protocol such as pre-post-pre might be performed. If the spike timing intervals interact in a linear fashion one would expect that a pre-post-pre protocol where there is a 10ms gap between the first paring (pre-post) and an equal 10ms gap between the second paring (post-pre) should lead to no overall change, and indeed this appears to be the case (Wang et al., 2005). On the other hand the alternative protocol post-pre-post, should also lead to no overall change due to an identical argument. In this case however, significant LTP was observed. Thus there cannot be a simple summation of the influences of extra spikes. A similar result was obtained in protocols consisting of four spikes, showing asymmetrical activation of LTP and LTD processes where a linear summation model would predict equal acti-

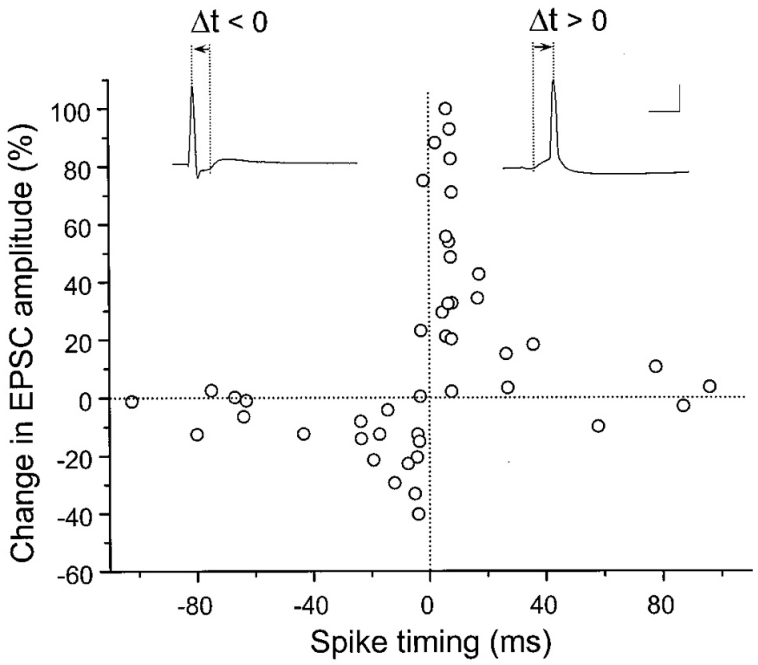

Figure 1.3: Spike timing dependent plasticity in cultured hippocampal pyramidal cells. Two cells were selected that made synaptic contact with one another. A spike is then induced in the pre-synaptic cell. After a time elapse a spike is induced in the post synaptic cell. This protocol is repeated and the average plotted. When the time elapse is negative $\Delta t < 0$ the synapse is depressed. When the time is positive $\Delta t > 0$ then the synapse is potentiated. Figure reproduced from Bi & Poo 1998.

vation of LTP and LTD. Evidence for the non-linear summation of spike interactions has been obtained by other authors including the case of natural spike trains in vivo (Froemke and Dan, 2002).

1.4.1 Theories of spike interactions

From a theoretical perspective, the issue of summation of individual STDP spike pair contributions is complex. There is no one model yet that can provide a general solution to this problem. Some models permit heuristic calculations based upon rules about which spike pairs can interact (Burkitt, Meffin, and Grayden, 2004; Izhikevich and Desai, 2003). Spike triplet data can be modeled (Pfister and Gerstner, 2006) but already the data tell us that the complexity does not stop with the third spike. Perhaps most relevantly, bursts of spikes and naturalistic spike trains can be modelled in a simple model that takes into account the previous synaptic weight change (Froemke and Dan, 2002; Froemke et al., 2006). This approach is perhaps the most promising, for while it might satisfy theoreticians, there is little point in exerting disproportionate effort in deriving a model that can account for spike interactions that are highly unlikely to occur in vivo (as is the implicit aim of solving for a completely arbitrary spike train).

Still another approach is to build a specific system of equations based upon proposed dynamics of cellular chemicals such as calcium. This has met with some success but generates models of high complexity (Rubin et al., 2005). One model proposes a more general solution, and indeed the model is capable of qualitatively reproducing much of the observed phenomenology. But it is mathematically intense and as yet untested against real data (Appleby and Elliott, 2006; Appleby and Elliot, 2005). In the STDP investigation in this thesis, these complexities are not considered. Instead two popular models of STDP that are directly based upon experimental data are investigated (van Rossum, Bi, and Turrigiano, 2000; Song, Miller, and Abbott, 2000) (see chapter 4).

1.4.2 STDP in the visual cortex

STDP has been successfully elicited in slice preparations of visual cortex (Sjöström, Turrigiano, and Nelson, 2001), suggesting that it might be physiologically relevant to visual development. There is also indirect evidence that STDP occurs in the adult visual cortex in vivo. It has been found that the optimum response angle of orientation selective cells in V1 of the cat can be shifted by repetitive presentation of stimuli that

cause neighboring columns to fire (Yao and Dan, 2001). When the stimuli are appropriately timed a window like function, similar to that discovered by Bi and Poo in hippocampal neurons can be plotted, relating the magnitude of the shift in optimum angle of the orientation column, to the relative timing between shift inducing stimuli. In the same study the authors demonstrate an analogous psychophysical effect in humans. One possible explanation for this data is that STDP is occurring in the lateral excitatory connections between adjacent columns. Subsequent high level modelling work reproduced the effect (Yao, Shen, and Dan, 2004). All together this was taken to suggest that STDP is potentially relevant to computation within V1 in vivo.

Recently there has been further evidence of the existence of STDP in the visual system. Upon retinal lesion the receptive fields of the visual cortex can reorganise. In cats, this reorganisation is highly convergent; receptive fields move in to plug the gap in the area having low activity. It has been shown that this convergent reorganisation process is compatible with a model that takes causal spike interactions in to account (i.e. STDP) but is incompatible with a model based purely upon correlations in spiking activity (Young et al., 2007). This again suggests that STDP might play a role in the adult visual cortex in-vivo.

1.5 Summary

In this chapter the key elements of the synaptic plasticity and memory hypothesis (SPM) have been set out. The elements are:

1. Synaptic plasticity underlies the initial encoding of the memory trace
2. Memory traces are stored as the values of an ensemble of synaptic weights
3. Memories are directly mediated by memory traces

Memory is an aspect of cognition and therefore touches upon a large proportion of the areas of the brain. For this reason the discussion of memories was mainly restricted to the hippocampus, which plays a pivotal role in the formation and initial storage of declarative memories. If the SPM holds then the processes that are responsible, at the low level, for the formation of memory are processes of synaptic plasticity such as STDP, LTP and LTD.

At the biophysical level changes to synaptic efficacy are mediated by a variety of processes. It is possible that these processes give rise to the multiple dynamic

timescales of LTP/D that are observed in experiments. In this thesis two processes are considered to allow direct potentiation or depression of the synapse: 1) The addition or removal of AMPAR: Processes such as exocytosis and capture of AMPAR by the PSD could increase the number of AMPAR in the postsynaptic zone. Conversely AMPAR endocytosis and disassociation of AMPAR from the PSD could lead to a decrease in the number of AMPAR in the postsynaptic zone. 2) Retention of AMPAR for longer periods: Consolidation (or deconsolidation) of synaptic weights is considered to be due to the addition or removal of PSD protein 'slots'. The effect of these slots is to capture and hold on to AMPAR receptors that are freely moving in the plasma membrane. This extends the lifetime of synaptic potentiation.

The aim of this thesis is to explore the stability of biological synaptic plasticity. To do this, a number of models of synaptic plasticity are analysed. Before this however, chapter 2 more clearly defines the problem to be considered: The plasticity stability dilemma. Chapter 3 discusses the mathematical methods used in attacking the problem as defined in chapter 2. In chapter 4 the stability of memory traces in two popular models of STDP is considered. Chapter 5 extends chapter 4 to STDP in recurrent networks whose function resembles orientation selectivity in V1, where there is evidence that STDP plays a role.

Novel models of hippocampal LTP/D are formulated and analysed in chapters 6 and 7. In chapter 6 the implications of multi-timescale LTP/D for the memory trace are demonstrated. In chapter 7 amnesia is induced in the models, leading to prediction of a novel phenomenon. Using the framework developed in chapters 3, 6 and 7, a novel model of synaptic tagging is formulated and presented in chapter 8. Finally, chapter 9 discusses the results and suggests future work.

Chapter 2

The plasticity stability dilemma

In the previous chapter the synaptic plasticity and memory hypothesis (SPM) was introduced (Martin and Morris, 2002). In the SPM it is proposed that the encoding of memory is mediated by synaptic plasticity via processes such as long term potentiation (LTP) and spike timing dependent plasticity (STDP) and that the storage of memories is mediated by the synaptic efficacy. One system that has been heavily studied within the SPM paradigm is the hippocampus. We saw how the hippocampus might be regarded as an associative memory device, where memories are stored as relationships between synaptic efficacies connecting the hippocampal layers. This implies that synapses must be used for both learning, (when modified by ongoing experience) and for storage (when required to retain a previously learned memory trace). These dual functions are antagonistic, giving rise to the 'plasticity versus stability dilemma' (Grossberg, 1987; Abraham and Robins, 2005): A trade off between strong learning that creates an easily detectable memory trace (thus modifying many of the synapses) and long term storage (through protection of previously modified synapses).

One form of the plasticity stability dilemma can be understood intuitively with the following analogy. Imagine that we were to write a dairy upon a single sheet of paper by writing entries at random locations on the sheet. The entries are written with permanent pen and so cannot be removed. When the first entry is written, it is completely readable. As further entries are added however the probability that a new entry overlaps an old entry grows. Depending upon the size of the paper and the length of the entries there will come a point where less and less of the diary is readable. Furthermore, new entries will be unreadable and so our memories 'decay' instantly. At this stage the diary is completely useless. An implication of the SPM is that an identical situation might apply in the hippocampus, with synaptic weights

as the page and memory traces as entries in the diary. This raises the question of how synapses should behave such that they can both store old information and accept incoming information.

Interesting in its contrast to the biological problem, is the engineering solution: Indexing. For example, digital computers use indexes (or 'filesystems') to enable files to be stored on a disk such that their bits do not interfere. This allows the computer to lay down new information while retaining a perfect copy of the old information. This method works very well until the capacity of the disk is reached. At this point it completely fails. There is of the order of 10^6 cells in the hippocampus each having order 10^4 synapses (Rolls, 1996; Megias et al., 2001). Assuming that synapses are binary and that they are all used in memory storage, this equates to a storage capacity of approximately 1.25 Gigabytes.[1] However 'capacity' in this sense, refers to the maximum possible size of one single binary pattern (or some concatenation of multiple patterns). Our intuitions about the brain imply that it operates in a different regime, in which random collections of synapses are used to store information with an occasional overlap with old information. This explains why we do not observe human memory becoming full, but we do observe a gradual degradation in some previously acquired information (forgetting). Thus in the context of the brain, a more natural measure of capacity is the maximum amount of time until a memory can no longer be retrieved, given the ongoing storage of new information.

In this chapter the plasticity stability dilemma, is introduced. First, previous work aimed at ameliorating the plasticity stability dilemma (sometimes referred to as catastrophic forgetting) in neural networks is reviewed. Next it is argued that synaptic weight evolution in models of synaptic plasticity can be regarded as a stochastic process. This allows stability to be studied in terms of the statistics of the stochastic process. Existing models that adopt this viewpoint are reviewed. Finally the problem to be addressed by this thesis is stated.

[1] The Shannon information content of such a disc is proportional to the probability of obtaining a *particular disk* rather than the number of bits on the disc. Therefore just because a disc is big, it is not guaranteed to be informative. This is because by reading every bit on a 1.25Gb disc, we can only be as surprised as the number of possible discs we could have been given (i.e. many of the bits might be redundant). Alternatively, the algorithmic information measures the information of the instance of a single disk, but is not guaranteed to be computable!

2.1 Catastrophic forgetting

In the cognitive science and neural networks literature the stability plasticity dilemma has been recognised for some time. It is often referred to as catastrophic forgetting in this context. There are two meanings that are attributed to catastrophic forgetting: Catastrophic forgetting of the first kind occurs when an artificial neural network (typically a back propagation network) suffers very rapid performance reduction on a previously learned task when new data occurs. For example a multilayer perceptron might be trained to classify some set of patterns A_1. The learning algorithm terminates when a certain level of classification error is reached. The network is now presented with a new set of patterns to learn A_2. After learning of A_2 is complete, members of A_1 are presented to the network and it is found that the network can no longer classify or recognise patterns in A_1. This forgetting is catastrophic because it typically involves removal of all memory of A_1, it occurs rapidly, and A_2 need not be a large set of patterns (French, 1999). Training a connectionist network on only one or two novel patterns still completely disrupts the initial memory (Grossberg, 1987; French, 1999). This is a significant problem because catastrophic forgetting does not appear to occur in experiments (French, 1999). For example, it is not the case that learning a new word leads to the complete disruption of one's vocabulary!

Catastrophic forgetting of the second kind occurs when patterns are stored in a neural network (typically a Willshaw or Hopfield network) when the storage capacity of the network has been exhausted. In this case we subject the network to on going learning of patterns. When the capacity is breached the network forgets its previously learned patterns rapidly.

If neural networks are to be regarded as models of memory then catastrophic forgetting is an embarrassing problem. Solutions to the problem of catastrophic forgetting are now summarised.

2.1.1 Adaptive resonance

If the neural network is trained with all examples that it must learn and then the weights are frozen, no catastrophic forgetting can occur. However this strong demarcation between training and learning phases is unbiological, a point that was forcefully made by Grossberg (Grossberg, 1987). There are areas of the brain, particularly those involved in memory that must continue learning at some useful rate. If this were not the case, our memories would be static. The solution proposed by Grossberg is the principle

of adaptive resonance whereby the network first assess the novelty of new input before any learning occurs (Grossberg, 1987; Carpenter and Grossberg, 2003). Should there be a high degree of similarity (or resonance) between the established coding implemented in the network and the novel input, then the pattern is learned by modification of existing weights. On the other hand if the pattern appears highly dissimilar to learned patterns then it is incorporated into the memory store via the recruitment of new nodes. Thus the previously learned weights are not catastrophically disrupted by new learning. This method of addressing the problem of catastrophic forgetting requires that the network be composed of different sub networks performing different roles, such as novelty detection. It also requires that memory traces themselves are stored in different parts of the network. The ART networks and derived models fall into a category of so called 'localist' approaches that aim to address the stability plasticity dilemma by storing learned representations within different parts of the neural network, hence strongly reducing the potential overlap of patterns to be stored (French, 1999).

2.1.2 Rehearsal and psuedorehearsal

Catastrophic forgetting of the first kind can be significantly reduced with the application of another method known as rehearsal (French, 1999; Robbins, 2004). Rehearsal mixes previously learned patterns with new patterns to be learned, by holding several old patterns with each new pattern in a 'rehearsal buffer'. The network (typically a backpropagation network) is then taught the patterns held in the rehearsal buffer. Intuition as to how this works can be gained by first remembering that a backpropagation network fits a function to the input/output mapping in some high dimensional space. When relearning is induced without rehearsal, the learned function is globally adjusted to fit the new data, even if only one new data point is presented. The repeated learned patterns have the effect of preventing a global remapping of the function encoded by the weights. Instead the new item leads to a local change to the input/output mapping. Continuous presentation of new patterns with rehearsal thus prevents disruption of the previously learned memory. Rehearsal is effective but suffers from the problem that all of the previous memories must be held such that they can be randomly selected and added to the rehearsal buffer. This requires that all input patterns be stored somewhere for later use. Biologically this seems implausible, but worse still, it begs the question of how the memories are stored and retained in the buffer. Psuedorehearsal removes

this caveat (Robbins, 2004). Since the network encodes the mapping between inputs and outputs that is to be preserved, inputs that are randomly generated and passed through the weights result in input/output pairs characterising the existing mapping. Thus rehearsal items can be generated by the network, requiring no additional storage of input patterns.

2.1.3 Associative memory networks

The models of hippocampal function described in chapter 1, all rely heavily on associative and autoassociative networks. In an autoassociative network (Hopfield nets) with N units and Hebbian learning, catastrophic forgetting of the second kind occurs when the number of patterns stored reaches the region of 0.14N, whereupon the network quickly degenerates into meaningless 'spin glass' activity states (Amit, Gutfreund, and Sompolinsky, 1985). Catastrophic forgetting of the second kind also occurs in associative networks (Willshaw nets). The point at which this occurs depends upon many details of the network, such as the degree of connectedness of the units, the sparseness of the coding and the readout scheme (Dayan and Willshaw, 1991; Sterratt and Willshaw, 2008). Catastrophic forgetting can be prevented in both Willshaw and Hopfield networks by allowing the weights to decay to some baseline point over time whereupon the network functions as a *palimpsest* (Nadal et al., 1986; Sterratt and Willshaw, 2008). In this case old memories fade gradually, ensuring that there is always capacity for new memories and that the network never suffers catastrophic forgetting. This is identical to the diary example stated at the beginning of this chapter if the ink were to fade with time. The palimpsest solution is at the expense of the number of patterns that can be stored and reliably recalled, which is reduced for standard learning rules.[2]

One way to mitigate against catastrophic forgetting of the second kind is to *sparsify* the input pattern coding (Dayan and Willshaw, 1991; Golomb, Rubin, and Sompolinsky, 1990). This can be achieved by casting the input pattern in a higher dimensional space than its initial description. For example binary patterns consisting of 10 bits might be stored in a network having 10000 binary synapses. This reduces the probability that stored patterns overlap. This can be combined with a process of *orthogo-*

[2]There is an exception to this caveat for Hopfield networks. A high capacity learning rule has been formulated by Storkey that gives a 0.25N capacity for a palimpsest-like network. This learning rule selectively decays patterns that interfere with recall rather than forcing all old patterns to decay. This is in contrast to typical Hebb like rules which deliver a capacity of only 0.05N in the palimpsest regime (Storkey and Valabregue, 1997).

nalisation that recodes the patterns such that they interfere with each other to the minimum extent possible. Both of these processes have been suggested as storage strategies that might be employed by the hippocampus (O'Reilly and McClelland, 1994; Guzowski and Knierim, 2004). The combination of sparsification and orthogonalisation much reduce the overlap of stored patterns and so lessen the interference between them. Hence more patterns can be stored before forgetting occurs. Returning to our diary analogy: Sparsification would reduce the size of the text (or perhaps code it in a new alphabet such that each entry is shorter). Orthogonalisation would ensure that entries are arranged on a grid on the paper, thus preventing overlap of the writing.

2.2 The stability plasticity dilemma of stochastic biophysical synapses

Neural activity is both stochastic, and as we saw in chapter 1, is capable of modifying synapses. Thus synaptic modification in-vivo is stochastic. The origins of this stochasticity are myriad, some important contributions are: Neural activity is stochastic. Stimuli triggering activity leading to potentiation and depression events is likely to occur with a degree of randomness in the natural environment. Synaptic transmission at central synapses is highly stochastic with low quantal content (Stevens and Wang, 1995), meaning that 'failures' to transmit a spike often occur. Therefore, in the central nervous system even if all other conditions permit, pre and post synaptic activity do not necessarily coincide at every possible opportunity, because the synaptic transmission can fail, preventing excitation of the post synaptic neuron by that particular spike.

Furthermore, there is evidence that synapses do not hold continuous real values, but rather are potentiated by all or none events (Petersen et al., 1998; O'Connor, Wittenberg, and Wang, 2005). This implies that the synaptic weight has discrete values. Assuming discrete synapses and random neural activity, synapses jump randomly from weight state to weight state with time. If there are a finite number of states then the synapses must at some point form a steady state distribution across those states[3]. Individual synapses fluctuate between these states and this gives rise to the

[3]Assuming that the timescale of equilibration of the weights is not longer than the lifetime of the organism. It is possible that for some synaptic populations, equilibration might never be achieved: Consider a 'Taj Mahal ensemble' in my brain. Being an Englishman, I might only ever see the Taj Mahal once in my lifetime. However, in principle I could visit it an infinity of times, each time viewing it in a subtly different way. For example, on some occasions it might be undergoing maintenance work and be underneath scaffolding. Would the synapses representing my memory trace of the Taj Mahal be

decay of the memory trace. This viewpoint is in contrast to many approaches in artificial neural networks, where synapses typically have one or several of the following properties: Continuous valued, unbounded, no hidden variables (i.e. there is only a single real number representing 'weight') (Amit and Fusi, 1992; Amit and Fusi, 1994; Fusi, Drew, and Abbott, 2005; Fusi and Senn, 2006).

Once a set of synaptic states has been defined, the state based approach assumes that the transitions between the synaptic states are stochastic and governed by transition probabilities. The transition probabilities are dependent on the neural model, the 'learning rule' and the stimulus. The evolution of the synaptic weight thus defined is assumed to be a Markov process (i.e. memoryless, see chapter 3). Viewing synapses in a Markovian light demands a distinction between equilibrium and non equilibrium regimes.

2.2.1 The stationary and non stationary stability plasticity dilemma

Two forms of the plasticity stability dilemma can be defined: The stationary plasticity stability dilemma (SPS) and the non stationary plasticity stability dilemma (NPS).

In the SPS, the system is taken to be at equilibrium. The probability distribution amongst the states can be denoted by a vector $\mathbf{p}(t_0+t)$. After an infinite time $t_0 \rightarrow \infty$ the equilibrium state $\mathbf{p}^\infty$ is established. In this case the mean synaptic weight is constant on average. In the case that synaptic modification is activity dependent, this implies that the activity is also stationary and hence the transition probabilities between the synaptic states are constant.

In the NPS the transition probabilities between the states are permitted to be changing, giving rise to synaptic weights that are not stationary. Alternatively, the transition probabilities may be constant but $\mathbf{p}(t) \neq \mathbf{p}^\infty$ so that the mean weight is not constant. The NPS also encompasses situations in which changing concentrations of neuromodulators or drugs are effecting the internal processes in cells, thus altering plasticity.

Situations in which the input ensemble is constant, and the synaptic weights are governed by stationary processes are examples of the SPS. In this case, the degradation of memory is due to ongoing stochastic transitions in the weights around the ensemble equilibrium point. A palimpsest associative network with uniform stored pattern statistics, that has reached the steady state (i.e. when the mean weight and variance of the weight are constant) and that is subject to ongoing learning belongs to the SPS.

---

different in the standard me (who has only seen it once) as compared to the infinite me (who has seen it under an infinity of conditions)?

Alternatively, situations in which the variance or mean of the synaptic weights is varying, belong to the NPS. The NPS applies to a Willshaw network with ongoing learning, no depression and no weight decay, because the mean weight is always increasing. The number of potentiated weights shall inexorably increase until the learning either stops or the memory fails.

2.2.2 State based synaptic models

In the previous literature dealing with state based models it has typically been assumed that the Markov process has attained equilibrium and that the transition probabilities are constant (Amit and Fusi, 1992; Amit and Fusi, 1994; Fusi, Drew, and Abbott, 2005; Fusi and Abbott, 2007; Ben Dayan Rubin and Fusi, 2007; Leibold and Kempter, 2008). The resulting Markov process is stationary. Thus the SPS applies.

In state based models, the memory trace can be quantified with the 'ideal observer' approach (Fusi, Drew, and Abbott, 2005; Ben Dayan Rubin and Fusi, 2007). It is imagined that we measure the weights directly after equilibrium has been established. The memory trace is then the ratio of the 'signal', proportional to the sum of weights modified at any instant, to the noise (see chapter 3 for more detail). The signal to noise ratio (SNR) is a measure of the detectability of the memory trace,

$$\frac{S(t_0+t)}{N} = \frac{|\mu_I(t_0+t)-\mu_N|}{\sqrt{\frac{1}{2}(\sigma_N^2+\sigma_I^2)}} \tag{2.1}$$

where μ_I is the mean value of the input (for example the current out of a cell, or the inner-product of inputs and weights) upon presentation of some stored pattern I and μ_N is the mean value of the input upon presentation of some unstored random pattern, having identical statistics to I. The variances in the signal and the noise are denoted σ_I^2 and σ_N^2 respectively and are assumed constant. The memory capacity can be profiled by considering the initial SNR, $S_0/N_0 = S(t=0)/N(t=0)$ calculated with Eq. (2.1) at the instant of memory storage $t=0$ and the timescale, t_{max} over which the signal to noise falls to 1. When the signal to noise ratio falls to 1, the signal is deemed undetectable and the memory trace has been forgotten. Although a standard technique, this is somewhat arbitrary[4]. There are alternatives, for example an information theoretic approach[5]. Below, existing state based models of synaptic plasticity are reviewed and

[4] Use of the signal to noise tacitly assumes that fluctuations are uncorrelated and Gaussian.

[5] Adam Barrett & Mark van Rossum (2008) pre-print

the findings stated, a complete discussion of the analysis techniques is left until chapter 3.

2.2.2.1 2 state model *(Fusi, Drew, and Abbott, 2005):*

We first consider the simplest case: The synapse can occupy 2 states having synaptic efficacy $w \in \{0,1\}$, Fig. 2.1. The synapse is potentiated with the transition $w = 0 \rightarrow w = 1$ (by LTP) with probability α per unit time and depressed, $w = 1 \rightarrow w = 0$ with probability α per unit time (LTP and LTD are balanced). It is assumed that the system has evolved for an infinite time such that equilibrium has been established and there is no longer any change to the mean weight with time. (The 2 state model is analysed in full chapter 3 and in chapter 6, but a sketch is provided here that follows Fusi and Abbott 2005 (Fusi, Drew, and Abbott, 2005).)

The initial signal scales with the probability that the synapse is modified α, multiplied by the total number of synapses in the ensemble Ω, $S_0 \backsim \alpha\Omega$. Ongoing storage tends to overwrite preexisting memory traces by flipping the state of synapses at random relative to their current state. This leads to exponential decay in the initial memory signal, $S(t) = S_0 exp(-t/\tau)$. The noise is constant and is proportional to the standard deviation of the binomial distribution, $N \backsim \sqrt{\Omega}$. Thus the initial signal to noise ratio scales as the square root of the number of synapses for some fixed transition rate α, $S_0/N_0 \backsim \alpha\sqrt{\Omega}$. The signal to noise ratio is,

$$\frac{S}{N} \backsim \alpha\sqrt{\Omega} exp(-2\alpha t) \tag{2.2}$$

and decays exponentially with $\tau = 1/2\alpha$. Importantly, Eq. (2.2) demonstrates the link between the initial signal $\alpha\sqrt{\Omega}$ and the rate of decay of the memory trace, τ. Increasing α increases plasticity thus enhancing the initial detectability of the memory (because more synapses are changed when a pattern is stored). However this is at the expense of a decreased SNR decay timescale, i.e. the memory is less stable. This is the essence of the SPS.

Let t_{max} be the time that it takes for the SNR to reach 1, at which point the memory is irretrievable. Taking the logarithm of Eq.(2.2) reveals that $t_{max} \sim ln(\alpha\sqrt{\Omega})/2\alpha$, and the maximum memory lifetime scales as the logarithm of the number of synapses recruited into the memory trace.

The 2 state process illustrates two worrying points from the point of view of memory storage: 1) The SPS introduces a strict and antagonistic correspondence between the strength of the memory trace S_0/N_0 and the amount of time that the memory trace

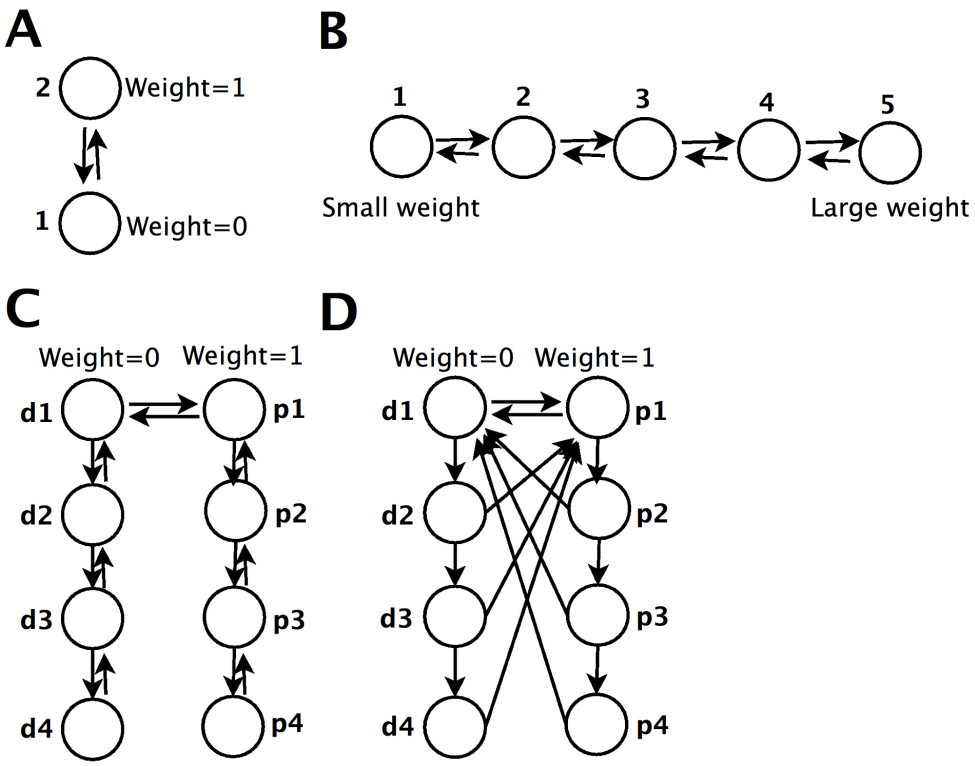

Figure 2.1: Existing state based models of synaptic plasticity. A: The simplest 2 state binary case consists of a depressed state (1) and a potentiated state (2). The arrows are transitions between the states. Each arrow has a transition rate associated with it. B: The linear bounded model has many states, each with a different value of synaptic weight, ranging from a lower bound (1) to a higher bound (5). C: The multistate model, is a binary model where there are many potentiated (p) and depressed states (d), but synapses can only be depressed or potentiated in the top level p1 and d1 states. D: The cascade model is a binary multistate model but every potentiated or depressed state can transition to the top level state (d1 or p1) in the opposite cascade. In the model of Fusi and Abbott, the transition probabilities were arranged such that as synapses move deeper to higher state numbers in the p and d cascades, the probability of making transitions further down the cascade or to the top level of the opposite cascade, is reduced.

can persist for, t_{max}. 2) The maximum lifetime of the memory scales as the logarithm of the number of synapses. Thus, in the long term adding more and more synapses to the memory trace does very little to make it more distinguishable. This is a pity since the one thing that we potentially have at our disposal is a large number of synapses.

The memory lifetime t_{max} can be extended by reducing the value of the transition probability α. The value of α that maximises t_{max} is $\alpha = e/\sqrt{\Omega}$ (Fusi, Drew, and Abbott, 2005). By introducing this scaling between the transition probability and the number of synapses, the memory lifetime can now be significantly improved, $t_{max} \propto \sqrt{\Omega}$. However the initial signal to noise ratio becomes independent of the number of synapses meaning that the detectability of the memory cannot be improved by adding more synapses.

The initial signal to noise ratio can be maximised by setting $\alpha = 1$. In this case the initial signal to noise ratio scales with the square root of the number of synapses $SNR_0 \propto \sqrt{\Omega}$. Thus, in this regime, memories can be made more vivid by modifying additional synapses. Unfortunately the maximum memory lifetime now only scales with the logarithm of the number of synapses $t_{max} \propto ln(\Omega)$. Thus all memories are rapidly forgotten even if huge numbers of synapses are recruited.

2.2.2.2 Linear bounded model *(Fusi and Abbott, 2007):*

In the case of synapses with many weight states, linked in linear increments, the signal decay timescale can be improved relative to the initial signal. However the decay process of the signal is still exponential and so the logarithmic scaling with Ω remains.

Let the synapse occupy one of ζ states, each having a differing efficacy, $w \in \{x_1, x_2, x_3, ..., x_\zeta\}$, Fig. 2.1B. Plasticity events occur at some rate r and some fraction of these events f_+ lead to potentiation while some fraction f_- lead to depression. When a potentiation or depression event occurs the weight is modified with $w \rightarrow w + q_+(w)$ for potentiation and $w \rightarrow w - q_-(w)$ for depression. In the simplest case, the magnitude of the weight updates are independent of the current weight $q_+ = q_- = \alpha$. Boundary conditions are arranged such that any transition that would potentiate the synapse above x_ζ or depress the synapse below x_1 is truncated (hard bounds). The number of potentiation and depression events can be balanced, $f_+ = f_-$ or unbalanced, $f_+ \neq f_-$.

When potentiation and depression events are balanced the initial signal to noise scales with α in the same way as the 2 state case $S_0/N_0 \propto \alpha$ and the memory retention time scales as $t_{max} \propto 1/\alpha^2$. Thus the initial detectability of the memories remains the same, but the addition of weight states improves the signal decay time. In the case that

LTP and LTD are unbalanced, the signal to noise ratio becomes independent of α.

Alternatively soft bounds can be imposed. Whereas hard bounds explicitly forbid potentiation and depression outwith a particular range, soft bounds arise when potentiation and depression force weights toward a central point. Soft bounds can be achieved by setting $q_+(w) = \alpha(1-w)$ and $q_-(w) = \alpha w$. Now, on average, the synaptic weights all 'collapse' on to a single point at $\langle w \rangle = f_+/(f_+ + f_-)$. In this case $S_0/N_0 \propto \sqrt{\alpha}$ while $t_{max} \propto 1/\alpha$. This scaling is independent of whether LTP and LTD are unbalanced.

Generalised softbounds can be imposed on the model by raising the weight dependence of the potentiation and depression to some power γ, $q_+(w) = \alpha(1-w)^\gamma$ and $q_-(w) = \alpha w^\gamma$. Again, in this case it is found that $t_{max} \propto 1/\alpha$ while the initial signal to noise ratio $S_0/N_0 \propto \sqrt{\alpha}$, regardless of the balance between LTP and LTD.

2.2.2.3 Multistate model *(Amit and Fusi, 1994):*

In the multistate model the synapses are again binary $w \in \{0,1\}$, Fig. 2.1C. Now however there is a chain of states where for half of the states $w = 0$ and for the other half $w = 1$. Synapses that are potentiated or depressed can transition deeper and deeper down this chain, however synapses can only make the transition $w = 0 \rightarrow w = 1$ or $w = 1 \rightarrow w = 0$ when at the top of the chain. Hence synapses can diffuse away from the weight changing transition. In the chain there are many distinguishable states that do not effect the weight, these are the *metaplastic* states. In the language of Markov models, the metaplastic states are *hidden* states.

In the multistate model, the initial SNR is similar to the 2 state case, but is reduced $S_0/N_0 \propto 1/\zeta$ as the number of states in the chain, ζ are increased. This is a result of the synapses now being spread across the whole chain, rather than being concentrated in the states that are capable of modifying the weight. In return for this reduced signal, the memory lifetime is increased, $t_{max} \propto 1/\zeta^2$ (Amit and Fusi, 1994; Ben Dayan Rubin and Fusi, 2007).

2.2.2.4 Cascade model *(Fusi, Drew, and Abbott, 2005):*

The linear bounded model demonstrates that by adding more weight states, modest improvements to the scaling between S_0/N_0 and t_{max} can be obtained. The multistate model can extend t_{max} by a potentially significant margin but S_0/N_0 is still reduced. This goes some way to addressing the stationary plasticity stability dilemma,

but does nothing to improve the scaling of the memory trace lifetime with the number of synapses Ω.

In both the multistate and the linear bounded models, the maximum memory lifetime scales as $t_{max} \propto ln(\sqrt{\Omega})$. The cascade model of synaptic plasticity, developed by Fusi & Abbot (Fusi, Drew, and Abbott, 2005) overcomes this problem and provides a memory decay timecourse that is close to optimal in that it combines performance that is nearly as good as the best initial signal and the best memory retention time offered by the 2 state model. In the cascade model, synapses are binary $w \in \{0,1\}$. Again there is a chain of metaplastic states having $w = 0$ or $w = 1$. However in contrast to the multistate model, the synapses are allowed to undergo the transition $w = 0 \rightarrow w = 1$ in all states having $w = 0$ and $w = 1 \rightarrow w = 0$ in all states having $w = 1$. The probability of potentiation or depression depends upon the current state of the synapse. Should a synapse be potentiated or depressed many times, then it will transition to a state that is 'deeper' in the cascade, making subsequent modification less likely, thus the plasticity is consolidated, Fig. 2.1D.

In the simple 2 state model described above, the best initial signal to noise ratio that can be obtained scales as the square root of the number of synapses $S_0/N_0 \backsim \sqrt{\Omega}$, but this is at the expense of a maximum memory lifetime $t_{max} \propto ln(\sqrt{\Omega})$ that only scales as the logarithm of the number of synapses. Alternatively the memory decay can be made long, such that $t_{max} \propto \sqrt{\Omega}$ but this is at the expense that the initial signal is independent of the number of synapses recruited.

These problems with the 2 state model are caused by the presence of the exponential function. Thus the strict antagonism between the initial signal strength and the maximum memory lifetime can be attacked by removing the exponential decay timecourse of the 2 state model. The introduction of multiple synaptic states into the cascade model allows many exponential processes, all with differing timescales, to be superimposed. When the transition probabilities in the model - and hence each of the exponential timescales - are suitably adjusted, an overall decay timecourse that is a powerlaw can be achieved. This is enormously advantageous because the signal to noise now decays as $S/N \propto \sqrt{\Omega} t^{-k}$ having a maximum memory lifetime of $t_{max} \propto \Omega^{1/2k}$. When the transition probabilities are arranged appropriately, the cascade model comes to within a logarithmic factor of obtaining the best scaling possible for a binary synapse. The initial signal improves as synapses are added, $S_0/N_0 \backsim \sqrt{\Omega}$ *and* the memory lifetime improves as synapses are added $t_{max} \propto \sqrt{\Omega}$ (Fusi, Drew, and Abbott, 2005).

2.3 Summary

The synaptic plasticity and memory hypothesis demands that synapses change in order that new memories are created. However the same synapses are required to store those memories. This introduces a trade off between the need for synapses to be plastic and the need for synapses to be stable. This trade off is the plasticity stability dilemma (Grossberg, 1987; Abraham and Robins, 2005).

The plasticity stability dilemma was first explicitly formulated by authors studying connectionist neural networks. As discussed, in this context it is also often referred to as catastrophic forgetting. Catastrophic forgetting occurs when a connectionist network is trained to do some task, for example a classification task. After the task has been learned the network is presented with novel data. If the connections in the network are still able to change, i.e. if plasticity is present in the synapses both before and after the initial learning, then the performance of the network in the learned task is rapidly disrupted. Methods of dealing with this problem in connectionist neural networks (rehearsal and adaptive resonance) were reviewed. These solutions prevent the global remapping of the learned weights by new data instances (French, 1999; Robbins, 2004).

Although adaptive resonance and rehearsal address the problem of catastrophic forgetting, it is not clear how fundamental catastrophic forgetting is in the context of biological networks. In the connectionist literature we commonly see catastrophic forgetting in back propagation networks. In fact the fundamental cause of catastrophic forgetting, the global remapping of the input-output function learned by the neural network, is a result of the chosen neural net paradigm. There is no empirical evidence that back propagation networks are of direct biological relevance. Thus, while catastrophic forgetting is certainly a problem for cognitive modelers, it is not clear whether it is a biophysical problem for the brain. In this chapter it was argued that by regarding biological synapses as stochastic processes, we find the stability plasticity dilemma is potentially a problem confronted by biological synapses as well.

In the state based approach, the plasticity stability dilemma is cast in terms of the detectability (initial signal to noise) of a memory trace (stochastic process) and the amount of time that the memory trace takes to decay (for example, the time for the SNR to reach 1). In the 2 state model we find that increasing the initial signal decreases the decay time of that signal and visa versa. The cascade model allows a high initial signal to be achieved with a long decay time course, where both the decay time course

and the initial signal scale favorably with the number of synapses participating in the memory trace (Fusi, Drew, and Abbott, 2005; Ben Dayan Rubin and Fusi, 2007).

Viewed from the state based perspective, the plasticity stability dilemma has two forms. When the transition probabilities are constant (when for example, the inputs to be learned are statistically uniform) and when the weights have attained equilibrium, the stationary plasticity stability dilemma applies. This is the situation sometimes referred to as 'ongoing learning'. Alternatively, if the transition probabilities are able to change or if the weights are not at equilibrium the non-stationary plasticity stability dilemma applies.

Previous state based models have studied the initial signal to noise and time for initial signal to noise to fall to 1 when a memory trace is stored at equilibrium, by a spontaneous fluctuation in the equilibrium state. Therefore these models have been concerned with the stationary plasticity stability dilemma. In order to overcome the SPS, these models have studied the memory strength and lifetime of memories as a function of the steady state distribution of synapses amongst the states. The best performance can be achieved by optimising the steady state itself, by means of adjustment of the transition probabilities between states.

This thesis explores the plasticity stability dilemma from the perspective of biologically motivated state based models of synaptic plasticity. The aim is not to 'solve' the problem and predict an optimal solution to the SPS in terms of the steady state, but rather to build models based upon experimental data and then asses the impact of the plasticity stability dilemma. To achieve this, transition probabilities are not optimised but instead are constrained by data.

How do models constrained against experiment perform from the point of view of the SPS and the NPS? We shall find that exploration of this question points to some possible solutions that nature may deploy in ameliorating the plasticity stability dilemma in biological synapses.

Chapter 3

Methods of solution of state based models

In order to study the plasticity and stability dilemma, one approach is to regard the dynamics of the synaptic weight as being the result of a discrete stochastic process across some set of states. The lifetime of the memory trace can then be quantified by calculating properties of the time evolution of this process. In this chapter the analysis techniques used to calculate the memory trace lifetime are discussed. Firstly methods of solution for state based models are explained, which allow the evolution of the probability density with time to be found. Once the system is solved in this manner, the initial signal of a typical memory trace can be found and the timescale of decay of the signal can be extracted by calculating the autocorrelation. Finally the relationship between these abstract measures of the memory trace and the observable dynamics of synapses (typically quantities related to the mean synaptic conductance) is discussed.

3.1 Methods of solution of state based models of synaptic plasticity

The models treated in this thesis are Markovian. Markovian systems are said to be 'memoryless'. More accurately: In a Markovian system, the probability of a change to the stochastic variable is dependent only upon the current state of the system and is independent of the history of the process. Thus for a set of samples of the stochastic variable x_n at successive times $t_n; n \in \{1,2,...\}$ the Markov condition states that

$$P(x_n,t_n|x_1,t_1;...;x_{n-1},t_{n-1}) = P(x_n,t_n|x_{n-1},t_{n-1}) \tag{3.1}$$

illustrating that in Markovian systems the conditional probability density at time t_n is dependent on the value x_{n-1} at time t_{n-1} only (van Kampen, 1992).

Markovian systems are of such utility because they allow the probability density of a stochastic process in the next time instant to be completely specified by a probability density at the current time instant and the conditional probability of moving between the states. Consider a non-Markovian stochastic process with some initial realisation x_1 at t_1. Given the joint probability density for this initial realisation we wish to calculate the joint probability densities for all successive future realisations, thus obtaining the evolution of the system. We can extract the sequence of probability distributions

$$P(x_1,t_1,x_2,t_2) = P(x_1,t_1)P(x_2,t_2|x_1,t_1) \tag{3.2}$$

$$P(x_1,t_1,x_2,t_2,x_3,t_3) = P(x_1,t_1,x_2,t_2)P(x_3t_3|x_1t_1,x_2,t_2) \tag{3.3}$$

$$P(x_1,t_1,...,x_n,t_n) = P(x_1,t_1,...,x_n,t_n)P(x_n,t_n|x_1,t_1,...,x_{n-1},t_{n-1}). \tag{3.4}$$

We see that a complete description of the system demands that we know a joint probability density and a joint conditional probability for every sample point of the system in terms of all previous sample points. However the picture is considerably simplified in a Markovian system by application of the Markov condition in Eq. (3.1)

$$P(x_1,t_1,x_2,t_2,x_3,t_3) = P(x_1,t_1,x_2,t_2)P(x_2,t_2|x_1,t_1) \tag{3.5}$$

and so

$$P(x_1,t_1,x_2,t_2,x_3,t_3) = P(x_1,t_1)P(x_2,t_2|x_1,t_1)P(x_3,t_3|x_2,t_2). \tag{3.6}$$

We only need knowledge of the initial probability density and the conditional probability for moving between states. Conditional probabilities such as those in Eq. (3.6) are referred to as transition probabilities.

3.1.1 Numerical integration of Markov models of synaptic plasticity

State based LTP models can be described as Markov chains, which are Markov processes across a finite number of discrete states at discrete points in time. In the case of a Markov chain with ζ states the probability density at time $\mathcal{T} \in \{1,2,...\}$ is a ζ component vector $\mathbf{p}(\mathcal{T})$ and the transition probabilities are expressed as an $\zeta \times \zeta$ stochastic matrix M whose components $0 \leq M_{ij} \leq 1$ are independent of time and are the transition probabilities of jumping from state j to i during 1 timestep. The Markov property

in this case implies that the transition probability between any two successive probability vectors goes as the transition matrix raised to the power of the timestep, $\mathcal{T}$ (van Kampen, 1992). Thus

$$\mathbf{p}(\mathcal{T}) = M^{\mathcal{T}}\mathbf{p}(0) \tag{3.7}$$

allowing the Markov chain to be iteratively calculated with knowledge of only the initial condition $\mathbf{p}(0)$ and the transition matrix. Eq. (3.7) provides a robust and fast method of integrating the evolution of the models[1].

In this thesis the theory of Markov chains is applied to the storage of patterns (memories) within an ensemble of synaptic weights governed by a transition matrix. The interval Δt between each timestep $\mathcal{T}$ can therefore be interpreted as the time between the storage of each pattern, and probabilities in the matrix M are then probabilities of each transition occurring within this interval. It is assumed that only one memory storage 'trial' occurs in each pattern storage interval. Thus it is assumed that the system can only jump between adjacent states in the chain when each pattern is stored. All of the transitions occurring within each storage interval are assumed to contribute to the storage of that pattern, i.e. all plasticity is meaningful, but patterns will interfere between intervals.

3.1.2 Method of Eigenvectors: General solution of state based models

Under certain conditions a solution can be constructed directly from the transition matrix. In this section we briefly outline a method for obtaining solutions for the model that is semi-analytic, in that we can anticipate the form of the solutions, although eigenvectors and eigenvalues must be numerically computed.

One approach for obtaining solutions for the time dependent probability densities of stochastic systems is to first formulate a master equation. The master equation is a gain-loss equation for the probability densities of each state. The rate of change of probability of occupancy of each state can be expressed as a sum of the outgoing and incoming flow of probability to and from the other states, where the total probability (i.e. sum of probability of occupancy of all states) is conserved. In the discrete, case the probability density is p_i,$i \in \{1,...,\zeta\}$ and the master equation has a particularly

[1] The theorem of Perron & Frobenius proves that this converges to the steady state solution for all well formed stochastic matrices. Perron & Frobenius is a more general result than the method of eigenvectors in §3.1.2 which is why this method is more robust.

simple form (van Kampen, 1992)

$$\frac{dp_i(t)}{dt} = \mathfrak{n}\sum_i \left(M_{ij}p_i(t) - M_{ji}p_j(t)\right) \tag{3.8}$$

where every second $\mathfrak{n} = 1/\Delta t$ patterns are stored and the elements of M are now the probabilities of making transitions from i to j *per pattern storage interval* Δt. We shall assume from now on that 1 pattern is stored in the weights per second, and so $\mathfrak{n} = 1s^{-1}$, $\Delta t = 1s$ without loss of generality, since this is equivalent to scaling the transition rates. If a higher rate of storage were required, then the rate constants would be uniformly increased. Alternatively if a lower rate were required then the rates would be uniformly decreased. Define the matrix R:

$$R_{ij} = \mathfrak{n}\left(M_{ij} - \delta_{ij}\left(\sum_{l\neq j} M_{lj}\right)\right) \tag{3.9}$$

so that the index l runs over all rows in column j excepting row j and consequently,

$$R_{ij} \geq 0 \text{ for i } \neq \text{ j}$$

$$R_{ij} < 0 \text{ for i = j} \tag{3.10}$$

such that the columns of R sum to zero. Each element R_{ij} has units s^{-1} and is the rate of the transitions from state i to j. Since the columns sum to zero the number of states in the system is conserved. Now we can rewrite Eq. (3.8)

$$\frac{d\mathbf{p}(t)}{dt} = R\mathbf{p}(t) \tag{3.11}$$

having solution,

$$\mathbf{p}(t) = e^{Rt}\mathbf{p}(0). \tag{3.12}$$

Without placing conditions on R, Eq.(3.12) can be manipulated no further. In general R need not be diagonalisable, but if it is then Eq.(3.12) can be used to find a solution in terms of the eigenvalues and eigenvectors of R (under the condition that the eigenvalues are not degenerate). In practice R may not be symmetric but remains diagonalisable if it describes a state diagram that possesses the property of detailed balance (van Kampen, 1992). We shall examine detailed balance in due course, but for now we assume that R satisfies this condition and hence is diagonalisable. We also assume that the eigenvalues of R are not degenerate (we deal with the degenerate case in §3.1.2.1). Eigenvectors of R satisfy

$$R\Phi_k = -\lambda_k\Phi_k \tag{3.13}$$

where λ_k is the eigenvalue associated with the kth eigenvector Φ_k. Assuming that the state diagram of the process contains no isolated disconnected regions (i.e. is irreducible) and that there are no degenerate eigenvalues, we shall find one eigenvalue $\lambda_0 = 0$ corresponding to the eigenvector Φ_0 that when normalised gives the steady state probability distribution and $\zeta - 1$ other eigenvalues and eigenvectors.

For diagonalisable R with non-degenerate λ_k the spectral decomposition of the matrix is,

$$R = VDV^{-1} \tag{3.14}$$

where V is the eigenvector matrix, having Φ_k on each column and D has the eigenvalues λ_k along the diagonal with all other elements 0. We assume that V is invertible. It follows from Eq.(3.14) that

$$\mathbf{p}(t) = e^{VDV^{-1}t}\mathbf{p}(0). \tag{3.15}$$

From the matrix exponential identity $e^{XYX^{-1}} = Xe^{Y}X^{-1}$, we find

$$\mathbf{p}(t) = Ve^{Dt}V^{-1}\mathbf{p}(0). \tag{3.16}$$

From Eq.(3.16) it follows that the general solution to Eq. (3.11) is

$$\mathbf{p}(t) = \sum_k c_k \Phi_k exp(\lambda_k t) \tag{3.17}$$

where the eigenvectors Φ_k are sometimes referred to as *eigenmodes* of the probability distribution. The eigenmodes are the characteristic forms that when linearly combined can describe the approach to steady state from any initial distribution.

The timecourse associated with each eigenmode is dependent upon the corresponding eigenvalue. For some initial condition $\mathbf{p}(0)$ it is therefore necessary to choose the constants c_k such that

$$\mathbf{p}(0) = \sum_k c_k \Phi_k. \tag{3.18}$$

i.e. the c_k are the column vector $V^{-1}\mathbf{p}(0)$ appearing in Eq.(3.16).

Once the evolution of the probability density has been determined, other quantities can be derived from it, for example the mean weight $\langle w(t) \rangle = \sum_i w_i p_i(t)$, where $\mathbf{w}$ is a vector containing the weight value w_i for each state i.

The property of detailed balance is sufficient to ensure that the Φ_k are complete and can thus provide solutions for any initial condition (van Kampen, 1992). Thus, once we have defined the matrix R we can solve for the temporal evolution of the model as follows: First we extract the eigenvalues and eigenvectors of the matrix R numerically.

We shall find one 0 eigenvalue with a corresponding eigenvector that has the form of the steady state distribution $\mathbf{p}^{\infty}$ of synapses amongst the states, in the limit $t \to \infty$.

The solution, $\mathbf{p}(t)$ is then constructed according to Eq. (3.17) and consists of a superposition of the equilibrium eigenvector and $\zeta - 1$ exponentially decaying eigenmodes. If each eigenvalue is non-degenerate, the solution can in principle display $\zeta - 1$ timescales. However in practice, depending upon degeneracy and the precise initial condition, some of these timescales may not be visible and some subset of them will dominate.

3.1.2.1 The degenerate case

So far we have assumed that the eigenvalues of R are non degenerate. If this is not the case then the spectral decomposition of R need not lead to a complete set of solutions because the eigenvectors of R are not guaranteed to form a basis. In this case we require a decomposition of R that supplies a complete set of eigenvectors even under degeneracy of the eigenvalues. Take the Jordan normal form of R,

$$R = QJQ^{-1} \tag{3.19}$$

where J is a Jordan matrix and Q is a matrix of generalized eigenvectors of R. Solution of Eq. (3.19) now proceeds as before, giving

$$\mathbf{p}(t) = Qe^{Jt}Q^{-1}\mathbf{p}(0). \tag{3.20}$$

The Jordan matrix J can be expressed as the sum of a diagonal matrix containing the eigenvalues of R and a nilpotent matrix[2] N, $J = D + N$. This gives,

$$\mathbf{p}(t) = Qe^{(D+N)t}Q^{-1}\mathbf{p}(0). \tag{3.21}$$

A nilpotent matrix is a matrix for which $N^q = 0$ given some positive integer q. Here N is a matrix that is everywhere zero excepting those superdiagonal elements, corresponding to the superdiagonal elements of the Jordan blocks of J. The matrix exponential of this nilpotent matrix can be expressed as a series,

$$e^N = I + N + \frac{1}{2!}N^2 + \frac{1}{3!}N^3 + \ldots + \frac{1}{(q-1)!}N^{q-1} \tag{3.22}$$

[2]Strictly speaking N is not nilpotent for arbitrary J. This is due to the fact that the Jordan matrix can have superdiagonal elements that are 0 whereas a nilpotent matrix is defined as having *all* superdiagonal elements set to unity. However Eq. 3.23 still applies in this case all be it with a smaller value of q (i.e. if the nilpotent matrix has some superdiagonal elements that are not 1 then its q value is reduced).

where I is the identity matrix having identical dimensions to R. From this it follows that

$$\mathbf{p}(t) = Qe^{Dt}[I + Nt + \frac{1}{2!}(Nt)^2 + \frac{1}{3!}(Nt)^3 + ... + \frac{1}{(q-1)!}(Nt)^{q-1}]Q^{-1}\mathbf{p}(0). \quad (3.23)$$

Eq. (3.23) implies that solutions in the degenerate case can contain terms that are polynomial in time. For the degenerate state based models of LTP in chapters 6 and 7 it was found that these terms were negligible in practice and that Eq. (3.23) gives solutions that are numerically very nearly identical to those given by application of the non degenerate method[3]. This is because the eigenvalues are small. The value of q is also low, because the models have only a small number of states. However in state based models having degeneracy one cannot discount the possibility of particular cases exhibiting a departure from pure exponential dynamics of the probability density; although this is not important in practice for any model proposed in this thesis using the parameter values explored.

3.1.2.2 Detailed balance

Previously it was stated that the model should obey detailed balance in order to guarantee that the solution method applies. Systems are often not symmetric, in the sense that the probability of making a transition between two states in one direction is not always equal to the probability of making the reverse transition. However, closed isolated physical systems are symmetric in the sense of microscopic reversibility. Microscopic reversibility states that after equilibrium has been established, no transition should occur *more frequently* in one direction than in the opposite direction (van Kampen, 1992). Thus if there are asymmetries in the transition probabilities, they must be compensated by the equilibrium probability density. Detailed balance can be stated as:

$$M_{ij}p_j^\infty = M_{ji}p_i^\infty \quad (3.24)$$

where $\mathbf{p}^\infty$ is the probability density of the equilibrium state after infinite time has elapsed and Eq. (3.24) must apply to each pair of transition probabilities.

In state based models that have no closed loops other than those formed by precise reverse transitions (i.e. that are one dimensional) it turns out that detailed balance automatically applies. Thus detailed balance applies to all of the state based models that were mentioned in chapter 2 excepting the cascade model. In the case of the

[3]The model of synaptic tagging in chapter 8, does not obey detailed balance and so these methods were not applied there and the numerical method alone was employed.

cascade model, explicit irreversible transitions are introduced meaning that Eq. (3.24) cannot hold.

For topologies containing closed loops, such as ring structures, detailed balance does not always hold for an arbitrary choice of transition rates, even after the steady state has been established. This can be intuited by envisaging that a closed loop in a state diagram can be at a steady state in which there are constant or oscillating flows around the loop.

Consider N states arranged in a ring. Detailed balance requires that the product of the transition probabilities going in one direction around the ring must be equal to the product of the transition probabilities when cycling in the opposite sense (Hille, 2001; Colquhoun et al., 2004), in this case set one transition probability in terms of the others in the ring:

$$M_{12}M_{23}...M_{N1} = M_{21}M_{32}...M_{1N} \tag{3.25}$$

$$M_{12} = \frac{M_{21}M_{32}...M_{1N}}{M_{23}...M_{N1}}. \tag{3.26}$$

Eq. (3.26) must apply to each closed cycle for detailed balanced to apply. The application of detailed balance to specific models is discussed further in chapter 6.

3.2 The memory trace

An effective approach to study the stability of the memory trace is to adopt the 'ideal observer' viewpoint (Fusi, Drew, and Abbott, 2005). We do not consider the specifics of encoding or decoding of memories stored in synapses. This simplifies the problem because in specific schemes of memory storage and retrieval the appropriate measure of the memory trace is dependent upon the chosen encoding scheme. Rather, we imagine that some encoding process alters the weights such that their values at that time instant denote the information to be stored. We also permit that some decoding processes subsequently examines the weights and extracts the stored information. In this thesis the signal to noise ratio is taken to be the measure of possible extraction of information from the memory trace. We make minimal assumptions about the process used to retrieve information from the weights, only that their *values* somehow store that information. Therefore the memory trace is based upon how rapidly the values of the weights are erased by ongoing storage after initial storage of the desired information.

3.2.1 The autocorrelation

To measure the lifetime of the memory trace we wish to quantify how long the precise arrangement of synaptic weights can persist. For this the autocorrelation of the synaptic weights is a suitable measure. Imagine that a memory is stored by weight transitions of the synapses. At the instant of storage, t_0, we take a snapshot of the weights. Next, let the weights evolve for some further time t. After this time the weights still reside in the equilibrium distribution but individual weights have moved due to random plastic state transitions. Now, another snap-shot of the weights is taken; there are now two lists of weights. The *normalised* autocorrelation of the weights is defined as

$$\kappa(t_0, t_0+t) = \frac{\langle w(t_0)w(t_0+t)\rangle - \langle w(t_0)\rangle \langle w(t_0+t)\rangle}{\sqrt{\sigma_w^2(t_0)\sigma_w^2(t_0+t)}} \tag{3.27}$$

where the average, indicated by the angular brackets, is over all realisations of plasticity and σ_w^2 is the variance of the weights. Sometimes the *unnormalised* $\langle w(t_0)w(t_0+t)\rangle$ autocorrelation, is useful. In this thesis, 'autocorrelation' refers to the normalised version, for which $\kappa(t_0,t_0) = 1$. When the unnormalised version is referred to, this shall be made explicit.

Of special interest is the *equilibrium autocorrelation*, calculated when $t_0 \to \infty$. Assume that the system settles into a steady state having $\sigma_w^2(t_0+t) = \sigma_{w,0}^2 = \sigma_{w,\infty}^2$ and $\langle w(t_0+t)\rangle = \langle w_0\rangle = \langle w_\infty\rangle$, where $\langle w_0\rangle$ indicates the mean weights at $t = 0$, $\langle w_\infty\rangle$ indicates the mean of the weights when $t \to \infty$ and $\sigma_{w,\infty}^2$ indicates the variance of the weights when $t \to \infty$. The equilibrium autocorrelation is defined as

$$\kappa_\infty(t_0, t_0+t) = \frac{\langle w(t_0)w(t_0+t)\rangle - \langle w_0\rangle^2}{\sigma_{w,0}^2}. \tag{3.28}$$

The value of $\kappa(t_0, t_0+t)$ gives us a measure of how much of the original trace remains after the elapsed time t and can be considered as the strength of the memory that the system has of its initial state at time t_0. The autocorrelation function is typically a sum of exponentials, with one exponential decaying the slowest. The timescale of this slowest decay is the measure of the memory trace retention time. For brevity we shall henceforth drop explicit inclusion of t_0. From now we assume that the system is already in the steady state at $t = 0$ unless otherwise stated.

One method for calculating the autocorrelation function follows from the Markov formalism. The weight is discretised into ζ states, w_i, $i \in \{1, 2, ..., \zeta\}$ of width δw allowing the unnormalised autocorrelation $\langle w_0 w(t)\rangle$ to be expressed as, (where we

adopt the notation $\langle w_0 w(t)\rangle$=$\langle w(t_0)w(t)\rangle$)

$$\langle w_0 w(t)\rangle = \sum_{ij} w_i(t=0)w_j(t)p(i,t=0)p(j,t|i,t=0). \tag{3.29}$$

In the case of binary weights $w_i \in \{0,1\}$. As we have seen, the probability of occupancy of each state p_i given some initial condition can be expressed as a linear combination of the eigenvectors of the rate matrix,

$$p(j,t|i,t=0) = \sum_k e^{\lambda_k t} C_{ik} s_j^{(k)} \tag{3.30}$$

where $\sum_k C_{ik} s_j^{(k)} = \delta_{ij}$, and $s_j^{(k)}$ is the k-th eigenvector of R (i.e. C is the inverse of the eigenvector matrix). Rearranging we find that

$$\langle w_0 w(t)\rangle = \sum_k \mathrm{e}^{\lambda_k t} \left(\sum_i p(i,t=0)C_{ik}w_i\right)\left(\sum_j s_j^{(k)} w_j\right) \tag{3.31}$$

from which the normalised version $\kappa(t)$ is easily calculated. To find the equilibrium autocorrelation one can insert $p(i,0) = s_i^{(1)}/\sum s_i^{(1)}$, where $s_i^{(k)}$ is the eigenvector with zero eigenvalue (the steady state). The longest timescale of the survival of correlations, Eq. (3.31) is the reciprocal of the subdominant eigenvalue[4] of R_{ij}.

3.2.2 The Signal to Noise Ratio

The signal to noise ratio has been used by other authors in the study of pattern storage in neural networks (Dayan and Willshaw, 1991; Sterratt and Willshaw, 2008). As mentioned in chapter 2 it has also been used in the study of memory trace survival within state based plasticity models (Fusi, Drew, and Abbott, 2005; Fusi and Abbott, 2007).

In state based models of synaptic plasticity, the aim of learning is to store a pattern Y presented to the inputs $\mathbf{x}$ using the weights $\mathbf{w}$. Patterns are retrieved from the weights by later presenting the stored pattern on the inputs $\mathbf{x}_Y$ and examining the signal (calculated from the the inner product of the weights and the pattern). The signal is elevated for stored patterns as compared to unstored random patterns N, presented on the inputs in an identical manner, $\mathbf{x}_N$.

When the pattern to be stored $\mathbf{x}_Y$ is presented to the weights at $t = 0$, we assume that all of the stochastic transitions between potentiated and depressed states happen so as to increase the inner product of the pattern to be stored $\mathbf{x}_Y$ and the synaptic

[4] In cases where degeneracy has little impact and R is diagonalisible.

weights at the instant after storage, $\mathbf{w}_0$. It is assumed that patterns are uncorrelated. Therefore storage of subsequent patterns causes random transitions with respect to the initial pattern $\mathbf{x}_Y$. Hence these patterns are considered noise patterns $\mathbf{x}_N$ and they degrade the signal due to the original pattern. The maximum capacity of a memory system thus defined is the longest time for which memory of the initial pattern $\mathbf{x}_Y$ can be distinguished from $\mathbf{x}_N$.

The memory trace can be quantified by considering the separation of the signal values (obtained upon presentation of $\mathbf{x}_Y$) from the noise values (obtained upon presentation of $\mathbf{x}_N$). The signal to noise ratio is used for this purpose,

$$\frac{S}{N} = \frac{|\mu_Y - \mu_N|}{\sqrt{\frac{1}{2}(\sigma_Y^2 + \sigma_P^2)}} \tag{3.32}$$

where μ_Y is the mean of the signal distribution upon presentation of the stored pattern $\mathbf{x}_Y$ and μ_N is the mean of the noise distribution upon presentation of some unstored noise pattern $\mathbf{x}_N$. The variances of signal and noise distributions are denoted by σ_Y^2 and σ_N^2 respectively. We shall asssume that $\mathbf{x}_Y$ and $\mathbf{x}_N$ are drawn from a single statistically uniform ensemble. In this thesis, signal to noise analysis is applied to models with binary synapses. In this case the mean values of the signal and noise distributions are given by the overlap between the binary input vector and the binary weight vector, namely $\mu_Y(t) = \Omega\langle \mathbf{x}_Y \mathbf{w}(t)\rangle$ and $\mu_N(t) = \Omega\langle \mathbf{x}_N\rangle\langle \mathbf{w}(t)\rangle$ for an ensemble of Ω synapses, giving

$$\frac{S}{N} = \frac{\Omega\,|\langle \mathbf{x}_Y \mathbf{w}(t)\rangle - \langle \mathbf{x}_N\rangle\langle \mathbf{w}(t)\rangle|}{\sqrt{\frac{1}{2}(\sigma_N^2(t) + \sigma_Y^2(t))}} \tag{3.33}$$

where $\sigma_Y^2(t) = \Omega\langle \mathbf{x}_Y \mathbf{w}(t)\rangle(1 - \langle \mathbf{x}_Y \mathbf{w}(t)\rangle)$ and $\sigma_N^2(t) = \Omega\langle \mathbf{x}_N\rangle\langle \mathbf{w}(t)\rangle(1 - \langle \mathbf{x}_N\rangle\langle \mathbf{w}(t)\rangle)$ since we are dealing with binary weights. It is assumed that the noise and the weights are uncorrelated. Often the SNR is reduced to the case in which the patterns and the weights have the same mean $\langle \mathbf{x}_Y\rangle = \langle \mathbf{x}_N\rangle = \langle \mathbf{w}\rangle$ and the same constant variance $\sigma_Y^2 = \sigma_N^2 = \sigma_w^2$. It is also typically assumed that the weights have equilibrated, such that $\langle \mathbf{w}(t)\rangle = \langle \mathbf{w}_\infty\rangle$. In this case

$$\frac{S}{N} = \frac{\Omega\,|\langle \mathbf{x}_Y \mathbf{w}(t)\rangle - \langle \mathbf{w}\rangle\langle \mathbf{x}_N\rangle|}{\sigma_Y}. \tag{3.34}$$

The initial signal is defined as

$$S_0 = \Omega\,|\langle \mathbf{x}_Y \mathbf{w_0}\rangle - \langle \mathbf{x}_Y\rangle\langle \mathbf{w_0}\rangle| \tag{3.35}$$

which is simply the absolute covariance of the weights and the stored pattern, scaled by the number of weights.

As mentioned above, when the pattern is stored, it is assumed that some process causes the random transitions of the weights to act so as to reflect the pattern as closely as possible, i.e. they cause the weights and pattern to become correlated. Thus $\langle \mathbf{x}_Y \mathbf{w}_0 \rangle$ is expressible in terms of the probability that synapses move into potentiated states and the probability that synapses move into depressed states during the time interval of pattern storage. To calculate this we define two matrices[5]; one matrix M_+ contains the transition probabilities for potentiating transitions, for which $w = 0 \rightarrow w = 1$, while the other matrix M_- contains the transition probabilities for depressing transitions for which $w = 1 \rightarrow w = 0$. These matrices are recovered from the transition matrix M as follows: Define the projection matrix, Γ which is everywhere zero excepting diagonal elements corresponding to *potentiated* states, and the projection matrix Γ' which is everywhere zero except for states corresponding to *depressed* states. Using these matrices, $M_+ = \Gamma M \Gamma'$ and $M_- = \Gamma' M \Gamma$, where for each state $\mathfrak{s}$, $\Gamma_{\mathfrak{s}\mathfrak{s}} = 1$ if $w_{\mathfrak{s}} = 1$ and $\Gamma = 0$ otherwise and $\Gamma'_{\mathfrak{s}\mathfrak{s}} = 1$ if $w_{\mathfrak{s}} = 0$ and $\Gamma' = 0$ otherwise. For example in a 4 state model where state 2 and state 3 are potentiated states, Γ would be a matrix that is everywhere zero except for $\Gamma_{22} = 1$ and $\Gamma_{33} = 1$ and Γ' would be everywhere 0 except $\Gamma'_{11} = 1$ and $\Gamma'_{44} = 1$.

The inner product of the weights $\mathbf{w}_0$ just after pattern storage with the pattern that was stored $\mathbf{x}_0$, is composed of an uncorrelated component that arises due to coincidence between the weights and the pattern, $\langle \mathbf{x}_Y \rangle \langle \mathbf{w}_0 \rangle$ and a component due to the transitions of the weights,

$$\langle \mathbf{x}_Y \mathbf{w_0} \rangle = \mathbf{w}^T M_+ \mathbf{p}_\infty + \mathbf{w}^T M_- \mathbf{p}_\infty + \langle \mathbf{x}_Y \rangle \langle \mathbf{w}_0 \rangle \tag{3.36}$$

where $\mathbf{p}_\infty$ is a column vector representing the steady state probability distribution and $\mathbf{w}$ is column vector of the synaptic weight for each state. From Eq.(3.36) the signal is calculated,

$$S_0 = \Omega_+ + \Omega_- \tag{3.37}$$

where $\Omega_+ = \Omega(\mathbf{w}^\mathbf{T} M_+ \mathbf{p}_0)$ and $\Omega_- = \Omega(\mathbf{w}^\mathbf{T} M_- \mathbf{p}_0)$. The initial signal to noise ratio S_0/N_0 is found by calculating the initial noise N_0

$$N_0 = \sqrt{\frac{1}{2}(\sigma^2_{Y,0} + \sigma^2_{N,0})}. \tag{3.38}$$

In the case that the variances of the signal distribution and the noise distribution are equal, the initial signal to noise is

$$\frac{S_0}{N_0} = \frac{\Omega S_0}{\sqrt{\mu_Y(1-\mu_Y)}}. \tag{3.39}$$

[5]Note that the *probabilities* and not the *rates* are used to calculate the SNR.

3.2.3 Relationship between the autocorrelation and the SNR

It is useful to know the link between the autocorrelation Eq. (3.27) and the SNR. Recall that when a pattern is stored the synaptic weight transitions act so as to increase the correlation between the weights and the patterns. We assume that all of the transitions act so as to increase the correlation, i.e. all transitions occurring at that instant lead to synaptic weights that match the binary values of the corresponding input bits. In the limit that all of the weights are permitted to change, this would lead to a perfect copy of the input pattern in the weights. However in useful models, all weights are usually not permitted to change because this causes very fast forgetting. Rather the weights are governed by the transition rates of R and so only a subset of them are permitted to change at the instant when the pattern is stored (on average). This is what gives rise to the link between the initial signal and the transition rates, Eq. (3.37).

Since the pattern stored at instant $t = 0$, $\mathbf{x}_0$ is fixed for all time, the covariance of the weights with the pattern is just a scaled version of the autocovariance of the weights, because the weights $\mathbf{w}_0$ are an imperfect 'copy' of the pattern $\mathbf{x}_0$. Hence,

$$\langle \mathbf{x}_Y \mathbf{w}(t) \rangle - \langle \mathbf{x}_Y \rangle \langle \mathbf{w}(t) \rangle = \eta \left(\langle \mathbf{w}_0 \mathbf{w}(t) \rangle - \langle \mathbf{w}_0 \rangle \langle \mathbf{w}(t) \rangle \right) \tag{3.40}$$

with η the constant of proportionality between $cov(w_0, w(t=0))$ and $cov(x_0, w(t=0))$.

Rearranging Eq. (3.33) and making use of Eqs. (3.40+3.27) gives,

$$\frac{S}{N} = \frac{\Omega \left| \eta \sqrt{\sigma^2_{w,0} \sigma^2_w(t)} \kappa(t) + \langle \mathbf{w}(t) \rangle \left(\langle \mathbf{x}_Y \rangle - \langle \mathbf{x}_N \rangle \right) \right|}{\sqrt{\frac{1}{2}(\sigma^2_Y(t) + \sigma^2_N(t))}}. \tag{3.41}$$

At $t = 0$, $\kappa(t)$=1, $\langle \mathbf{w}(t) \rangle = \langle \mathbf{w}_0 \rangle$, $S/N = S_0/N_0$, $\sigma^2_Y(t) = \sigma^2_{Y,0}$, $\sigma^2_N(t) = \sigma^2_{N,0}$, $\sigma^2_w(t) = \sigma^2_{w,0}$, therefore

$$\eta = \frac{1}{\sigma^2_{w,0}} \left(\frac{S_0}{\Omega} - \langle \mathbf{w}_0 \rangle \left(\langle \mathbf{x}_N \rangle - \langle \mathbf{x}_Y \rangle \right) \right). \tag{3.42}$$

Assuming that the system is initially in the steady state state such that $\langle \mathbf{w}(t) \rangle = \langle \mathbf{w}_\infty \rangle$ and $\sigma^2_{w,0} = \sigma^2_{w,\infty}$, and assuming $\langle \mathbf{x}_Y \rangle = \langle \mathbf{x}_N \rangle$, Eq. (3.41) reduces to

$$\frac{S}{N} = \frac{\Omega \eta \sigma^2_{w,\infty} \kappa(t)}{\sqrt{\frac{1}{2}(\sigma^2_Y + \sigma^2_N)}} \tag{3.43}$$

with

$$\eta = \frac{S_0}{\Omega \sigma^2_{w,\infty}}. \tag{3.44}$$

From this it follows that at equilibrium,

$$\frac{S}{N} = \frac{S_0}{N_0}\kappa(t) \tag{3.45}$$

the signal to noise ratio is the autocorrelation function scaled by the initial signal.

3.3 Relationship of the memory trace to observable synaptic timescales

The signal to noise ratio and the autocorrelation would be extremely difficult to obtain from an ensemble of synapses by experiment. Thus we must consider how they relate to quantities that are experimentally observable. As was noted in chapter 1, contemporary LTP experiments typically measure the EPSP slope. This is thought to be proportional to the synaptic conductance (weight , $w(t)$). Since these experiments sample from a large population of synapses simultaneously and are usually averaged over several trials, this can be thought of as a time dependent ensemble average of the weight w across the synaptic population $\langle w(t)\rangle$. Therefore in state based models of synaptic plasticity, abstract memory trace measures can be related to experimental observations by relating the autocorrelation of the models Eq. (3.27) to the synaptic dynamics, or the *response*, $\langle w(t)\rangle$.

For synaptic state diagrams that observe detailed balance, the timescales of the synaptic dynamics calculated by finding the eigenvalues of the transition matrix apply to both the response functions and the unnormalised autocorrelation functions. This does not rule out the possibility that state diagrams that do not obey detailed balance as defined by Eq. (3.24) can still have matching response and autocorrelation timescales. For example, this is presumably the case in the cascade model because spectral methods can be used in its analysis (Fusi, Drew, and Abbott, 2005; Leibold and Kempter, 2008), and is likely to be due to the high degree of symmetry in the cascade model. One model analysed in chapter 6 also falls into this category because its deviation from detailed balance is sufficiently small that it makes little numerical difference.

When the timescales of the response and unnormalised autocorrelation match, then both the response functions and the unnormalised autocorrelation functions are superpositions of the same timescales. However the *mixture* of timescales in this superposition is not identical unless the system obeys the *fluctuation dissipation theorem*.

3.3.1 The Fluctuation Dissipation theorem

In §§3.3.1-3.3.3 mention of the autocorrelation refers to the *unnormalised* autocorrelation.

The fluctuation dissipation theorem (FDT) relates the microscopic state transitions of a system at static equilibrium, to the macroscopic dynamics of the same system when it is close to equilibrium (Weber, 1956; Kubo, 1966; Berman, 1975; Felderhoff, 1978; Kubo, Toda, and Hashitsume, 1998). Here 'close' means that the system is close enough to its equilibrium that its dynamics can be considered as a linear perturbation from the equilibrium point of the system. In practice this means that the equations governing the perturbed system must be identical to the equations governing the system at equilibrium[6].

The idea behind the FDT is that the same processes that bring the *ensemble averaged* quantities of a stochastic system to equilibrium, are also responsible for the decay of any *specific microscopic* equilibrium state. Intuitively this can be understood by visualising an array of binary 2 state synapses (see chapter 2). At time $t = 0$ we initialise the bistable synapses in some state other than the steady state. For example (assuming that LTP and LTD are balanced) we may choose to initialise $11/20$ of the synapses to be in the depressed state $w = 0$. Given infinite averaging and random transitions we shall then observe that the mean weight tends back to a value of $\langle w \rangle = 1/2$. Consider how this occurs: The rate at which the ensemble mean can approach the equilibrium value of $\langle w \rangle = 1/2$ is limited by the rate at which *individual* synapses are permitted to make the transition $w = 0 \rightarrow w = 1$. Hence the macroscopic properties of the ensemble cannot move faster than the underlying microscopic transition rates between the states.

Now after equilibrium has been established, we record a specific state of the system at some time instant $t = t_0$ by writing down the state of each synapse. As time unfolds we make a note of the state of the system at each step. By calculating the correlation of the system at each timestep with itself at the initial time t_0 we determine the autocorrelation, Eq. (3.28). In the same manner that the macroscopic properties of the system can only move as fast as the underlying microscopic transitions, so too the system can only decorrelate with itself as fast as the microscopic transitions. Thus there is a fundamental link between the near equilibrium dynamics of the system and

[6]Often in the physical world taking some system far from equilibrium introduces significant additional dependencies between variables, or variations in what were previously considered to be constants. Thus the assumptions that held in the equations at equilibrium are violated, even if the form of the equations remains valid (which it may not) when additional dependencies are taken into account.

the autocorrelation decay timescale [7]. In this section the link between the response and the autocorrelation is made more explicit for synaptic ensembles at equilibrium.

Assume that we observe a state based ensemble of synapses having some time independent transition matrix that has come to equilibrium such that the probability density is stationary, $\mathbf{p}(t) = \mathbf{p}^{\infty}$. We are interested in the evolution of the mean weight with time (or response) after some small equilibrium fluctuation due to the storage of a memory at $t = t_0$. The response is governed by the underlying dynamics of the probability density as determined by the master equation, Eq. (3.8). Since the expectation of the number of synapses in each state is $n_i(t) = \Omega p_i(t)$, where $i \in 1, ..., \zeta$, the evolution of the number of synapses in each state is,

$$\dot{\mathbf{n}}(t) = R\mathbf{n}(t) + \mathbf{g}(t) \tag{3.46}$$

where $\mathbf{n}$ is a ζ element column vector, and R is the $\zeta \times \zeta$ rate matrix defining the synaptic model. Since we are dealing with an observable within a finite stochastic system, fluctuations are present. To model the fluctuations we apply the Langevin assumption and add the noise term $\mathbf{g}(t)$ which is a column vector containing independent Gaussian noise processes $g_i(t)$(Berman, 1975; van Kampen, 1981). It is assumed that the total number of synapses Ω is constant. Thus, there is a conservation law $\sum_{i=1}^{\zeta} n_i = \Omega$. Now the system of equations represented by Eq. (3.46) reduces to $\zeta - 1$ equations, namely

$$\dot{n_i}(t) = \sum_{j=1}^{\zeta-1} R_{ij} n_i(t) + R_{i\zeta} \left(1 - \sum_{j=1}^{\zeta-1} n_j(t) \right) + g_i(t) \text{ for } i \in \{1, .., \zeta - 1\}. \tag{3.47}$$

To establish a link between the response and the autocorrelation it is necessary to recast Eq. (3.47) in frequency space by Fourier transforming $n_i(t)$. At the moment however this is not possible because the $n_i(t)$ are not square integrable[8]. This situation arises because of the conservation law that applies to Ω. It is not possible for $n_i(t)$ to decay to zero. This situation can be remedied by forcing Eq. (3.47) to be homogeneous by subtracting $R_{i\zeta}$ from the right hand side. This has the effect of shifting the equations

[7] This example provides an opportunity to describe why the FDT only applies when close to equilibrium. Imagine that the transition probabilities are not really constant, but that they depend upon $\langle w \rangle$. However when at equilibrium, $\langle w \rangle$ is constant. In this situation it is valid to make calculations about the system under the assumption that $\langle w \rangle$ is constant and consequently the transition rates are constant. It is not valid however to extrapolate this assumption about the system to the non-equilibrium case (when $\langle w \rangle$ is no longer constant). In this example it might be that $\langle w \rangle$ and the rates of the transition, depends upon postsynaptic firing rate. If this were the case then if the system is perturbed to the extent that the postsynaptic firing rate is altered, then the link between the dynamics of the mean and the autocorrelation breaks down.

[8] That is to say that their integrals over all time are not finite.

Eq. (3.47) such that when $\dot{n}_i(t) = 0$, $n_i(t) = 0$ rather than some fraction of the total number of synapses,

$$\dot{n}'_i(t) = \dot{n}_i(t) - R_{i\zeta}. \tag{3.48}$$

Now Eqs. (3.48) can be Fourier transformed with $n'_i(t) = \int_{-\infty}^{+\infty} n'_i(\omega) e^{i\omega t} d\omega$ and $g_i(t) = \int_{-\infty}^{+\infty} g_i(\omega) e^{i\omega t} d\omega$, yielding

$$i\omega n'_i(\omega) = \sum_{j=1}^{\zeta-1} R_{ij} n'_j(\omega) - R_{i\zeta} \sum_{j=1}^{\zeta-1} n'_j(\omega) + g_i(\omega). \tag{3.49}$$

The simultaneous equations, Eqs. (3.49) can be solved by standard methods. The set of solutions $n'_i(\omega)$ thus obtained represents the frequency spectrum of the variations of the occupancies of the states. The power spectral density of the synaptic *weight* fluctuations *at equilibrium* is

$$\Phi(\omega) = \left|\sum n_k(\omega)\right|^2 \text{ for } k \in X \tag{3.50}$$

where X are the states i for whom $w = 1$. By the Wiener-Khinchine theorem, the inverse Fourier transform of Eq. (3.50) yields the unnormalised weight autocorrelation,

$$\langle w_0 w(t) \rangle = \mathcal{F}^{-1}[\Phi(\omega)]. \tag{3.51}$$

To demonstrate the FDT we wish to compare the autocorrelation $\langle w_0 w(t) \rangle$, Eq. (3.51), to the response of the system $\langle w(t) \rangle$ to a small *perturbation*. To find the response a similar strategy is employed but now the Laplace transform of Eqs. (3.47) is used, $n'_i(t) = \int_0^{+\infty} n_i(s) e^{is} ds$ yielding

$$s n'_i(s) = \sum_{j=1}^{\zeta-1} R_{ij} n'_j(s) - R_{i\zeta} \sum_{j=1}^{\zeta-1} n'_j(s) + n'_{i,0} \tag{3.52}$$

Note that there is no noise term because we wish to obtain the precise weight evolution in the case of an average over an infinite ensemble subject to the perturbation. The term $n'_{i,0}$ is the small perturbation to be applied. Solving Eqs. (3.52) simultaneously yields the Laplace transformed dynamics of the occupancies of the states, $n'_i(s)$. Now, the inverse Laplace transform of the $n'_i(s)$ for the states having $w = 1, X$, yields the dynamics of the synaptic weight after the perturbation,

$$\langle w'(t) \rangle = \mathcal{L}^{-1}[\sum_k n'_k(s)] \text{ for } k \in X \tag{3.53}$$

where $\langle w'(t) \rangle$ denotes that this is the transformed synaptic weight, centered on $\langle w \rangle = 0$. When the FDT applies, and as $n'_{i,0} \to 0$ (or alternatively, as the $n_{i,0} \to n_{i,\infty}$ tend to the steady state) then the timecourse of $\langle w'(t) \rangle$ converges with the timecourse of $\langle w_0 w(t) \rangle$.

If the synaptic ensemble is in the steady state then the normalised autocorrelation $\kappa(t)$ has an identical timecourse to $\langle w_0 w(t) \rangle$ because the mean and variance are unchanging. If the system is not at the steady state, then this correspondence does not necessarily occur, and $\kappa(t)$ also includes the timescales of variation in the mean and variance.

3.3.2 When does the fluctuation dissipation theorem apply?

There are several conditions that must apply to a state based model for the FDT to apply:

- Linear dynamics: The equations governing the evolution of the probability density should be linear, such as Eq. (3.8). In physical systems this can often be assumed when the system is near equilibrium, but does not necessarily apply when it is far from equilibrium.

- Close to equilibrium: The conditions at t_0 appear in the Laplace transform of the dynamics Eq. (3.52). Thus for arbitrary initial conditions the superposition of timescales obtained in the solution to the dynamics (the inverse Laplace transform, Eq. (3.53)) does not necessarily match the superposition of timescales in the autocorrelation (although the values of the timescales themselves do match). It is only when the initial conditions tend towards the steady state that the two mixtures of timescales tend toward each other (assuming all other conditions apply). However, some systems maintain correspondence between the response and the autocorrelation even if they are away from the equilibrium state.

- Additive noise: If this is not the case then the Langevin assumption cannot be made (van Kampen, 1981) and this assumption is the basis of Eq. (3.46).

- Mean weight evolution must be Markovian: The observable (for example the mean weight) whose response and autocorrelation is being calculated must be Markovian. Although we have stipulated that the evolution of the probability density is Markovian, this does not imply that the evolution of some *observable,* linearly coupled to the probability density, is Markovian. Often in state based models the mean weight is found with a linear projection of the total probability density onto a subspace (for example when the mean weight is directly proportional to the sum of the number of synapses occupying some subset of states).

This potentially introduces non-Markovian behavior into the evolution of the mean weight, because the value of the weight can be determined by other hidden variables (Kubo, Toda, and Hashitsume, 1998)[9]. In general this breaks the direct correspondence between the autocorrelation and the response, although versions of the FDT can be found (Berman, 1975).

If the FDT applies then the memory trace decay timecourse at the steady state is identical to the decay timecourse of a small perturbation to the mean synaptic weight. In principle this might allow experimental measurements of small disturbances to the weights, to uncover the timescales present in the decay of the memory trace. In practice it is difficult to measure subtle disturbances to the synaptic weight experimentally. More fundamentally, it seems doubtful that biological synapses would obey the FDT. This is due to the fact that the values of biological weights depend upon many hidden variables, that cannot be extracted by the experimenter and so manifest themselves as non-markovian behaviors.

Nevertheless the FDT is a useful tool in reasoning about models of synaptic plasticity. If a model does not obey it then we should not expect the decay timescales of the response to carry directly over into the decay timescales of correlations or the signal to noise.

The methods outlined above are often not analytic for complex state based models, usually because either the inverse Fourier transform or inverse Laplace transform cannot be found. In this situation the model can be solved numerically for very small random perturbations about the steady state. When the FDT applies, the response to the perturbation and the equilibrium autocorrelation are identical under some linear rescaling.

Finally, linear bounded state based models having no hidden states (see chapter 2) all obey the FDT (subject to the other conditions above) (Berman, 1975). Thus any model of synaptic plasticity that can be expressed in this form implies that small perturbations to the synaptic weight should have an identical decay profile to the equilibrium autocorrelation (we shall examine this further in the next chapter).

3.3.3 Application to the 2 state model

As an example of the FDT we shall demonstrate its application to the simple 2 state binary model. The FDT is particularly straightforward here because there are no hidden

[9]This is *metaplasticity.*

states. It is also applicable to the models of STDP in chapters 4+5 for the same reason and more generally to any linear bounded model with no hidden states, such as those mentioned in chapter 2.

In the bistable model with the depressed state labelled 1 and the potentiated state labelled 2 the rate matrix is

$$R = \begin{pmatrix} -r_{12} & r_{21} \\ r_{12} & -r_{21} \end{pmatrix} \tag{3.54}$$

Eq. (3.49) yields one equation for the frequency spectrum of the number of synapses in the potentiated state,

$$i\omega n_2'(\omega) = -r_{12}n_2'(\omega) - r_{21}n_2'(\omega) + g_2. \tag{3.55}$$

Since this is the only potentiated state, the power spectral density, Eq. (3.50) is simply the Lorentzian

$$\Phi(\omega) = \frac{g_2^2}{\omega^2 + (r_{12}+r_{21})^2} \tag{3.56}$$

which is sometimes termed 'telegraph noise'. When Fourier transformed Eq. (3.56) provides the autocorrelation function for an ensemble of bistable weights at equilibrium, Eq. (3.51)

$$\langle w_0 w(t)\rangle = \frac{g_2^2}{(r_{12}+r_{21})} e^{-(r_{12}+r_{21})t} \tag{3.57}$$

for $t > 0$. The autocorrelation (and indeed the signal) therefore decay with timeconstant $\tau = 1/(r_{12}+r_{21})$. For an ensemble of Ω bistable synapses in which Ω_+ synapses have $w = 1$ at $t = 0$, $\langle w_0 w(0)\rangle = \Omega_+$. Hence $g_2^2 = \Omega_+/\tau$ and the autocorrelation is simply,

$$\langle w_0 w(t)\rangle = \Omega_+ e^{-(r_{12}+r_{21})t}. \tag{3.58}$$

The rate matrix Eq. (3.54) yields the Laplace transformed dynamics, Eq. (3.52),

$$sn_2'(s) = -r_{12}n_2'(s) - r_{21}n_2'(s) + n_{2,0}' \tag{3.59}$$

having solution,

$$n_2'(s) = \frac{n_{2,0}'}{s + r_{12} + r_{21}} \tag{3.60}$$

which by Eq. (3.53) provides the decay of the perturbation

$$\langle n_2'(t)\rangle = n_{2,0}' e^{-(r_{12}+r_{21})t} \tag{3.61}$$

as expected this is a single exponential. We see that the timescale of the autocorrelation is identical to the timescale of decay of the perturbation. The bistable model has

the property that so long as r_{12} and r_{21} are constant, the correspondence between the equilibrium autocorrelation and response function holds regardless of the initial state of the system. This is for the simple reason that if only one timescale exists, there can be only one member of the set of possible superpositions.

3.4 Summary

In this chapter the technical underpinnings of what is to follow were explained. The evolution of probability densities of all of the models in this thesis is Markovian. This allows the evolution of the models to be expressed in closed form as a master equation. To solve the master equation, a number of techniques can be used. In this thesis, the main technique deployed is the 'method of eigenvectors' or *eigenvector decomposition* as it is sometimes known. This is a useful technique when it applies because it allows us to think about the evolution of the probability density as a linear superposition of its eigenmodes.

To explore the plasticity stability dilemma it is necessary to extract information about the memory trace from the models of synaptic plasticity. Once a solution to the evolution of the probability density is within our grasp, the memory trace can be quantified by calculating the autocorrelation. In itself, the autocorrelation informs us about the survival time of correlations between synapses directly after memory storage (the initial time) and subsequent states of the synapses. The signal to noise ratio is directly related to the autocorrelation and tells us about the detectability of the memory. We saw that if the synapses are at equilibrium then the signal to noise ratio is simply the autocorrelation scaled by the initial signal to noise ratio. In chapter 2 we saw that in state based models the stability plasticity dilemma is manifest as a trade off between the initial signal and the memory survival time. This is quantified by calculating the initial signal of the model and the time taken for the SNR to fall to 1.

One aim of this thesis is to link the abstract measures of the memory trace, the autocorrelation and the SNR, which cannot be experimentally probed, to quantities that can be easily experimentally probed. One quantity that is often measured in in-vitro experiments is the EPSP slope, which is assumed to be proportional to the mean weight of the synaptic ensemble. To relate these quantities in the models we must understand the relationship between the mean weight of the ensemble $\langle w(t)\rangle$ and the autocorrelation $\kappa(t)$. The fluctuation dissipation theorem predicts that if the conditions outlined in §3.3.2 are obeyed then the response precisely matches the autocorrelation:

i.e. the decay of the memory trace is identical to the timecourse of decay of LTP/LTD itself. In biology it is unlikely that this is the case however. It is more plausible that the palette of timescales available to the decay of induced LTP/D and the decay of the memory trace is similar. This would mean that the decay of LTP/D as induced in the lab displays some subset of the timescales available to the memory trace decay[10]. However it should be borne in mind that in a system such as a synapse, which is an open system that is able to use energy liberated by the cell, that there is no necessary correspondence between the timescales visible in the weight decay and those at play in the memory trace. This is dictated by the principle of detailed balance. Despite these caveats, this thesis explores the memory trace under the assumption that the timescales of synaptic dynamics discussed in chapter 1 are representative of the timescales of synaptic dynamics that are employed by the memory trace in vivo.

[10]I believe that this is the most common implicit assumption

Chapter 4

Spike Timing Dependent Plasticity in single units

As was described in chapter 1, spike timing dependent plasticity (STDP) is the name given to the observation that synapses change their efficacy depending on the precise timing difference between presynaptic and postsynaptic spikes (Levy and Steward, 1983; Markram et al., 1997; Bi and Poo, 1998; Sjöström, Turrigiano, and Nelson, 2001). STDP has been observed in many systems (Abbott and Nelson, 2000) including cultured hippocampal cells and cortical slices (Bi and Poo, 1998; Tsukada et al., 2005; Aihara et al., 2007) and has been argued to be crucial for receptive field development (Young et al., 2007; Mu and Poo, 2006), and adult visual plasticity (Yao and Dan, 2001; Dan and Poo, 2006).

The stability plasticity dilemma is pertinent to STDP because STDP implies that individual spike pairings are capable of modifying the synaptic weight. Since neural activity is stochastic it thus follows that under STDP, synaptic weights are always fluctuating due to random pre-post spike pairings. In this chapter the plasticity stability dilemma is analysed in the case of spike timing dependent plasticity (STDP) with isolated neurons.

The argument was made in Chapter 1 that regardless of the details of the implementation of memory within the hippocampus, one idea that is almost universal is that the modification of excitatory synaptic weights mediates the memory trace. In this chapter the memory trace within a single homogenous synaptic population implementing STDP is examined. Firstly, the STDP models to be investigated are justified and introduced. Next it is shown using simulations that even when synapses are calibrated to have constant mean and intensity of fluctuation, the precise dynamics of the STDP

learning rule has an enormous impact upon the stability of the memory trace. Specifically, under otherwise identical conditions, we see that bistable weight dynamics lead to greater robustness of the memory trace in the face of the ongoing fluctuations predicted by STDP. Once the computational results have been presented, analytic results regarding the memory trace retention times are given. One method of analysing STDP is to regard it as a linear bounded state based model such as that discussed in chapter 2. Using this analysis the large difference in memory trace retention time in the two STDP learning rules is explained.

Material from this chapter appeared at SfN 2006 in abstract form (Billings and van Rossum, 2006) and is currently under review for publication.

4.1 Models of STDP

Models of STDP describe how timing differences between pre and post synaptic spikes map on to synaptic modifications. Most of these models achieve this by means of a 'learning rule', a deterministic function that directly returns synaptic weight changes from pre and post synaptic spike timing differences (Gerstner et al., 1996; Song, Miller, and Abbott, 2000; Song and Abbot, 2001; van Rossum, Bi, and Turrigiano, 2000; Kistler, 2002; Pfister and Gerstner, 2006; Toyoizumi et al., 2007). Alternatively other authors have considered spike timing dependent plasticity as a stochastic switching process (Appleby and Elliot, 2005; Appleby and Elliott, 2006).

An early STDP model proposed a learning rule that is a literal interpretation of data from Bi and Poo (Bi and Poo, 1998), modifying the synaptic weight as an exponential function of the time difference of pre and post synaptic spikes alone, independent of the synaptic weight itself (Song, Miller, and Abbott, 2000). This rule is henceforth referred to as non-weight-dependent STDP (nSTDP). For random pre and post synaptic spike trains, nSTDP leads to divergent weights and so requires that upper and lower bounds be imposed preventing unlimited potentiation or depression. When hard bounds are imposed and pre and post synaptic spike trains are Poisson like, nSTDP causes strong competition between inputs to a neuron, which is reflected in a bimodal synaptic weight distribution (Rubin, Lee, and Sompolinsky, 2001). nSTDP thus selects certain inputs above others even in the absence of large degrees of correlation between input spike trains.

In the nSTDP rule, the weight change due to a pre-synaptic and post-synaptic spike

pairing is

$$\Delta w = \begin{cases} A_+ \exp(-s_{mn}/\tau_+) & s_{mn} > 0 \\ -A_- \exp(s_{mn}/\tau_-) & s_{mn} < 0 \end{cases} \tag{4.1}$$

where $s_{mn} = t_{post}^{(m)} - t_{pre}^{(n)}$ is the time difference between post and pre synaptic spikes with times labeled m and n, Fig. 4.1A. The constants A_+ and A_- set the amount of potentiation and depression, respectively, while τ_+ and τ_- set the duration of the potentiation and depression plasticity windows. The plasticity windows are exponential with $\tau_+ = \tau_- = 0.02s$ unless otherwise stated. Furthermore, we set $A_+ = 1pS$ and take a slightly larger value $A_- = A_+(1+\varepsilon)$, where $\varepsilon = 0.05$. These parameters were calculated by Song, Abbott and Miller to match the Bi and Poo data (Song, Miller, and Abbott, 2000). The bounds are imposed at a minimum value of 0pS and a maximum value of $w_m = 200pS$.

In contrast, a different learning rule, also based upon the data of Bi and Poo, incorporates the observation that strong synapses have been observed to be harder to potentiate than weak ones (Bi and Poo, 1998; Debanne, Gähwiler, and Thompson, 1996; Debanne, Gähwiler, and Thompson, 1999; Montgomery, Pavlidis, and Madison, 2001). This learning rule is henceforth referred to as weight-dependent STDP (wSTDP). The weight dependence of wSTDP gives rise to weight dynamics that are not divergent, having a central fixed point. Thus in the case of wSTDP, no hard bounds need to be imposed. The wSTDP weight distribution closely matches the weight distributions observed experimentally (O'Brien et al., 1998; Turrigiano et al., 1998; Song et al., 2005) and thus wSTDP is perhaps more realistic, although this should be qualified by the consideration that a bimodal distribution could appear unimodal if it is sampled from a population of synapses having a stochastic upper bound, especially if the measuring technique has a limited resolution for small synaptic weights.

Since it has a stable fixed point to which all synaptic weight trajectories flow (on average), wSTDP in its raw form has no competition between weights. This is in contrast to the strong competition between weights in nSTDP. The dichotomy between nSTDP and wSTDP is not strict and intermediate models have been proposed that combine stronger competition with stable learning (Gutig et al., 2003; Meffin et al., 2006; Toyoizumi et al., 2007; Morrison, Aertsen, and Diesmann, 2007). One way of achieving this is to raise the weight dependence in Eq. (4.2) to some arbitrary power (Gutig et al., 2003; Meffin et al., 2006; Morrison, Aertsen, and Diesmann, 2007), although conclusive experimental justification for this is left wanting and so these rules are neglected in this thesis.

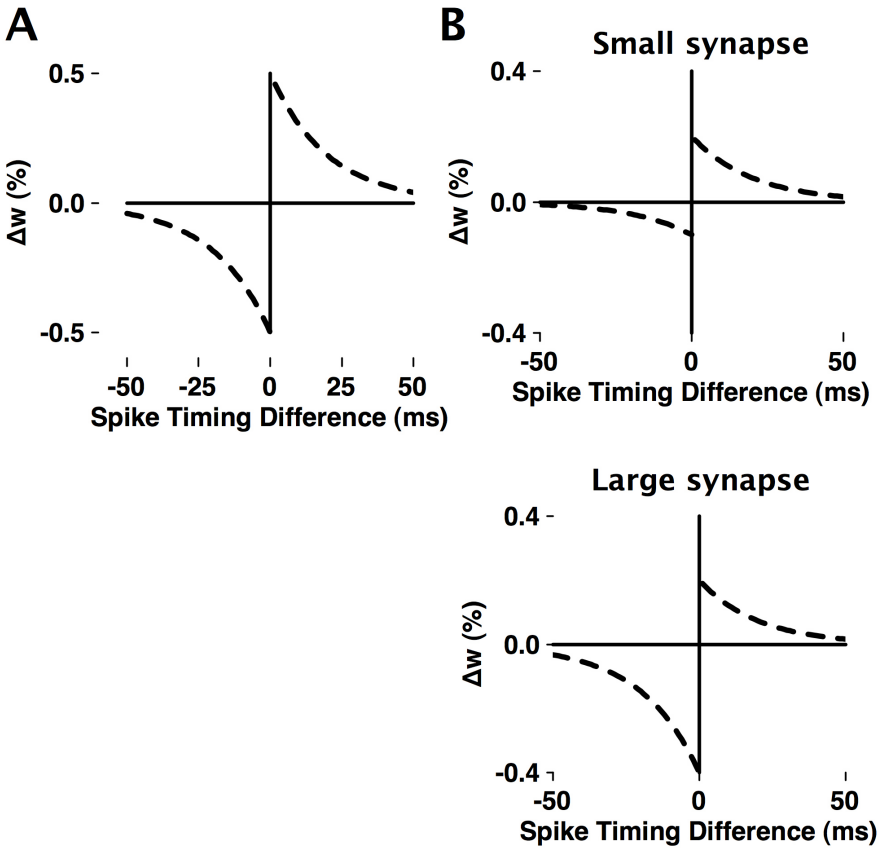

Figure 4.1: Illustration of the nSTDP and wSTDP learning rules with the parameter values stated in the text. A: In nSTDP the weight is modified as a function of the pre/post synaptic spike timing difference alone. B: In wSTDP, the magnitude of the exponential depression plasticity window is modulated by the synaptic weight such that large synapses are depressed more than weak synapses.

There is a number of ways to introduce weight dependence into STDP. One could arrange for both potentiation and depression to depend upon the synaptic weight, a scheme that is termed 'multiplicative' (Kepecs et al., 2002). Alternatively, potentiation (depression) could be made to depend upon the synaptic weight while depression (potentiation) does not, a scheme that is termed 'mixed'. (The situation in which neither depends upon the synaptic weight, as is the case for nSTDP, is termed 'additive'). In STDP experiments it has been observed that the potentiation as a fraction of the total weight is less for strong synapses, while depression shows no such dependence (Bi and Poo, 1998; Montgomery, Pavlidis, and Madison, 2001; Debanne, Gähwiler, and Thompson, 1999). Bi and Poo found a linear dependence between the potentiation and the synaptic weight such that $\Delta w_+ \propto (1-w)$, but no dependence of the depression on the synaptic weight. This implies that STDP is mixed. Furthermore, since $\Delta w_{total} = \Delta w_+ - \Delta w_-$ this is equivalent to introducing a linear dependence between depression and the synaptic weight such that $\Delta w_- \propto w$. This leads to the weight dependent STDP rule (wSTDP), Fig. 4.1B, (van Rossum, Bi, and Turrigiano, 2000),

$$\Delta w = \begin{cases} a_+ \exp(-s_{mn}/\tau_+) & s_{mn} > 0 \\ -a_- w \exp(s_{mn}/\tau_-) & s_{mn} < 0 \end{cases}. \tag{4.2}$$

Here the potentiation increment is taken as $a_+ = 1pS$ and the dimensionless depression constant is $a_- = 0.0114$. Since we wish to compare nSTDP and wSTDP, these values of a_+ and a_- were chosen such that the mean weight is identical for wSTDP and nSTDP ($93pS$). This is essential to ensure that the nSTDP and wSTDP learning leads to the same postsynaptic firing rate, ν_{post}.

Controlling the mean alone is not sufficient to ensure a fair comparison between nSTDP and wSTDP, because the mean weight of wSTDP can be set independently of the absolute plasticity step size, since $\langle w_\infty \rangle = a_+\tau_+/a_-\tau_-$ (Burkitt, Meffin, and Grayden, 2004) (see also §4.3.1). It is therefore necessary to ensure that the weight fluctuations are comparable between nSTDP and wSTDP. These values of a_+ and a_- ensure that the numerically determined modification rate, $\langle \nu_{post} |\Delta w| \rangle$ of the synaptic weights is equal for nSTDP and wSTDP ($7 \times 10^{-3} pSs^{-1}$) as well as the mean weight. Matching the nSTDP and wSTDP processes in this way is equivalent to constraining their initial signal to noise ratios to be equal. In this case, the equilibrium autocorrelation allows a direct comparison of the memory retention time at the steady state.

As mentioned in Chapter 1, to implement the STDP rules one needs to specify how multiple spikes interact. There is a large variety of possible rules, e.g. nearest pre and post synaptic spikes only, all possible pre and post synaptic spike pairs, or some

heuristic choice of spikes, such as; all pre synaptic spikes between the last two post synaptic spikes interact with the last post synaptic spike (Sjöström, Turrigiano, and Nelson, 2001; Froemke and Dan, 2002; Burkitt, Meffin, and Grayden, 2004; Wang et al., 2005; Pfister and Gerstner, 2006). Here the situation in which all spike pairings contribute to the change in synaptic weight is considered.

Even though nSTDP and wSTDP are relatively simple learning rules, they can be seen as limiting cases of unimodal and bimodal STDP and so understanding them is important. In this chapter, the properties of nSTDP and wSTDP and hence bimodal and unimodal STDP are compared from the point of view of the plasticity stability dilemma.

4.1.1 Single neuron simulations

So far we have taken the existence of pre and post synaptic spike trains for granted. In the simulations presented here, post synaptic spike trains are generated by a leaky integrate and fire (LIF) neuron with membrane potential $V(t)$ dynamics governed by: $\tau_m \frac{dV(t)}{dt} = -V(t) + V_r + R_{in}I(t)$, where $I(t)$ is the input current to the neuron, Fig. 4.2A. The neuron fires when the membrane potential reaches a threshold value V_{thr} and upon firing resets to its resting value V_r. The parameters are: Membrane time constant $\tau_m = 20ms$, threshold potential $V_{thr} = -54mV$, resting potential $V_r = -74mV$, input resistance $R_{in} = 100M\Omega$. The neuron receives current inputs through 800 excitatory synapses. These excitatory AMPA-like synapses have an exponential time course with a time constant of 5ms and a reversal potential $V_0 = 0mV$. The input to the neuron at any time is the sum of the current contributions from all of the inputs $I(t) = \sum_i w_i g_i(t)(V_0 - V(t))$, where $g_i(t)$ is an exponential function representing the synaptic time constant and w_i is the synaptic weight.

The pre synaptic spike trains inputing the LIF neuron have Poisson statistics. Each input has a firing rate drawn from a Gaussian distribution of (10 ± 4)Hz. At the end of a random time interval, drawn from an exponential distribution with a mean of $\tau_c = 20ms$, the rates are re-drawn from the Gaussian distribution. This ensures that the correlation between any two inputs $\nu_i(t)$ and $\nu_j(t')$ is proportional to $\exp(-|t-t'|/\tau_c)$. This correlation was chosen in a previous study in rough analogy with input to the visual system (Song and Abbot, 2001).

4.2 Retention of the memory trace

In order to compare the retention time of synaptic weights with nSTDP and wSTDP, a single integrate-and fire neuron receiving stationary Poisson inputs was simulated. After an initial period, the synaptic weights reach an equilibrium distribution, Fig.4.2B+C, in which individual weights fluctuate, Fig.4.2D+E, while the overall distribution remains stationary. As has been shown previously, nSTDP and wSTDP give rise to two very different equilibrium weight distributions (Song, Miller, and Abbott, 2000; van Rossum, Bi, and Turrigiano, 2000).

In Fig. 4.2F+G the autocorrelation $\kappa(t)$ of the weights of a single neuron with 800 Poisson inputs is plotted for nSTDP and wSTDP learning. For nSTDP learning the autocorrelation decays exponentially at large timescales with a time constant of 18 hours. Under comparable conditions, the wSTDP autocorrelation falls rapidly with a time constant of 29s, Fig. 4.2G. For comparison, the nSTDP autocorrelation has been replotted on this timescale in Fig. 4.2G, emphasizing the difference; the nSTDP autocorrelation decay is more than 2000 times slower than the wSTDP decay. Thus learning dynamics giving rise to bistability in the weights provides a much longer memory trace retention time than dynamics with no bistability but identical intensity of weight modification. Analysis of the autocorrelation timescales (see section §4.4) for nSTDP and wSTDP match the simulations well, Fig. 4.2F+G.

4.2.1 Forgetting and the autocorrelation timescale

Intuitively, the autocorrelation seems a reasonable measure of the memory trace lifetime, and in chapter 3 we saw how it relates to the signal to noise ratio. However in this section the effect of fluctuations upon a stored pattern is explicitly demonstrated. To show this, patterns are instantaneously embedded within the STDP weights of a LIF neuron that has reached equilibrium. First, a pattern is stored where a group of 10 weights is set to $200pS$ (about twice the mean weight) using one LIF unit. Next, in a separate simulation, 10 weights are set to 0, Fig. 4.3A+B (the case where weights are set to $200ps$ is illustrated in the upper graphs with one small group of weights at $200ps$, while the lower graphs illustrate the case when the weights are set to $0pS$).

After the intervention the simulation is continued with random inputs and the evolution of the mean values of these depressed and potentiated groups of weights is tracked. To retrieve this simple 'memory' we measure the mean weight of each group, which can be considered as the signal mean. Over time the signal mean decays expo-

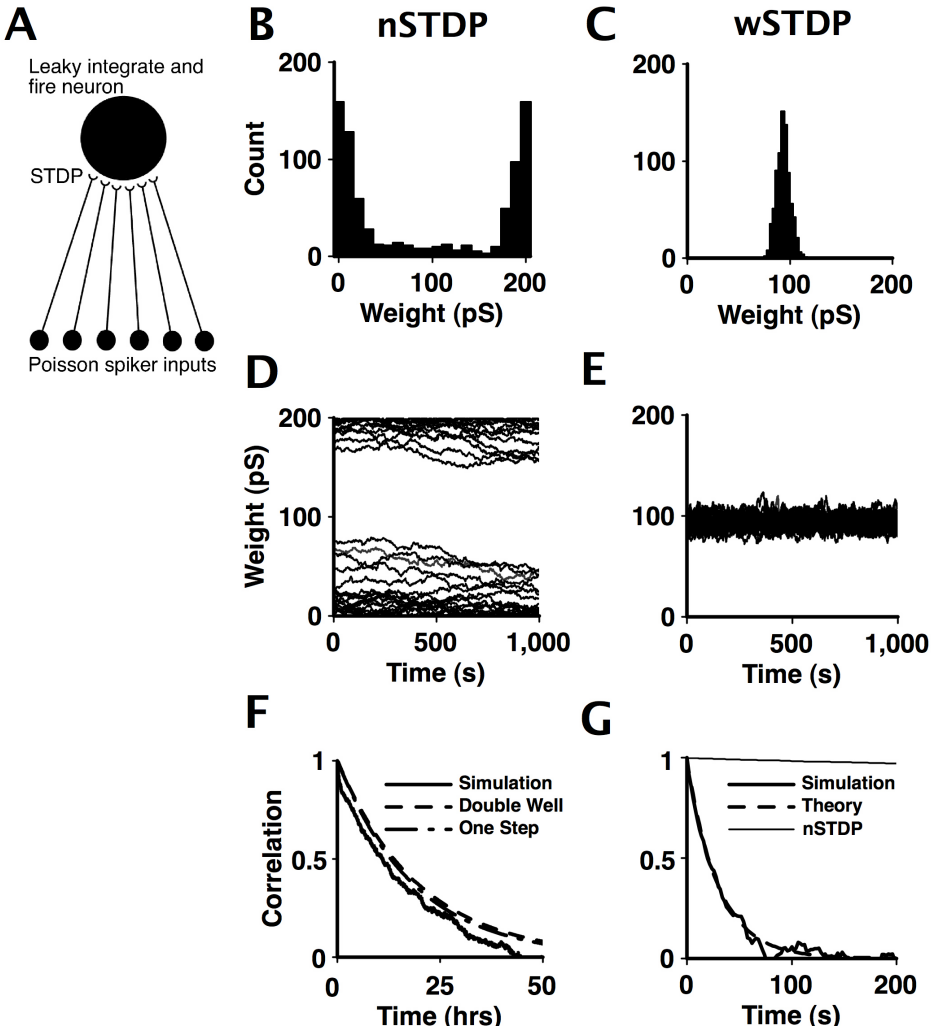

Figure 4.2: Weight distributions and weight persistence in nSTPD and wSTDP. A: Diagram of the single neuron simulation. B: The equilibrium weight distribution of nSTDP. C: As B but for wSTDP. D: The evolution of nSTDP weights sampled at random from the distribution in B. E: As D but for wSTDP. F: The equilibrium autocorrelation $\kappa(t)$ of the nSTDP weights. The solid line is the simulation. The line labelled double well is an approximate calculation found by considering nSTDP as diffusion in a double well potential. The curve labeled one step is an approximate calculation of the autocorrelation based upon approximating nSTDP with a linear bounded state based model. G: The autocorrelation of the wSTDP weight vector versus time. The solid line is the simulated autocorrelation for wSTDP synapses. The nSTDP simulation data is replotted on this timescale (curve labelled nSTDP); the wSTDP autocorrelation decays 2235 times more rapidly. Also included on the graph is the analytical autocorrelation function (curve labelled theory).

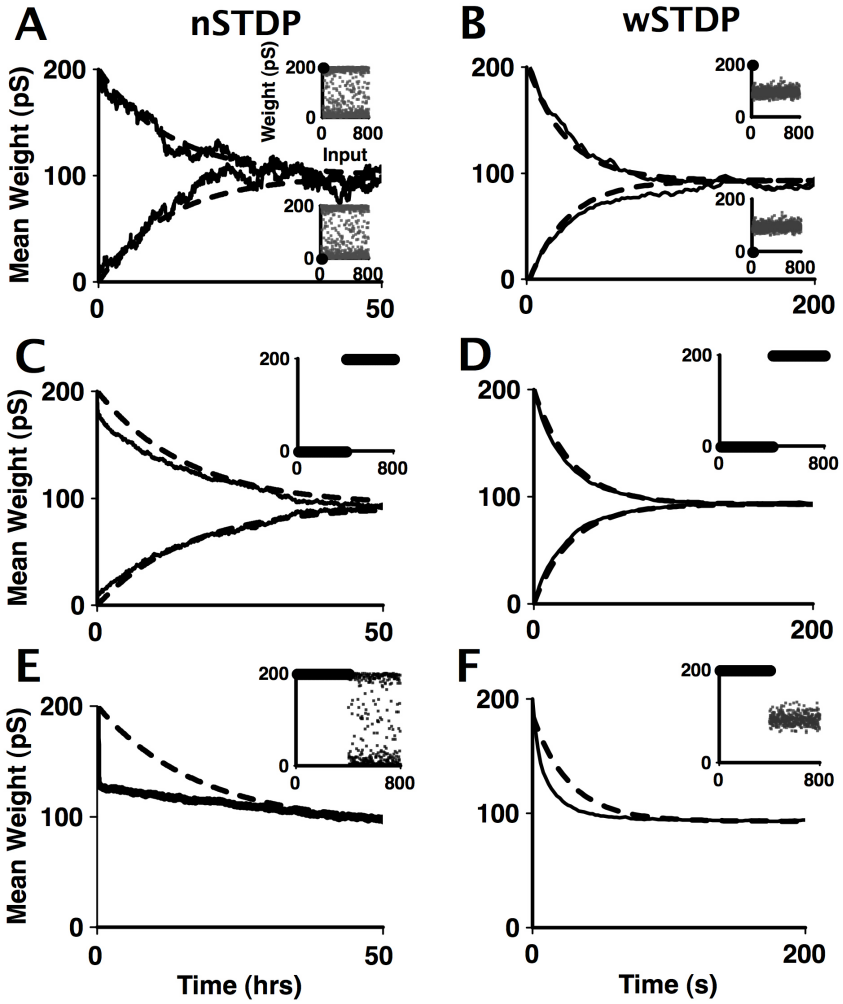

Figure 4.3: The relationship between the retention timescale for a stored pattern and the autocorrelation timescale. A: In nSTDP, 10 weights are set to either hard bound in two separate simulations as shown in the inset plots of the feedforward weights. The undisturbed weights are in grey while the small groups of perturbed weights are black and highlighted. The temporal evolution of the mean values of the subgroups of weights was tracked in the two separate simulations and then plotted on the same axis (solid black lines). Included on the plot are two exponentials with the autocorrelation timescale (dashed lines). B: Same as A, but for wSTDP. C: The case that all of the weights are disturbed but the mean weight is kept the same by balancing the pattern with equal potentiation and depression. The inset graph shows the pattern stored in the weights. The evolution of the mean value of both of these groups of weights is plotted (solid black lines). Also included on the plot are two exponentials with the autocorrelation timescale (dashed lines). D: The identical case to C, but for wSTDP. E: Storing a pattern that causes a large deviation to the mean weight in nSTDP weights. The evolution of the mean value of this group of weights is plotted (solid black lines). An exponential with the autocorrelation timescale is also plotted. F: The same as E, but for wSTDP.

nentially back to the baseline, which can be considered the noise mean. Also plotted are exponentials with their timescales extracted from the autocorrelation for nSTDP and wSTDP determined above.

For both nSTDP and wSTDP, the timescale of relaxation of the means back to the equilibrium value matches the autocorrelation timescale. In the case of wSTDP, the analysis in §4.4.1 shows that the timescale of evolution of the mean weight is identical to the autocorrelation timescale, and this is also demonstrated by Fig. 4.3B. In the nSTDP case and at the longest timescales, nSTDP can be considered as a 2 state switching process which obeys the fluctuation dissipation theorem (chapter 3), §4.4.2. This explains the correspondence between the evolution of the signal mean (which in this case is a response function) and the autocorrelation.

4.2.2 Large Perturbations

The fluctuation dissipation theorem tells us to expect that when patterns are stored in the weights at steady state, the resulting small perturbation to the mean of the ensemble dies out at the autocorrelation timescale. This raises the question of what happens if the perturbations are larger. To test this, consider a single LIF unit in which all weights are changed during the imposed learning process by placing half of the weights at $200pS$ and the other half of the weights at $0ps$, so that the mean weight and output firing frequency is maintained. Despite this large perturbation of the weight distribution, the pattern retention timescale again matches the autocorrelation timescale, Fig. 4.3C+D. In the case of nSTDP, there is an initial rapid decay in the memory strength, which is caused by the weights diffusing near the hard bounds (a similar effect can be observed in Fig. 4.2F). However, at later times the retention time is again dominated by the process of weights jumping between the upper and lower bounds; as we have seen this timescale is long. In the case of wSTDP, the decay matches the exponential autocorrelation decay very well.

The resistance of the LIF unit with STDP to large perturbation by balanced patterns is a special case property of this combination of models and is not to be expected in general. The intuitive reason for this behavior is that the neuron model in this case is not affected by variation in the higher moments of the synaptic weight distribution. Rather, the LIF output firing is dominated by the mean weight.

4.2.3 Unbalanced patterns

Finally, Figs. 4.3E+F show the result when half the weights only are set to $200pS$. In this case the mean weight is not preserved by the learning process. Consequently the output firing frequency increases and as expected the retention timescale no longer matches the equilibrium autocorrelation timescale, but the decay is faster, in particular right after the storage. If in contrast, the pattern reduces the mean weight (and the output firing frequency), the pattern retention timescale is longer than the autocorrelation timescale (not shown). By allowing the mean weight and firing frequency to change, the system is in a regime where there is coupling between the output firing frequency and the rate of change of the weight. This disrupts correspondence between the equilibrium autocorrelation and the response (see §4.4.3.2).

4.2.4 Retention time and the plasticity windows

So far we have examined cases for which the depression and potentiation plasticity windows are equal. However, experiments suggest that the STDP depression time window is approximately twice as long as the potentiation time window, e.g. $\tau_- = 34 \pm 13ms$ and $\tau_+ = 17 \pm 9ms$ (Bi and Poo, 1998). This raises the question of how the timewindow alters the pattern retention time. Given the experimental data, of particular importance is the case where τ_- is changed while τ_+ is kept fixed. For wSTDP learning, the dependence of the autocorrelation time on τ_- is given by Eq. (4.23). However, changing the time window τ_- also changes the mean steady state weight $w_\infty = \tau_+ a_+ / \tau_- a_-$ (see §4.3.1 and (Burkitt, Meffin, and Grayden, 2004)), which changes the output firing frequency, thus affecting the autocorrelation timescale, Eq. (4.23). In this case, as τ_- is increased, the weights are 'cooled' leading to an overall increase in the autocorrelation timescale, Fig. 4.4B. Alternatively, this effect can be compensated by scaling a_+ by the same factor, as τ_- is varied. The simulation results match the theory well Eq. (4.23), Fig. 4.4D and we find that when compensation is provided, the autocorrelation timescale decreases with increasing τ_-.

In contrast to wSTDP, the dependence of the retention time on the parameters in nSTDP learning is more complicated as the shape of the nSTDP weight distribution changes as τ_- is varied. As τ_- is reduced, potentiation dominates, and the weights cluster at the upper bound and the output firing rate saturates, Fig. 4.4A (inset graphs show variation in the mean and output firing rate). In this case the bi-modality of the nSTDP weight distribution is completely lost and the autocorrelation timescale

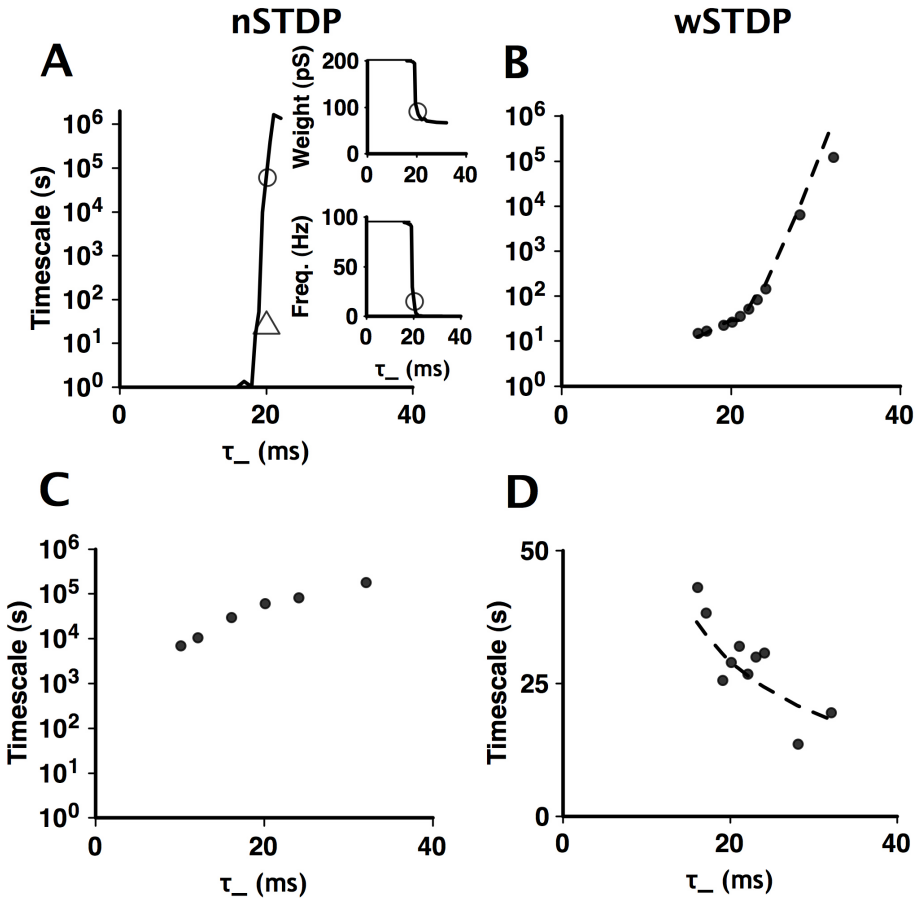

Figure 4.4: The consequences of changing the size of the depression window τ_-, while the potentiation window was kept at $\tau_+ = 0.02s$. A: The autocorrelation timescale as τ_- is varied for nSTDP. Insets show the mean weight and firing frequency in the same simulations. The balanced scenarios are highlighted by the circle. The triangle is the autocorrelation timescale of the equivalent wSTDP weights. B: The autocorrelation timescale in the wSTDP case when the mean weight is not conserved by compensating a_+ (see text). In this case the mean weight and output firing frequency are reduced as τ_- is increased, leading to an overall increase in the autocorrelation timescale. C: The autocorrelation timescale for nSTDP as τ_- is varied and A_- is compensated. D: The wSTDP case when a_+ is compensated. The change in autocorrelation timescale as determined by simulation (points) is well matched by the change in autocorrelation timescale as predicted by Eq. (4.23) (dashed curve). Here $\nu_{post} = 25Hz$ was taken to be the average output firing frequency.

becomes short, Fig. 4.4A. The weight distribution has become unimodal, resulting in the fast de-correlation also seen in the wSTDP case.

Conversely, as τ_- is increased, depression dominates and the synaptic weights congregate near the zero bound. However, at a certain time post synaptic firing ceases, thus freezing the weights. At that point, the distribution is still somewhat bimodal but has more weights at the zero bound than the upper bound, Fig. 4.4A. The autocorrelation timescale is longer than for the balanced case, but only at the expense of a strongly decreased output firing rate.

The strong dependence of the autocorrelation timescale upon the output firing frequency can be compensated for in the nSTDP case, by reducing (increasing) A_- by the same factor that increases (reduces) τ_-. By compensating changes in τ_- by adjusting A_- so that $\tau_- A_-$ is constant, the synaptic weight distribution remains bimodal. In this case, the mean weight and output rate are fixed, and as a result the retention time varies much less as τ_- is changed (although the dependence is still substantial) Fig. 4.4C. These results show that while bistable dynamics in nSTDP lead to a longer memory retention time, the parameters of the plasticity must be precisely tuned such that approximately half of the synapses are in the upper stable state and approximately half of the synapses are in the lower stable state.

4.3 Synaptic weight dynamics in nSTDP and wSTDP

In the last section simulations demonstrated that the dynamics of nSTDP and wSTDP differ markedly, and that this exerts a large effect upon the memory trace retention time. In this section the dynamics of wSTDP and nSTDP are analysed. Specifically, the dynamics refers to how the average weight of a large ensemble of synapses evolves over time. In all cases the analysis here concerns STDP in which all spike pairs are included and where uncorrelated Poisson spike trains are provided as input.

4.3.1 Weight dynamics in the wSTDP case:

For wSTDP and in the all to all spike implementation case we can consider the dynamics of the synaptic weights as being contributed to by a potentiation process w_+ and a depression process w_-. Each time a pre synaptic spike occurs, variable w_+ is updated by addition of the potentiation constant a_+. Conversely, each time a postsynaptic spike occurs variable w_- is updated by addition of the depression constant multiplied by the

current synaptic weight $a_- w(t)$. While this process is ongoing with the activity of the inputs and the LIF unit, the potentiation variable w_+ decays with timeconstant τ_+ and the depression variable w_- decays with timeconstant τ_-. In this case the dynamics of the mean values of these variables are

$$\frac{d\langle w_+(t)\rangle}{dt} = -\frac{1}{\tau_+}\langle w_+(t)\rangle + \nu_{pre} a_+ \tag{4.3}$$

$$\frac{d\langle w_-(t)\rangle}{dt} = -\frac{1}{\tau_-}\langle w_-(t)\rangle + \nu_{post} a_- \langle w(t)\rangle \tag{4.4}$$

where ν_{pre} and ν_{post} are the pre and post synaptic firing rates. The rate of change of the synaptic weight at any given time is the difference of these variables gated by the post and pre synaptic firing rates

$$\frac{d\langle w(t)\rangle}{dt} = \nu_{post}\langle w_+(t)\rangle - \nu_{pre}\langle w_-(t)\rangle \tag{4.5}$$

implying,

$$\frac{d\langle w(t)\rangle}{dt} = \nu_i\nu_j[\tau_+ c_+ - \tau_- c_- \langle w(t)\rangle] + \nu_i\tau_- \frac{d\langle w_+(t)\rangle}{dt} - \nu_j\tau_+ \frac{d\langle w_-(t)\rangle}{dt}. \tag{4.6}$$

Eq.(4.6) describes the macroscopic behavior of wSTDP with Poisson inputs and all-to-all spike implementation. We refer to this expression for $\frac{d\langle w(t)\rangle}{dt}$ as the drift. At equilibrium with constant input frequency $\frac{d\langle w_+(t)\rangle}{dt} = 0$ and $\frac{d\langle w_-(t)\rangle}{dt} = 0$ and consequently the drift reduces to $\nu_i\nu_j[\tau_+ c_+ - \tau_- c_- \langle w(t)\rangle]$. Thus we find that the fixed point of the dynamics (occurring where the drift is equal to zero) is $w_\infty = \frac{\tau_+ c_+}{\tau_- c_-}$, a result that has been obtained previously by a differing route (Burkitt, Meffin, and Grayden, 2004).

At equilibrium the evolution of the weights is governed by,

$$\frac{d\langle w(t)\rangle}{dt} = \nu_i\nu_j[\tau_+ c_+ - \tau_- c_- \langle w(t)\rangle] \tag{4.7}$$

having solution

$$\langle w(t)\rangle = (w_0 - w_\infty) exp\left(\frac{-t}{\tau}\right) + w_\infty \tag{4.8}$$

where $\tau = 1/(\nu_i\nu_j\tau_- c_-)$. Thus for any initial condition w_0 the mean weight decays back toward the steady state value with timescale τ, Fig. 4.3B.

4.3.2 Weight dynamics in the nSTDP case:

The non weight dependent rule is harder to deal with analytically because the boundaries are somewhat ill defined. One way in which the boundaries can be explicitly introduced is to discretise the weight and express nSTDP as a linear bounded state based

model. The system can then be solved with the method of eigenvectors introduced in chapter 3.

This method exploits the idea that nSTDP can be described by the Fokker-Planck formalism. The Fokker-Planck equation is,

$$\frac{\partial P(w,t)}{\partial t} = -\frac{\partial[\mathcal{A}(w)P(w,t)]}{\partial w} + \frac{1}{2}\frac{\partial^2[\mathcal{B}(w)P(w,t)]}{\partial w^2} \tag{4.9}$$

where $P(w,t)$ is the probability density of the synaptic weight distribution. Fokker-Planck equations provide an approximation to the full solution of a system governed by some master equation and can be thought of intuitively as diffusion equations for probability. Fokker-Planck equations express the evolution of a probability distribution $P(w,t)$ in terms of two processes: Firstly a drift process that determines the movement of the centroid of the probability distribution, $\mathcal{A}(w)$. (This process is identical to the drift equation described above provided that the fluctuations are taken as independent of the synaptic weight (van Kampen, 1981; Risken, 1996); an assumption that is adopted here because the variation in the fluctuations as a function of weight is small.) Secondly, in addition to the drift there are fluctuations that give rise to a diffusive process, $\mathcal{B}(w)$ which tends to flatten the probability distribution.

To solve the nSTDP dynamics Eq. (4.9), discretise w into ζ states, w_i, $i \in \{1,2,...,\zeta\}$ of width δw. After this discretisation has been applied the nSTDP is formally equivalent to a linear bounded state based model (chapter 2), Fig. 4.5B, and the tools developed in chapter 3 can be applied (see also §4.4.2). The right hand side of Eq.(4.9) is now expressed as

$$\frac{\partial \mathcal{A}(w_i)P(w_i,t)}{\partial w} = \frac{1}{\delta w}[\mathcal{A}(w_i+\delta w)P(w_i+\delta w,t) - \mathcal{A}(w_i)P(w_i,t)] \tag{4.10}$$

for the drift, and similarly

$$\begin{aligned}\frac{\partial^2 \mathcal{B}(w_i)P(w_i,t)}{\partial w^2} = \\ \frac{1}{2(\delta w)^2}[\mathcal{B}(w_i-\delta w)P(w_i-\delta w,t) \\ +\mathcal{B}(w_i+\delta w)P(w_i+\delta w,t) - 2\mathcal{B}(w_i)P(w_i,t)]\end{aligned} \tag{4.11}$$

for the diffusion. For each weight state w_i, the rate of change of the occupancy of that state in terms of the drift and diffusion into the nearest neighboring states, $w_{i+1} = w_i + \delta w$ and $w_{i-1} = w_i - \delta w$ is calculated with Eqs. (4.10+4.11). The rate of change of occupancy of w_i due to flow into states that are more than one step away is set to zero. Furthermore, flow contributions into w_i from states that are more than one step away

are also set to zero. Thus the rate of change of the probability of occupancy of states away from the boundaries is

$$\begin{aligned}\frac{\partial P(w_i,t)}{\partial t} = \\ -\frac{1}{\delta w}\left(\mathcal{A}(w_{i+1})P(w_{i+1},t)-\mathcal{A}(w_i)P(w_i)\right) \\ +\frac{1}{2(\delta w)^2}\left(\mathcal{B}P(w_{i-1})+\mathcal{B}P(w_{i+1})-2\mathcal{B}P(w_i)\right)\end{aligned} \tag{4.12}$$

while for states at the boundaries where $i=1$ and $i=\zeta$, the rate of change of occupancy is

$$\begin{aligned}\frac{\partial P(w_1,t)}{\partial t} &= \frac{1}{2(\delta w)^2}\left(\mathcal{B}P(w_2)-2\mathcal{B}P(w_1)\right) \\ \frac{\partial P(w_\zeta,t)}{\partial t} &= \frac{1}{2(\delta w)^2}\left(\mathcal{B}P(w_{\zeta-1})-2\mathcal{B}P(w_\zeta)\right)\end{aligned} \tag{4.13}$$

where at the boundaries we take $\mathcal{A}(w)=0$ since the weights experience no drift in either direction when in the lowest or highest states. The diffusion process is taken to be independent of the weight, $\mathcal{B}(w)=\mathcal{B}$. In light of Eqs. (4.12+4.13) the rate matrix R_{ij} can now be defined,

$$\begin{aligned}R_{ij} &= -\frac{1}{\delta w}\mathcal{A}(w_i)+\frac{1}{2(\delta w)^2}\mathcal{B} \text{ for j=i+1} \\ &= \frac{1}{2(\delta w)^2}\mathcal{B} \text{ for j=i-1} \\ &= \frac{1}{\delta w}\mathcal{A}(w_i)-\frac{1}{(\delta w)^2}\mathcal{B} \text{ for j=i}\end{aligned} \tag{4.14}$$

for states, i, away from the boundaries. For states at the boundaries,

$$\begin{aligned}R_{k_\mathfrak{b}l_\mathfrak{b}} &= -\frac{1}{(\delta w)^2}\mathcal{B} \\ R_{k_\mathfrak{b}l_\mathfrak{b}} &= \frac{1}{(\delta w)^2}\mathcal{B}\end{aligned} \tag{4.15}$$

where $\mathfrak{b}\in\{1,2\}$ is one of two boundaries so that $\{k_1=1,l_1=2\}$ or $\{k_2=\zeta,l_2=\zeta-1\}$. The matrix is constructed such that the total probability is conserved, i.e. $\sum_j R_{ij}=0$. Defining the state vector of the system as $\mathbf{p}(t)=\left\{p(w_1,t),p(w_2,t),...,p(w_\zeta,t)\right\}$ we can therefore write the evolution of the system in the standard way as $\dot{\mathbf{p}}=R\mathbf{p}(t)$. The mean value of the weight in this case is given by, $\langle w(t)\rangle=\sum_i w_i p(w_i,t)$. As desired we can expand the evolution of the probability distribution in terms of the eigenvectors of R , and from this the evolution of the mean weight is found to be multiexponential

$$\langle w(t)\rangle = \sum_i w_i \sum_k c_k \Phi_k exp(-t/\lambda_k). \tag{4.16}$$

The drift and diffusion terms, $\mathcal{A}(w)$ and $\mathcal{B}$, can be calculated by virtue of a Taylor expansion of the master equation. What follows is a sketch of the derivation due to van Rossum (van Rossum, Bi, and Turrigiano, 2000). The master equation can be written as

$$\frac{1}{\rho_{in}}\frac{\partial P(w,t)}{\partial t} = -p_p P(w,t) - p_d P(w,t) + p_p P(w-w_p,t) + p_d P(w+w_d,t). \tag{4.17}$$

Eq.(4.17) sums the losses of synapses flowing from some particular weight value w with the gains of synapses flowing into w from the adjacent states, just below $w-w_p$ and just above $w+w_d$. ρ_{in} is the input firing frequency. The w_p and w_d are the magnitudes of the potentiation and depression steps respectively. To simplify, we assume that the synapse is potentiated by a fixed amount $w_p = A_+$ if a post synaptic spike follows a pre synaptic spike within a fixed time window of t_w. The same assumption applies to depression and it is assumed that if a pre synaptic spike follows a post synaptic spike within t_w then the synapse is depressed by $w_d = A_-$. If the spikes fall outside t_w, then the synapse is not modified. This assumption thus removes the exponential nature of the plasticity window and is justified because the average behavior of the synaptic weight will depend upon the total cumulative depression and potentiation. In calculations $t_w = \tau_+ = \tau_-$ since this choice preserves the area underneath the plasticity windows.

The Fokker-Planck equation is easily obtained from Eq. (4.17) by Taylor expanding $P(w-w_p,t)$ and $P(w+w_d,t)$ up to second order, yielding

$$\frac{1}{\rho_{in}}\frac{\partial P(w,t)}{\partial t} = -\frac{\partial[\mathcal{A}(w)P(w,t)]}{\partial w} + \frac{1}{2}\frac{\partial^2[\mathcal{B}(w)P(w,t)]}{\partial w^2} \tag{4.18}$$

where $\mathcal{A}(w)$ and $\mathcal{B}$ can be approximated by

$$\begin{aligned} \mathcal{A}(w) &= p_d(w/W_{tot} - \varepsilon)A_- \\ \mathcal{B}(w) &= 2p_d A_-^2 \end{aligned} \tag{4.19}$$

where $p_d = \rho_{in} t_w$ is the probability of depression. The total 'drive' of the plasticity is described by $W_{tot} = t_W \rho_{in} \Omega \langle w\rangle$ (van Rossum, Bi, and Turrigiano, 2000). The drive accounts for the fact that if the the number of inputs Ω, or the mean synaptic weight, or the input firing frequency increases, then the number of spikes falling within t_w also increases. Alternatively if t_w is increased, then the number of spikes falling within

the window is also increased. When the number of spikes falling within the plasticity window changes, then the overall accumulation of potentiation and depression also changes. Thus the magnitude of the drift depends crucially upon the balance between depression and potentiation, ε and the total drive, W_{tot}.

The fixed points of the rate of change of the average weight occur at the points where the drift is zero. These points occur at $w = w_1$, $w = w_\zeta$ and at $w = A_- \varepsilon W_{tot}/p_d$. As long as $\varepsilon > 0$ then $\frac{dw}{dt}$ is negative at $w = w_2$ and is positive at $w = w_{\zeta-1}$. Thus weights that stray from these boundaries are driven back and consequently the fixed points at the boundaries are stable. When $w = A_- \varepsilon W_{tot}/p_d$, the drift is negative in the direction of decreasing weight and positive in the direction of increasing weight. Hence synaptic weights are repelled from this point and it is thus an unstable fixed point of the dynamics.

4.4 Autocorrelation functions in nSTDP and wSTDP

The autocorrelation of the synaptic weights is the chosen measure of the lifetime of the memory trace in this chapter (see chapter 3). In this section the autocorrelation functions for wSTDP and nSTDP are calculated.

4.4.1 Autocorrelation for the weight dependent case:

We can easily analyse the weight dependent case in order to determine the autocorrelation timescale explicitly. Weights implementing STDP can be regarded as stochastic processes approximated by the Fokker-Planck and Langevin treatments. To calculate the autocorrelation of wSTDP we shall make use of the Langevin approach, from which we obtain an approximation to the time evolution of a *particular realisation* of the stochastic process underlying wSTDP (van Kampen, 1992), rather than a description of the time evolution of the mean of the whole *ensemble* (as was the topic in §4.3.1). This is achieved by assuming that the macroscopic dynamics which govern the evolution of the mean, also apply microscopically to individual weights. Thus the macroscopic equation, Eq.(4.6) is added to microscopic fluctuations in the form of a term that successively samples from a constant Gaussian distribution (this is the 'white noise' assumption). Although, in general, the noise term of wSTDP is weight dependent (van Rossum, Bi, and Turrigiano, 2000), we assume here that the noise is constant because for the choice of parameters in our simulations and at the scale of

synaptic weight under consideration, the fluctuation term varies only negligibly with the weight as compared to the drift. Thus wSTDP with Poisson inputs and an all to all spike implementation can be approximated with the following Langevin equation,

$$w(t+dt) = w(t) + \alpha[w_\infty - w(t)]dt + N(0,1)\sqrt{cdt} \tag{4.20}$$

where the drift is identical to Eq. (4.6), but for brevity we write the drift term as $\alpha[w_\infty - w(t)]$ with $w_\infty = \frac{\tau_+ a_+}{\tau_- a_-}$ and $\alpha = (\tau_- a_- \nu_i \nu_j)$. Here α is the gradient of the drift with respect to weight. In Eq.(4.20) $N(0,1)$ denotes a Gaussian distribution with zero mean and unit variance and c denotes the variance of the white noise.

To extract the autocorrelation from Eq. (4.20), we first multiply by the weight at time zero w_0 and take the ensemble average,

$$\langle w_0 w(t+dt)\rangle = \langle w_0 w(t)\rangle + \alpha[\langle w_0\rangle w_\infty - \langle w_0 w(t)\rangle]dt \tag{4.21}$$

where we note that $\langle w_0\rangle = \langle w_\infty\rangle$ since we are asserting that the system is at equilibrium in the initial state. The diffusion constant c has no direct impact here because the Gaussian distribution averages to zero $\langle N(0,1)\rangle = 0$[1]. Consequently we find that

$$\frac{d\langle w_0 w(t)\rangle}{dt} = \alpha[w_\infty^2 - \langle w_0 w(t)\rangle] \tag{4.22}$$

having the solution $\langle w_0 w(t)\rangle = \sigma^2 \exp(-\alpha t) + w_\infty^2$. Hence the autocorrelation is given by

$$\begin{aligned} \kappa(t) &= \frac{1}{\sigma^2}\left[\langle w_0 w(t)\rangle - \langle w^2\rangle\right] \\ &= \exp(-\tau_- a_- \nu_i \nu_j t) \end{aligned} \tag{4.23}$$

The autocorrelation decays exponentially with a time constant $\tau = 1/(\tau_- a_- \nu_i \nu_j)$ which is the reciprocal of the gradient of the drift with respect to weight. Using values for the parameters from the simulation we find the time-constant to be $29s$ (solid curve in Fig. 4.2G). We see from Eq. (4.8) and Eq. (4.23) that the autocorrelation timescale and the timescale of the response are identical for wSTDP. This is due to the linearity of Eq. (4.20) and is an example of the fluctuation dissipation theorem. However, if the linearity of Eq.(4.21) is disrupted by a varying output frequency this correspondence will no longer hold.

[1] This is perhaps quite surprising, but is verified when tested with simulations. The decay timescale of the autocorrelation of wSTDP with a single LIF neuron is invariant if Gaussian noise is added to the weights. This is due to the fact that addition of noise increases the dispersion of the unimodal wSTDP synaptic weight distribution. Therefore, although the weights are 'hotter' they are spread over a larger weight range.

4.4.1.1 Dependence of the memory trace retention upon the plasticity parameters

Although Eq. (4.23) seems independent of τ_+ and a_+, it does depend on them indirectly via the post synaptic firing rate ν_{post}. To understand this we make use of the relationship between the autocorrelation timescale and the rate of change of the mean weight (drift). We assume that we change τ_+ or a_+ but make a compensatory change to the complimentary potentiation constant a_+ or τ_+ such that ν_{post} (i.e. the mean steady state weight $\langle w_\infty \rangle$) is unchanged. Alternatively, we can compensate with the depression constants a_- or τ_- such that $\langle w_\infty \rangle$ and ν_{post} are unchanged. The weights are taken to be at equilibrium, having mean value $\langle w_\infty \rangle$.

When we make these changes to the plasticity constants, then if the drift is unchanged by them, so too is the autocorrelation timescale because the autocorrelation timescale is the reciprocal of the gradient of the drift. Since ν_{pre} and ν_{post} are constant, the variation in the gradient of the drift is determined by $[\tau_+ a_+ - \tau_- a_- \langle w_\infty \rangle]$. Consider the case where we alter the potentiation constants, but demand that $\langle w_\infty \rangle$ remains unaltered. It is clear that if we multiply τ_+ (a_+) by some factor Δ and we compensate by multiplying a_+ (τ_+) by $1/\Delta$ then there will be no change to the drift, and hence the autocorrelation timescale is also unaltered. If however the compensation is made with the depression constants by altering τ_- or a_- then the drift is altered and therefore the autocorrelation timescale is also altered (but $\langle w_\infty \rangle$ is maintained). Thus we realise that non-trivial alterations to the potentiation constants (i.e. alterations made to τ_+ or a_+that are not compensated by a reciprocal change to a_+or τ_+) do in fact exert an effect upon the autocorrelation either through the output firing frequency (which shall be changed if no compensation is applied) or through necessary changes to τ_- or a_- if the mean weight is to be maintained at $\langle w_\infty \rangle$.

4.4.2 Autocorrelation for the non-weight dependent case:

One method for calculating the autocorrelation function for nSTDP follows from the Markov formalism. As in §4.3.1 the weight is discretised into ζ states, w_i, $i \in \{1, 2, ..., \zeta\}$ of width δw. As mentioned previously, this turns nSTDP into a linear bounded state based model whose transition matrix is calculated from the Fokker Planck equation. The method of calculating the autocorrelation of a state based model was derived in

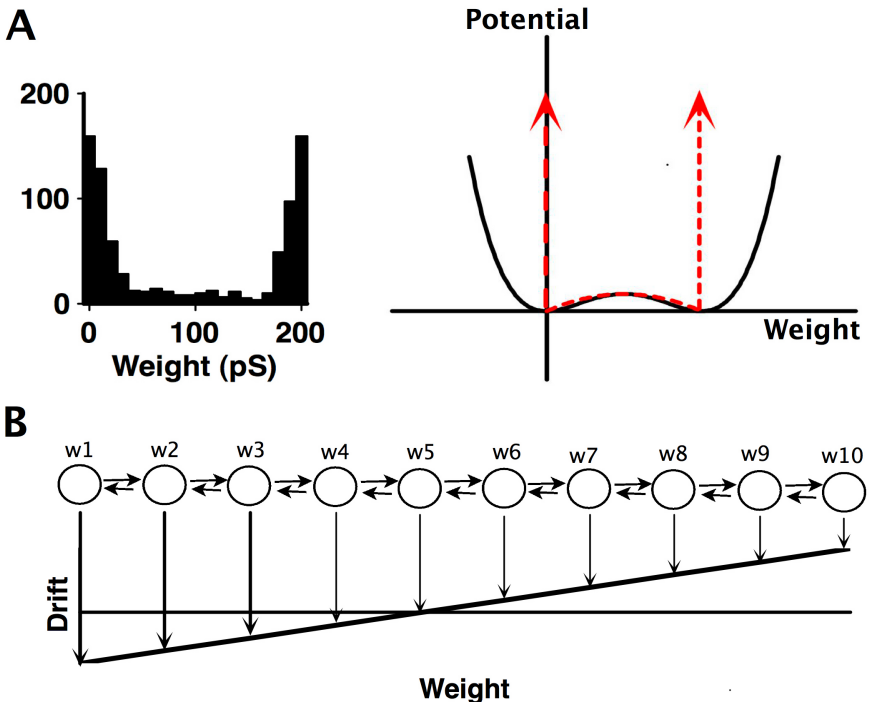

Figure 4.5: The two methods of solving the nSTDP dynamics and autocorrelation. A: Diffusion within a double well potential. The equilibrium distribution of nSTDP (left) can be regarded as arising due to diffusion in a quadratic potential (right, red dashed line). The boundaries of the quadratic are not closed, making solution difficult. The problem is solvable if the quadratic is approximated by a quartic (right, black solid line). B: The 'one step' approximation. nSTDP can be discretised into some number of weight states. The evolution of the plasticity is then a stochastic process across these states. The transition rates between the weight states are estimated using the diffusion $\mathcal{B}$ and drift $\mathcal{A}(w)$ terms in the Fokker Planck equation for nSTDP, Eq. (4.19). nSTDP thus described is a linear bounded state based model (chapter 2).

chapter 3 and can thus be directly applied here

$$\langle w_0 w(t)\rangle = \sum_k e^{\lambda_k t}\left(\sum_i p(i,t=0)C_{ik}w_i\right)\left(\sum_j s_j^{(k)}w_j\right) \tag{4.24}$$

where $\sum_k C_{ik}s_j^{(k)} = \delta_{ij}$, and $s_j^{(k)}$ is the k-th eigenvector of R (i.e. C is the inverse of the eigenvector matrix). Since we are investigating the equilibrium case, one can insert $p(i,t=0) = s_i^{(1)}/\sum s_i^{(1)}$, where $s_i^{(k)}$ is the eigenvector with zero eigenvalue. The longest timescale of Eq. (3.31) is the reciprocal of the subdominant eigenvalue of R_{ij} and this is the autocorrelation timescale, 'one step' in Fig. 4.2F.

4.4.3 'Double well' approximation: nSTDP as a 2 state switching process

An alternative way to approximate the nSTDP autocorrelation timescale exploits the fact that in nSTDP weights congregate near 0 and w_{max}. At long timescales this means that the autocorrelation depends on how quickly the weights randomly move from one side to the other. Thus nSTDP can be regarded as a stochastic escape process and can therefore be recast in terms of diffusion within a potential, Fig. 4.5A. Definition of an appropriate potential allows us to approximate the hard boundaries. In the case that the Fokker-Planck equation for a stochastic process has an asymptotic steady state solution $p_\infty(x)$, the potential for that process is (Miguel and Toral, 1997)[2]

$$V(x) = -\frac{\sigma^2}{Z}\ln(p_\infty(x)) \tag{4.25}$$

where σ^2 is the variance of the fluctuations and Z normalises $p_\infty(x)$. If we imagine some particle diffusing in a potential then Eq.(4.25) says the particle will tend to have peaks in its steady state displacement distribution whose locations are determined by the minima of the potential of its motion. For nSTDP the equilibrium weight distribution is described by $p_\infty(w) = Z\exp[(-\varepsilon w + \frac{1}{2}w^2/W_{tot})/A_-]$ (van Rossum, Bi, and Turrigiano, 2000). Thus

$$V(w) = \frac{\sigma^2}{A_-}(\varepsilon w - \frac{1}{2}w^2/W_{tot}) \quad (0 < w < w_m) \tag{4.26}$$

$$= \infty \text{ (otherwise)} \tag{4.27}$$

where $\sigma^2 = \mathcal{B} = 2p_d A_-^2$. Outside the limits the potential is infinitely high due to the imposition of hard bounds.

[2] For this to be true the process must obey microscopic reversibility and must be closed and isolated. This is indeed the case here.

The potential in Eq. (4.26) can be approximated with a quartic 'double well' potential, Fig. 4.5A, whose minima coincide with 0 and w_{max} and that provides a potential barrier between the two stable points, of the same height as the original potential. To achieve this we fit the quartic according to the conditions that $V_{aprx}(0) = V(0)$, $V_{aprx}(w_m) = V(w_m)$, $V_{aprx}(w_{D0}) = V(w_{D0})$, where w_{D0} is the point at which the drift vanishes and $V'_{aprx}(0) = 0$, $V'_{aprx}(w_m) = 0$. This potential is,

$$V_{aprx}(w) = \frac{\sigma^2}{2w_m^2 A_- W_{tot}} \left[w^2(4w^2 - 6ww_m + w_m^2 - 4\varepsilon w W_{tot} + 6\varepsilon w_m W_{tot}) \right]. \quad (4.28)$$

Under the assumption that the potential is symmetric (i.e. LTP and LTD are balanced) and the barrier between the wells is sufficiently high so as to separate the timescale of diffusion within the well from the timescale of diffusion between wells (so that the system is not critical), the mean first passage time (MFPT) of a diffusing particle across the centre of the double well potential can be approximated as (van Kampen, 1992)

$$\tau_X = \frac{2\pi}{\sqrt{V''_{aprx}(X)\left|V''_{aprx}(B)\right|}} exp\left(\frac{V_{aprx}(B) - V_{aprx}(X)}{\sigma^2}\right) \quad (4.29)$$

where B is the central maximum of the potential and X is either of the minima of the potential A or C. If we follow a weight at either A or C then the MFPT tells us how long we need to wait on average to see the weight 'switch' to the other well. Thus a course grained description of nSTDP is created in which the weights are switching between two stable states with some rate. The autocorrelation function for such a random two-state 'telegraph noise' system is $A(t) = exp[-t/\tau_A - t/\tau_B]$ where $1/\tau_A$ and $1/\tau_B$ are the transition rates between A and B and B and A respectively (this was derived in chapter 3). The autocorrelation time is thus given by

$$\tau_c = \frac{\tau_A \tau_B}{\tau_A + \tau_B} \quad (4.30)$$

this gives $\tau_c = 20hrs$ for the values of parameters taken in the simulation, 'double well' in Fig. 4.2F. In this course grained description the continuous nSTDP weights are behaving as simple binary 2 state weights.

On long timescales the autocorrelation will be exponential and dominated by this switching process as long as the distribution is strongly bimodal. Hence at this longest timescale, when the nSTDP synapses are behaving like 2 state binary synapses, we would expect the fluctuation dissipation theorem to apply.

4.4.3.1 The influence of the depression constant

The effects of altering the depression step size A_- upon the autocorrelation timescale are sufficiently simple as to permit an analytical treatment. We assume the balanced case where the double well approximation holds. This is ensured by the imposed scaling between A_- and A_+ (§4.1). We regard all other parameters as constants and assume that the nSTDP potential is balanced so that $\tau_A = \tau_B = \tau$ and $\tau_c = \tau/2$ where τ is the MFPT associated with crossing from one well to the other. We express the double well potential as $V_{aprx}(w) = \frac{\sigma}{2w_m^2 A_- W_{tot}}\phi(w)$. Since we have asserted that everything other than A_- is constant, the function $\phi(w)$ is also constant. Defining $\alpha = \sqrt{\phi''(A)\phi''(B)} = \sqrt{\phi''(C)\phi''(B)}$ and $\beta = \phi(B) - \phi(A) = \phi(B) - \phi(C)$, the autocorrelation timescale is

$$\tau_c = \frac{\pi w_m^2 W_{tot}}{p_d A_- \alpha} exp\left(\frac{\beta}{2A_- w_m^2 W_{tot}}\right) \tag{4.31}$$

which is good agreement with the one-step calculation method. Eq. (4.31) shows that in the LTP-LTD balanced regime, the size of the plasticity step size very strongly influences the memory retention time.

4.4.3.2 Unbalanced pattern decay

Violation of the correspondence between the autocorrelation timescale and the response timescale can be understood using the two state approximation to nSTDP in which weights are always at either of the nSTDP boundaries. We can construct a linear equation with a timescale of evolution that is identical to the longest timescale of the autocorrelation. This is the case because the two state approximation to nSTDP implied by calculating the rate of transfer between potential wells, has linear dynamics if the wells are well separated (i.e. the potential barrier is high), and the rates of transfer are constant in time. The nSTDP must be approximately balanced for the double well description to work well. In this case we can write a stochastic differential equation for the number of weights in the well at $w = 0$, as

$$\frac{dn_0(t)}{dt} = -\frac{n_0(t)}{\tau_c} + \frac{n_T}{\tau_B} + \Gamma(t) \tag{4.32}$$

where n_0 is the number of weights in the well near $w = 0$, and we define n_{w_m} as the number of weights near $w = w_m$ and $n_T = n_0 + n_{w_m} = \text{const}$ is the total number of weights. $\Gamma(t)$ is the Gaussian noise process[3]. The rate at which weights jump from w_0

[3]It is well known that the Gaussian noise process as stated here, within a differential equation, is mathematically pathological (it is not differentiable). However here it is intended as short hand for

to w_m is $1/\tau_A$, while the rate of jumping from w_m to w_0 is $1/\tau_B$ and $\tau_c = 1/\tau_A + 1/\tau_B$, Eq. (4.30). We can extract both the autocorrelation timescale and the macroscopic timescale (in the limit of $n_T \to \infty$) from Eq. (4.32) and they are of course identical to τ_c. However, as with wSTDP, the linearity of Eq. (4.32) is disrupted if the output firing frequency varies and hence τ_A and τ_B are no longer constant. Then we should expect to see a pattern retention timescale that differs markedly from the equilibrium autocorrelation timescale. In Fig. 4.3E we see this disruption when we store an unbalanced pattern and the rates of transfer between the two stable states are no longer constant but depend on the output firing rate.

4.5 Discussion

In this chapter it has been demonstrated that for single LIF neurons with plastic synapses, the memory trace retention time for wSTDP is orders of magnitude shorter than for equivalent nSTDP. Thus there are orders of magnitude difference in the ability of the weights implementing these learning rules to retain patterns, under otherwise identical conditions. This difference is attributable to the bimodal weight distribution of nSTDP.

The bimodal synaptic weight distribution engendered by nSTDP can retain correlations between successive realisations of the weights for long periods of time, as has been suggested by other authors (Rubin, Lee, and Sompolinsky, 2001). The long retention time of the weights is dependent on this bi-modality and hence the balance between potentiation and depression. This behavior generalizes to the case where we measure a memory signal based upon some subset of the weights as long as the memory only weakly perturbs the ensemble mean weight. This is a result of the fluctuation dissipation theorem (see chapter 3).

In the linear bounded models of chapter 2 we saw that the values of the transition probabilities between the weight states exert an effect upon the resulting steady state distribution amongst the states and upon the memory lifetime. There are some parallels between the linear bounded models and wSTDP/nSTDP. It was shown that nSTDP and wSTDP can be regarded as linear bounded models, where the transition probabilities between the states are determined by the drift and diffusion terms of their Fokker-Planck equations. We saw that nSTDP gives rise to a bimodal weight distribution that out performs the unimodal wSTDP weight distribution as a memory store. Finally, it

the equivalent forward Langevin equation. Since the equation is not manipulated, this form makes the argument clearer conceptually.

should be noted that memory traces in both nSTDP and wSTDP decay exponentially. Thus the scaling of the maximum memory lifetime as the logarithm of the number of synapses (see chapter 2) occurs in both of these models.

Previous investigations of the stability of spike timing dependent plasticity have concentrated upon the dynamic stability of the synaptic weights rather than upon the memory retention time (van Rossum, Bi, and Turrigiano, 2000; Song, Miller, and Abbott, 2000; Izhikevich and Desai, 2003; Burkitt, Meffin, and Grayden, 2004). As we have seen it is possible for distributions to be stable but provide radically different autocorrelation times under identical conditions. For this reason a more complete understanding of the learning rule requires that we also study the correlation properties of the weights (or perform a signal to noise analysis, which is equivalent at equilibrium).

Recently it has been shown that optimality considerations lead naturally to a spike timing dependent plasticity like learning rule (Toyoizumi et al., 2007). This learning rule gives a unimodal distribution and hence unstable memory traces when there is little correlation between inputs. However when the correlations are strong the learning rule gives rise to a bimodal weight distribution that provides retention times of the memory trace of the order hours when the training correlations are removed. The greater the degree of correlations, the greater the bimodality and the more effectively the memory trace is stored. While Toyoizumi et al differs from this work in that the authors considered the memory stability under a change of stimulus statistics (i.e. they were studying the non-stationary rather than the stationary plasticity stability dilemma) it nevertheless supports the idea that bistability might be important for the long term storage of memory traces.

In this chapter a number of assumptions have been made. Firstly, the STDP rules used an all-to-all spike implementation, i.e. all spikes are included in the synaptic modifications and the contributions from each spike pairing sum linearly. However, as mentioned in chapter 1 there is evidence that non-linear corrections exist (Sjöström, Turrigiano, and Nelson, 2001; Froemke and Dan, 2002; Wang et al., 2005). Some of this data has been modeled heuristically (Pfister and Gerstner, 2006), but a unified model is still lacking, making it difficult to give more general predictions about the memory lifetime of STDP. However the nSTDP and wSTDP rules provide limiting examples of the bimodality of the synaptic weight distribution. The differences between these cases are so great that whatever the details of the underlying rule, the distinction will likely remain important. Furthermore it is important to understand the stability of nSTDP and wSTDP in their own right because they are still widely used as working

approximations to STDP. An important recent study showed that receptive field plasticity in adult cats is compatible with STDP rather than with non-causal covariance based learning (see chapter 1) (Young et al., 2007). This study assumes nSTDP in synaptic connections.

The second important approximation is that the temporal and correlation structure of actual input and output spike trains is likely much more complicated than assumed here, i.e. this chapter deals almost exclusively with the stationary plasticity stability dilemma. The more general non-stationary case is more difficult to deal with, not least because the number of possible scenarios becomes infinite. Nevertheless the case of random spike trains studied here is important because it is generally understood that neurons are stochastic and thus spike with random statistics when at rest, giving rise to so called 'background activity'. It is therefore reasonable to wonder how weight traces that have been learned previously are affected when subjected to background activity and this is the question that was addressed here.

Chapter 5

Spike timing dependent plasticity in networks

In the previous chapter it was shown that isolated single neurons with weight dependent STDP synapses forget their weights rapidly in comparison to units with nSTDP synapses. This rapid forgetting occurs due to the lack of bistable weight dynamics. Another way to phrase this is that because the wSTDP learning rule is not competitive, weights are not segregated into two stable groups. In this chapter the stability of nSTDP and wSTDP within spiking networks is examined. Despite its lack of competition in the single unit, wSTDP networks can perform input selection when neurons are linked by lateral inhibition. Thus interactions between neurons can introduce competition and stabilise weights that otherwise use unstable learning.

Material from this chapter appeared at SfN 2006 in abstract form (Billings and van Rossum, 2006) and is currently under review for publication.

5.1 Network model

In this chapter we study a single layer network with all to all lateral inhibitory connections and plastic feed-forward excitatory connections that receive the input spike trains. This model can be interpreted as a simple model for orientation selectivity if each output unit is considered as operating in analogy to a tuned cell within an orientation selective cortical column, (Ben-Yishai, Bar-Or, and Sompolinsky, 1995; Song and Abbot, 2001; Shapley, Hawken, and Ringach, 2003; Yao, Shen, and Dan, 2004). The network consists of one layer of 60 integrate and fire neurons with parameters as above. The network has periodic boundary conditions to eliminate edge

effects and ensure that all neurons operate under comparable conditions. The neurons receive feed-forward input from a layer of 600 Poisson inputs through STDP synapses and receive all-to-all inhibition through lateral connections. The neurons do not self-inhibit. In this case the dynamics of the membrane potential of each neuron receives two current contributions: $\tau_m \frac{dV(t)}{dt} = V_r - V(t) + R_{in}(I_{\mathrm{ff}}(t) - I_{inhib}(t))$ where $I_{\mathrm{ff}}(t)$ is the feed-forward input current and $I_{inhib}(t)$ is the inhibitory current. Feed-forward excitatory synapses are identical to the single neuron case in chapter 4. In all simulations the feedforward weights are intialised to be uniformly distributed at random between $0pS$ and $200pS$. Inhibitory synapses (conductance-based) are exponential with a time constant of 5ms and have a reversal potential of -74mV. The inhibitory synapses are not plastic and are uniform across the inhibitory population.

Inputs to the network are again Poisson trains, but the firing rate is spatially modulated as follows: input a has a rate $\nu_a = \nu_0 + \nu_1(e^{-(s-a)^2/2\sigma^2} + e^{-(s+\lambda-a)^2/2\sigma^2} + e^{-(s-\lambda-a)^2/2\sigma^2})$, where the stimulus is centered at input s, the background rate is $\nu_0 = 10Hz$ and peak rate $\nu_1 = 80Hz$, λ is the width of the network, and σ is the width of the stimulus set to be one tenth of the number of inputs. The 2nd and 3rd term ensure the periodic boundary conditions. The center of the stimulus was randomly chosen at time intervals drawn from an exponential distribution with a mean of 20ms, Fig. 5.1A. This input structure was chosen to be comparable to a previous study of nSTDP in which receptive fields formed successfully (Song and Abbot, 2001).

In one set of simulations, the stability of the receptive fields when a blank stimulus is presented is tested. In this case the input stimulus consists of unstructured, uncorrelated Poisson spike trains, i.e. $\nu_a = \nu_0$.

5.1.1 Receptive field stability

Receptive fields of the neurons in the network are quantified as follows: At given times the synaptic weights are frozen and the same input stimulus as described above is swept across the inputs. The tuning curve of each neuron is measured at $m = 24$ stimulus locations (25 stimuli around each location; response measured for 20ms). The tuning curve is plotted in a polar plot and the vector average is calculated. Thus the receptive field of each neuron is characterized by the two dimensional vector defined as $\mathbf{p} = \{\frac{1}{m}\sum_{i=k}^{m} \nu_k \sin(\frac{2\pi k}{m}), \frac{1}{m}\sum_{l=k}^{m} \nu_k \cos(\frac{2\pi k}{m})\}$, where k indicates the stimulus location, and ν_k is the average firing rate at that location.

The autocorrelation is one measure of the memory trace retention (chapter 3). In

chapters 3 and 4 the autocorrelation measured how long *weight* correlations last. Now, a similar measure is defined to quantify retention of the *receptive fields* using the receptive field vectors. For a network with N neurons $\mathbf{p}$ is a 2N component vector $\{p_n^x, p_n^y\}$ where $n = \{1,...,N\}$. The autocorrelation of this vector is calculated in exactly the same way as for the weight vector. If the autocorrelation is one, the receptive fields have not changed from their initial state and have remained in their initial locations. If, in contrast, the autocorrelation is zero, their receptive field locations have become independent of their initial positions.

In addition to measuring the persistence of the receptive fields it is useful to quantify how peaked they are around the optimal stimulus. The selectivity S of a neuron is calculated as, $S = 1 - \frac{1}{m\nu_{max}} \sum_{l=1}^{m} \nu_l$, where ν_{max} is the maximum firing rate of the neuron (occurring at the optimal stimulus position) (Bienenstock, Cooper, and Munro, 1982). In the case that the tuning curve is flat we find a selectivity of 0. If the tuning curve is more peaked, the selectivity increases. In the limit that the tuning curve is a delta function, we find a selectivity of $S = 1$.

5.2 Receptive fields in STDP networks

While the development of input selectivity in a network using a non-competitive learning rule (wSTDP) is surprising, the formation of receptive fields in competitive nSTDP networks has been demonstrated previously. It was found that neurons develop receptive fields even in the absence of recurrent connections (Song and Abbot, 2001; Delorme et al., 2001). This is the result of the strong competition intrinsic to the nSTDP rule in the single unit, which selects one group of inputs above another, the winner being determined by the initial conditions. Since the initial weights are random, the map of the receptive fields is also random. When local recurrent excitatory connections are added, all neurons in the network become selective for the same area of the input range (like a single column). When, in addition, all to all inhibition is included, a map forms in which the receptive fields of the neurons tile the input in a locally continuous manner. In nSTDP networks with lateral inhibition only, disordered maps develop.

To compare receptive field development in wSTDP and nSTDP networks, the networks were trained from random initial conditions on the input stimulus shown in Fig. 5.1A (and see §5.1). One group of networks has lateral inhibitory connections while the other has no lateral inhibition. As in previous studies receptive fields form read-

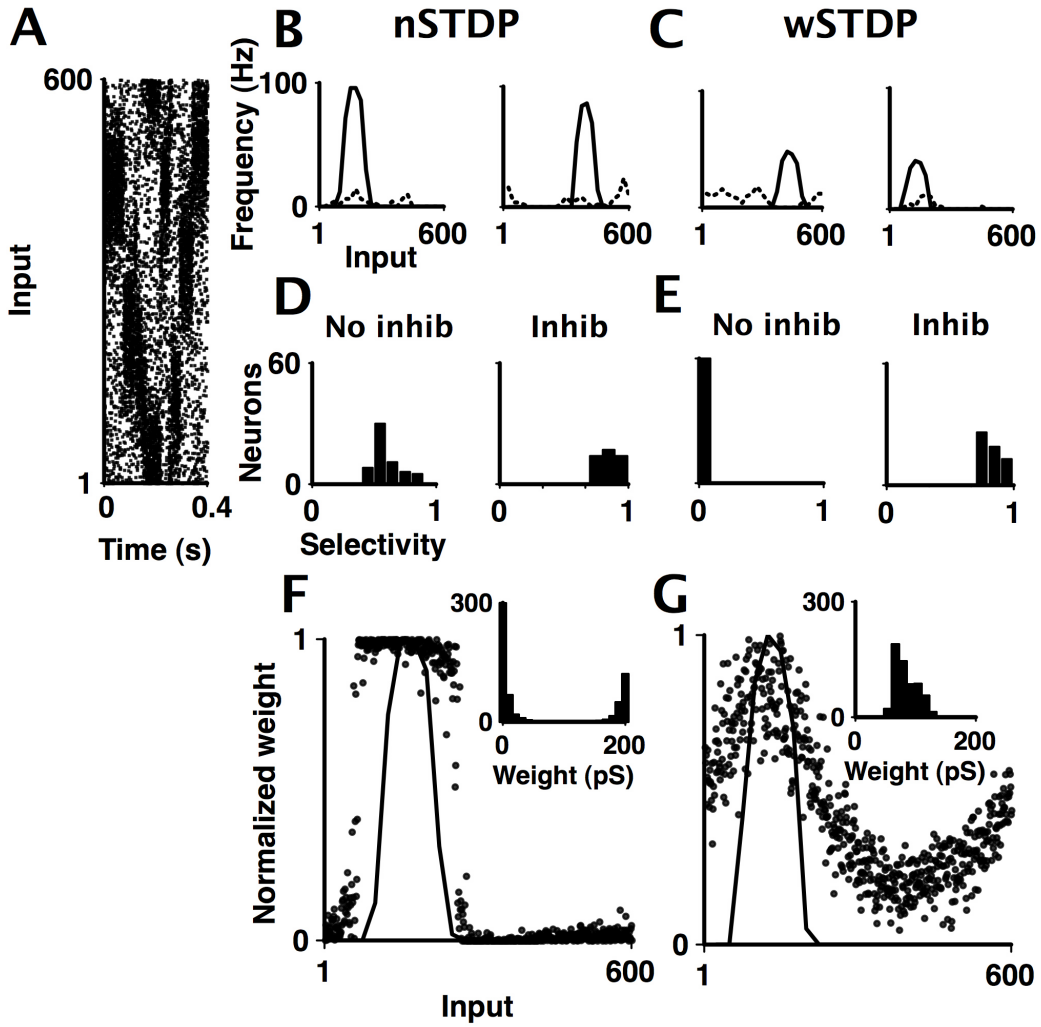

Figure 5.1: Receptive field development in nSTDP and wSTDP networks (with 60 units and 600 inputs) A: Raster plot of a short sample of input to the network. B: Tuning curves for two neurons in the nSTDP network, before (dashed line) and after (solid line) training. C: Same as B but for the wSTDP network with lateral inhibition. D+E: Histogram of the selectivities of the neurons after training with no inhibition (left) and with 7nS lateral inhibitory connections (right). For the wSTDP network inhibition is essential to develop selectivity. F: The leftmost nSTDP tuning curve of panel B is plotted along with the feed-forward weights. Inset is the weight distribution for this neuron. G: Same as F but for wSTDP. wSTDP forms receptive fields that are similar to nSTDP receptive fields when sufficient lateral inhibition is present, although the underlying weights are more centrally distributed.

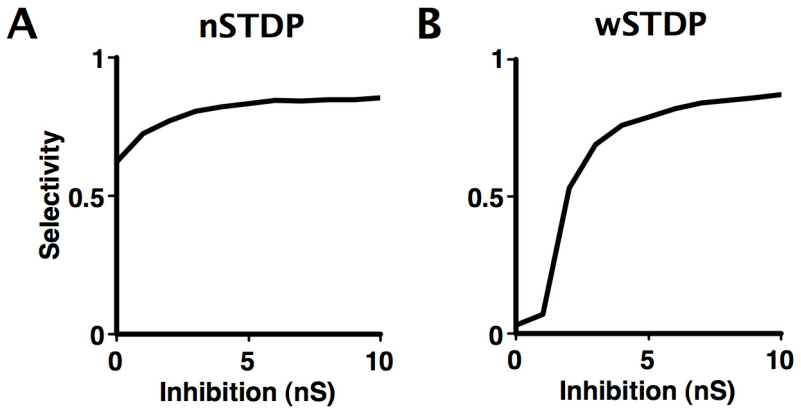

Figure 5.2: The development of selectivity in nSTDP and wSTDP networks as a function of the lateral inhibitory connection strength. A: In the nSTDP networks, receptive fields always form regardless of inhibitory connection strength although increasing inhibition leads to a sharpening of receptive fields. B: In contrast, for wSTDP networks, inhibition is necessary for receptive fields to form. Additional inhibition sharpens the selectivity.

ily in the absence of lateral inhibition in the nSTDP network, Fig. 5.1D. The average selectivity increases from around 0 to 0.65 during training in the nSTDP network.

In the wSTDP case without inhibition there is no receptive field development and no increase in the mean selectivity of the neurons in the network, Fig. 5.1E. However, with inhibition present, input selectivity does also develop in the wSTDP network, Fig. 5.1E. The receptive fields sharpen and in some cases receptive fields develop where there was little initial structure. The mean selectivity increases from 0.65 to 0.8 during training, a change that is as pronounced as it is for the nSTDP network (Fig. 5.1D+E), where the mean selectivity increases from around 0.65 to 0.85. This demonstrates that the development of selectivity that is intrinsic to the nSTDP learning rule, but that is absent from the wSTDP learning rule, can nevertheless occur in wSTDP networks with lateral inhibition.

Note that in both networks some selectivity already exists before training. This is due to the random initial conditions of the feedforward weights §5.1. However these initial receptive fields are generally different from those obtained after training: 1) They are small; the peak firing rate is typically less than 1/3 of the final one for wSTDP learning (the difference is even larger for nSTDP learning). 2) The tuning curves often have multiple peaks and are irregular. In contrast, after training the receptive fields

are smooth with only one peak and the receptive fields tend to evenly distribute across input space, Fig. 5.1B+C.

The underlying structure in the feed-forward weights is shown in Fig. 5.1F+G. Associated with the receptive fields in nSTDP networks is a region of weights at the maximum weight value, while all other weights are zero, Fig. 5.1F. The characteristic bimodal distribution of nSTDP is still present, but the weights are spatially inhomogeneous. This is a result of the strong competitive behavior of nSTDP explored in chapter 4, driving elevation of the correlated input group at the stimulus location. In wSTDP networks however the underlying feed-forward weight structure corresponding to the receptive fields remains unimodal, Fig. 5.1G.

Having established that receptive fields can form in wSTDP networks, the strength of the lateral inhibition is now varied Fig. 5.2. As stated above, neurons in nSTDP networks are selective after training, regardless of the inhibitory connection strength. Nevertheless, the tuning curves sharpen with increasing inhibition. In contrast, wSTDP does not form receptive fields, unless the lateral connections reach a critical value in the region of $2nS$, Fig. 5.2B. A further increase of the inhibition, leads to further sharpening and selectivities that are comparable to those formed by nSTDP.

In all cases the mean excitatory input current to each unit is around 0.5pA (averaged during 20s over both selective and non-selective stimuli). As the inhibitory conductance is increased from 0nS to 10nS, the mean inhibitory current rises from 0pA to 0.25pA, around 50% of the average excitatory input. This inhibition does not have to be unrealistically strong to stabilize the receptive fields, instead the inhibition is of the same order of magnitude as the excitation.

An analytic treatment of the formation of receptive fields in wSTDP networks is difficult. However the underlying processes of the formation of the receptive fields in wSTDP networks and the relationship of their development to lateral inhibition can be described. Recall that wSTDP can be defined in terms of separate depression w_- and potentiation w_+ processes (chapter 4). Mathematically the equations governing w_- and w_+ resemble the description of a filter. The dynamics of wSTDP with an all to all implementation can thus be regarded as a combination of two filters acting upon the input and output frequencies of the neuron. In chapter 4 it was shown that the dynamics of the mean weight of wSTDP with the all to all spike implementation for a single unit can be expressed as

$$\frac{d\langle w(t)\rangle}{dt} = \nu_i\nu_j[\tau_+c_+ - \tau_-c_-\langle w(t)\rangle] + \nu_i\tau_-\frac{d\langle w_+(t)\rangle}{dt} - \nu_j\tau_+\frac{d\langle w_-(t)\rangle}{dt} \qquad (5.1)$$

where the variable $\langle w_-(t)\rangle$ is a filtered version of the mean input firing frequency and the variable $\langle w_+(t)\rangle$ is a filtered version of the mean output firing frequency of the neuron multiplied by the mean weight. Clearly, at equilibrium the last two terms of Eq.(5.1) are zero and this is the situation for the single unit simulations studied in chapter 4. However when the input stimulus is varying the last two terms in Eq. (5.1) can be non-zero, thus allowing a modification to the dynamics of the mean weight. Thus a persistently varying input stimulus leads to a persistently varying mean weight. In the network studied here, the stimulus is varying and this allows the last two terms of Eq.(5.1) to be non-zero and thus opens the possibility that the mean weight can vary with time even after training.

As we have seen the development of receptive fields in wSTDP networks is dependent upon the presence of lateral inhibition. If the mean weight is fluctuating due to the process described above, there is no reason for that fluctuation to lead to the persistent elevation or depression of the feedforward weights for any one neuron in the network. Thus input selectivity shall not arise. Instead the output firing frequencies of the neurons fluctuate randomly (for random input stimulus statistics). However if lateral inhibition is introduced, competition between neurons can occur. Dominant neurons are now able to suppress the firing of other neurons with smaller input from their feedforward weights. In the case that suppressed neurons stop firing all together, their weight evolution freezes completely. This allows any variations in the weights that have accumulated due to the varying input stimulus to be retained. Since such weight variations can be retained between stimulus presentations, neurons then mutually reinforce their differences. This process introduces competition and leads to the structure in the feedforward weights shown in Fig. 5.1.

5.2.1 Receptive field stability in STDP networks

Having established that receptive fields can develop in wSTDP networks, the stability of those receptive fields is now quantified. The network is presented again with the stimulus of Fig. 5.1A. As in the single neuron case of chapter 4, there is no separate learning and testing phase, instead we measure the persistence under the continued stimulation with the same stimulus ensemble. We track the receptive fields of the neurons, by plotting the tuning curves on a polar plot and taking the vector sum of the responses as described in §5.1.1. The direction of this receptive field vector gives a measure of the preferred direction, while its length depends on both the selectivity and

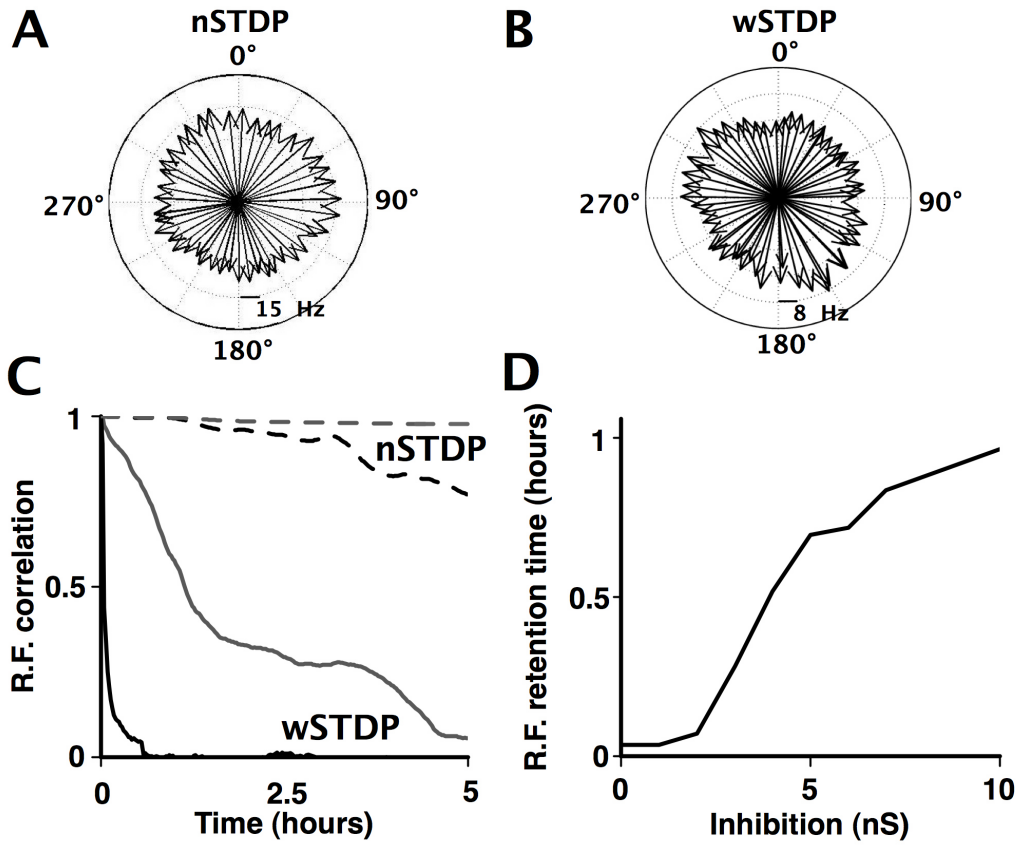

Figure 5.3: The stability of receptive fields in nSTDP and wSTDP networks. A: Receptive fields for the nSTDP network with no lateral inhibitory connections. B: The same as A but for a wSTDP network with 7nS lateral inhibitory connections. C: The autocorrelation of the receptive field for wSTDP and nSTDP networks with 0nS (black) and 7nS (grey) lateral inhibition strength. The receptive fields in nSTDP networks are stable even when there is no lateral inhibition. The receptive fields in wSTDP networks are less stable but their stability can be varied by altering the lateral inhibition. D: The receptive field retention time in the wSTDP network as a function of lateral inhibition.

the firing rate. The nSTDP receptive fields form an unordered map, i.e. neighboring cells don't necessarily have neighboring receptive fields (Song and Abbot, 2001), Fig. 5.3A. In wSTDP an unordered map forms as well provided that there is sufficient lateral inhibition, Fig. 5.3B, but the selectivity of the neurons is more variable.

To measure the stability of the receptive fields the autocorrelation of the receptive fields is calculated. If the receptive field autocorrelation is one, the receptive fields have not moved. If the receptive field autocorrelation falls to zero, the receptive fields have no relation to their initial locations. The nSTDP network with no lateral inhibition gives rise to a receptive field autocorrelation that decays with a timescale of 11 hrs, Fig. 5.3C. With inhibition this increases to 93 hrs but an accurate fit is difficult in this case because the very slow decay means that an enormous simulation time would be needed to see substantial decay of the memory.

The wSTDP receptive fields de-correlate quickly in comparison to the nSTDP network. Importantly however, the de-correlation timescale depends on the strength of the lateral inhibitory connections: When lateral connection strength is zero, no stable receptive field vectors exist (because, as we have seen, no receptive fields form) and the correlation time is simply that of filtered noise. However, as the strength of the inhibitory connections is increased, and the receptive fields sharpen, the correlation timescale of the receptive field vectors increases, Fig. 5.3D. Thus the stability of the receptive fields in wSTDP networks can be varied by altering the level of lateral inhibition. When the inhibition is sufficiently large, the receptive fields remain correlated with their initial positions for more than one hour. Although this is shorter than for the nSTDP network, the persistence in the network is much longer than the wSTDP single neuron persistence which is only 29s. Although in the nSTDP network inhibition also stabilizes the receptive fields, the improvement is much less dramatic than in the wSTDP case.

These results show that lateral inhibition introduces competition in the wSTDP network. This inhibition can be varied, hence varying the competition and the readiness with which receptive fields form in the network. Conversely strong competition is already present in the nSTDP learning rule itself. Thus while inhibition does sharpen the receptive fields, its absence does not prevent them from forming.

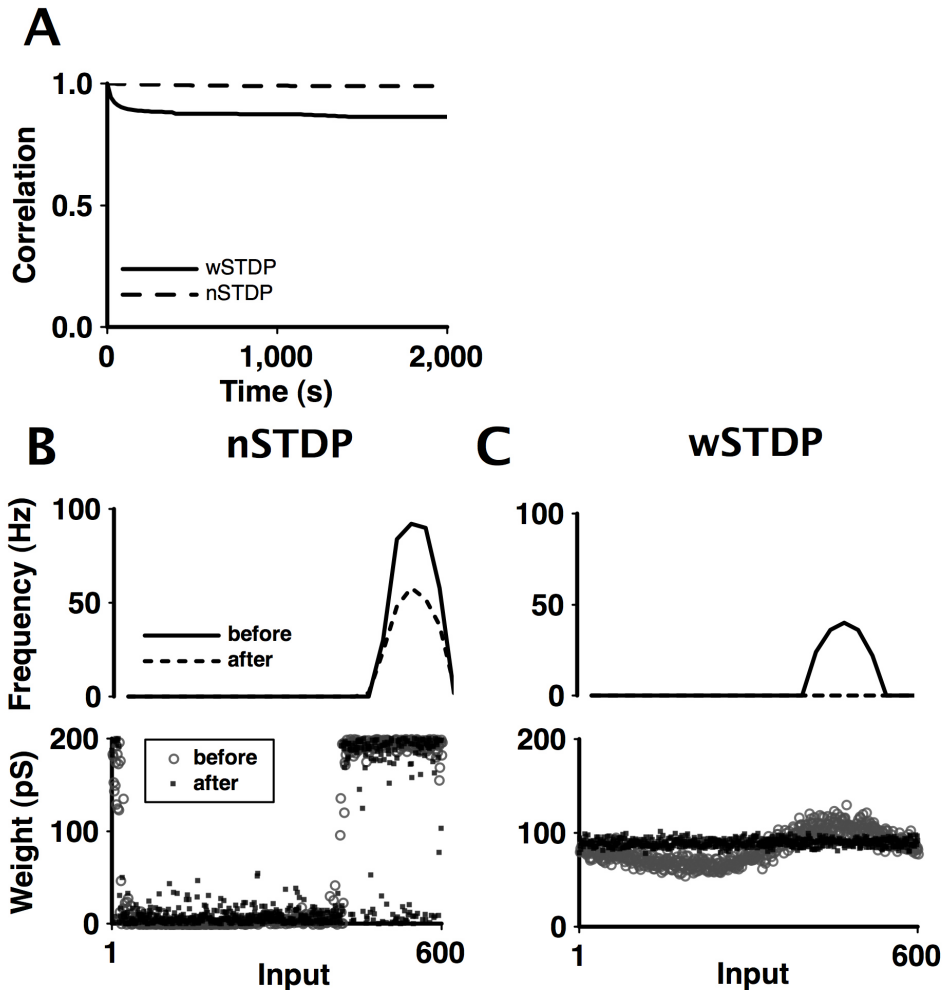

Figure 5.4: Untraining receptive fields in nSTDP and wSTDP networks with lateral inhibition. After the completion of training and formation of receptive fields, the networks are presented with a blank stimulus where all inputs fire at a uniform rate of 50Hz. A: The autocorrelation of the receptive field vectors for both wSTDP and nSTDP, when the blank stimulus is presented. B+C: The weights and tuning curve for a neuron in the wSTDP and nSTDP network. Top: Example receptive fields before and after untraining. B: The receptive field from a neuron in the nSTDP network before and after presentation of the blank stimulus. In the nSTDP case the peak firing rate of the receptive fields is equally reduced for all neurons. C: A receptive field from the wSTDP network before and after presentation of the blank stimulus, from a neuron that fires relatively rapidly at $\sim 10Hz$ during presentation of the blank stimulus. In the wSTDP case, receptive fields of rapidly firing neurons such as this one are wiped out, while the receptive fields of more slowly firing neurons take much longer to be forgotten. Bottom: The weights underlying the receptive fields plotted above before and after untraining.

5.2.2 Forgetting of receptive fields

So far, we have considered a situation in which there is no distinction between the learning phase and the test phase, as an identical stimulus ensemble was presented throughout and learning was ongoing. Thus we were considering the stationary stability plasticity dilemma (chapter 2). The non stationary case poses the equally important question of how quickly the learned receptive fields are forgotten when a different stimulus is presented. If the stimulus is absent altogether, no pre- or post-synaptic spikes are generated and the weights are maintained indefinitely. Therefore forgetting was assessed using an unstructured Poisson stimulus with the same firing rate for all inputs ($\nu_a = \nu_0$, §5.1). The forgetting is strongly dependent on the firing rate. When all inputs fire at the 10Hz background rate, no significant post-synaptic firing results in either the nSTDP or wSTDP network ($\nu_{post} < 0.5$Hz), and the receptive fields are retained for a very long period.

For an input of $\nu_a = \nu_0 = 50Hz$ the postsynaptic firing rate in the nSTDP network with lateral inhibition is about 4Hz, and the receptive fields do not decorrelate appreciably, Fig. 5.4A. The locations of the receptive fields remain fixed, because the depressed weights are so weak that they cannot drive the target neuron and the competitive property of nSTDP ensures strong inputs remain strong. Eventually the weak weights can become strong by chance but this takes place on a very long timescale comparable to the nSTDP single unit autocorrelation timescale at $\nu_{post} = 4$Hz.

In the case of the wSTDP network with $7nS$ lateral inhibitory connections, a 50Hz stimulus leads to some forgetting as reflected in the quick initial decay of the correlation, on a timescale of $\sim 50s$ Fig. 5.4A. Note however, that the correlation does not decay to zero. The reason is heterogeneity in the firing rates: some neurons fire at high rates ($\sim$10Hz) and these neurons forget the receptive field quickly, e.g. Fig. 5.4C, while other neurons fall silent in response to the unstructured stimulus, and hence retain their weights. The rapid and substantial de-correlation of wSTDP receptive fields is due to the loss of selectivity in the fastest firing neurons. Note that this effect does not occur when the network is stimulated with the original training input, Fig. 5.3. In that case, none of the neurons falls completely silent, Fig. 5.3C and the autocorrelation falls to 0.

When lateral inhibition is removed and the stimulus is set to be a uniform firing rate of 50Hz, the receptive fields in the wSTDP network are rapidly destroyed, Fig. 5.5. Removal of inhibition removes the suppression of firing of weakly driven neu-

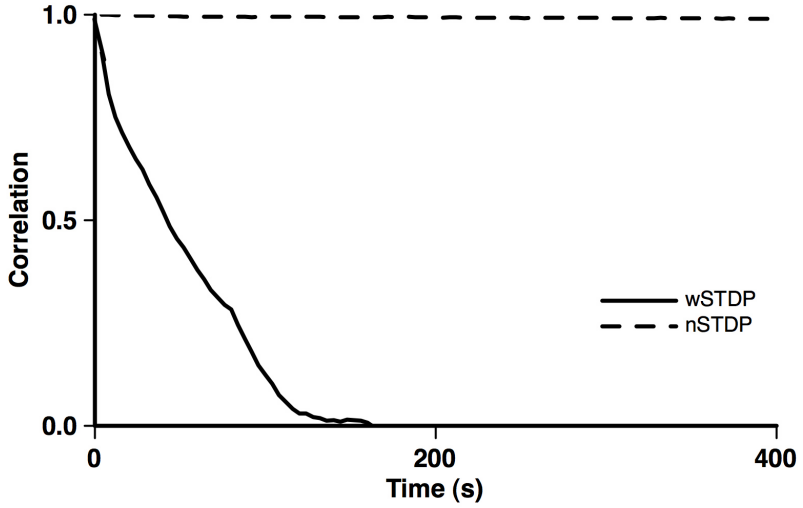

Figure 5.5: Untraining of receptive fields in nSTDP and wSTDP networks as in Fig. 5.4, but with acute removal of lateral inhibition. The wSTDP (solid line) network experiences a rapid loss of receptive fields. The nSTDP (dashed line) network retains its receptive fields.

rons, by strongly driven neurons. In this case all neurons loose their receptive fields and the structure in the weights of all neurons evolves in an identical fashion to Fig. 5.4C. The nSTDP network does not rapidly loose its receptive fields. This is because the existence of receptive fields in nSTDP networks does not depend upon lateral inhibition. When the blank stimulus is presented, the nSTDP weights gradually evolve toward their equilibrium distribution when the input is unstructured. This is identical to the single unit distribution discussed in chapter 4. Hence the reason that the nSTDP receptive fields survive even when lateral inhibition is removed is that the synaptic weights are intrinsically bistable.

5.3 Discussion

In this chapter, the study of the plasticity stability dilemma in STDP was extended to the case of networks where the neurons interact via lateral inhibition. This reveals two ways in which receptive fields in a network can be stabilised in the face of ongoing learning. 1) Synaptic plasticity with a high degree of intrinsic stability (nSTDP) gen-

erates receptive fields with a similarly high level of stability. 2) Lateral inhibition can stabilise receptive fields in a network with synaptic plasticity that is not intrinsically stable (wSTDP).

The intrinsic stability of nSTDP allows it to produce receptive fields that are far more stable than those in the wSTDP network. However, relearing the nSTDP receptive fields would require that the intrinsic stability of the learning rule be temporarily reduced, either by large increases in the firing rates of neurons or by transient alterations to the parameters of the plasticity (for example the step size). While neuromodulation might account for such changes, we saw in contrast that wSTDP is able to forget (and hence relearn) receptive fields as a result of simple variation in the level of lateral inhibition in the network.

Although receptive fields in wSTDP networks are less stable than those in nSTDP networks, the addition of lateral inhibition vastly slows down the decorrelation of the receptive fields, such that they remain correlated with their initial positions on a behaviorally relevant timescale of the order of an hour. This represents a substantial improvement over the correlation time of 29s for the weights in the single unit case. Thus in the context of the stationary plasticity stability dilemma, the lateral inhibition can increase the timescale of decorrelation of the receptive fields.

The networks were also tested in the context of the non stationary plasticity stability dilemma when an unstructured stimulus was presented. When lateral inhibition was present both nSTDP and wSTDP retained the previously learned structure. When lateral inhibition is removed, the wSTDP network suffers a rapid loss of the previously learned receptive fields, where all receptive fields were destroyed within 200s. Thus the lateral inhibition protected wSTDP against the substantial background firing rates of the 'blank' stimuli.

Several experimental studies suggest that spike timing dependent plasticity plays a role in plasticity in the visual cortex of various animals (Yao and Dan, 2001; Mu and Poo, 2006; Young et al., 2007; Dan and Poo, 2006). Modeling studies predict that spike timing dependent plasticity can lead to the development of input selectivity in populations of spiking neurons (Song and Abbot, 2001; Wenisch, Noll, and van Hemmen, 2005). In this chapter it was shown that wSTDP too can lead to the development of receptive fields.

Interestingly a recent study in rat auditory cortex demonstrated a process resembling the dependence of wSTDP receptive fields on lateral inhibition seen here (Froemke, Merzenich, and Schreiner, 2007). The nucleus basalis was stimulated, leading to re-

lease of acetylcholine into primary auditory cortex. The stimulation of the nucleus basalis was paired with tones to activate selective cells. It was found that the release of acetylcholine rapidly reduced the inhibitory input to the activated cells in the auditory cortex, leading to a period of large excitatory input. After a period of around 30 mins the auditory receptive fields had shifted to prefer the paired input. This shift coincided with a return of the inhibition to pre-pairing levels over a timecourse of around 2 hours. The results in this chapter suggest that one way in which this inhibition related remapping might be achieved is by using synaptic plasticity lacking strong intrinsic correlation stability.

Chapter 6

State based models of Long Term Potentiation

It is the prevailing paradigm in neuroscience that memories are stored by patterns of efficacy within populations of excitatory synapses. As discussed in chapter 2, this raises the possible risk that synapses will forever overwrite their previous learning due to ongoing changes. In chapter 4 this scenario was analysed in the case of two popular models of STDP. It was found that the dynamics of the learning rule determines the effect of fluctuations on the autocorrelation of the synaptic weights. In this chapter, the implications of synapses that vary on more than one timescale is analysed using a novel model of LTP.

Long term potentiation is the process whereby synapses increase and maintain their efficacies over long periods. Conversely long term depression enables synapses to decrease their efficacies over similarly extended timescales. In hippocampal slice preparations several decay timescales of LTP/D have been observed (Bliss and Lomo, 1973; Dudek and Bear, 1992; Frey and Morris, 1997; Bashir and Collingridge, 1994; Abraham, 2003; Alpermann et al., 2006; Reymann and Frey, 2007) (see chapter 1). These decay timescales are associated with different 'phases' of LTP/D. Changes to synapses that decay within a timescale on the order of tens of minutes or several hours are typically referred to as early phase. With more sustained stimulation the synapse can be made to retain changes on a far longer timescale, on the order of days, and even longer. This longer lasting form of LTP/LTD is termed late phase and is protein synthesis dependent (Otani et al., 1989; Frey and Morris, 1997). Interestingly the magnitude of the synaptic weight change can be similar for early and for late LTP and the synaptic dynamics observed in LTP experiments typically appear multi-exponential.

Previous authors have explored the persistence of memory traces within a pool of synapses that are taken to occupy a discrete number of states (Fusi, Drew, and Abbott, 2005; Fusi and Senn, 2006; Fusi and Abbott, 2007). The state based approach is a promising new method to construct models of synaptic plasticity. Its graphical basis allows for intuitive reasoning to produce models that can be easily transformed into standard equations. As mentioned in chapters 1&2, there have been notable models using the state based approach to calculate the effects of spike interaction in STDP (Appleby and Elliott, 2006) and calculate memory trace lifetimes in ensembles of synapses (Fusi, Drew, and Abbott, 2005; Senn and Fusi, 2005; Fusi and Senn, 2006; Fusi and Abbott, 2007).

In this chapter models of Long Term Potentiation and Depression are developed that are based on postulated synaptic states. Electrophysiological measurements are assumed to represent the ensemble mean of a large pool of synapses each making stochastic transitions between the postulated states. These electrophysiological measurements are used to determine the transition rates between the states within the model. Firstly the general modeling approach is outlined and assumptions stated. Next, the methods used for calculating with the models and evaluating the memory trace lifetime are explained. This allows 2, 4 and 8 state models to be described and justified. Finally, the models are compared to experimental data and are used to calculate the memory trace strength and lifetime in an ensemble of putative hippocampal CA1 synapses.

6.1 Modeling approach

There have been experimental observations suggesting that synapses are capable of discrete state transitions (Petersen et al., 1998; O'Connor, Wittenberg, and Wang, 2005). Although these experiments await further confirmation, they support the state based modeling approach that was introduced in chapter 2.

One way that synapses might change state is by altering their molecular composition. Addition or removal of a synaptic component (such as an AMPAR), or the phosphorylation of some synaptic component (such as CAMKII), are two plausible candidate mechanisms whereby the synapses might change state. In this chapter, synapses are assumed to occupy one of a small number of discrete states. These states are based upon the subcellular machinery as identified in the literature discussed in chapter 1. Synapses are permitted to move between these states with some probability.

It is assumed that the fundamental difference between early and late LTP/LTD is that AMPAR receptors are more firmly anchored within the postsynaptic membrane of synapses that have undergone a late phase change. Experiments suggest that the mobility of AMPARs is linked to how readily the AMPARs associate with the post synaptic density (PSD) (Choquet and Triller, 2003; Triller and Choquet, 2003; Bredt et al., 2004; Triller and Choquet, 2005; Kim et al., 2007). For this reason it is assumed that the maintenance of LTP/D is postsynaptic and is regulated by the association between receptors and the PSD.

Binary synapses having capacity for a single 'unit' of AMPA receptors to be accommodated are considered. The synaptic weight $w \in \{0,1\}$ is 0 when receptors are absent and 1 when they are present. Synaptic stability is mediated by a postsynaptic 'slot' structure which is a simplified form of the 'hyperslot' suggested in a recent qualitative model (Lisman and Raghavachari, 2006). When the slot variable $s \in \{0,1\}$ is 0, the AMPARs are not anchored and have a short lived occupancy in the post synaptic membrane. On the other hand, when the slot variable is 1, the AMPARs are anchored and remain in the synapse for a longer duration.

We already met the simplest model, the 2 state binary model, in chapter 2 and in chapter 3. In the 2 state model, there is only one variable w associated with the synapse, giving only two possible states, Fig. 6.1A. In the other models here, the synapses are associated with both the weight variable w and the slot variable s. In this case, the possible combinations of AMPAR and slot in the synapses is, Fig. 6.1B:

1. Empty ($w = 0$, $s = 0$): The synapse contains neither AMPARs nor the apparatus for stabilising them. Thus the weight is depressed and the synapse will not retain AMPAR receptors should they be placed in the synapse.

2. Unanchored ($w = 1$, $s = 0$): The synapse contains AMPAR receptors, but no apparatus required to anchor the receptor. Thus the weight is potentiated but liable to switch back to $w = 0$ rapidly.

3. Anchored ($w = 1$, $s = 1$): The synapse contains both an AMPAR receptor and the underlying apparatus required to anchor that receptor to the PSD. Thus the weight is potentiated and stable[1]

4. Depleted ($w = 0$, $s = 1$): The synapse contains apparatus for stabilising the AMPAR receptor, but no receptor. Thus the weight is depressed but will be both

[1] Stable in the sense that the fluctuations of a weight in this state are comparatively small and so the correlation timescale of a synapse in this state is longer than other states.

potentiated and stable should an AMPAR receptor be placed in the synapse.

Each of these combinations defines a state. States will be referred to using the numbers listed here. Two, four and eight state models are presented. The two state model consists of states $\{1,2\}$ only, Fig. 6.1A. The four state model utilises states $\{1,2,3,4\}$, Fig. 6.1B. The eight state model is formed by the addition of a switch variable, $h \in \{0,1\}$ that determines whether the synapses are potentiated or depressed on long timescales. The switch could be a process such as autophosphorylation of CAMKII which has been shown to be a plausible bistable switch (Hayer and Bhalla, 2005). The 8 state model thus incorporates: The states $\{1,2,3,4\}$ with $h = 0$ for which the transition probabilities between the states are arranged so that synapses tend to 'flow' towards the empty state (1), Fig. 6.1C (front ring). In addition the states $\{5,6,7,8\}$ are identical to the four states described above, but with $h = 1$ which switches the transition probabilities such that synapses tend to flow toward the anchored state (7), Fig. 6.1C (rear ring).

Synapses are able to move between the states in each model with probabilities associated with the respective transitions as shown in Fig. 6.1. Synapses cannot undergo transitions between states that are not adjacent. For example transitions cannot occur that alter both the weight and the slot variable simultaneously because no single transition between adjacent states allows this to happen in Fig. 6.1. This 'one-step' constraint simplifies the analysis of the model and amounts to the assumption that the system can only make one transition within a memory storage interval. The models are completely specified by a transition matrix containing the transition probabilities between the states.

The induction of plasticity is modelled as a temporary change to the transition probabilities. This causes synapses to change their states and leads to a deflection in the mean weight of the ensemble. The dynamics of the deflection can be matched to electrophysiological data. In this chapter the models are matched to unpublished experimental data obtained by Roger Redondo in Edinburgh. This data has decay timescales of early and late LTP that are comparable to those quoted in chapter 1. The following interpretations of the experimental data are made:

- To prevent anomalous results, slices should be left for several hours so that biochemical equilibrium is attained. Typically slices are left for at least 4 hours (Sajikumar and Frey, 2004). It is assumed that at the start of classical LTP experiments, the ensemble of synapses in the hippocampal slice is in the steady state. We shall refer to this as the steady state and the transition probabilities in

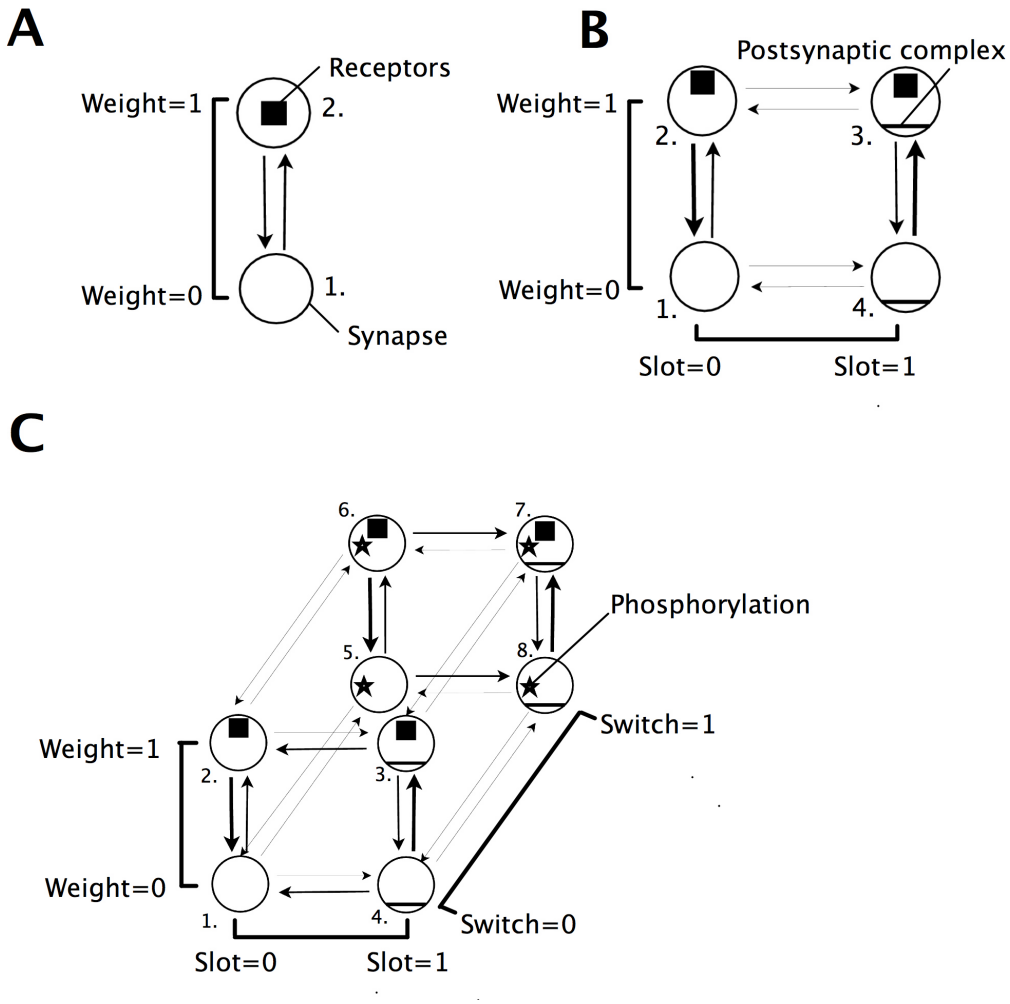

Figure 6.1: State based models of the synapse. Each transition, marked with an arrow has an associated probability that the synapse undergoes the transition from one state to the other in the direction of the arrow. The thickness of the arrows indicates the magnitude of the transition probability: Thin arrows represent transitions that occur seldomly, while thick arrows represent transitions that occur frequently. A: Simple two state model in which the synapse can be potentiated or depressed. B: A four state model in which the synapse can be potentiated or depressed. In addition to this, the receptors can be anchored. C: The 8 state model consists of two ring diagrams such as in B, where one ring (front) has its transition probabilities configured such that at equilibrium the synapses tend to congregate in the empty state (1) while the second ring (rear) has transition probabilities configured such that at equilibrium the synapses tend to congregate in the anchored potentiated state (7). In this framework late phase LTP/LTD corresponds to changing the switch variable causing a transition between the two rings. Late phase LTP is the transition of synapses from the depressed ring into the potentiated ring. Late phase LTD is the transition of synapses from the potentiated ring into the depressed ring.

this state as the baseline transition probabilities. The steady state gives rise to the stable weight trace that is often referred to as 'baseline' in the experimental literature. In the models presented here, baseline occurs when the model is at equilibrium and the probability density across the states is stationary.

- It is assumed that the synaptic ensemble within an equilibrated hippocampal slice is governed by a process for which all transition probabilities are constant in time. When an early LTP/D or late LTP/D protocol is applied to the slice, then the probability of AMPAR insertion is altered during the duration of the protocol. It is assumed that the transition probabilities are constant for the duration of induction protocols. When the protocol ends it is assumed that the transition probabilities for AMPAR and slot insertion return to their constant baseline values.

- When an early LTP induction protocol is applied to a hippocampal slice, biochemical processes are activated that lead to an increase in the synaptic efficacy by insertion of additional AMPARs. This can be interpreted as a temporary increase in the probability of AMPAR insertion. After the LTP induction protocol is complete we assume that the probability of AMPAR insertion instantaneously returns to the baseline probability as it was before the application of the protocol.

- When a late LTP protocol is applied, biochemical processes are activated that lead to an increase in the probability of insertion of AMPARs and the insertion of slots that stabilise those AMPARs. When the protocol ends, the transition probabilities instantaneously return to their initial baseline values. Again, the synaptic weight trace is deflected and then decays. However it decays at a rate that is far slower than in the early LTP case, because the synapses that contain slots have a lower probability that their AMPARs are removed.

- It is assumed that LTP and LTD are symmetric. Thus early or late LTD corresponds to the exact opposite of early or late LTP, having identical induction timescales and decay timescales.

First the state based models are specified and matched to the decay and induction timescales of LTP protocols (the decay of a perturbation from equilibrium). Then the models shall be used to calculate the lifetime of a memory trace within a synaptic ensemble conforming to the experimentally observed synaptic dynamics (in an SPS sense: The equilibrium autocorrelation/SNR).

6.2 Description of the models

The methods discussed in Chapter 3 are used to analyse the state based models in this chapter. In particular the dynamics of the models is solved using Eq. (3.7) and expansion in eigenfunctions Eq. (3.17). The memory trace is quantified in terms of the signal to noise ratio Eq. (3.32).

6.2.1 Plasticity induction in state based models

In Markov models of synapses, individual synaptic weights are stochastic processes described by a master equation derived from the transition matrix specifying the model (see chapter 3). Regardless of the initial condition, the ensemble of synapses will reach some fixed point of the dynamics (provided that the model satisfies the constraints mentioned in chapter 3). In the homogeneous case, the transition probabilities do not vary. Thus once the mean synaptic weight of the ensemble has reached the fixed point it fluctuates about that point for all time.

The models proposed in this thesis assume that the dynamics of synapses during LTP can be described using the transition rates between states. In this case, although individual synapses shall undergo random fluctuations, it is very unlikely that large numbers of synapses shall undergo any one transition simultaneously at equilibrium. Thus a large deflection to the mean synaptic weight as observed in LTP experiments requires that the transition rates be altered[2]. For this reason, it is proposed that LTP induction changes the transition rates between the synaptic states and that the model describing this process is therefore an inhomogeneous Markov process.

In an inhomogeneous process the transition probabilities can alter over time. In the state based models, plasticity is induced by making changes to the transition probabilities, Fig. 6.2. We shall consider the case that changes to the transition probabilities are piecewise, i.e. at some instant in time the matrix of transition probabilities in the model is replaced by a new matrix, that remains constant for some duration. In this case the model remains solvable as described above but in a piecewise sense: The initial condition for each segment is the final state of the preceding segment at the preceding time instant.

Consider the case in which the model is intialised at t_0 with a transition matrix M and the probability vector associated the stable state of the matrix, $\mathbf{p}_\infty$. At a later

[2]Given that the model has linear dissipative dynamics and that the fluctuations are Gaussian. For non-linear dynamics or more exotic noise sources, this would not necessarily be the case.

time $t_1 > t_0$ an instantaneous substitution of the initial transition matrix M for a new matrix E is made. The matrix E remains for some duration of time $\Delta t = t_2 - t_1, t_2 > t_1$. The set of probabilities E imply a different steady state $\mathbf{p}_E$ for the system and so for the duration Δt the system follows a trajectory toward $\mathbf{p}_E$. At t_2 the original transition matrix M is substituted back into the model causing a relaxation toward $\mathbf{p}_\infty$ along some trajectory.

6.2.2 Two state model

The solution to the dynamics of the two state model shown in Fig. 6.1A can be found easily. The dynamics of the mean synaptic weight of the ensemble is governed by a single exponential having timescale

$$\tau = \frac{1}{R_{12} + R_{21}} \tag{6.1}$$

where R_{12} and R_{21} are the transition rates between the states. When R_{12} and R_{21} are the baseline transition rates, this is the timescale of decay of LTP. On the other hand, when the baseline transition matrix has been replaced by a protocol matrix such that $R_{12} = R_{12}^{protocol}$ and $R_{21} = R_{21}^{protocol}$ representing the altered rates of AMPAR receptor insertion and removal respectively, Eq.(6.1) would be the timescale associated with the induction of the protocol. The fixed point of the dynamics is

$$\langle w \rangle = \frac{R_{12}}{R_{12} + R_{21}}. \tag{6.2}$$

For any initial condition the mean weight of the ensemble converges upon the fixed point in Eq. (6.2) with an exponential weight trajectory having timeconstant Eq. (6.1). As was demonstrated in chapter 3, the 2 state model has an unnormalised autocorrelation timescale that is identical to its response timescale regardless of the initial condition, Eq. (6.1). Thus, in the two state model with constant transition rates at equilibrium, there is only ever one timescale of weight relaxation and signal decay (the autocorrelation timescale). (However this timescale does change when the transition rates are altered during LTP induction protocols.)

Consider the strength of a memory trace that is stored at the steady state of the 2 state model when $\mathbf{p}(t) = \mathbf{p}_\infty$. The magnitude of the strength of the signal is proportional to the number of synaptic weights that transition in the memory storage interval Δt (chapter 3). The initial signal of a memory trace stored within an ensemble of 2 state synapses is calculated with Eq. (3.37) (chapter 3). For the 2 state

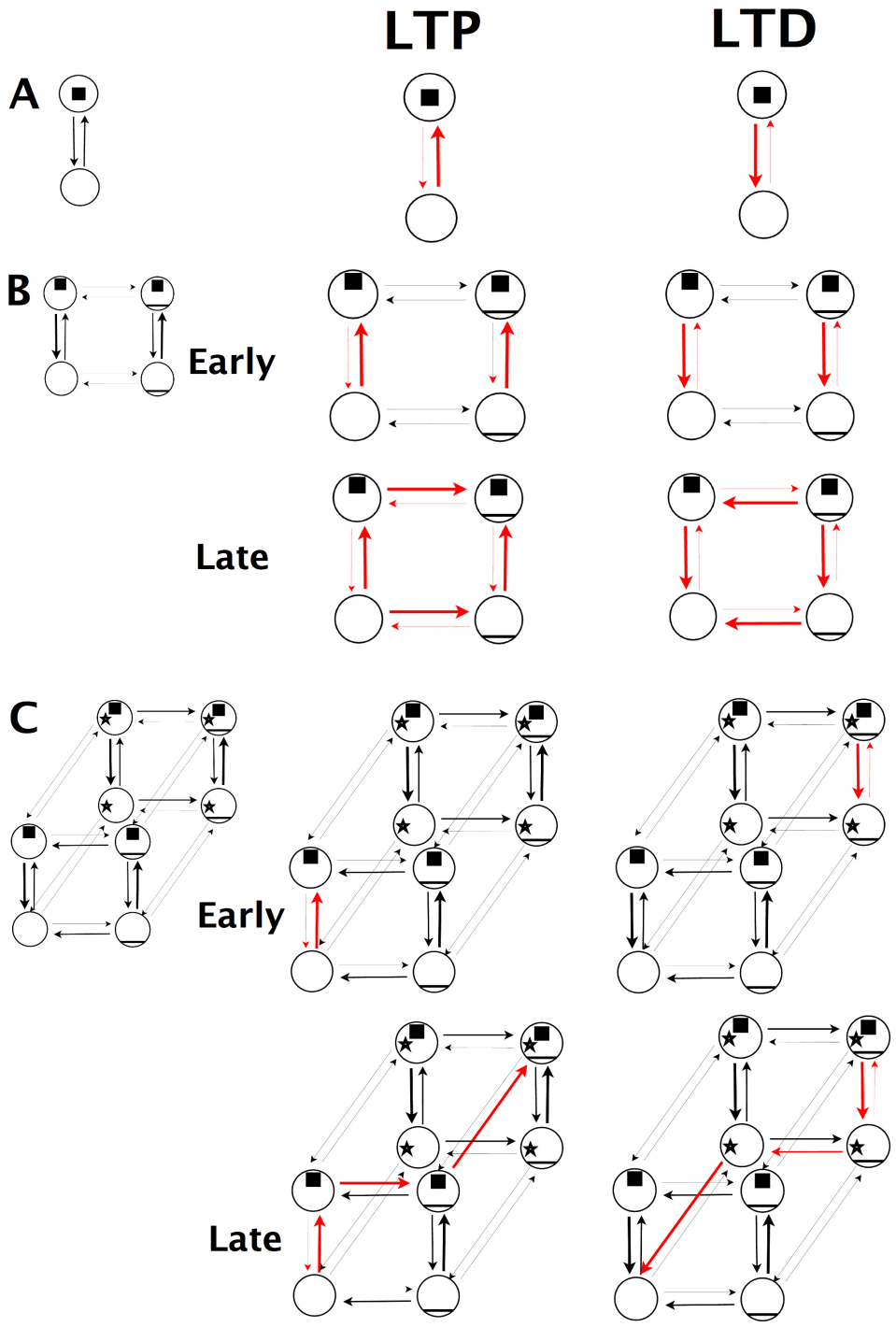

Figure 6.2: Plasticity induction protocols in state based models. Red arrows indicate transitions where rates are altered by application of the protocol. The size of the red arrows indicates the magnitude of the rate during the application of the protocol. A: LTP and LTD in the 2 state model. In the two state model there can only be one timescale of LTP/D. B: LTP and LTD in the 4 state model. In the 4 state model, separate protocols are defined which lead to the decay of the weight change on two possible timescales, early LTP/D and late LTP/D. C: LTP and LTD in the 8 state model. Late LTP causes synapses to be transferred between rings due to elevation of $3 \rightarrow 7$, while late LTD causes synapses to transfer via $5 \rightarrow 1$.

model $S_0 = \Omega(\mathbf{w^T} M_+ \mathbf{p}_\infty + \mathbf{w^T} M_- \mathbf{p}_\infty) = \Omega(M_{12} p_1 + M_{21} p_2)$, since $M_+ = \Gamma M \Gamma'$ and $M_- = \Gamma' M \Gamma$, where Ω is the number of synapses in the ensemble. There are only 2 states and so $p_1 = 1 - p_2 = \langle w \rangle$. At the steady state $p_2 = R_{12}/(R_{12} + R_{21}) = \Delta t M_{12}/[\Delta t (M_{12} + M_{21})]$, where $\Delta t = 1s$ is the pattern storage interval and 1 pattern is stored per second (chapter 3). Hence

$$\frac{S_0}{\Omega} = \frac{2 M_{12} M_{21}}{M_{12} + M_{21}}. \tag{6.3}$$

where the transition probabilities M_{12} and M_{21} are the baseline transition probabilities of the model. Eq. (6.3) gives an estimate of the number of synaptic weight transitions expressed as a fraction of the total number of synapses, as a function of the transition probabilities between the states, when the model is at the steady state.

In Eq. (6.3) we see the important ingredient of the stationary plasticity stability dilemma, namely that we cannot boost the initial signal of the memory trace by increasing M_{12} or M_{21} without a concomitant reduction to the memory lifetime in Eq. (6.1).

Since there is only one timescale in the two state model there cannot be both early and late phase potentiation and depression. We shall take the decay timescale of LTP/LTD in the two state model to be 4800s. This is the decay timescale of a single exponential fit to the averaged experimental data in Fig. 6.5A obtained by Roger Redondo [3] from CA1-CA3 synapses in acute hippocampal slices and is compatible with values in the literature (see chapter 1).

LTP is performed in the two state model by increasing R_{12} relative to R_{21}, Fig. 6.2A. We take the early LTP induction timescale found from the data in Fig. 6.5A of 150s for 150% potentiation. In the experiments producing this data, plasticity was induced by means of a theta burst protocol of 4 pulses at 100Hz with 200ms interpulse interval. Assuming that if LTP were saturated, 99% of the synapses would be potentiated, rates corresponding to this induction timescale can be found by solving the synaptic weight increase during induction with a single exponential, yielding $R_{12}^{LTP} = 0.0044 s^{-1}$ and $R_{21}^{LTP} = 8.97 \times 10^{-5} s^{-1}$. The choice of what proportion of synapses should be potentiated if LTP were saturated, makes no difference to the results presented here and is relevant only so that LTP and LTD in the models can be induced at approximately the same rate as the data (i.e. so that the initial condition prior to LTP decay matches the initial weight before LTP decay in the experiment).

LTD is induced by increasing R_{21} relative to R_{12}. As stated previously it is assumed

[3] Morris lab Edinburgh, unpublished data

that LTD is symmetric with LTP. Thus LTD is induced in the two state model with $R_{21}^{LTD} = 0.0044s^{-1}$ and $R_{12}^{LTD} = 8.97 \times 10^{-5}s^{-1}$.

To further constrain the transition rates, $\langle w \rangle$ must also be fixed. Assuming that $w \in \{0,1\}$ and if half of the synapses are potentiated and half are depressed at baseline such that $\langle w \rangle = 1/2$, then the magnitude of EPSP slope increase upon LTP should be identical to the magnitude of EPSP slope decrease upon saturation of LTD, i.e. the asymptotes of LTP and LTD should be symmetric. The data on this issue are a little unclear, some studies show an asymmetry in the LTP and LTD asymptotes (O'Connor, Wittenberg, and Wang, 2005) while in other studies they appear to be symmetrical (Dudek and Bear, 1992). Here, in all models it is assumed that $\langle w \rangle = 1/2$, this being the simplest option from the point of view of analysing the models. This implies $R_{12} = R_{21} = 1.04 \times 10^{-4}s^{-1}$.

Since patterns are stored at the rate of 1 per second the transition probabilities during the memory storage interval are $M_{12} = M_{21} = 1.04 \times 10^{-4}$. Therefore the values of the rate constants and the transition probabilities are equal. This is the case for the rest of this chapter. But what are the implications for other values of the number of patterns stored per second?: If $\mathfrak{n} > 1s^{-1}$ the pattern storage interval Δt would be reduced. Thus the rates R would be increased and the decay timescale of LTP would be reduced. To match the decay timescale of LTP/D to data, the rates R would need to be readjusted to their previous values. However this would require a *reduction* in the transition probabilities M. This would lead to a decrease in the initial signal of a pattern stored at the steady state. Thus in the state based models here, more patterns could be stored per second in the steady state, but this would be at the expense of their initial signal if the decay timescales of LTP/D are constrained against the data.

6.2.3 Four state ring model

The 2 state model is useful in order to illustrate the principle of state based models and as a benchmark against the performance of more complex models. However it is poorly suited to being a model of early and late phase LTP because it has only one weight decay timescale. Thus to produce a model that can be utilised to explore the effect of early and late LTP upon the memory trace, another timescale of the weight decay must be introduced. One way of doing this is to introduce more states into the state diagram taking into account more of the synaptic variables stated in §6.1. To achieve this a four state model is proposed whose states enumerate the possible

combinations of the AMPAR variable w and post-synaptic slot variable s.

The 4 state model is arranged in a ring because this is a simple structure that is capable of reproducing the phenomenology of LTP/D. Specifically, the typical sequences of LTP/D induction are themselves cyclic: early LTP occurs *prior* to late LTP, suggesting that in order to reach the stable potentiated state, the synapse must *pass through* an unstable potentiated state; once late LTP is established, early LTD temporarily depresses the synapse such that it returns to the previous elevated baseline; further depression of the synapse induces late LTD, which returns the synapse to its initial strength.

For arbitrary transition probabilities, the dynamics of the four state ring in Fig. 6.1B cannot be found by analysis. If detailed balance is imposed on the ring using Eq. (3.26) (see chapter 3) then solution is possible, however the solutions are lengthy. Manageable solutions can be obtained in the case that simplifying assumptions are made. It is again assumed that in the steady state, half of the synapses in the ensemble are potentiated and that half are depressed and that early LTP and early LTD are symmetric. Thus about half of the synapses should be in state 1 and half in state 3 (because the empty and anchored states are the most stable states), Fig. 6.1B. Hence the transitions $\{1 \to 2, 2 \to 1\}$ should be as likely as $\{3 \to 4, 4 \to 3\}$. We shall assume that at baseline it is extremely unlikely that postsynaptic slots are added and removed. Hence the transition probabilities for transitions $\{1 \to 4, 3 \to 2\}$ must be small. We can satisfy these constraints and impose detailed balance by specifying $R_{34} = R_{12}$, $R_{43} = R_{21}$ and $R_{14} = R_{41} = R_{23} = R_{32} = \varepsilon$ where ε is the rate of slot addition and removal. The rate matrix for the four state model thus specified has four eigenvalues, three of which are non-zero. The non-zero eigenvalues are,

$$\begin{aligned} \lambda_1 &= -2\varepsilon \\ \lambda_2 &= -R_{12} - R_{21} \\ \lambda_3 &= -2\varepsilon - R_{12} - R_{21} \end{aligned} \tag{6.4}$$

where the timescales of weight decay are $\tau_1 = -1/\lambda_1$, $\tau_2 = -1/\lambda_2$ and $\tau_3 = -1/\lambda_3$. The asymptotic steady state is given by the eigenvector associated with the zero eigenvalue,

$$\mathbf{p}_\infty = Z\left\{\frac{\varepsilon + R_{21}}{\varepsilon + R_{12}}, 1, \frac{\varepsilon + R_{21}}{\varepsilon + R_{12}}, 1\right\} \tag{6.5}$$

where $Z = 1/[2(1 + \frac{\varepsilon + R_{21}]}{\varepsilon + R_{12}})$ ensures $\sum_i p_i = 1$. Thus starting from some arbitrary initial distribution, the ensemble of synapses approaches the steady state with weight dynamics that is a linear combination of exponentials each having a timescale chosen from Eqs. (6.4). Since ε is small only two of these timescales are distinct by a large mar-

gin (τ_1 and $\tau_{2/3}$) and so there are two timescales in practice. These two timescales are the timescales associated with the decay of early LTP/D, $\tau_{2/3}$ and late LTP/D, τ_1. The precise mixture of exponentials Eq. (3.17) (chapter 3) having these timescales is determined by the initial condition. For certain initial conditions - such as those straight after an induction protocol - the weight dynamics might display only a single timescale.

For the four state model a formula for the initial signal in terms of the steady state and the decay timescale can be derived. In the four state model, the initial signal is $S_0 = \Omega(p_1M_{12} + p_2M_{21} + p_3M_{34} + p_4M_{43}) = 2(p_1M_{12} + p_2M_{21})$ due to symmetry. Substituting from Eq. (6.5) gives[4].

$$\frac{S_0}{\Omega} = \frac{3M_{12}M_{21} + \rho(2M_{12} + M_{21})}{2(2\rho + M_{12} + M_{21})} \tag{6.6}$$

where $\rho = \varepsilon\Delta t$.

The restrictions made upon the transition probabilities mean that there is only one independent occupancy at baseline in Eq. (6.5), since $1 = 2p_1 + 2p_2$. Thus the baseline distribution amongst the states is completely described by $p_2 = 1/[2(1 + \frac{\varepsilon+R_{21}}{\varepsilon+R_{12}})]$, implying

$$M_{12} = \frac{\rho + 4\rho p_2 + 2p_2M_{21}}{2p_2 - 1}. \tag{6.7}$$

Since ε is small we assume $\tau_3 = \tau_2 = \tau$ and match τ to the eLTP timescale. From Eq. (6.4), $\tau = \Delta t/(M_{12} + M_{21})$, substituting for M_{12} from Eq. (6.7) yeilds,

$$M_{21} = \frac{\Delta t - 2p_2\Delta t - \rho\tau}{\tau}. \tag{6.8}$$

Substituting Eqs. (6.7+6.8) into Eq. (6.6) gives,

$$\frac{S_0}{\Omega} = \frac{3p_2\Delta t - 6p_2^2\Delta t - \rho\tau + \rho p_2\tau}{\tau}. \tag{6.9}$$

Since the addition of slot proteins is very improbable and ε is small, take $\rho = 0$,

$$\frac{S_0}{\Omega} = \frac{\Delta t(3p_2 - 6p_2^2)}{\tau}. \tag{6.10}$$

Eq. (6.10) cannot be negative because we have asserted that the transition probabilities between states 1 and 2 and states 3 and 4 are symmetric ($M_{34} = M_{12}, M_{43} = M_{21}$). This implies that $0 \leq p_2 \leq 1/2$. Another consequence of the simplifying assumptions made

[4] Again, recall that $R = \mathfrak{n}M$ with $\mathfrak{n} = 1s^{-1}, \Delta t = 1/\mathfrak{n}$ (chapter 3).

about the transition probabilities is that the four state model has a mean weight that is always $\langle w \rangle = 1/2$. This follows from $\langle w \rangle = p_2 + p_3$, giving

$$\langle w \rangle = \frac{\frac{\varepsilon+R_{21}}{\varepsilon+R_{12}}}{2\left[1+\frac{\varepsilon+R_{21}}{\varepsilon+R_{12}}\right]} + \frac{1}{2\left[1+\frac{\varepsilon+R_{21}}{\varepsilon+R_{12}}\right]} = \frac{1}{2}.$$

Despite the independence of the mean weight $\langle w \rangle$ and the transition rates in the 4 state model, the initial signal is dependent on the transition rates according to Eq. (6.10). In order to compare the four state model with the 8 state model, the transition probabilities are set such that the initial signal of the four state model is identical to the that of the 8 state model, $S_0^{(4)} = S_0^{(8)}$ with an eLTP timeconstant $\tau_2 = 4800s$.

From a single exponential fit to data from Roger Redondo in which late LTP was induced in CA1 of hippocampal slices, the late LTP decay timescale is taken to be $3 \times 10^5 s$ (83 hours). Thus $\varepsilon = 1.7 \times 10^{-6} s^{-1}$. When the initial signal of the four state model is matched to that of the 8 state model (see next section) by using Eq.(6.10) and for an early LTP/LTD decay timescale of 4800s, $R_{21} = R_{43} = 1.9 \times 10^{-4} s^{-1}$ and $R_{12} = R_{34} = 1.3 \times 10^{-5} s^{-1}$.

Early LTP is performed in the four state model by switching the direction of flow of states between state 1 and 2 such that synapses tend to make transitions toward the two states having $w = 1$, Fig. 6.2B. To match the eLTP induction timescale of $150s$ for 150% potentiation set $R_{12}^{eLTP} = 0.0044 s^{-1}$ and $R_{21}^{eLTP} = 8.97 \times 10^{-5} s^{-1}$ where we take the same probabilities as in the two state case by virtue of the smallness of ε (i.e. ε has little effect upon early LTP induction). In order that detailed balance be preserved it is also necessary to set $R_{43}^{eLTP} = R_{12}^{eLTP}$ and $R_{34}^{eLTP} = R_{21}^{eLTP}$. This has no impact upon the number of synapses occupying states $\{3,4\}$ although it does increase their flux, but this is of no consequence in this study.

Late LTP is performed in the four state model by setting $R_{12}^{lLTP} > R_{21}^{lLTP}$ but in addition R_{23}^{lLTP} is larger than R_{23}, causing a flow of synapses into state 3, Fig. 6.2B. Detailed balance is preserved by setting $R_{14}^{lLTP} = R_{23}^{lLTP}$, $R_{41}^{lLTP} = R_{32}^{lLTP}$. Thus in order to preserve detailed balance synapses flow from state 1 to state 4 in addition to flowing from state 2 to state 3. In the case of late LTD we set $R_{34}^{lLTD} > R_{43}^{lLTD}$ and in addition R_{32}^{lLTD} is larger than R_{32} where again $R_{14}^{lLTD} = R_{23}^{lLTD}$, $R_{41}^{lLTD} = R_{32}^{lLTD}$.

Late LTP was induced in area CA1 of acute hippocampal slices by Roger Redondo using a theta burst stimulation protocol consisting of 3 trains of 100 pulses at 100 Hz with a 200ms intertrain interval, Fig. 6.5B. This yields $\sim 180\%$ potentiation in 150s. The data can be matched by setting $R_{43}^{lLTP} = R_{12}^{lLTP} = 0.0071 s^{-1}$, $R_{34}^{lLTP} = R_{21}^{lLTP} =$

$1.45 \times 10^{-4} s^{-1}$ and $R_{23}^{lLTP} = 0.0071 s^{-1}$ for lLTP and $R_{21}^{lLTD} = R_{34}^{lLTD} = 0.0071 s^{-1}$, $R_{12}^{lLTD} = R_{43}^{lLTD} = 1.45 \times 10^{-4} s^{-1}$ and $R_{32}^{lLTD} = 0.0071 s^{-1}$ for lLTD. The application of late LTP/D causes not only the redistribution of synapses amongst the weight states but also redistributes them amongst slot states, Fig. 6.2B. Thus the decay of late LTP/D has a significant component that decays with the slow timescale τ_1.

6.2.4 Eight state model

The four state model has only one stabilisation process, AMPARs are either anchored to the post synaptic density or not. As we have seen, the four state model has only two distinct decay timescales in practice. However, as discussed in chapter 1, there are several candidate stabilisation processes at the synapse. In particular it seems that phosphorylation plays an important role in LTP, complimenting the synthesis and regulation of structural proteins. For example knockout mice have shown that activation of ERK pathway due to activity dependent phosphorylation is necessary for late LTP in the hippocampus (Kelleher et al., 2004) and models have shown that it is plausible that persistent CAMKII autophosphorylation could act as a molecular switch at the synapse (Hayer and Bhalla, 2005). Thus in addition to structural changes, persistent phosphorylation might change the biochemical conditions of synapses for long periods.

To model the role of an additional stabalisation process such as phosphorylation, the 8 state model consists of two rings such as those in the four state model, Fig. 6.1B that are joined at the vertices. One ring, consisting of states $\{1,2,3,4\}$ has transition probabilities set such that synapses in that ring tend to occupy state 1 (the empty state). This is the depressed ring. Synapses in the depressed ring lack activation of some underlying persistent phosphorylation (the switch variable is not set). The effect of this is that synapses tend to be empty of AMPAR due to a low rate of production/insertion of AMPAR or PSD slots, or perhaps due to a reduction in the strength of interaction between AMPARs and the PSD (Kim et al., 2007).

The other ring, consisting of states $\{5,6,7,8\}$ has its transition probabilities set such that synapses tend to occupy state 7. This is the potentiated ring. In the potentiated ring, phosphorylation leads to the activation of biochemical processes such that the quantity of associated slots and AMPARs in the synapse is upregulated. Hence the synapses tend to be potentiated.

At baseline the transitions between the vertices of these rings are made small. Thus

synapses within each ring remain within that ring for a long period of time. In the 8 state model, late LTP is a process whereby a bistable biochemical switch is flipped leading to the alteration of transition rates such that synapses are potentiated by the addition of stable slots and AMPARs.

Solving the eight state model in Fig. 6.1C can be achieved in the case that we specify two four state rings as above but join their vertices with allowed transitions. However, as we shall see, this yields a model that is not suitable. Justifications for the eight state model are thus forced to be somewhat more heuristic since solutions representing dynamics of interest in this study cannot be compactly dealt with analytically.

6.2.4.1 Detailed balance

In order to ensure that detailed balance applies in the 8 state model it is necessary to enforce it for every possible cycle within the state diagram. In a state diagram such as Fig. 6.1C this can be achieved by adjusting one transition probability from five of the six cycles surrounding each face of the cube, using an equation such as Eq.(3.25) (Colquhoun et al., 2004). We first choose transition probabilities for the depressed ring and set detailed balance in that ring according to the transition probabilities $\{M_{12}, M_{21}, M_{23}, M_{32}, M_{34}, M_{43}, M_{41}\}$ with,

$$M_{14} = M_{41}\frac{M_{12}M_{23}M_{34}}{M_{21}M_{32}M_{43}}. \tag{6.11}$$

Transition probabilities in the potentiated ring are determined by transition probabilities in the depressed ring because we are assuming symmetry of plasticity as stated previously,

$$\begin{aligned} M_{78} &= M_{12} & M_{87} &= M_{21} \\ M_{85} &= M_{23} & M_{58} &= M_{32} \\ M_{56} &= M_{34} & M_{65} &= M_{43} \\ M_{67} &= M_{41} & M_{76} &= M_{41}\tfrac{M_{12}M_{23}M_{34}}{M_{21}M_{32}M_{43}}. \end{aligned} \tag{6.12}$$

Next, impose detailed balance upon top ring in Fig. 6.1C choosing the transition $2 \rightarrow 6$,

$$M_{26} = M_{62}\frac{M_{23}M_{37}M_{14}}{M_{32}M_{73}M_{41}} \tag{6.13}$$

where M_{62} is a free probability. Next the left most ring in Fig. 6.1C is balanced,

$$M_{15} = M_{51}\frac{M_{12}M_{26}M_{65}}{M_{21}M_{62}M_{56}} \tag{6.14}$$

where M_{51} is another free probability. Finally the bottom ring is balanced,

$$M_{48} = M_{84}\frac{M_{15}M_{41}M_{32}}{M_{51}M_{14}M_{23}} \tag{6.15}$$

where M_{84} is a free probability. Calculation verifies that the right most ring in Fig. 6.1C now also obeys detailed balance as would be expected for the final ring of an 8 state cube model (Colquhoun et al., 2004).

The following assumptions are made:

1. In order that the magnitude of late LTP and late LTD be symmetric, an equal number of synapses should occupy the depressed ring as occupy the potentiated ring when the ensemble is at equilibrium. Thus the transition probabilities from the depressed ring to the potentiated ring should be equal to the transition probabilities from the potentiated ring to the depressed ring.

2. The transition probabilities between the depressed and potentiated rings should all be individually equal to one constant value at equilibrium. If we permit unequal transition probabilities between rings, we typically find that synapses within the depressed ring at steady state do not tend to occupy the empty state or that synapses within the potentiated ring at steady state do not tend to occupy the anchored state. In other words the coupling between the rings changes the desired steady state within each ring. This disrupts the model.

Requirement 1 implies that

$$\begin{aligned} M_{15} &= M_{51} \\ M_{26} &= M_{62} \\ M_{37} &= M_{73} \\ M_{48} &= M_{84} \end{aligned} \tag{6.16}$$

and so if Eqs. (6.16) are imposed, requirement 2 follows trivially if we set $M_{51} = M_{62} = M_{73} = M_{84}$. From 6.16 it follows that,

$$\frac{M_{12}M_{65}}{M_{21}M_{56}} = \frac{M_{23}M_{14}}{M_{32}M_{41}} = 1 \tag{6.17}$$

reducing to

$$\frac{M_{12}M_{43}}{M_{21}M_{34}} = \frac{M_{23}M_{14}}{M_{32}M_{41}} = 1 \tag{6.18}$$

by substitution from Eqs. (6.12). Eq. (6.18) is identical to the balance condition for the four state ring model as described in the previous section. Thus we find (as we would intuitively expect) that two four state rings as specified previously that are joined at each vertex by transitions that are equally likely in either direction automatically obey detailed balance. In this case both the depressed and potentiated rings behave in an identical way to the four state model. Half of the synapses occupying each ring are

potentiated, in states $\{2,3,6,7\}$ or depressed, in states $\{1,4,5,8\}$ and there is no difference in mean weight between the rings. Consequently, although the small transition probabilities separating the rings potentially introduce another decay timescale in to the model, this timescale is not strongly expressed in the dynamics of the synaptic weight and so cannot be a timescale of early or late LTP/D. Thus, this form of the eight state model would add nothing over the 4 state model and so detailed balance cannot be imposed upon a suitable model obeying requirements 1 and 2.

6.2.4.2 Transition probabilities in the 8 state model

In order to create an eight state model that we can fit well to the data, and that displays an additional timescale of decay we must set the transition probabilities in a more heuristic fashion, described below. Unfortunately this violates detailed balance. In practice however, we find that as long as detailed balance holds in the depressed and potentiated rings individually and that as long as the transition probabilities between the rings Eq.(6.16) are kept small, then the real part of the solution in terms of eigenfunctions matches the numerical solution extremely closely. Thus in effect the system remains solvable in terms of eigenvectors. This can be understood intuitively by realising that as long as the transition probabilities between the rings are small, then each of the joined rings is a weakly perturbed version of an identical single ring. Since the single ring obeys detailed balance and can be solved with the method of eigenvectors, then so too the joined ring should be approximated by the single ring solutions.

The eight state model is adjusted such that the behavior of the model matches the electrophysiological probes of LTP/D. Qualitatively this can be done by altering the balance of flux into and out of each state. For a quantitative match it is then necessary to be able to vary the magnitudes of the transition rates. Unfortunately there are many possible configurations of parameter values in the eight state model. One approach is to attempt to vary the transition rates using only a small number of parameters. This can be achieved by introducing a scaling law determining the scaling between the transition probabilities.

Intuitively it seems likely that if one were to measure the reaction rates of biochemical processes in the cell one could plot a graph of the transition rates versus the rank order of their magnitudes, i.e. fastest rates to slowest rates. The form of that graph determines how reactions that occur frequently relate to reactions that occur infrequently: This is the scaling law of the transition rates. For example, the scaling law could be linear, in which case as we look at reactions in turn from the most frequently occurring

to the most infrequently occurring, we would find that the transition rates are related to their rank by some linearly decreasing function. Searching the literature reveals no data regarding what this function should be, but one possibility is that it should be a power law. Another possibility is that it might be exponential. Both possibilities were tested and it was found that a power law gave a better fit to data. For the purposes of the 8 state model, we imagine that we take four rates from this powerlaw scaling, from most frequently occurring to least frequently occurring $\{1,2,3,4\}$. The rates are,

$$R_{ij} = \alpha x^{-\gamma} \tag{6.19}$$

for $x \in \{1,2,3,4\}$. The scaling parameters are α and γ. Elements of R consequently adopt one of 4 values $\{a,b,c,\varepsilon\}$, Fig. 6.4. The magnitude of the transition rates can be adjusted by altering α while their spacing can be adjusted by altering γ, Fig. 6.3. In fitting the model we are interested primarily in the the spacing between the timescales of early and late LTP/D and in their absolute values. The purpose of introducing the power law scaling is to enable us to vary both the absolute size and relative spacing of the transition probabilities by varying only two parameters.

Fig. 6.3A demonstrates that by increasing the power of the scaling law the timescales in the weight dynamics become more separated. The absolute values of the timescales also increase greatly, growing approximately exponentially for this range of γ. The fixed points for the depression and potentiation rings are plotted in Fig. 6.3C, the upper curve shows the fixed point weight of the depression ring as γ is increased and the lower curve shows the fixed point weight of the potentiation ring as γ is increased. The sum of these two fixed points is the mean weight for the whole ensemble and this always remains constant at $1/2$. The significance of the fixed point for each separate ring is that this is the value of weight at which a group of synapses would sit if they were all transferred in to either ring. Thus as γ increases and these two fixed points diverge, the ensemble of synapses is becoming more polarised in to a depressed and a potentiated group[5].

To match the timescales of decay of late and early LTP/LTD seen in the experimental data, a power law of $\alpha = 0.01s^{-1}$, $\gamma = 6.1$ is taken, Fig. 6.3A. Fig. 6.4 contains

[5]This may seem a puzzling statement given that the synapses in this model are binary. However the synapses can become more polarised in terms of the amount of time they spend occupying the stable states. In a population of synapses where the fixed points of the depressed and potentiated rings were identical, synapses would spend equal amounts of time in equivalent states in each ring. If however the stable states in each ring are very different then synapses in each ring will spend significant amounts of time in non-equivalent states, for example depressed states in one ring or potentiated states in the other. Thus a temporal average would reveal significant polarity between rings.

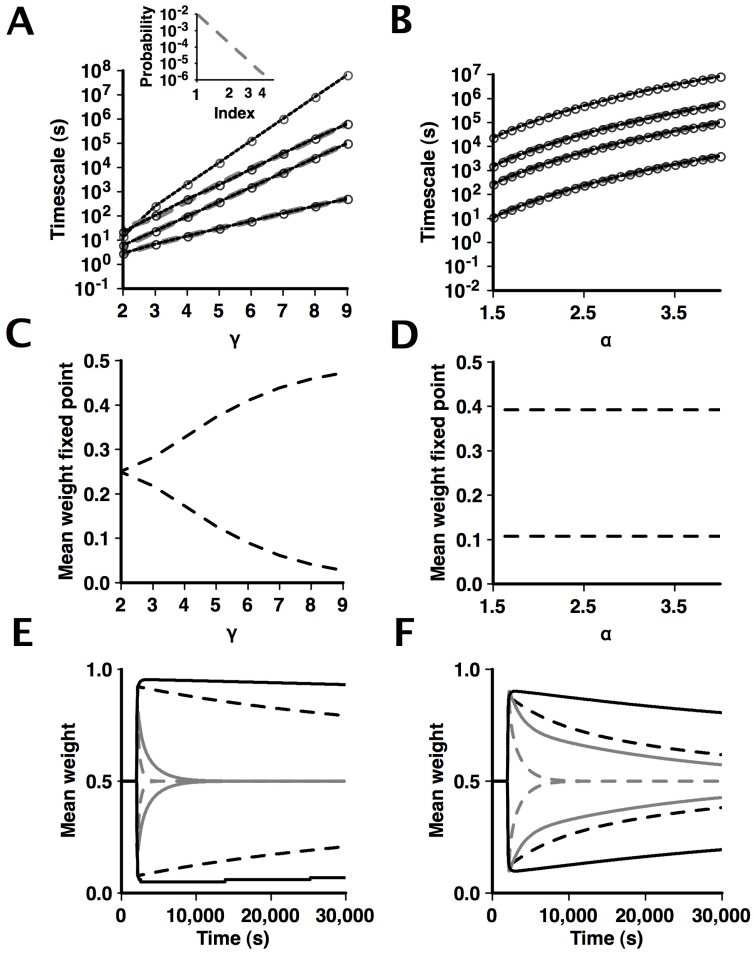

Figure 6.3: The effect of changing the transition rates of the 8 state model according to a scaling law Eq. (6.19). A: The timescales ($1/\lambda_k$) of the equilibrium rate matrix plotted as a function of the power of the scaling law on a logarithmic scale. Although there are 7 eigenvalues, 3 of them are degenerate (grey dashed lines) due to symmetries in the transition matrix. Increasing the power of the scaling law has the effect of causing the timescales to grow nearly exponentially and to become more spaced apart. The powerlaw chosen in this chapter, where $\alpha = 0.01s^{-1}$ and $\gamma = 6.1$ is shown on a log-log plot (inset). B: The timescales of the baseline matrix plotted as a function of the maximum rate, α. Alteration of α allows all of the timescales to be varied without changing their relative spacing. C: The effect of changing the power of the scaling law, γ upon the mean steady state weight of each ring where the bottom line is the depressed ring and the top line is the potentiated ring. Note that the sum of their average values is always 0.5, so the mean of the whole ensemble remains constant at 0.5. Increasing the power of the scaling law increases the polarity of the synapses in the ensemble. D: Altering the minimum rate has no effect upon the fixed points of the depression and potentiation rings. E: Late LTP/D (solid lines) and early LTP/D (dashed lines) for scaling laws of power $\gamma = 8$ (black lines) and of power $\gamma = 4$ (grey lines). F: Late LTP/D (solid lines) and early LTP/D (dashed lines) for scaling laws with minimum rate $\alpha = 0.04s^{-1}$ (black lines) and $\alpha = 0.08s^{-1}$ (grey lines).

a visualisation of the baseline rate matrix of the eight state model. The entries in the matrix marked β are set with Eqs.(6.11+6.12) so that detailed balance holds within the depressed and potentiated rings. Since detailed balance is not set in every ring however, the model does not obey detailed balance overall but in practice this does not matter for small transition rates between rings.

6.2.4.3 Applying plasticity protocols

In addition to the baseline rate matrix for the eight state model R, four other matrices are defined, each one allowing the model to approximate an experimental protocol. The protocol matrices are; early LTP (E_P), early LTD (E_D), late LTP (L_P) and late LTD (L_D). Each of the protocol matrices is identical to the baseline rate matrix, Fig. 6.4, but for the modification of certain transition rates such that experimentally observed plasticity induction times can be matched, Fig. 6.2C.

In the case of the $\{E_P, E_D, L_P, L_D\}$ matrices, the precise transition rates that are altered are analogous to the four state model, but for the fact that late phase LTP and LTD increases the transition rates $3 \rightarrow 7$ and $5 \rightarrow 1$ in L_P and L_D respectively, Fig. 6.2C. The timescales for induction of early and late LTP/D are taken to be the same values as stated previously. The modifications to the baseline transitions are listed for each protocol matrix: (in all cases, detailed balance is set by adjusting transitions $1 \rightarrow 4$ and $7 \rightarrow 6$ according to Eqs. (6.11+6.12).)

- E_P: $R_{12}^{eLTP} = 0.0055s^{-1}$, $R_{21}^{eLTP} = 1 \times 10^{-7}s^{-1}$, $R_{23}^{eLTP} = 1 \times 10^{-4}s^{-1}$
- E_D: $R_{78}^{eLTD} = 0.0055s^{-1}$, $R_{67}^{eLTD} = R_{87}^{eLTD} = 1 \times 10^{-7}s^{-1}$, $R_{85}^{eLTD} = 1 \times 10^{-4}s^{-1}$
- L_P : $R_{12}^{lLTP} = 0.0055s^{-1}$,$R_{21}^{lLTP} = R_{41}^{lLTP} = 1 \times 10^{-7}s^{-1}$,$R_{23}^{lLTP} = R_{37}^{lLTP} = 0.1s^{-1}$
- L_D: $R_{78}^{lLTD} = 0.0055s^{-1}$, $R_{87}^{lLTD} = R_{67}^{lLTD} = 1 \times 10^{-7}s^{-1}$, $R_{85}^{lLTD} = R_{51}^{lLTD} = 0.1s^{-1}$

Plasticity protocols are applied to the model as described in §6.2.1: The model begins in the equilibrium state of the baseline rate matrix. At the beginning of the protocol, the baseline rate matrix is instantaneously substituted for one of the protocol matrices $\{E_P, E_D, L_P, L_D\}$. The system now moves toward the steady state of the protocol matrix, Fig. 6.2C. Since the protocol matrices are constructed so as to give rise to steady states that have a very different synaptic weight, this causes a deflection in the mean weight of the ensemble (i.e. depression or potentiation). After the duration of the protocol, the baseline matrix is instantaneously substituted back in place of the protocol

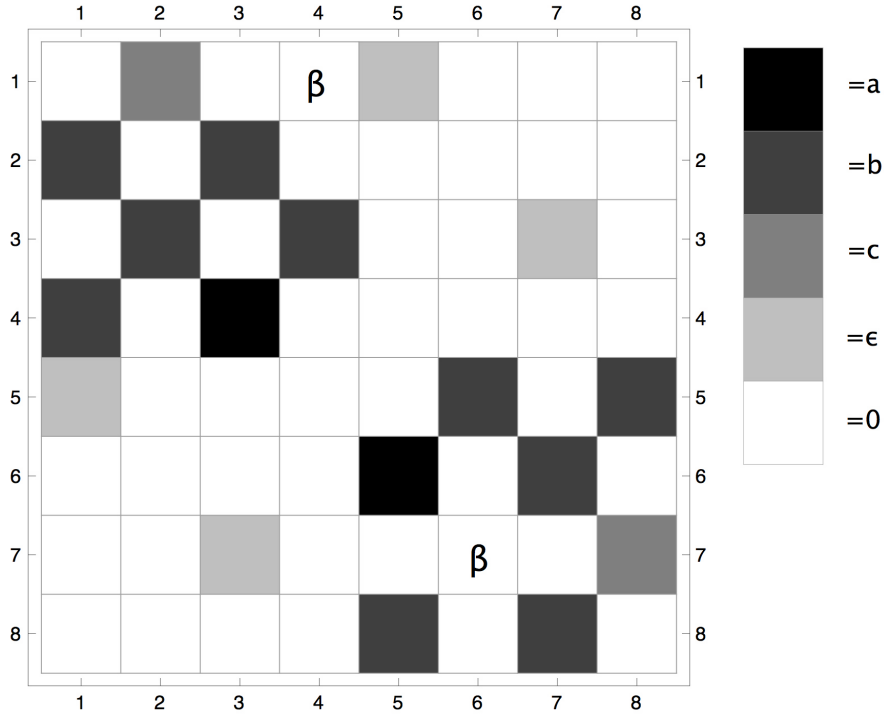

Figure 6.4: The baseline transition rates for the 8 state model. Each value in the matrix is the transition rate from the state on the horizontal scale to the state on the vertical scale. Each value in the rate matrix is one of $\{a, b, c, \varepsilon, 0\}$ chosen from a powerlaw (see text). All elements have been assigned a greyscale value. Black denotes a, while white denotes 0. Diagonal elements have been set to 0 for clarity although they are defined according to Eq. (3.9). The transitions marked β are set such that detailed balance holds within the depressed and potentiated rings.

matrix, Fig. 6.1. At this point the system is not in a steady state of the baseline matrix and it relaxes back to that steady state (i.e. this is the decay phase of potentiation or depression). The relaxation occurs at some timescale that is dependent upon where the protocol drove the synaptic ensemble to in the state diagram, i.e. early or late LTP can result depending upon the induction protocol.

6.3 Synaptic dynamics and the memory trace

Early LTP can modeled in the 2, 4 (black dashed line, identical solution for 2 and 4 state models) and 8 state (solid black line) models by substitution of transition probabilities

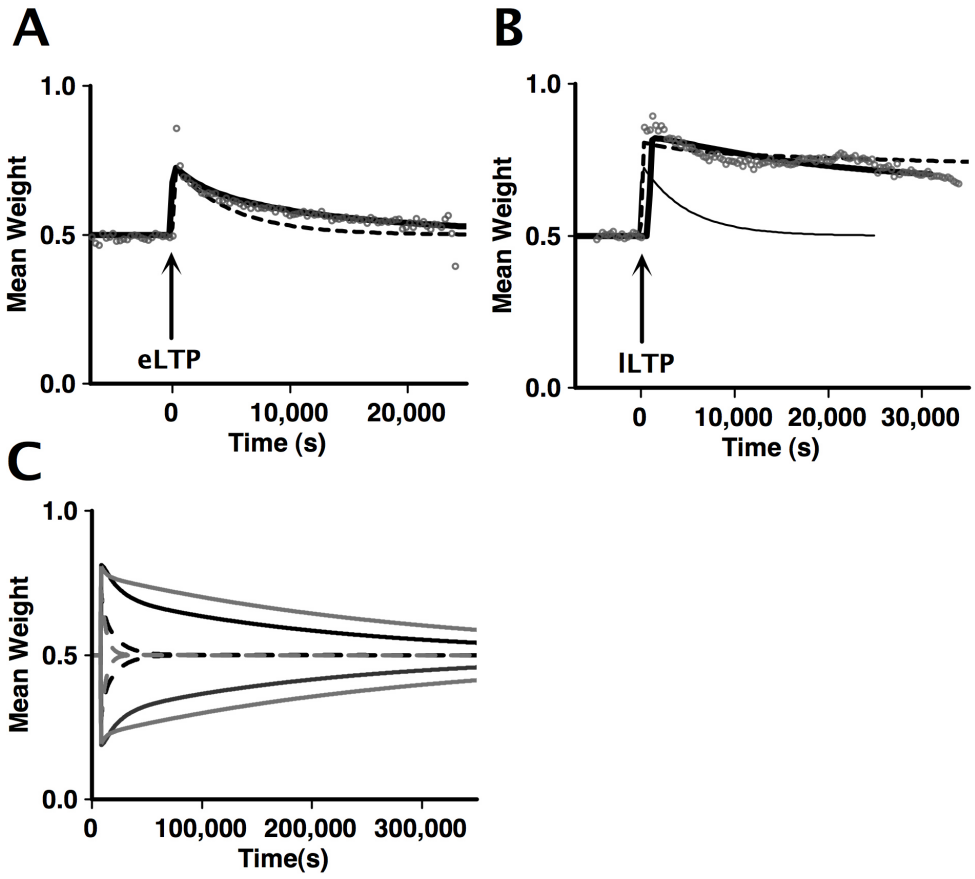

Figure 6.5: Early and late phase LTP/LTD in state based models. A: 2,4 and 8 state models fitted to experimental data from Roger Redondo (grey circles). The experimental data is an average EPSP slope from synapses in CA1 of acute hippocampal slice preparations in which early LTP has been elicited. The experimental data has been linearly rescaled such that it ranges from a baseline of 0.5 to 1 (rather than 100% to 200%) for direct comparison to the models. The 2 and 4 state models give rise to early LTP curves that completely overlap (black dashed line). The 8 state model is the thick black line. B: Same as A but for slices in which late LTP has been elicited. The 2 state model is plotted with a thin black line for comparison but cannot incorporate late LTP. C: LTP and LTD in the 2, 4 and 8 state models on an extended timescale. In this chapter it is assumed that LTD is precisely symmetric to LTP. Both early (dashed lines) and late (solid line) LTP and LTD is plotted for the 4 state (grey lines) and 8 state (black lines) models. Early LTP in the 2 state model overlaps precisely with that in the 4 state model (grey dashed line).

as previously described for a duration of 150s, Fig. 6.5A. Also plotted is experimental data from Roger Redondo showing the average of 12 records of the EPSP slope of a population of synapses in CA1 over the course of early LTP induction. The experimental data has been rescaled for comparison to the output of the models. Late LTP can be modeled in the 4 and 8 state models, Fig. 6.5B, where the experimental data (grey circles) is an average of 6 induction trials. In Fig. 6.5C early and late LTP/D is plotted on a longer timescale. The 2 state model only accommodates one timescale (which here we have set to be the timescale of eLTP). Both the 4 and 8 state models can represent early and late phase LTP in a manner consistent with the data.

6.3.1 Depotentiation of early LTP

Depotentiation is the name given to the observation that recently induced synaptic potentiation can be removed by application of low frequency stimulation soon after initial induction. Crucially, low frequency stimulation is of a magnitude such that it would have little or no effect on a synapse that has not been recently modified. Typically low frequency stimulation consists of a stimulus at 2Hz for 250s-500s (Bashir and Collingridge, 1994; Sajikumar et al., 2005). Depotentiation of early LTP has been observed in hippocampal slice preparations (Sajaykumar and Frey, 2004). Fig. 6.6A is a reproduction of depotentiation of early LTP in CA1 of Hippocampal slices from Sajaykumar & Frey 2004. We see that early LTP has a 'memory'. If LFS is applied soon after the induction of early LTP, 5 mins in this case, then the potentiation is obliterated. If however the gap is longer, for example 15 mins, then the initial potentiation recovers transiently before decaying away. Depotentiation of late LTP has also been observed and operates in much the same way (Bashir and Collingridge, 1994; Staubli and Chun, 1996; Woo and Nguyen, 2002).

Depotentiation only occurs in synapses that have undergone recent modification. In the case of early LTP this implies that depotentiation removes excess AMPAR (relative to the steady state), that were added by the early LTP induction protocol. In the 4 and 8 state models this corresponds to causing synapses that occupy state 2 (AMPAR but no slot), to decay rapidly back to state 1 (no AMPAR, no slot). This AMPAR removal can be achieved by increasing the transition rates from state 2 to 1. Note that if no previous early LTP induction has been applied, then increasing the rate of transition $2 \rightarrow 1$ has a negligible effect on the synaptic weight because at the steady state there is only a small number of synapses in state 2.

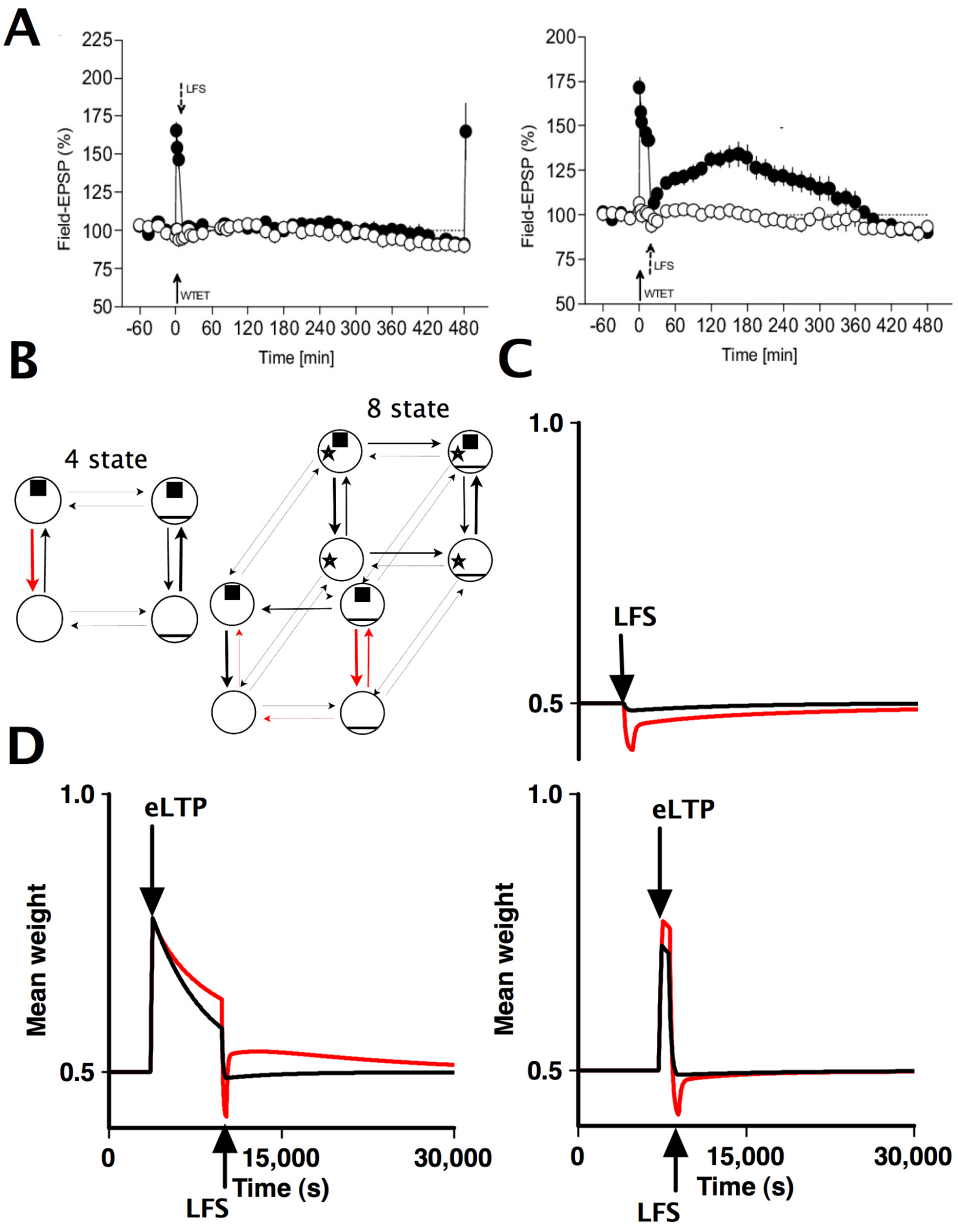

Figure 6.6: Depotentiation of early LTP. A: Reproduction of Fig. 1 b+d from Sajaykumar & Frey (Sajaykumar and Frey, 2004). In the left most graph a synaptic pathway connecting with synapses in CA1 of the hippocampal slice is first potentiated with an early LTP inducing protocol (WTET) and then subjected to low frequency stimulation (LFS) 5 mins later (black filled circles). The control pathway is also shown (open circles). In the rightmost graph the experiment is repeated except LFS is now delivered 15mins after the induction of early LTP. B: The state diagrams illustrate the LFS protocol in the 4 and 8 state models. C: LFS delivered to naive synapses that have not experienced eLTP has little effect on the synaptic weight in the 8 state model (red) or 4 state model (black) (top panel). Early LTP can be obliterated by LFS in the 8 state (red) and 4 state (black) models when the LFS follows only 300s after the LTP (bottom panel). D: However if early LTP is followed by LFS 6000s (100 mins) later there is a partial recovery of the early LTP in the 8 state model (red line) but not in the 4 state model (black line).

In the 4 state model, AMPAR removal is modelled with a low frequency stimulation (LFS) protocol matrix. This opposes early LTP by increasing R_{21} only, Fig. 6.6B. This is not the same as early LTD because early LTD increases the rate of $3 \rightarrow 4$, and $2 \rightarrow 1$. To perform LFS in the 4 state model set $R_{21}^{LFS} = 0.01s^{-1}$.

In the 8 state model an LFS matrix, F, is defined. In F the transition rates of the transition $2 \rightarrow 1$ and $3 \rightarrow 4$ are elevated, Fig. 6.6B. Thus set $R_{21}^{LFS} = R_{34}^{LFS} = a$, $R_{43}^{LFS} = b$ and $R_{41}^{LFS} = \varepsilon$, Fig. 6.6B. LFS in the 8 state model thus forces all synapses in the $w = 1$ states 2 and 3 in the front (depressed) ring, Fig. 6.1, into states 1 and 2 having $w = 0$. Note that analogously with the 4 state case, LFS in the 8 state model is not the same as early LTD. Early LTD increases the probability of $7 \rightarrow 8$ in the potentiated ring. LFS increases the probability of $2 \rightarrow 1$ and $3 \rightarrow 4$ in the depressed ring. LFS was not quantitatively matched to data. For this reason the rates in F were sampled from the same set of probabilities composing R rather than being matched to experimentally observed rates.

Upon substitution of the LFS protocol matrix for the baseline matrix, depotentiation of early LTP can be elicited in the 8 state model, but not in the four state model, Fig. 6.6D. Firstly we verify that LFS does not greatly effect the synaptic weight of an ensemble when applied to the 8 state model when it has not previously undergone eLTP (i.e. is in the steady state), Fig. 6.6C. When LFS is similarly applied to the naive 4 state model the weight deflection is even smaller, only around 1%.

In both the 8 state and the 4 state models, application of LFS rapidly, 300s after eLTP obliterates the potentiation, Fig. 6.6C. However if the gap is more significant at 6000s then the 8 state model gives rise to a partial recovery of the initial potentiation that does not occur regardless of the time gap in the 4 state model. The four state model is not able to accommodate the late and partial recovery of the synaptic weight after depotentiation of early LTP.

Depotentiation of early LTP in the 8 state model occurs due to the presence of two separate potentiation pathways. One of these pathways is the addition of AMPAR only (transition $1 \rightarrow 2$). The other pathway is the association of slots with AMPAR (transition $4 \rightarrow 3$). The LFS protocol in the 8 state model has no effect upon synapses that are protected by the phosphorylation process (i.e. synapses that occupy state 7 in the second ring). Thus LFS has little effect upon the weight of synapses that have not recently undergone LTP.

In 8 state synapses that have undergone recent early LTP however, there is a surplus of receptors associated with slot proteins (in state 3) provided that sufficient time has

elapsed after application of the eLTP protocol and synapses have had time to diffuse into state 3. In this case the LFS protocol does significantly reduce the synaptic weight. There is a subsequent rebound because at baseline there is a tendency for receptors to reassociate with slots (i.e. flow of states is from state 4 to state 3). But on a longer timescale these receptors are depleted by synapses flowing back in to state 1 (i.e. the overall flow is from state 4 to state 1, but this flow is slower than that from state 4 to state 3). In contrast, when early LTP is induced in the 4 state model there is only one state transition involved, the addition of AMPAR via $1 \rightarrow 2$. Reversal of this single process can only lead to the permanent removal of the potentiation regardless of the time elapse.

6.3.2 The memory trace lifetime

We have seen that remarkably simple models that are designed to exhibit multi-timescale behavior can reflect early LTP, early LTD and depotentiation of early LTP (in the 8 state case). Having constrained the models against experimentally observed synaptic dynamics we now use them to estimate the lifetime of a memory trace within the synaptic population from which the data was gathered, assuming that the synapses occupy the steady state.

Fig. 6.7A shows the steady state distributions across the synaptic states of the three models in this chapter. In the steady state, synapses fluctuate between the states (analogously to the STDP weights in chapter 4). We can calculate the autocorrelation of the synaptic weight of the ensemble for each model with Eq.(3.31). As expected, the autocorrelation of the 2 state model is a single exponential with a timescale identical to that of eLTP as extracted from the data (4800s) grey line Fig. 6.7B. For our choice of parameters, based upon matching the experimentally observed $\langle w(t) \rangle$ decay timescales, the autocorrelation of the 8 state model falls more rapidly initially than the autocorrelation of the 4 state model (solid and dashed black lines respectively). Thus the signal to noise ratio of the 8 state model is always below the signal to noise ratio of the 4 state model for an identical initial signal to noise.

Ideally we should like a high initial strength and a slow timecourse of decay. To quantify this we calculate the initial signal to noise ratio with Eq. (3.37) (see chapter 3). Since the synapses are binary, the variance in the number of synapses that are potentiated at any one time is $\sigma^2 = \Omega P_{w=1}(1 - P_{w=1})$, where $P_{w=1}$ is the total probability that a synapse is a potentiated state once the system has reached the asymptotic steady

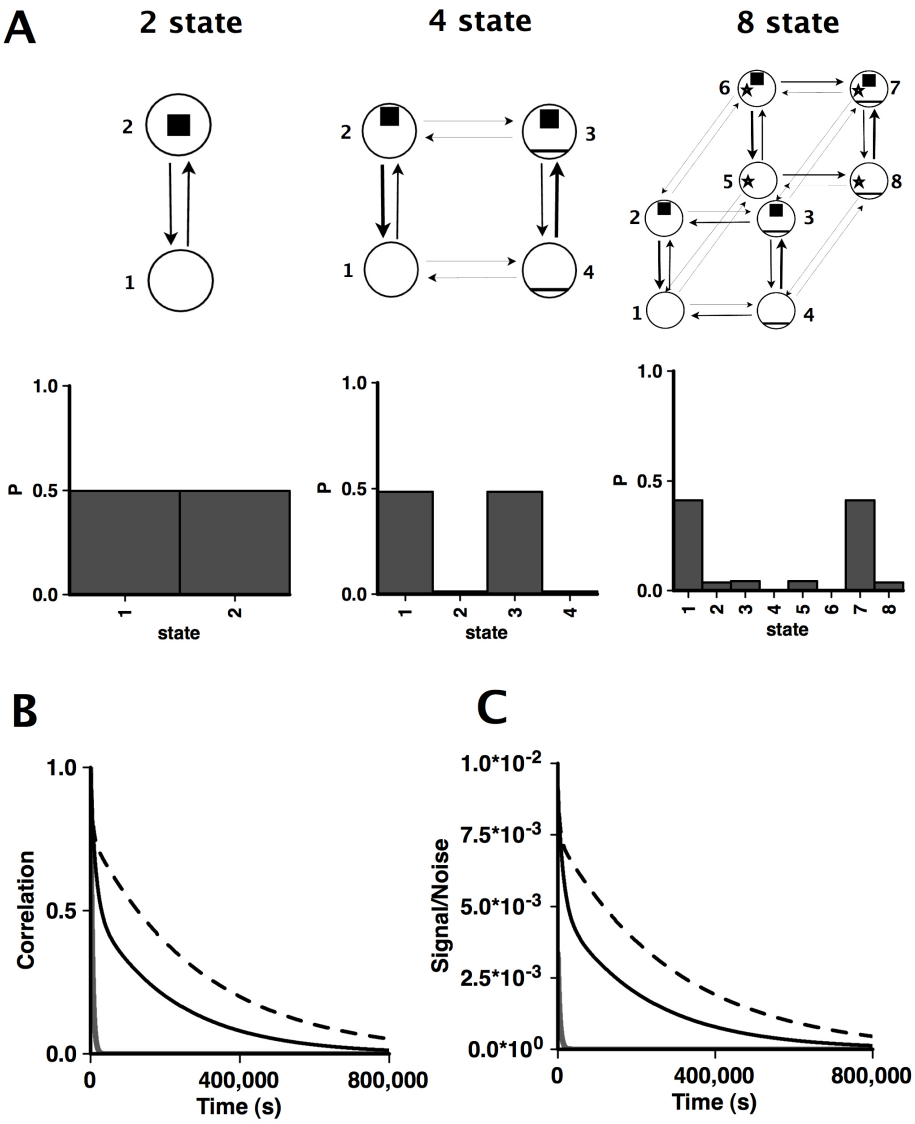

Figure 6.7: State based models of LTP in the steady state. A: The steady state distributions across the states for the 2, 4 and 8 state models. B: The steady state autocorrelation functions for the 2 (grey line), 4 (dashed line) and 8 state (solid line) models. C: The signal to noise ratio of the memory trace for the 2, 4 and 8 state models using 30,000 synapses

state. It is assumed that pattern storage does not alter the variance of the weights. As discussed in chapter 3, the SNR is the autocorrelation scaled by the initial signal to noise ratio, calculated for one cell based upon 30,000 excitatory synapses in Fig. 6.7C.

There are around 4.6×10^6 cells in CA1 of humans (Rolls, 1996). Again, assuming $30,000$ excitatory synapses per cell and if we permit all of the cells in the CA1 region to participate in the ensemble then $\Omega = 3.8 \times 10^{11}$. In this case the initial signal to noise ratio is 32.5. Conventionally a signal is assumed to be undetectable when the SNR reaches 1. This occurs after 267 hours for the 4 state model. For the 8 state model, the signal to noise ratio falls to 1 after around 167 hours. In contrast it takes only 6.9 hours for the SNR to reach 1 in the 2 state model.

In chapter 2 we discussed the antagonism between initial SNR and the memory decay timescale engendered by the plasticity stability dilemma. This trade off also applies to these models. Eq. (6.10) derived in §6.2, shows that the signal can be boosted in the 4 state model while keeping the timescale constant. Plotting the quadratic dependence of the initial signal on the equilibrium state of the 4 state model (i.e. the occupancy of p_2) reveals that the initial signal is optimised when the occupancy of p_2 is 0.25. This implies that when the ring is in the diffuse state (i.e. all 4 states are equally likely), then the initial signal is at its highest possible value, while the decay timescale is still 4800s, Fig. 6.8C. However there is a cost: The diffuse state amounts to removing the distinction between early LTP and late LTP and reduces all timescales to a single degenerate timescale of 4800s. In this case the autocorrelation function decays at the early LTP timescale alone, Fig. 6.8E and the four state model looses its extended memory trace lifetime in comparison to the binary model. Further to this, the 4 state model would no longer match experimental observations by exhibiting both early LTP and late LTP.

The same considerations apply to the 8 state model. Decreasing the decay timescales leads to an increase in the initial signal (although at the expense of loss of agreement with the data). Since the eigenvectors of the 8 state model are independent of α, the steady state can be held constant while the timescale is varied by altering α, Fig. 6.8B. All four distinct timescales are affected identically when α is scaled. Unfortunately it is not possible to analytically find combinations of α and γ that preserve the decay timescales of the 8 state model while simultaneously altering the steady state. Attempts at this result in transcendental equations. However the initial signal to noise can be calculated as the scaling parameter γ is varied, Fig. 6.8D. In this case both the decay timescales and the steady state are simultaneously altered. We find that the

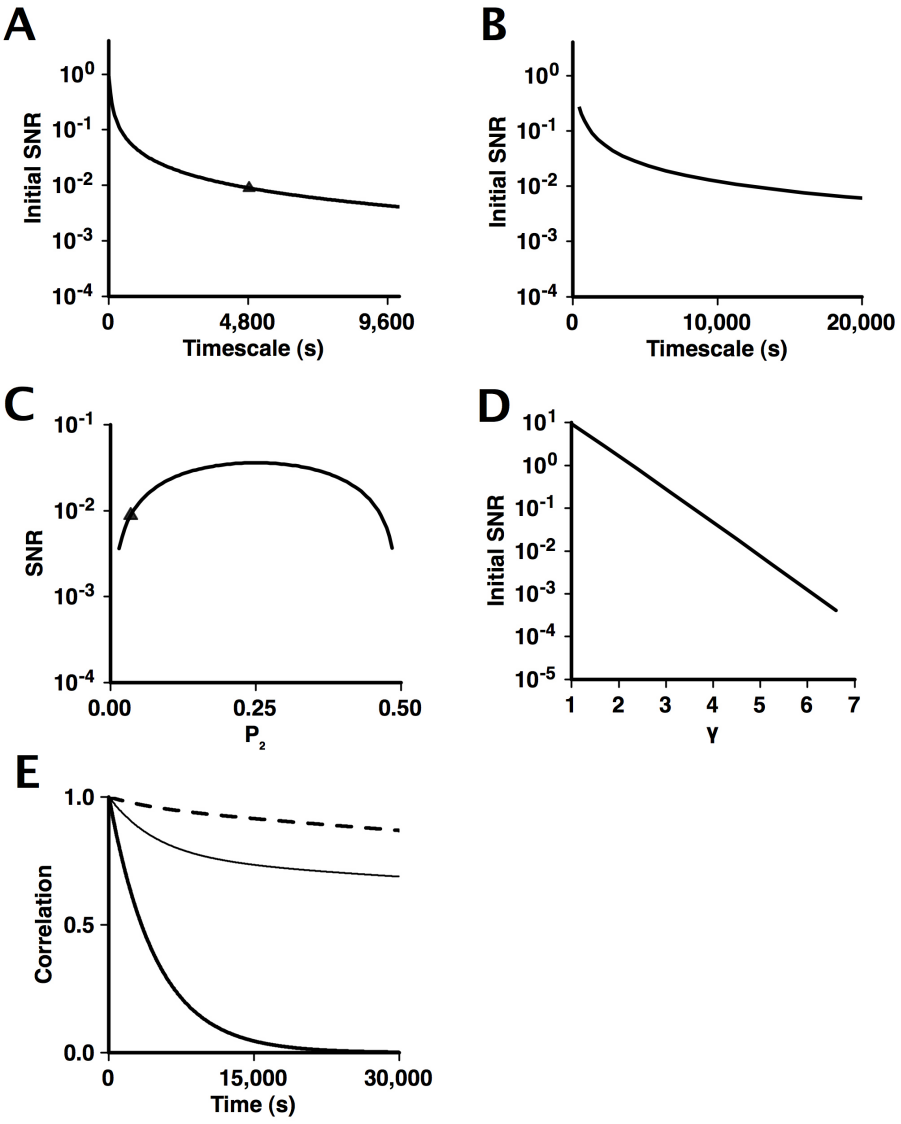

Figure 6.8: Effects of altering the decay timescale and steady state on the initial SNR. A: 4 state model: The early LTP timescale is varied as the steady state distribution is held constant. The triangle marks the initial signal to noise ratio for the 8 state model as in Fig. 6.7. This is identical to the initial signal of the 8 state model when the early LTP timescale is 4800. B: 8 state model: Initial signal to noise ratio as a function of one of the four separate decay timescales. To vary the timescale, the powerlaw is adjusted so as to preserve the steady state (i.e. α is varied). C: The initial signal to noise ratio as a function of the steady state of the four state model as indicated by the probability of occupancy of state 2, p_2, see (Eq. 6.10). The triangle is the initial signal of the 4 state model D: Initial signal to noise ratio of the 8 state model as γ is varied. E: The autocorrelation of the four state model when the steady state is highly polarised (i.e. $p_2 \approx 0$) (dashed line) and when the steady state is diffuse ($p_2 \approx 0.25$) (solid line). The thin line is the autocorrelation of the four state model when it is matched to experimental data.

initial signal to noise ratio of the 8 state model falls exponentially as γ is increased. In contrast we saw in Fig. 6.3A that the timescales of decay of the memory trace increase exponentially as γ is increased. Thus, as we have come to expect from the plasticity stability dilemma, in the 8 state model too, there is direct antagonism between the initial signal and the memory lifetime.

6.4 Discussion

In this chapter, two multitimescale models of long term potentiation were proposed and compared to the simple single timescale 2 state model. Assuming that synaptic plasticity as elicited in experimental hippocampal slice preparations can be considered as piecewise homogeneous, the two state model is unable to account for even the most basic observations relating to LTP: Namely the existence of early and late LTP. The addition of biologically motivated hidden states in the 4 and 8 state models allows early LTP and late LTP to be accommodated within the framework of the piecewise homogenous Markov process. These models were adjusted so as to be compatible with the experimentally observed dynamics of an ensemble of weights in CA1 of hippocampal slice preparations. Once the timescales of the models had been thus constrained the signal to noise ratio of a memory trace stored at equilibrium within the synaptic ensemble was found for each model. The 8 state model incorporates both timescales of addition and removal of AMPAR and slots and a timescale associated with a bistable phosphorylation switch. The 4 state model incorporates only timescales associated with the addition and removal of AMPAR and slot proteins. It was found that, due to the presence of more timescales, the 8 state model can qualitatively account for the depotentiation of early LTP. However the 4 state model provides a longer time for the signal to noise to reach 1, and thus performs better as a memory store.

Here the approach was to formulate a state based model of LTP based upon experimental observations and then to infer the memory trace lifetime in those synapses assuming that the dynamics observed experimentally are an accurate reflection of the in vivo synaptic timescales (this being the widespread assumption). Previous work has been more abstract that this, with little attempt being made to link the models to the dynamics of synapses as observed electrophysiologicaly (Amit and Tsodyks, 1992; Amit and Fusi, 1994; Fusi, Drew, and Abbott, 2005; Senn and Fusi, 2005; Fusi and Senn, 2006; Fusi and Abbott, 2007; Ben Dayan Rubin and Fusi, 2007; Leibold and Kempter, 2008). Furthermore, no previous state based model of the mem-

ory trace[6] has been of the class of models examined in this chapter, i.e. models containing rings of bidirectional state transitions.

It was assumed that synaptic plasticity as elicited in slice experiments can be described as a piecewise Markov process. This implies that the synapses are Markovian, that plasticity protocols lead to discrete alterations to the transition rates of biochemical processes and that such discrete transitions can be made in a near instantaneous manner. If one were to watch a synapse within a brain as its owner moved around the world, then there is reason to doubt that it would appear Markovian, because we could not hope to have knowledge of all variables influencing its evolution. For the models here however, that consider synapses in the far more controlled conditions of the laboratory, it appears that it is reasonable to assume that synapses are Markovian because plasticity protocols, when applied with the correct controls, give reproducible behavior.

It is unlikely that the application of plasticity protocols gives rise to instantaneous discrete changes to the values of rate constants of processes within cells. Indeed it is known that LTP is composed of a gradual cascade of biochemical processes that can still be ongoing days or weeks after the initial plasticity event (Abraham and Williams, 2003). Nevertheless at the level of the synaptic weight it seems that the observed dynamics can be adequately explained with piecewise changes to the rate constants. This implies that processes having a direct impact upon the synaptic weight (rather than, say, an ultrastructural role) do indeed vary on a timescale that is short compared to the decay of early and late LTP. It is also possible that changes in gene expression that have been observed on longer timescales after plasticity induction, have some role in maintaining particular configurations of rates of biochemical process rather than having a direct impact upon the synaptic efficacy.

Detailed balance was imposed on the models containing rings. There is no reason to suppose that synapses obey detailed balance, being as they are not closed isolated systems. Thus detailed balance is imposed not for biological realism but for reasons of mathematical simplicity. It is easier to understand the behavior of the 8 state model in particular if we can think in terms of a superposition of eigenfunctions. For this to be the case detailed balance (or weakly distorted detailed balance) must apply. Since the models can match experimental data well when detailed balance is obeyed, there is little reason not to impose it at this stage.

[6]There are interesting state based models of subcellular plasticity dynamics (Shouval, Bear, and Cooper, 2002; Smolen, 2007) but these do not directly consider the memory trace.

In order to make the 8 state model tractable, a power law scaling was introduced between the transition rates. The parameters of the power law were then adjusted until the weight decay matched the experimental data. This worked surprisingly well. Nevertheless the models could be improved by formally extracting the magnitudes of the timescales present in the decay of early and late LTP given some assumptions about factors such as the number of timescales present. The expansion in eigenvectors could then be matched to the empirically determined timescales.

Chapter 7

Amnesia and synaptic overload

The hippocampus is crucial to the initial formation of episodic memories (Schoville and Milner, 1957; Morris et al., 2003). Evidence, primarily from hippocampal lesions, shows that destruction of the hippocampus prevents the formation of new declarative memories but does not disrupt existing long term memories (Alvarez, Zola-Morgan, and Squire, 1995; Teng and Squire, 1999; Corkin, 2002). This implies that after some time consolidation permits a subset of hippocampal traces to be completely mediated by circuits elsewhere in the cortex, forming a long term memory (chapter 1).

The hippocampus contains several subregions that are implicated in learning and memory. There are theoretical and anatomical reasons to believe that the hippocampus achieves its function of association of neocortical memory traces by means of a multilayer associative network across layers CA3-CA1 (Marr, 1971; Rolls, 1996; McNaughton and Morris, 1987; Rolls and Treves, 1998). Furthermore, rapidly acquired spatial learning has been shown to be highly dependent upon the dentate gyrus, which is a different sub network of the hippocampus (Moser et al., 1995).

Whatever the mechanistic details of hippocampal function, many of which are still not clear, it seems that the purpose of the hippocampus is to allow rapid, automatic storage of traces linking other neuronal groups such that the traces are either discarded or transferred to neocortex at a later stage. Perhaps the best possible explanation for the encoding and storage of memory traces in the hippocampus is by modification of ensembles of excitatory synapses on pyramidal neurons (Morris et al., 2003; Martin and Morris, 2002). If we assume this to be the case, then the stability plasticity dilemma is pertinent to hippocampal memory traces for the following reasons: 1) The population of cells and hence synapses is comparatively restricted. For example, there are 2.5×10^5 cells in CA1 in rats and 4.6×10^6 in humans (Rolls, 1996), each having

around 3×10^4 excitatory synapses[1] (Megias et al., 2001). 2) There is a necessity for a high initial strength of all memory traces (because all of them might be consolidated). 3) It must be possible for traces to be retained for long enough such that they can be transferred to long term storage should they be deemed to be of behavioral relevance after their initial formation.

The synaptic plasticity and stability hypothesis predicts that saturation of LTP in the hippocampus should lead to the disruption of learning in-vivo because the number of weights available for modification (and hence learning) would be restricted. Experiments where LTP was induced in the dentate gyrus of rats have been performed and resulted in behavioral amnesia (McNaughton et al., 1986; Castro et al., 1989), although the results were mixed with several labs unable to reproduce these findings (Sutherland, Dringenberg, and Hoesing, 1993; Jeffery and Morris, 1993; Cain et al., 1993; Barnes et al., 1994). More recent studies suggest that as long as the degree of saturation is sufficient then amnesia can be provoked (Moser and Moser, 1999; Moser and Morris, 1998). In this chapter, amnesia by saturated LTP induction is demonstrated in the state based models of LTP.

Firstly, amnesia inducing protocols are simulated. We imagine that a memory trace is stored in the equilibrium state of the system at $t = t_0$ and the memory trace is tracked. At a later time $t_1 > t_0$ early or late LTP is induced and its impact upon the memory trace is observed. This tests the extent to which LTP disrupts the memory trace. As is suggested by experiment, it is found that the extent of amnesia is dependent upon the extent of saturation of the LTP. Furthermore, these simulations show that early LTP induction leads to *reversible* disruption of the original memory trace. This implies that early LTP might allow temporary modification of synapses, such that a new memory trace can be stored without complete disruption of the original trace.

The observation that early LTP might allow the coexistence of memory traces superimposed upon different timescales is referred to as 'synaptic overload' here, but has been discussed by other authors (Gardner-Medwin, 1989; Rolls, 1996; Morris et al., 2003). In light of this, simulations were carried out in which multiple memories were stored within one population of synapses using the two different timescales of LTP. This is achieved by first storing a memory trace in the synapses by using the late LTP/D protocols defined in chapter 6. After this initial storage, further patterns are stored using early LTP/D protocols. The recall of both the initial pattern, stored with

[1] In his classic paper, David Marr assumes of the order of 1×10^4 synapses per pyramidal cell and he assumes 1×10^4 pyramidal, or 'output' cells (Marr, 1971; Willshaw and Buckingham, 1990).

late LTP/D and the subsequent patterns, stored with early LTP/D is then assessed.

If we consider hippocampal LTP as serving the function of 'automatic recording of experience' (Morris and Frey, 1997; Morris et al., 2003), then synaptic overload might alleviate the plasticity stability dilemma in the hippocampus by allowing rapid formation of new, strong memory traces without permanent destruction of pre-existing traces. This might allow both the rapid storage of high signal memories and the retention of previous, important memories for long enough that they can be consolidated at the systems level.

7.1 Amnesia and saturated LTP in the hippocampus

The dentate gyrus is an area of the hippocampus that has been shown to be crucial for spatial learning (Moser et al., 1995). Thus it follows from the synaptic plasticity and memory hypothesis, that it might be possible to disrupt spatial learning by preventing modification of the synapses between the perforant path and the dentate gyrus mossy fibres. One way of achieving this experimentally is to saturate LTP in the efferent synapses of the perforant path. By doing this it is supposed that any memory trace in the dentate gyrus is also saturated, thus preventing useful storage of information.

The effects of LTP induction upon memory can be studied from two perspectives. Firstly saturation of a synaptic population in vivo by induction of LTP within the dentate gyrus, can destroy previously learned memories, either by disrupting the memory trace or by preventing it from being retrieved. We refer to the case that a previously learned memory trace is disrupted as *retrograde* amnesia. Alternatively it might be that saturation of synapses prevents the acquisition of new memories. This is referred to as *anterograde* amnesia. Later in this chapter retrograde and anterograde amnesia are studied in state based models of synapses where LTP is saturated.

Experimental studies investigating disruption of memory by LTP typically record from within the dentate gyrus and stimulate the perforant path of the hippocampus. At least two groups of rats are used in each study: In the control group, only recording takes place and no LTP is induced. In the test group, LTP is induced to the maximum attainable level. Retrograde amnesia can be studied by training both groups of rats on a navigation task (after the synaptic baseline has been established) and then subjecting the test group of rats to LTP. A comparison of the control and test groups allows an assessment of the extent of retrograde amnesia. To test anterograde amnesia, both groups of rats are trained on a navigation task after, or during, a period in which the

test group receives LTP.

There is evidence that saturation of LTP in vivo readily leads to anterograde amnesia (Castro et al., 1989; McNaughton et al., 1986), disrupting the acquisition of new learning. Furthermore, retrograde amnesia can be produced provided that LTP is induced within hours of learning (McNaughton et al., 1986). However, four of the seven commonly cited studies of amnesia and LTP induction found no effects of LTP induction on learning (Sutherland, Dringenberg, and Hoesing, 1993; Cain et al., 1993; Jeffery and Morris, 1993; Barnes et al., 1994). In light of the study by Moser and Morris, it seems likely that this is because the outcome of these experiments is dependent upon the level of saturation attained by the LTP protocol (Moser and Morris, 1998; Moser and Moser, 1999).

In this chapter, synaptic saturation with LTP is simulated in the state based models. Retrograde amnesia is assessed by tracking the signal to noise ratio and autocorrelation of a memory trace when the trace is initially stored in the steady state of the model prior to application of the LTP. Anterograde amnesia is assessed by tracking the signal to noise ratio and autocorrelation of a memory trace stored in a non-steady state of the model, just after the application of the LTP protocol. In light of the experiments mentioned above, it is expected that the models should exhibit anterograde and retrograde amnesia in a manner that is dependent upon the level of saturation of the synapses.

In experiments, the time elapse between LTP induction and initial training (i.e. memory trace formation) is at least of the order of minutes. Here we shall assume that the memory is stored at the first time instant after the completion of LTP induction. It is also assumed that the memory trace is formed in a 'one-shot' manner. This is not the case in experiments, where training occurs over many trails. However if we imagine an ethically dubious experiment, where we are allowed to average over many rats, we would predict an improvement in the spatial memory task even after the first trial. The results here are equivalent to this imaginary experiment.

In chapter 6 and chapter 4 the memory trace was studied from the equilibrium perspective (the stationary plasticity stability dilemma). The synapses were settled into the equilibrium state and modifications to the synapses comprising the memory trace were therefore fluctuations around equilibrium, Fig. 7.1A. In this case a natural interpretation of those fluctuations (i.e. random transitions between weight states) is that they occur as a *consequence* of memory storage.

This picture must be expanded slightly when we wish to consider the effects of large perturbations (i.e. the induction of LTP) on the memory trace (the non-stationary

plasticity stability dilemma). Firstly, we must understand the effect of disruption of the steady state on a memory trace that was at first stored at equilibrium (in the retrograde case). Secondly we must consider the case where a memory trace is stored away from the steady state as is the case when memory is acquired directly after LTP induction (in the anterograde case), Fig. 7.1B. We shall find that when memories are stored away from the steady state, our intuitions must be modified.

Study of anterograde amnesia raises the possibility of a memory storage regime, in which individual synapses are not at the steady state[2]. In this case memories are not considered as resulting from spontaneous fluctuations near equilibrium. Rather, memory results from large LTP/D potentiation and depression events involving subsets of the synaptic ensemble. From this viewpoint, small fluctuations of the synapses governed by the transition probabilities between states are no longer the cause of the memory trace[3]. Now the memory trace is composed of large displacements of groups of synapses in weight space, and the dynamics of the trace are governed directly by the dynamics of eLTP/D and lLTP/D, Fig. 7.1C. We shall see that in this regime the distinct decay timescales of early and late phase plasticity could be beneficial to memory storage.

7.2 Calculation of the memory timecourse

Previously in chapter 6 the autocorrelation and signal to noise ratio (SNR) of the synaptic ensemble was used to measure the decay timecourse of the memory trace when the synapses were at *equilibrium* for all time. In this chapter we are interested in the evolution of the memory trace in two *non-equilibrium* cases: 1) The case that a memory trace is initially stored in the equilibrium state of the synaptic ensemble, but where the synapses are subsequently subject to LTP, inducing a non-equilibrium state. This tests retrograde amnesia. 2) The case that a memory trace is stored in the ensemble straight *after* LTP is induced when the mean weight is elevated. This tests anterograde amnesia.

In the models here LTP is induced in an identical fashion to chapter 6, by substitution of the baseline matrix for the LTP protocol matrix for some duration of time. We shall investigate amnesia as a function of the extent of saturation of the synapses (i.e.

[2]That is to say that the probability that a synapse occupies each state is a function of time.

[3]But such fluctuations are still crucial because they give rise to the decay timescales of eLTP and lLTP. However from this perspective the fluctuations are not seen as being usefully correlated with events in the outside world, but rather as the consequence of cellular metabolism.

the duration of substitution of the protocol matrix) and the timescale of decay of the LTP (i.e. whether LTP is early phase or late phase). The mean value of the synaptic weights in the ensemble with the baseline matrix or during the application period of a protocol is calculated by expansion in eigenvectors (chapter 3+4),

$$\langle w(t)\rangle = \sum_i w_i \sum_k c_i^{(k)} \Phi_i^{(k)} exp(-t/\lambda_k) \tag{7.1}$$

with

$$p_i(t_0) = \sum_k c_i^{(k)} \Phi_i^{(k)} \tag{7.2}$$

where $\Phi^{(k)}$ are the eigenvectors of the baseline or protocol rate matrices of the model and where $p_i(t_0)$ is the initial occupancy of each state. The expansion is always checked against the direct numerical method of raising the transition matrix to a power (chapter 3). Both methods agree as would be expected for models with detailed balance.

7.2.1 Case 1: Retrograde amnesia

To test retrograde amnesia we require the autocorrelation in the case that the transition probabilities change during the duration of interest (case 1 above). If the synaptic ensemble begins in some state $\mathbf{p}(w_i,t_0)$ at t_0 then as time evolves the autocorrelation follows (chapter 3),

$$\langle w(t_0)w(t')\rangle = \sum_{ij} w_j(t') p(j,t'|i,t_0) w_i(t_0) p(i,t_0). \tag{7.3}$$

We now intervene at some time t_1, where $t_0 < t_1 < t'$ and instantaneously substitute the baseline rate matrix for one of the LTP protocol matrices $\{E_P, L_P\}$. In this case the two point autocorrelation becomes

$$\left\langle w(t_0)w(t')\right\rangle = \sum_{ijk} w_k(t') p(w_k,t'|w_j,t_1) p(w_j,t_1|w_i,t_0) w_i(t_0) p(w_i,t_0) \tag{7.4}$$

where a new index and conditional probability must be introduced at the time when we intervene. Index i runs over the initial weight state at t_0, index j runs over the weight state at the point in time that the LTP is initiated (t_1) and index k runs over the weight state at the present moment in the time (t'). After some time interval $\Delta t = t_2 - t_1$, $t_1 < t_2 < t'$ the transition rates are returned to their baseline values. The weight deflection caused by the application of LTP now decays away. The autocorrelation proceeds

$$\left\langle w(t_0)w(t')\right\rangle = \sum_{ijkl} w_l(t') p(w_l,t'|w_k,t_2) p(w_k,t_2|w_j,t_1) p(w_j,t_1|w_i,t_0) w_i(t_0) p(w_i,t_0) \tag{7.5}$$

where the index k now runs over the weight state at the time that the LTP protocol is removed and the baseline weight matrix is replaced at t_2. Provided that we make no further intervention and as $t' \to \infty$, then $\langle w(t)\rangle \to \langle w_\infty\rangle$ and $\langle w(t_0)w(t')\rangle \to \langle w_\infty^2\rangle$. Eq.(7.5) describes that the autocorrelation now depends upon the initial state $p(w_i,t_0)$, the protocol matrix, the baseline matrix, the time at which we applied the plasticity protocol t_1, its duration Δt and the present moment in time t'. Thus the overall process is non Markovian by virtue of its dependence upon t_1 and Δt, however the autocorrelation is piecewise Markovian because at all times in between the instants when we perform the plasticity protocols, it is Markovian. This procedure can be repeated for cases where multiple interventions are made, for example if LTP is induced twice or more (see §§7.4.1+7.5). Thus, given the values of the various transition matrices and the times when the transition matrices are substituted, the autocorrelation can be recursively calculated. Eq. (7.5) is normalised

$$\kappa(t) = \frac{\langle w(t_0)w(t)\rangle - \langle w(t_0)\rangle \langle w(t)\rangle}{\sqrt{\sigma_w^2(t_0)\sigma_w^2(t)}} \tag{7.6}$$

such that the autocorrelation ranges between 0 and 1. From the autocorrelation the signal to noise can be found,

$$\frac{S}{N} = \frac{\Omega\left|\eta\sqrt{\sigma_{w,0}^2\sigma_w^2(t)}\kappa(t) + \langle \mathbf{w}(t)\rangle\left(\langle \mathbf{x}_Y\rangle - \langle \mathbf{x}_N\rangle\right)\right|}{\sqrt{\frac{1}{2}(\sigma_Y^2(t)+\sigma_N^2(t))}}. \tag{7.7}$$

where $\sigma_Y^2(t) = \Omega\langle \mathbf{x}_Y \mathbf{w}(t)\rangle(1-\langle \mathbf{x}_Y \mathbf{w}(t)\rangle)$ and $\sigma_N^2(t) = \Omega\langle \mathbf{x}_N\rangle\langle \mathbf{w}(t)\rangle(1-\langle \mathbf{x}_N\rangle\langle \mathbf{w}(t)\rangle)$. In the signal to noise calculations here it is assumed that the mean of the signal and noise distributions are identical to the steady state mean weight. That is to say that $\langle \mathbf{x}_Y\rangle = \langle \mathbf{x}_N\rangle = \langle \mathbf{w}\rangle_\infty$. This amounts to the following 2 assumptions: 1) That when there is no intervention, whatever patterns are stored by the animal have similar statistics to the steady state statistics of the weights (assuming binary weights and inputs) and 2) when LTP is induced, this does not alter the statistics of the patterns that are presented to the dentate gyrus, i.e. LTP does not alter the patterns of activity encountered by the weights. In the case of retrograde amnesia the system is initially in the steady state, $\sigma_{w,0}^2 = \sigma_{w,\infty}^2$ and Eq. (7.7) becomes

$$\frac{S}{N} = \frac{\Omega\eta\sqrt{\sigma_{w,\infty}^2\sigma_w^2(t)}\kappa(t)}{\sqrt{\frac{1}{2}(\sigma_Y^2(t)+\sigma_N^2(t))}} \tag{7.8}$$

from the initial condition at $t = t_0$,

$$\eta = \frac{S_0}{\sqrt{\sigma_{w,\infty}^2 \sigma_{w,0}^2}\Omega} \tag{7.9}$$

giving

$$\frac{S}{N} = \frac{S_0 \sigma_w(t)}{\sigma_{w,0}\sqrt{\frac{1}{2}(\sigma_Y^2(t) + \sigma_N^2(t))}}\kappa(t) \tag{7.10}$$

where we note that at the time $t = t_0$, $\sigma_w(t) = \sigma_{w,0}$ and $\sigma_Y(t) = \sigma_{Y,0}$, $\sigma_N(t) = \sigma_{N,0}$ such that the steady state relation is recovered, i.e. the SNR is the autocorrelation scaled by the initial SNR. This is to be desired for retrograde amnesia since the steady state condition should be recovered before any intervention is made.

The initial signal for the 2 and 4 state models was derived in chapter 6, recall that for the 2 state model the initial signal is

$$S_0 = \Omega(M_{12}p_1 + M_{21}p_2) \tag{7.11}$$

where M_{12} is the transition probability from the depressed state (state 1) to the potentiated state (state 2) and M_{21} is the transition probability from the potentiated to the depressed state. In the initial state the probability of occupancy of state 1 is p_1 and for state 2 is p_2. In the balanced case where $M_{12} = M_{21} = r$ this reduces to $S_0 = \Omega r$ and the initial signal is independent of the initial state.

For the 4 state model, the initial signal is

$$S_0 = \Omega(p_1M_{12} + p_4M_{43} + p_2M_{21} + p_3M_{34}) \tag{7.12}$$

which reduces to $S_0 = 2\Omega(p_1M_{12} + p_2M_{21})$ under the detailed balance conditions stated in chapter 6.

In the case of the 8 state model, the initial signal is calculated from the transition matrix with $S_0 = \Omega(\mathbf{w}^T M_+ \mathbf{p}_\infty + \mathbf{w}^T M_- \mathbf{p}_\infty)$ which is simply the fundamental formula from which Eqs. (7.11+7.12) originate (chapter 3). As in chapter 6, $\Omega = 3.8 \times 10^{11}$.

7.2.2 Case 2: Anterograde amnesia

Chapter 6 made use of the autocorrelation of the synaptic weights as they evolve from an initial state (which was taken to be the equilibrium state) with constant transition rates. To test anterograde amnesia we shall use this again, but now the initial state is not

the equilibrium state. The un-normalized autocorrelation, Eq. (7.3) can be expressed as (chapter 3),

$$\langle w(t_0)w(t)\rangle = \sum_k e^{\lambda_k t}\left(\sum_i p(i,t_0)C_{ik}w_i\right)\left(\sum_j \Phi_j^{(k)}w_j\right) \tag{7.13}$$

where C is the inverse of the eigenvector matrix of the baseline rate matrix. Again, Eq. (7.13) is normalised such that the autocorrelation ranges between 0 and 1. In the case of anterograde amnesia $\sigma_{w,0}^2 \neq \sigma_{w,\infty}^2$. Now Eq. (7.7) becomes

$$\frac{S}{N} = \frac{S_0\sigma_w(t)}{\sigma_{w,\infty}\sqrt{\frac{1}{2}(\sigma_Y^2(t)+\sigma_N^2(t))}}\kappa(t) \tag{7.14}$$

where we note that at the time $t=\infty$, $\sigma_w(t)=\sigma_{w,\infty}$ and $\sigma_Y(t)=\sigma_{Y,\infty}$, $\sigma_N(t)=\sigma_{N,\infty}$ such that the steady state relation is recovered. Again, this should occur because in the case of anterograde amnesia the steady state should be recovered in the limit $t\to\infty$ when all influence of the induced LTP is lost.

7.3 Effects of LTP induction on the memory trace in state based models

7.3.1 2 state model

The evolution of the memory trace is calculated when LTP is applied to the ensemble of synapses using the methods discussed in the previous section. To illustrate the effect of LTP on the memory trace, we first consider the 2 state model. We assume that at $t=t_0$, the ensemble of synapses is in the steady state and that at this instant a memory is stored (memory 1 in Fig. 7.2A). Some time later at $t=t_1$ LTP is induced. When LTP is saturated by applying the protocol for $1000s$, the initial memory trace is completely destroyed and the SNR and autocorrelation of memory 1 goes to 0, Fig. 7.2B. However if the protocol is applied for a shorter duration of $100s$, the memory is only partially disrupted, having a reduced SNR and autocorrelation after induction. Hence in the 2 state model, retrograde amnesia is induced by application of the LTP protocol.

Anterograde amnesia is also studied. At $t=t_2$, when LTP induction is at its peak value, we assume that another memory is stored (memory 2 in Fig. 7.2A). Memory 2 differs from memory 1 in that the ensemble of synapses is no longer at the steady state when the memory is stored. The timecourse of the SNR of memory 2 is identical

to the SNR timecourse in the case that no LTP is induced. Thus in the 2 state model anterograde amnesia does not occur if we consider the strength of the memory to be determined by the SNR. However if a memory decoder were reading the normalised autocorrelation, then anterograde amnesia would be experienced.

This disparity between the SNR and the autocorrelation might seem puzzling in light of the analysis of the 2 state model presented in chapter 3. The 2 state model obeys the fluctuation dissipation theorem (FDT). Furthermore we found that there is only one timescale of the *un-normalised* autocorrelation and the mean weight evolution regardless of the initial condition. But the autocorrelation timescale now appears to differ depending upon the initial condition, Fig. 7.2C. This is because the *normalised* autocorrelation is plotted here. Let $\langle w(t_0)w(t)\rangle \propto \phi(t)$. From the FDT we know that $\langle w(t)\rangle \propto \phi(t)$ also holds. Since we have binary synapses $\sigma_w^2(t) \propto \phi(t)(1-\phi(t))$. Thus the normalised autocorrelation, Eq.(7.6) gives[4] $\kappa(t) \propto \sqrt{\phi(t)}/\sqrt{(1-\phi(t))}$. In contrast to the autocorrelation, the SNR Eq. (7.14) does directly reflect $\phi(t)$, since

$$SNR \propto \frac{\sqrt{\phi(t)(1-\phi(t))}\sqrt{\phi(t)}}{\sqrt{(1-\phi(t))}} \propto \phi(t). \tag{7.15}$$

Finally, the initial signal of the 2 state model Fig. 7.2C is decreased by a small amount when LTP is saturated. While the initial signal S_0 is not greatly effected by LTP induction in the 2 state model, the signal and noise standard deviations, $\sigma_Y(t)$,$\sigma_N(t)$ and the standard deviation of the weight $\sigma_w(t)$ are altered during induction. Thus the signal to noise ratio is slightly modified.

7.3.2 4 & 8 state models

To test the 4 and 8 state models, an identical procedure to Fig. 7.2 is used. We imagine that at $t = t_0$ a memory is stored in the steady state of the system. Some time later at $t = t_1$ LTP is induced. At the peak of the LTP at $t = t_2$ another memory is stored. The SNR and the autocorrelation of both memories is calculated.

The behavior of the 4 and 8 state models is richer than the 2 state model and so five levels of saturation of early LTP and late LTP were induced in the models, Fig. 7.3A+B, top row, (grey levels, from light for low saturation to dark for higher saturation, heavy black line is the mean weight when LTP is saturated). In the remaining

[4]The normalised autocorrelation will not take this form if the variance and the mean are constant, i.e. if the system is at the steady state. In this case the normalised autocorrelation $\kappa(t)$ has an identical timecourse as the un-normalised autocorrelation $\langle w(t_0)w(t)\rangle$.

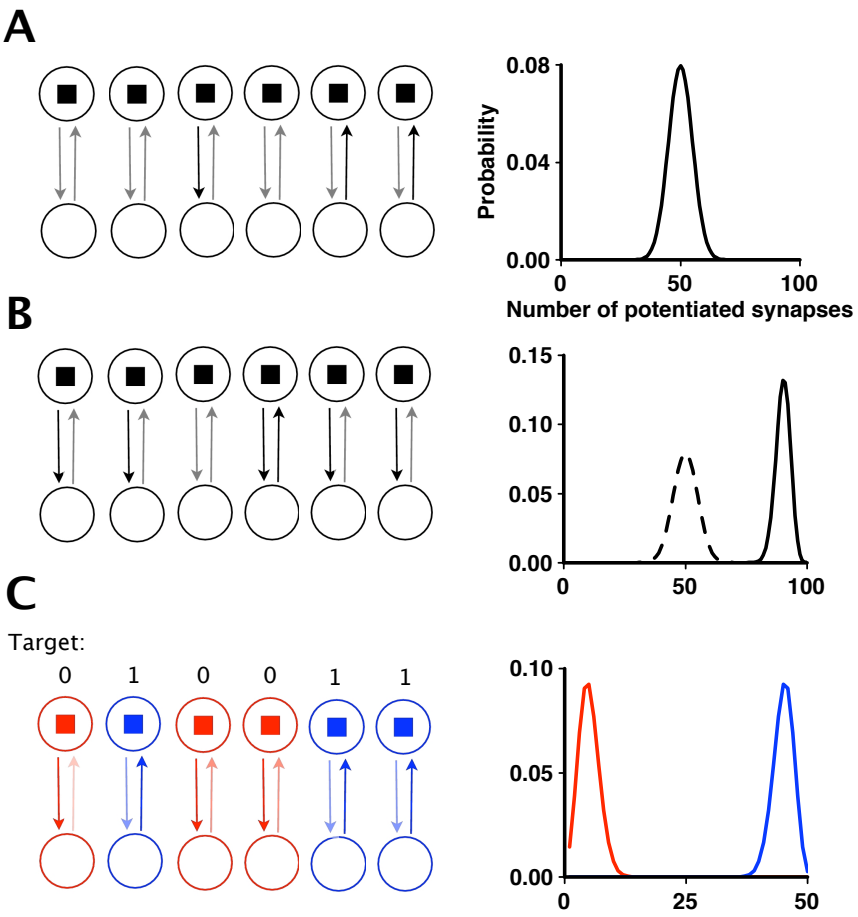

Figure 7.1: Three regimes of memory storage. A: Memory storage at equilibrium. Synaptic weights make stochastic transitions once equilibrium has been established. It is assumed that all of the instantaneous transitions contribute to the memory trace. On average the steady state distribution of the number of potentiated synapses is preserved at all times during storage and recall. This is an example of the SPS. B: Memory storage after saturation of LTP. An LTP protocol is applied such that the maximum number of potentiated synapses is achieved. At the instant that the protocol ceases a memory is stored and the distribution of the number of potentiated synapses relaxes back to the steady state distribution. As a result of this there are many synaptic weight transitions at the instant after LTP induction. This is an example of the NPS. C: Memory storage by driving the synaptic weights to non-steady state target values (top row). LTP (blue synapses) and LTD (red synapses) is applied such that all synapses participate in the pattern. This creates two distributions of the number of potentiated synapses, while the overall mean remains constant. If there is no further intervention these distributions relax back to the steady state distribution and the timescale of this relaxation determines the timecourse of the SNR.

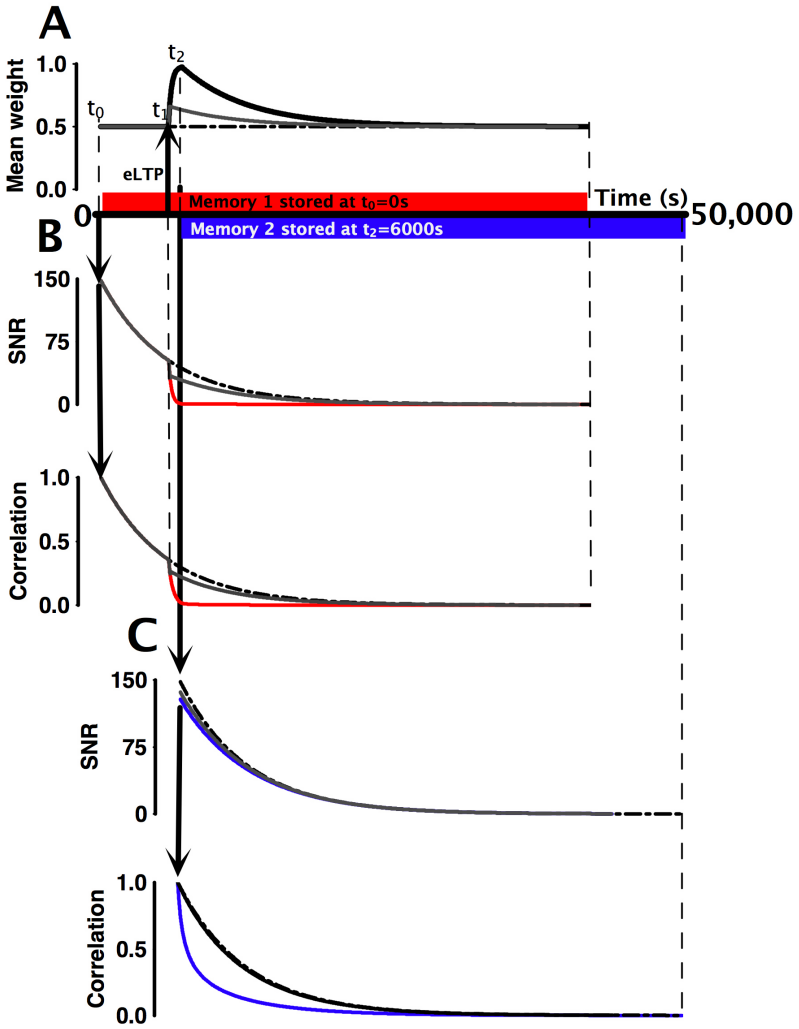

Figure 7.2: Retrograde and anterograde amnesia in the 2 state model. A: The effect of LTP induction on the synaptic weight. The heavy black line is the case where LTP is saturated and the protocol is applied for $1000s$. The thinner grey line is the case where LTP is not saturated and the protocol is applied for only $100s$. The dashed line is the mean weight of the undisturbed system. The time axis is a time line for the whole experiment. At $t_0 = 0s$ the synapses are in the equilibrium state and a memory is stored. The red box labelled 'Memory 1' shows the duration of testing of recall of this memory. At $t_1 = 5000s$ LTP is induced. When LTP induction is complete at t_2 a second memory is stored in the synapses. The blue box labelled 'Memory 2' indicates the duration of testing of this memory. B: Recall of memory 1. LTP at t_2 disrupts the memory leading to a reduction in the SNR and autocorrelation associated with memory 1 (retrograde amnesia). The dashed lines are the SNR and autocorrelation in the case the the system is undisturbed and remains at equilibrium. C: Recall of memory 2. Previous induction of LTP causes the autocorrelation to decay rapidly (anterograde amnesia), although the SNR is almost undisturbed.

panels of Fig. 7.3, red indicates maximum saturation in case of memory 1, blue indicates maximum saturation in case of memory 2.

Finally, note that in the case that LTP is saturated, the protocol matrix is applied for long enough that the model achieves the equilibrium point of the protocol matrix. This does not necessarily imply that all weights are saturated at $w = 1$. Thus even if the maximal LTP is achieved all of the weights are not saturated and in any case, they are still free to fluctuate upon removal of the protocol.

7.3.2.1 Case 1: Retrograde amnesia

Calculation of the autocorrelation, Fig. 7.3B (bottom row), reveals that late LTP leads to a progressively greater degradation of the correlations in the memory trace as the synapses are taken to higher levels of saturation. Maximum saturation of late LTP leads to very large disruption of the memory trace, with the correlation decaying away rapidly. The same is true during the induction period of early LTP when it is induced to progressively greater levels of saturation, Fig. 7.3A (bottom row). After the induction period of early LTP however, the behavior of the autocorrelation is quite different than in the late LTP case, compare Fig A+B (bottom row). Rather than mono-phasically decaying to zero, the autocorrelation first rebounds, increasing before it falls. Thus after induction of early LTP, the synapses are again correlated with their previous values prior to LTP induction. This is in contrast to the late LTP case where the correlation with the initial pattern is removed for all time. In all cases, the rebound returns the autocorrelation to the timecourse that it would have taken if no LTP had been induced.

The behavior of the signal to noise is dominated by the autocorrelation, which is driven towards zero and hence strongly attenuates the signal Eq. (7.10). The detectability of the memory is therefore dramatically reduced during the induction of both late LTP and early LTP.

7.3.2.2 Case 2: Anterograde amnesia

Anterograde amnesia in the state based models corresponds to the case in which the SNR of the memory trace is calculated from just after the completion of the induction of LTP at $t = t_2$, Fig. 7.3A+B (top row). For early LTP, the anterograde memory trace is plotted in Fig. 7.3C, while for late LTP the anterograde memory trace is plotted in Fig. 7.3D.

In the anterograde case, the initial memory trace is stored in a non stationary state

of the system. For this reason the autocorrelation of the memory trace decays more rapidly than in the steady state case when either early LTP or late LTP are induced, Fig. 7.3C+D bottom row. Interestingly the autocorrelation decays equally rapidly regardless of whether early or late LTP are induced in the four state model. In the eight state model the autocorrelation of the new memory trace decays more rapidly when early LTP was induced prior to its formation. In all cases the rate of decay of the new memories is more rapid when the initial saturation of the LTP induction is increased.

The difference between the signal to noise and the autocorrelation now becomes important. These measures are equivalent at equilibrium in the sense that the SNR can be thought of as a scaled version of the autocorrelation (chapter 3). When the system is not at equilibrium however, this is no longer the case because the timecourse of the mean and the variance of the weights have differential effects upon the autocorrelation and the SNR, as we saw previously in the case of the 2 state model, Eq. (7.14). Furthermore, the 4 and 8 state models do not obey the FDT because they incorporate hidden states, and therefore the autocorrelation $\langle w(t_0)w(t)\rangle$ does not have the same timecourse as the response $\langle w(t)\rangle$. Thus the SNR and the normalised autocorrelation $\kappa(t)$ can have more complex behavior because more mixtures of timescales are possible, Fig. 7.3C+D.

Another consideration is that away from the steady state, the initial signal S_0 is no longer constrained by the steady state $\mathbf{p}_\infty$ but varies largely. This means that even if the choice of initial state causes the weight autocorrelation to decay quickly, the memory trace can still remain detectable for a long period of time as long as the initial signal is large. This might look like we are getting something for nothing, but we are not. This is a consequence of the fact that if we are not constrained to be at equilibrium, the signal distribution can be a long way away from the noise distribution. Thus the signal can be detectable regardless of the temporal weight correlation. The price we pay is that if we try to store another non-equilibrium memory in this manner, by moving the signal distribution again, then we are in danger of obliterating the previous memory instantaneously. In §7.5 we examine this scenario.

In the case of early LTP, the initial signal of a memory trace stored just after induction increases as the saturation increases, Fig. 7.3C. This can be understood by realising that the effect of early LTP is to move an ever greater fraction of synapses that are initially in state 1 into state 2, of both the 4 state model and the 8 state model. Thus synapses are moved from a state having a low rate of potentiation to a state having a high rate of depression. This allows more synapses to make depressing transitions,

which still represents a change relative to the noise distribution, all be it a negative one. But in principle this change can be detected also (and this is why the signal is defined in terms of the absolute difference between signal and noise means in chapter 3).

In the case of late LTP, the initial signal at first increases as a function of the saturation. This is due to the fact that partial late LTP causes early LTP and so the initial signal increases for the reason stated previously. However when late LTP is saturated, this implies that the synapses that were occupying state 1 in the 4 state model are moved to state 4. In the 8 state model it implies that the synapses that were occupying state 1 are moved to state 7. Depression transitions from state 4 in the 4 state model and state 7 in the 8 state model are equally as likely as the potentiation transitions from the initial states in both models (this is a result of the symmetry in the transition rates that was enforced in chapter 6). Thus when late LTP is saturated, we would expect that the initial signal be nearly equal to the initial signal at equilibrium. Indeed this is the case, Fig. 7.3D (bottom row).

There is another aspect to the SNR in Fig. 7.3D that deserves mention. It appears to *increase* after the initial storage time when the late LTP is saturated. This is observed in both the 4 state and 8 state models. This occurs because the model is not at equilibrium and the average number of potentiated synapses is changing. Thus the variances of the signal and noise distributions, σ_Y^2, σ_N^2 shrink, leading to a transient increase in the signal to noise ratio.

In conclusion both the 4 state and 8 state models show anterograde amnesia when viewed from the point of view of the autocorrelation of the weights. However in general the SNR does not follow this trend. In fact LTP induction can lead to an increase in the memory strength of newly acquired memories when measured with the SNR. Only early LTP saturation in the 4 state model leads to anterograde amnesia. Thus we find that whether or not anterograde amnesia is observed depends upon the saturation of the LTP, the phase of the LTP and the state space of the model. Therefore there are in fact several non-trivial behaviors that are possible. On the other hand retrograde amnesia can be provoked more robustly.

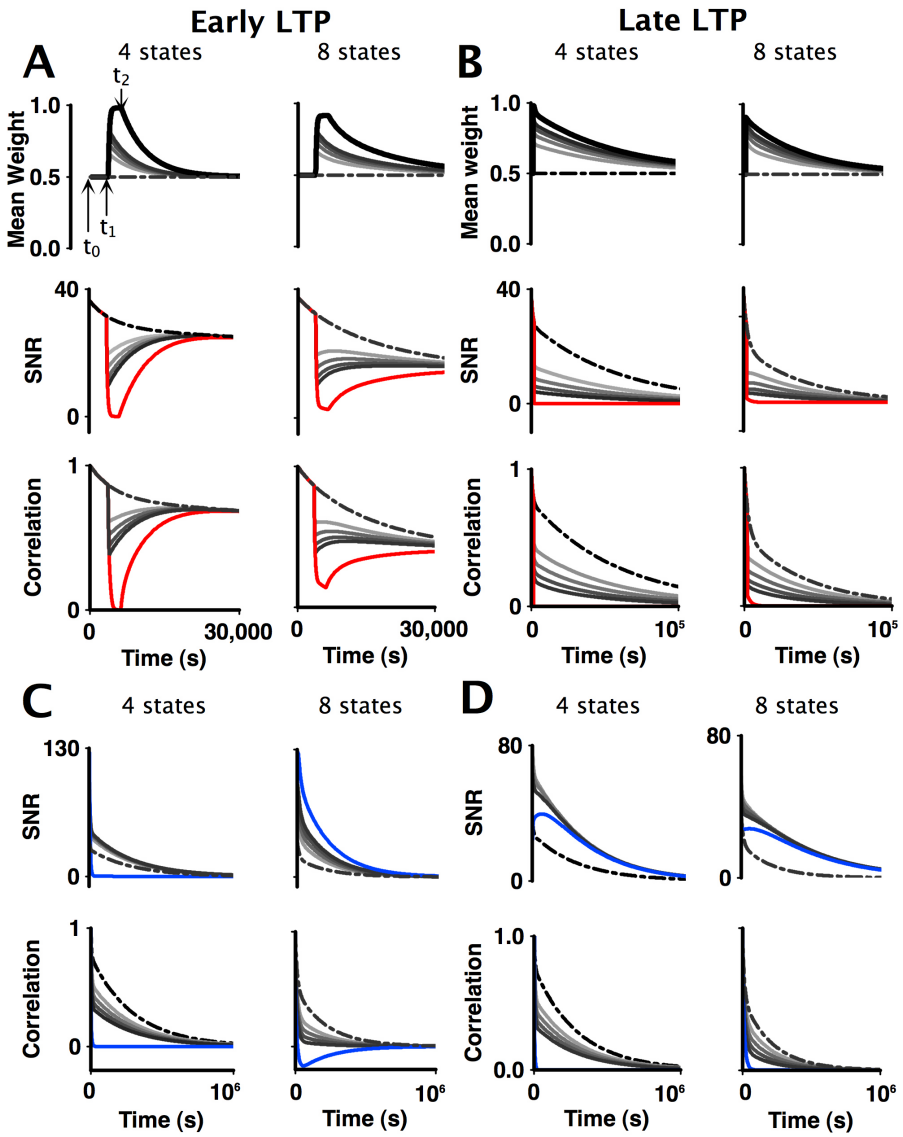

Figure 7.3: The autocorrelation and the SNR of weights in the 4 state (left) and 8 state (right) models during and after induction of LTP. A: Memory 1 is stored at $t = t_0$ while the system is at the steady state. Early LTP is induced at $t = t_1$ (top row). Each line is shaded with a differing grey level denoting the degree of saturation of LTP from light grey to black. The heavy black trace is maximum saturation. The dashed lines indicate the steady state. The red line indicates the timecourse of memory 1, when LTP is saturated. The SNR behaves in an analogous manner (middle row). B: Same as A, but in the case of late LTP is applied. C: Memory 2 is stored directly after induction of LTP is complete at the peak of LTP induction $t = t_2$ (when the system is in a non-steady state). The blue line indicates the timecourse of memory 2 when LTP is saturated. D: Same as C, but in the case that late LTP was induced.

7.4 Memory traces in the recognition unit with state based synapses

In the previous section it was demonstrated that after the induction of early LTP in an ensemble of state based synapses, the correlation of those synapses with their initial state can 'rebound' following the decay trajectory that would otherwise be followed if no LTP had been induced. This is a result of there being more than one timescale in the synaptic dynamics. Changes made to the synapses on the short timescale can completely decay, while changes made on more slowly decaying timescales remain. Next, it is demonstrated that this allows the memory traces to be superimposed. To show this, we first consider a perceptron. The aim of this is not to study the perceptron itself, or the perceptron learning rule. Indeed the perceptron learning rule is not used. Rather the aim is to demonstrate that in principle multitimescale synapses might be useful in a system performing a simple classification task. Implementation of this in a more sophisticated system is beyond the scope of this thesis, but is an obvious extension to this work.

The linear threshold unit (or Perceptron) can be used to classify two clusters within some data set (Hertz, Krogh, and Palmer, 1991) and has been used previously by other authors investigating the survival of memories in state based synapses (Fusi and Senn, 2006; Baldassi et al., 2007). Here we apply the perceptron to investigate memory storage using synapses implementing the state based LTP models.

The aim of the linear threshold unit is to classify patterns into two classes C_1 and C_2 by labelling each input pattern $\mathbf{x}$ with output $f(a) \in \{0,1\}$ such that $f(a) = 0$ for class C_1 and $f(a) = 1$ for class C_2. As input the linear thresholder is passed a pattern vector $x_i \in \{0,1\}, i \in \{1,...,\Omega\}$. In standard perceptrons using the percepetron learning algorithm, the weights are adjusted in the case of both positive and negative examples.

Here the perceptron learning algorithm is not used. Instead, weights of the perceptron are only adjusted in the case of positive examples. The perceptron is used to classify patterns as unseen (C_1) or seen (C_2) and the weights are only adjusted when a pattern is to be stored in C_1. The perceptron has stochastic binary weights, where the probability of the weights being 0 or 1 is controlled by a state based model of LTP. To store an input pattern in the seen class, LTP and LTD is performed on the weights. If an input pattern is in the unseen class, then no change is made to the weights.

In order to perform a classification, the activation is first calculated

$$a(\mathbf{x}) = \sum_i w_i x_i - \phi \tag{7.16}$$

where ϕ is the bias. The activation is then passed through the step function

$$f(a) = \begin{cases} 1 & a \geqslant 0 \\ 0 & a < 0. \end{cases} \tag{7.17}$$

The linear thresholder can only separate patterns that are linearly separable. The decision between attributing a given input vector to class C_1 or to class C_2 occurs when $a(\mathbf{x}) = 0$. This criterion is satisfied by an $\Omega - 1$ dimensional hyperplane within input space. This surface is the decision boundary of the classification. The bias ϕ determines the displacement of the decision boundary from the origin.

To understand how the decision boundary relates to the weight vector $\mathbf{w}$ giving rise to it, consider two locations on the decision boundary $\mathbf{x}_1$ and $\mathbf{x}_2$. Since $a(\mathbf{x}_1) = a(\mathbf{x}_2) = 0$ it is the case that $\mathbf{w}^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_2) = 0$ which can only be satisfied if the displacement from $\mathbf{x}_1$ to $\mathbf{x}_2$ is orthogonal to the weight vector (Bishop, 2006). Thus, components of the weight vector associated with some decision boundary are identical to the components of the normal to the decision boundary in input space.

In general, linear threshold units are permitted to have negative weights, although biological synapses cannot contain negative conductances. Inhibitory synapses contain positive conductances for hyperpolarising currents and these could be considered as being negative weights. However no known synapse is capable of being both excitatory and inhibitory as would be implied if we were to permit negative weights in the linear thresholder. In the present study we shall neglect inhibition and assume that only excitatory synapses are permitted. This is a necessary assumption because the synaptic model under investigation is a model of excitatory synapses.

7.4.1 Pattern storage

Allowing only positive weights limits the behavior of the linear threshold unit because now there are many decision boundaries that cannot be constructed (any hyperplane that implies a negative weight component). The linear thresholder without negative weights can only score patterns upon the basis of their total input (i.e. the total number of bits for which the weight and the input are both 1).

For each pattern to be stored, a decision boundary can be constructed by adjusting the weights to be equal to the pattern vector itself. This amounts to directing the norm

of the decision boundary toward the target pattern in input space, thus constructing a plane that is perpendicular to the line between the origin and the target pattern. Weights in the state based models are stochastic and so rather than directly setting each weight to be deterministically zero or one, the probabilities that the weights are zero or one are adjusted.

To match the pattern to be stored, some weights are potentiated with early or late LTP, and some are depressed with early or late LTD applied for a duration of 600s (causing the maximum possible potentiation or depression of the weights). The scheme here leads to operation of the linear thresholder that is similar to that of a simple matched filter (Turin, 1960).

The inputs to the perceptron are binary patterns drawn from three possible groups, learned patterns $\mathbf{x}_i^{(L)}$, unlearned patterns $\mathbf{x}_j^{(U)}$ and test patterns $\mathbf{x}_k^{(T)}$. The learned patterns are patterns that have previously been stored, unlearned patterns are patterns that have not been stored and test patterns are patterns that we wish the recognition unit to classify. The mean values of these patterns is always equal to the mean weight $\langle \mathbf{x}_i^{(L)} \rangle = \langle \mathbf{x}_j^{(U)} \rangle = \langle \mathbf{x}_k^{(T)} \rangle = \langle \mathbf{w} \rangle$, i.e. LTP and LTD are always balanced. The weights of the perceptron $\mathbf{w}$ are binary and are set to one with probability $\mathbf{p}$. For the balanced models in this thesis $\langle w \rangle = 0.5$ prior to the storage of any patterns. Each pattern in the unlearned group $\mathbf{x}_i^{(U)}$ has inputs that are 0 and inputs that are 1. Therefore some of the weights $\mathbf{w}$, those that fall within the LTP group, must be potentiated such that $w \rightarrow 1$. The other group of weights are in an LTD group and must be depressed, $w \rightarrow 0$. This is done simultaneously so that on average the mean weight of the whole ensemble, including both LTP and LTD weights, remains $\langle w \rangle = 0.5$. When multiple random patterns are stored, there are unique potentiation and depression groups for each pattern. Thus some weights will first undergo LTP, then perhaps LTD, then LTD again, and so on. For random patterns, as the number of inputs increases, the number of unique histories required rapidly converges on the number of inputs. For example, storing many 100 input patterns at random requires that 100 unique weight trajectories are calculated (at least it is very improbable indeed that two trajectories would happen to be identical as the number of patterns stored becomes large). Here, just two patterns are stored, giving four possible individual weight trajectories $\langle w(t) \rangle$, Fig. 7.4B.

7.4.2 Pattern recognition

The distributions of the total number of inputs *and* weights that are coincidentally 1 are binomial distributions of mean $\Omega\left\langle \mathbf{x}_i^{(L)}\mathbf{w}\right\rangle$ for learned patterns, $\Omega\left\langle \mathbf{x}_j^{(U)}\mathbf{w}\right\rangle$ for unlearned patterns and $\Omega\left\langle \mathbf{x}_k^{(T)}\mathbf{w}\right\rangle$ for test patterns. These quantities are identical in form to the mean of the signal distribution in chapter 3 and of course $\Omega\left\langle \mathbf{x}_i^{(L)}\mathbf{w}\right\rangle$is the mean of the signal distribution for a learned pattern in this case, but for the time being we consider the probability that the output unit is 1.

The probability that the unit gives an output $O = 1$ is the probability that the input exceeds threshold, $\sum_i w_i x_i > \phi$. Since the inputs and weights are discrete this amounts to the probability that some minimum number of weights and their corresponding inputs are both 1. This probability is found directly from the binomial distribution. Thus the probability that the threshold of the output unit is exceeded is

$$P(O=1,t) = \sum_{i=b}^{i=\Omega} \frac{\Omega!}{(\Omega-i)!i!}[\left\langle \mathbf{x}_i^{(\Gamma)}\mathbf{w}(t)\right\rangle]^i[1-\left\langle \mathbf{x}_i^{(\Gamma)}\mathbf{w}(t)\right\rangle]^{\Omega-i} \tag{7.18}$$

where $\Gamma \in \{L,U,T\}$ provides the probability that the unit fires in response to a learned pattern, an unlearned pattern or a test pattern respectively, and b is the minimum number of inputs and weights that must both be 1 in order for the unit to output 1. Note that $P(O=1,t)$ is time dependent due to the weight decay in the state based model used to calculate the weight trajectories $\langle w(t)\rangle$. The optimum value of b was set by plotting a receiver operator curve for the perceptron and minimising the total number of errors such that false positives equals the number of false negatives (Fawcett, 2006).

7.4.3 Super-imposing two patterns

Patterns can be stored on more than one timescale in the 4 and 8 state models. When a pattern is first stored using late LTP/D another pattern can be stored over the top of the first pattern using early LTP/D. This is in contrast to the 2 state model where patterns can only be stored on one timescale. Initially the synaptic weights are random at $t = t_0$, Fig. 7.4B+C. The first pattern, pattern A is stored at $t = t_1$ by applying late LTP/D to the synapses. Pattern B is stored at $t = t_2$, but this time weights are depressed or potentiated by applying early LTP/D. Since pattern B was stored with early phase transitions it decays away after approximately 5000s and the weights appear disordered again for a short time, $t = t_3$ (when the probability that they are potentiated is around 0.5). Eventually at $t = t_4$ pattern A returns and the perceptron responds clearly upon presentation of the original pattern A, but not to pattern B.

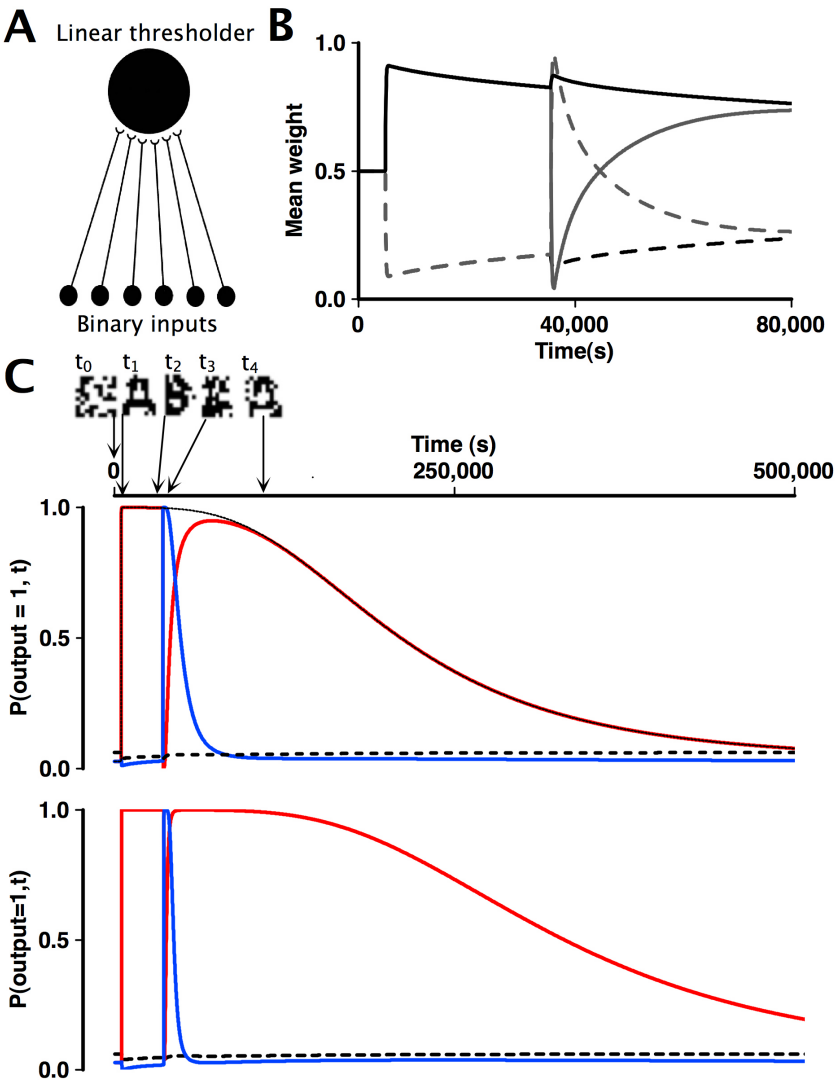

Figure 7.4: Pattern 'A' (red) is stored on the long timescale. Pattern 'B' (blue) is stored on the short timescale. A: The synapses use either the 4 state or the 8 state model. B: The weights implement the state based models of LTP/D and the patterns are stored with eLTP or lLTP for a duration of 500s. Each of the 90 synapses follows one of four possible histories. Solid black line: Late LTP followed by early LTP. Black dashed line: Late LTD followed by early LTP. Solid grey line: Late LTP followed by early LTD. Dashed grey line: Late LTD followed by early LTP. C: (Top) probability that the recognition unit with the 8 state model, recognises the stored patterns. Boxes contain diagrams of the values of the weights at several points of the 8 state simulation. Initially the weights are random. Pattern A is then stored in the weights with late LTP/D. After some time pattern B is stored in the weights using early LTP/D. Some time later pattern B has decayed. Finally there is a recovery of pattern A. On the plot is the response of the unit to random patterns (labeled 'noise'). The thin line is the response of the unit to pattern A in the event that pattern B is not stored. (Bottom) as C but for the 4 state model.

Each synapse has a trajectory that falls into one of four groups, Fig. 7.4A. Before any pattern is stored, the thresholder only very rarely fires for either of the patterns A or B or for patterns drawn at random, Fig. 7.4B+C. (Recall that patterns A and B are statistically identical to the random ensemble themselves.)

These results demonstrate that recognition of pattern B could be stored on top of the prior recognition of pattern A. This causes temporary disruption of pattern A, however on the long timescale, recognition of pattern A returns. This is a result of the ability to superimpose early LTP on late LTP.

7.5 Synaptic overload

The example in the previous section raises the possibility that patterns could be sequentially stored in the synaptic weights on a short timescale, while a previous pattern stored with more slowly decaying late phase plasticity is retained but partially obscured. The possibility that early phase LTP might allow the storage of a memory trace on a short timescale superimposed over a memory trace stored on a long timescale (i.e. stored with late LTP) has been suggested in non quantitative terms before (Morris et al., 2003; Rolls, 1996). Storage of memories with weights using dual timescales has also been shown to improve storage capacity in Willshaw networks (Gardner-Medwin, 1989). In this section, superposition of memory traces is demonstrated with state based models of LTP/D and the perceptron. We are still considering recognition, but we evaluate the case where many patterns are stored at random in an ongoing manner. In this section we return to signal to noise analysis.

In this section the SNR is calculated directly from the simulations of the synaptic weights. This is in contrast to §7.2 where the SNR is calculated from the transition matrix of the model via the autocorrelation. The reason for this difference is that in this section we are specifying the patterns to be stored in the weights and then driving the weights strongly toward those values with LTP/D, Fig. 7.1C, where as §7.2 assumed that unspecified patterns were stored in the stochastic fluctuations of the weights, Fig. 7.1A+B.

The signal to noise for binary weights is (chapter 3)

$$\frac{S}{N} = \frac{\Omega \left| \langle \mathbf{x}_Y \mathbf{w}(t) \rangle - \langle \mathbf{x}_N \rangle \langle \mathbf{w}(t) \rangle \right|}{\sqrt{\frac{1}{2}(\sigma_N^2(t) + \sigma_Y^2(t))}} \tag{7.19}$$

where $\mathbf{w}(t)$ is the synaptic weight vector, $\mathbf{x}_Y$ is the pattern to be classified as seen or unseen, $\langle \mathbf{x_N} \rangle$ is the mean of the unstored (noise) patterns. The simulations proceed

in an analogous way to §7.4.1: The first pattern to be stored is drawn at random and is stored with late LTP/D by driving the weights towards the values of the inputs (0 or 1). This produces a large increase in the SNR for that pattern. After some time interval drawn from an exponential distribution with mean μ_s another pattern is drawn at random and stored in the weights with early phase LTP/D. The SNR for this pattern now increases by a large amount, but decays away more quickly than for the first pattern. Next, another pattern is drawn at random and stored with early LTP/D. All subsequent patterns are stored at random intervals using early LTP/D.

7.5.1 Ongoing storage of patterns with early LTP

Random patterns $\mathbf{x}$ having $\langle \mathbf{x} \rangle = 0.5$ are stored at random intervals. The LTP/D induction time (i.e. the duration over which the protocol matrix is applied) was 300s. Reasonable choices of induction time do not strongly effect the conclusions. This timescale is chosen to be compatible with the induction times seen in experiments. The first pattern was stored with late LTP/D and all subsequent patterns are stored with early LTP/D (this assumption is removed in §7.5.2). 60 patterns were stored in a perceptron with 100 inputs. Each weight has an individual weight history, two examples are plotted in Fig. 7.5A. Storage of the first pattern leads to a slowly decaying deflection in the weight due to the fact that it was stored with late phase transitions. Subsequent patterns were stored with early LTP/D leading to more rapidly decaying deflections to the weight, Fig. 7.5A. The different timescales of pattern decay in the weights have a direct impact upon the dynamics of the SNR. We see that the SNR for the response of the recognition unit to pattern 1, stored with late phase transitions, decays slowly in comparison to the SNR for a subsequent pattern stored with early phase transitions Fig. 7.5B.

The average SNR over 6 isolated perceptrons for response to the initial pattern stored with late LTP/D, and storage intervals between subsequent patterns of $\mu_s = 7500s$ and $\mu_s = 22500$, is shown in Fig. 7.5C in the 2, 4 and 8 state cases. In the 2 state case (thin black line) the initial pattern is rapidly lost and the SNR falls below 1 in less than $5000s$ (84 mins). In the 4 and 8 state models (heavy grey and black lines respectively), despite the disruption to the weights engendered by ongoing storage of patterns, the average SNR for the initial pattern is elevated above 1 for 55hrs when $\mu_s = 22500s$. Further spacing of the storage of early phase patterns by increasing μ_s does not lengthen the decay of the initial pattern. Once the early phase patterns are

sufficiently spaced out such that the decay of the initial pattern is similar to the decay time in the case that no early phase patterns are stored, then there can be no further increase in the decay time of the initial pattern. This limit is reached when the decay time of the initial pattern is comparable to the decay time of late LTP itself. This makes intuitive sense, because it is the timescale of intrinsic decay of late LTP that ultimately limits the evolution of the mean of the signal distribution.

Now we consider patterns stored after the first pattern using early LTP. In the 2 state case the SNR of the patterns overlaps completely with the 4 state case, Fig. 7.5D. In the 4 and 8 state cases we see that in addition to retention of the first pattern (stored with late LTP, 7.5C) there is also a recognition signal for subsequent patterns (stored with early LTP), Fig. 7.5D. This demonstrates that indeed a response to both the initial pattern and newly stored patterns can be obtained simultaneously on average in the 4 and 8 state models, but not in the 2 state model. Thus one pattern has been stored on the late LTP timescale, while other patterns can be stored on the shorter early LTP timescale in an ongoing manner. The signal of the early phase patterns remains above 1 for 1050s (17.5 mins) when $\mu_s = 7500$ and 2250s (37.5 mins) when $\mu_s = 22500$. Over the duration taken for the SNR of the initial pattern to fall to 1, around 20 sequential temporary patterns can be stored and subsequently decay. Finally, note that these simulations use only $\Omega = 100$ synapses. The time taken for the signal to fall to 1 increases as $ln\sqrt{\Omega}$.

7.5.2 Sparse coding and synaptic overload

Synaptic overload might be useful for memory storage because it allows the simultaneous storage of patterns on a long timescale and patterns on a short timescale [5]. This increases storage capacity as compared to a 2 state model having only one available timescale. However, in the last section it was assumed that the initial pattern storage is through late LTP but that all subsequent events are mediated by early LTP. Unfortunately, if the second pattern is not stored with early LTP, but rather with late LTP, then the initial pattern will be instantaneously wiped out. Furthermore, in order to obtain the maximum decay time of the initial memory of 55hrs, the average spacing between

[5] Why is this useful? You may wish to remember where you put your cup of tea without permanently erasing your memory of an important meeting that morning. The memory of the meeting needs to be retained for longer than the memory about the cup of tea. But although the cup of tea memory only needs to be retained for 20mins (or for as long as it takes to drink it), it has to be formed close to the time that you put your cup down. The memory of the meeting needs to be retained for long enough that it can be consolidated at the systems level, but does not nessecarily have to be perfectly accessible as you drink your tea.

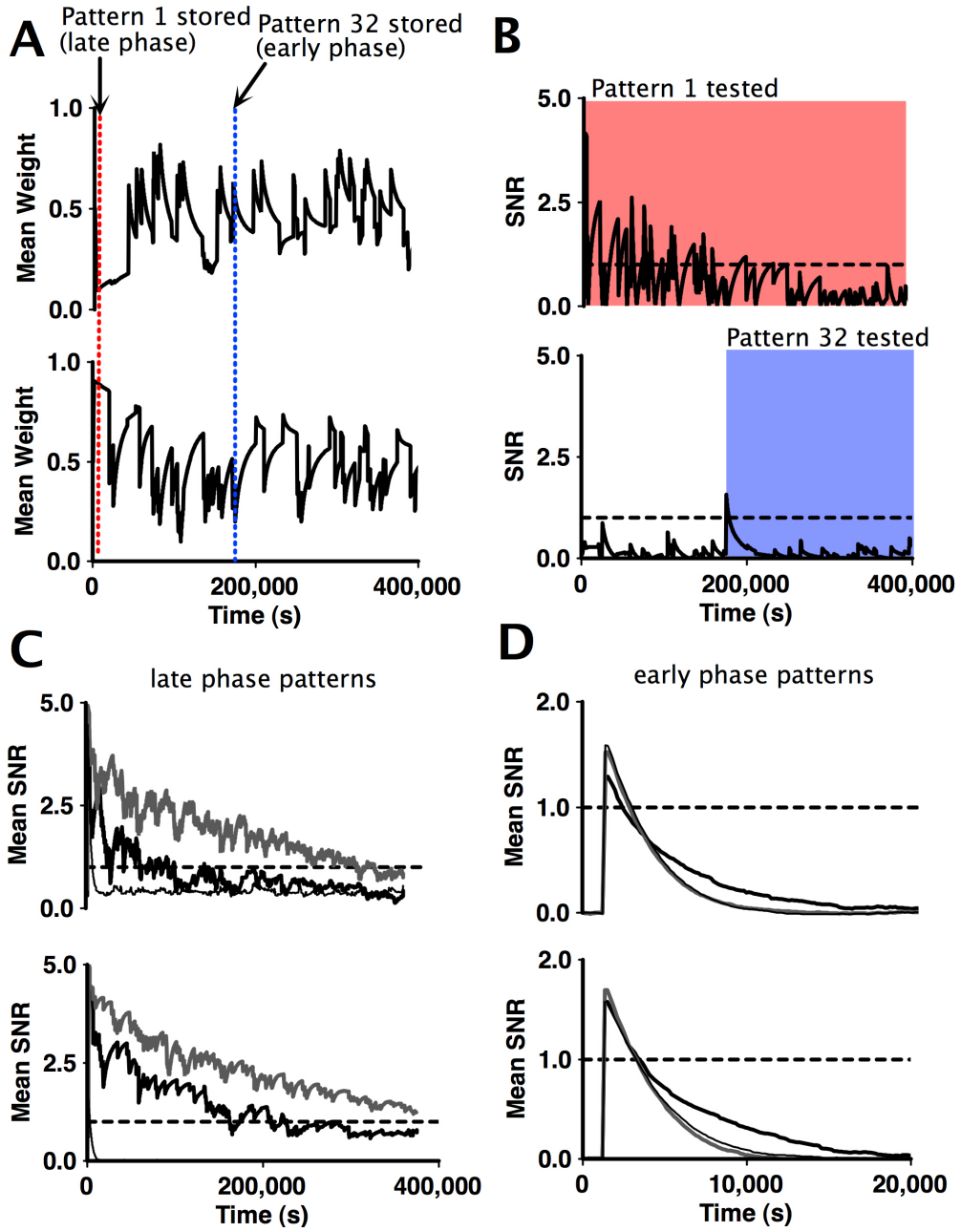

Figure 7.5: Patterns stored at random using the recognition unit. A: Two example synaptic weight histories from the 8 state model. The synaptic group initially undergoes lLTP (top) or lLTD (bottom). Subsequent patterns are stored with early phase transitions. The 4 state model is similar. B: The SNR of two stored patterns in the 8 state model. The top graph is the SNR for the response of the unit to the first pattern, stored with late phase LTP/D. The second graph shows the SNR in the case that a subsequent pattern (pattern 32) is stored with early phase LTP/D. The early phase SNR decays more rapidly than in the case that the pattern was stored with a late phase transition. The dashed line indicates SNR=1. C: The average of the SNR when $\mu_s = 7500s$ (top panel) and $\mu_s = 22500s$ (lower panel). The black line indicates the 8 state model, the grey line indicates the 4 state model. The thin black line is the 2 state model. D: As C but the average SNR is for patterns that were stored with early phase transitions for the two values of the mean storage interval.

storage events has to be in excess of 20000s. For memory to be useful, the spacing between storage events should ideally be much shorter. In this section we see how both of these problems are reduced in the limit of sparse coding.

7.5.2.1 Sparse codes

The coding fraction of an assembly of neurons f is the ratio of the number of neurons that are active during some information processing step to the total number of neurons (Barlow, 1989; Laurent, 2002; Földiák, 2002; Földiák and Endres, 2008). If the coding fraction is high, for instance 0.5, then the code is a dense code. Dense codes represent information using the combinatorial activity of many neurons simultaneously. On the other hand, a low coding fraction, implies a sparse code. Sparse codes represent information using the activity of only a small fraction of the population of available neurons. In the limit of very low coding ratios, such that only individual neurons are active, the code is local. The notion of the 'grandmother cell', a cell that is only active when one recognises one's grandmother, is the classic example of local coding.

The hippocampus, early visual system and cerebellum are all brain organs that have been hypothesised as coding incident information in a sparse code (Marr, 1971; Willshaw and Buckingham, 1990; Olshausen and Field, 1996). In caricature, this is achieved by re-representing the incoming activity within a larger population of neurons than at the initial stage [6]. Although sparse codes are less efficient than dense codes - X binary neurons can represent 2^X codewords in the limit of a dense code, but only X codewords in the limit of local codes - they prove to be far easier to use for the purposes of unsupervised learning and processing (Olshausen and Field, 1997). Furthermore, since the number of codewords increases exponentially with the coding ratio, sparse codes do not need to have high coding ratios in order to gain significant increases in representational capacity as compared to a local code.

Sparse codes have been shown to be highly beneficial to the preservation of memory traces because they can minimise the overlap between modified synapses (Dayan and Willshaw, 1991; Leibold and Kempter, 2008). Sparse coding applies to memory traces when the number of synapses involved in the memory trace is much less than the total available. In the treatment of the state based models of LTP in chapter 6 it was assumed that all synapses participate in the memory trace. It is this assumption

[6] There is much more to it than this. For the code to useful we must also ensure that it represents as much of the input information as possible, i.e. the code should be as complete as possible. However this does not impact on this discussion.

that gives rise to the logarithmic dependence of the maximum memory lifetime on the number of synapses in the ensemble. As was discussed in chapter 2, adding more synapses to the linear bounded, multistate and 2 state models increases the initial signal, but this only logarithmically increases the maximum memory lifetime (Ben Dayan Rubin and Fusi, 2007). If we permit the memory trace to be sparse, this is no longer the case (Amit and Fusi, 1994). This can be understood by seeing that the probability that a synapse is modified is proportional to the coding ratio. A low coding ratio implies a long timescale of decay of the memory trace. Therefore, adding more synapses in this case directly increases the timescale of decay itself and therefore breaks the logarithmic dependence of memory lifetime on the number of synapses.

Recently the interaction between sparse coding and state based synaptic models of synaptic plasticity, such as the multistate and cascade models has been explored (Leibold and Kempter, 2008). It was found that the linear dependence of the maximum memory lifetime upon the number of synapses allowed by the sparse code, outweighed the sublinear power-law dependence provided by the cascade model. Furthermore mixing state based synapses with sparse codes merely leads to a reduction in the efficacy of the sparse code. Sparse codes imply that relatively few synapses are activated upon recall. Hence, under sparse coding complex state based synapses perform poorly relative to the simplest 2 state model, because the initial signal scales less favorably with the number of synapses. This initial signal reduction occurs because all of the available states are more thinly spread in phase space. An example of this is the multistate model examined in chapter 2, where adding states increased the memory lifetime, but 'diluted' the initial signal.

The conflict between synapses with complex state diagrams and sparse coding results from constraining the synaptic population to be at equilibrium. If the memory can be stored away from equilibrium however, then the initial SNR can be more steeply proportional to the square root of the number of synapses, i.e. LTP events can tend to drive synapses toward saturation Fig. 7.5A. Another way of stating this is that away from equilibrium, by definition, we can choose where in state space the synapses lie. Therefore we can choose to place the synapses in a location that gives a high initial signal.

7.5.2.2 Sparse codes improve the performance of synaptic overload

In this section the effect of sparse coding on the performance of synaptic overload in the 4 and 8 state models of LTP is considered. When a target pattern is stored, a

randomly selected fraction of f synapses undergoes plasticity, Fig. 7.6A. This is in contrast to §7.5.1, where all synapses underwent plasticity to store patterns. We shall find that this improves the performance of synaptic overload such that the lifetime of patterns stored with early and late phase plasticity approaches the lifetimes of eLTP and lLTP in a single trial, despite rapid ongoing storage of other patterns.

To asses the improvement in memory storage as sparsity is increased, we first store a pattern with saturated late LTP. Subsequently, patterns are stored at intervals drawn from a exponential distribution having a mean value of $\mu_s = 750s$, where 1% of the patterns are late phase and all other patterns are stored with early phase plasticity. This is an order of magnitude shorter than the interval between patterns in §7.5.1. The induction time of the plasticity used to store subsequent patterns was 300s, and so plasticity was not saturated[7]. The time taken for the SNR of the initial stored pattern to reach one is the measure of memory trace survival time.

The initial SNR of patterns stored with early and late LTP in the state based models scales as $\sqrt{\Omega}$, as we would expect from the definition of the SNR (chapter 2). Storage of a pattern with saturated LTP/D and no sparse coding leads to a steep scaling of the initial signal to noise with the number of synapses,

$$SNR_0 = \frac{\sqrt{\Omega}(p_{in} - p_{in}^2)}{\sqrt{p_{in}(1-p_{in})}} \tag{7.20}$$

where p_{in} is the probability that a bit is on in the input patterns and noise patterns, Fig. 7.6B where $p_{in} = 0.5$. This steep scaling occurs because the saturated plasticity causes the strongest memory trace possible. When plasticity is not saturated the scaling is not so steep and depends upon the level of saturation, Fig. 7.6B (black open circles). This was the case for patterns stored after the first pattern.

Increasing the sparsity of the patterns (where sparsity is $(1-f)$) leads to a decrease in the SNR. In the case of saturated LTP/D (as is the case for the first pattern stored here), the relationship between the signal to noise and sparsity is approximated by

$$SNR_0 = (af + b)\sqrt{f\Omega} \tag{7.21}$$

where $a = 0.4$ and $b = 0.1$ are constants that were found by fitting a line to $SNR_0/\sqrt{f\Omega}$ for 50 values of f between 0 and 1 with $\Omega = 200$, Fig. 7.6C (sold black circles). Patterns stored with non-saturated LTP/D respond identically, but have a uniformly reduced SNR (open black circles).

[7] If plasticity is saturated the results presented here still hold, however a linearly greater degree of sparsification is required.

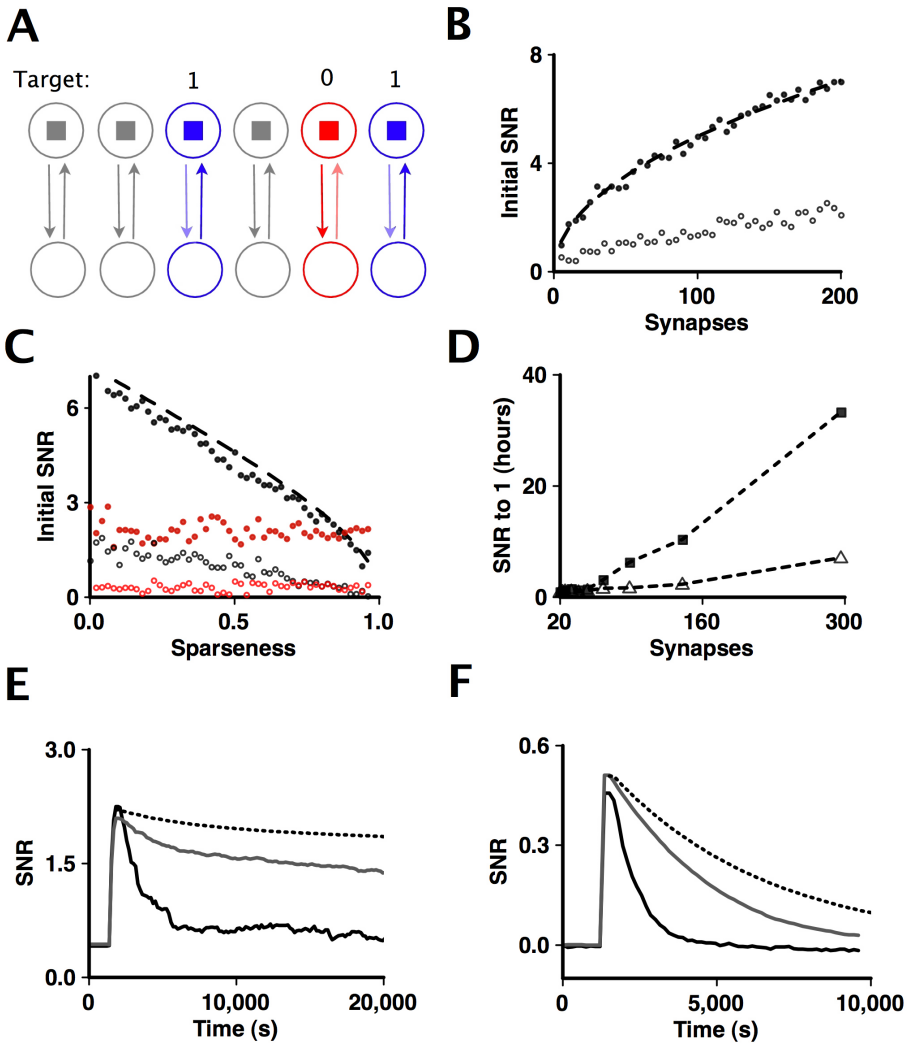

Figure 7.6: Scaling of the memory lifetime with sparse coding. A: Sparse coding uses only a fraction of the synapses. Grey synapses represent those not used by the storage of the target pattern: 1,0,1. B: The initial SNR of patterns stored. Solid black circles are patterns stored with saturated plasticity. The dashed line is $SNR_0 = 0.5\sqrt{\Omega}$, Eq. (7.20). Open circles are unsaturated. C: SNR_0 decreases with sparsity $(1-f)$. Black filled circles are saturated patterns and black open circles are unsaturated. The dashed line is $SNR_0 = (0.4f - 0.1)\sqrt{f\Omega}$, Eq. (7.21). The initial signal can be held constant by increasing the number of synapses Ω: Filled red circles for saturated patterns, open red circles for unsaturated patterns. D: The initial signal is held constant and the sparseness is increased. The time for SNR of the first pattern to reach 1 increases for both the 4 state (squares) and 8 state (triangles) model. E: Examples of the average SNR as a function of time for the first pattern stored with late phase plasticity in the 4 state model for $\mu_s = 750s$. Black line is the case where sparseness=0 (20 synapses). Grey line is the case where sparseness=0.9 (296 synapses). Dashed line is the case where only one pattern is stored (i.e. the unperturbed case) but the initial signal is identical. F: As E, but for the early phase patterns. Precisely the same phenomenon occurs in the 8 state model, but the increase in the time for the SNR to reach 1 scales more gradually with sparsity.

We have seen that the initial signal of a pattern stored increases with an increasing number of synapses but decreases with increasing sparseness. Thus it is possible to compensate for the reduced signal due to sparseness by increasing the total number of synapses in the system. This can be achieved by scaling the synapses Ω' such that,

$$\Omega' = \frac{S_0^2}{(af-b)^2 f} \tag{7.22}$$

which follows from rearrangement of Eq.(7.21), with S_0 the desired initial signal of the memory. Now the initial signal is independent of the sparseness, Fig. 7.6C (solid red circles). Patterns stored with non-saturated plasticity also have a constant (although smaller) SNR when the compensation is applied, Fig. 7.6C (open red circles).

Increasing the sparsity of the patterns while compensating the number of synapses such that the initial signal remains fixed leads to an increase in the survival time of the first pattern, Fig. 7.6D+E in both the 4 and 8 state models and an increase in the survival time of the patterns stored with early phase LTP, Fig. 7.6F[8]. Each data point in Fig. 7.6C is the time for the SNR to reach 1 for the first pattern stored averaged across 20 simulations where each point is a separate sparsity of $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ respectively, where the number of synapses was determined by the condition that the initial signal be constant and was $\Omega' = 20$ for $f = 1$, Eq.(7.22).

The increased survival time in Fig. 7.6D results from the decreasing interference between memory traces due to the sparse code, Fig. 7.6E+F. The marked increase in the memory lifetime seen in Fig. 7.6D only continues until the lifetime of the sparsely coded trace approaches the intrinsic decay time of early and late LTP/D, Fig. 7.6E+F (dashed lines). After this point the lifetime of the memory trace again scales as the logarithm of the number of synapses (chapter 2) due to intrinsic synaptic decay.

Sparse coding allows the survival time of patterns stored with late and early LTP/D to be increased such that they decay on the maximum possible timescale of the intrinsic decay of the LTP/D. This allows synaptic overload, even when the storage interval is small and when some subsequent patterns are late phase (1% in this case). The compensation regime above allows the interval between pattern storage to be arbitrarily small and the initial signal to be arbitrarily large as long as sufficient synapses are

[8]For this to occur in the 8 state model, the early phase plasticity protocols must be slightly adjusted from the form quoted in chapter 6. In chapter 6 early LTP(D) only affects transition rates in the front (rear) ring. Specifically LTP leads to the elevation of $\{R_{12}^{eLTP}, R_{23}^{eLTP}\}$ and the reduction of R_{21}^{eLTP} in the front ring. For overload to function when the interval between pattern storage $<$ early LTP/D decay time, the rates in the rear ring $\{R_{58}^{eLTP}, R_{87}^{eLTP}\}$ must be elevated and R_{78}^{eLTP} must be lowered. This step of making the rear ring symmetric with the front ring during LTP induction must also be applied to early LTD. This has no effect on any of the results quoted previously where the plasticity was never applied with an interval below that of the early LTP/D decay time.

available. This means that in principle an initial pattern, stored with late phase plasticity can survive for nearly as long as the plasticity itself even while other patterns are stored in the same synaptic population in an ongoing fashion. We have seen this provides the possibility of storing many short lived patterns temporarily on the timescale of early LTP/D, while retaining an initial pattern stored with late phase plasticity.

7.6 Discussion

In this chapter retrograde and anterograde amnesia by saturation of synaptic weights was reproduced in the 4 and 8 state models of LTP/D. It was found that previous memory traces are disrupted both by induction of early and late LTP in both models. The degree of disruption however is dependent upon the degree of saturation of the weights and whether early or late phase LTP is induced. Anterograde amnesia of the correlation with a pattern stored just after plasticity induction was observed. In this case correlations decay away more rapidly than they do in the steady state, because the memory traces are stored in an unstable state of the system. The rate of decay of newly stored correlations increases as the degree of saturation of the weights increases and is similar regardless of whether early or late LTP was previously induced. However since the initial signal can be increased when a memory is stored away from equilibrium, the signal to noise of the memory traces is not always harmed in the anterograde case.

Interestingly, there are differences in the dynamics of retrograde amnesia in both the 4 and 8 state models depending upon whether early or late LTP is induced. If late LTP is induced then the memory trace is permanently degraded until, when maximum saturation of the weights is achieved, the memory trace is completely obliterated. If instead early LTP is induced then the degradation of the memory trace recovers to the level of SNR that would have existed if the intervention had never taken place, even when maximum saturation is induced. This recovery occurs on a timescale that is identical to the timescale of decay of early LTP.

The above result supports the idea that different memory traces might exist on differing timescales within one synaptic population by virtue of the different phases of LTP (Gardner-Medwin, 1989; Rolls, 1996; Morris et al., 2003). To test this, recognition memories were stored using a Perceptron. It was found that providing that long term traces, stored with late phase LTP/D, are stored relatively infrequently, recognition memories survive while new memories are superimposed upon them with early phase LTP/D. Thus the perceptron can, on average, perform recognition of both the

original stored pattern and new stored patterns. For a single perceptron, this occurs at the cost of temporary disruption of recall of the initial long term trace. This could provide a mechanism by which neural ensembles in the hippocampus meet the need of having both long lasting traces (those awaiting consolidation) and strong immediate traces (of non-specific incoming information, i.e. 'automatic recording'). It was found that for this to be feasible with the models proposed in this thesis, the coding of the memory trace would have to be sparse. If this is not the case, a useful rate of memory storage and a long time period of retention of memories stored with late phase synaptic plasticity cannot be achieved with these models.

The results obtained with the state based models in this chapter lend further credence to the idea that disrupting the synaptic weights in the hippocampus should damage memory (Moser and Moser, 1999) if those memories are dependent upon correlations amongst synapses. However, the results also indicate that the signal to noise ratio is not necessarily harmed by LTP induction, because the initial state can in principle lead to a higher signal. In experimental terms this means that although the hippocampal synapses are saturated by LTP, this does not prevent them from *changing*, and in fact may even facilitate changes. It should be noted that the imperviousness of the SNR to anterograde amnesia in this chapter results from the definition of the SNR, in which depression transitions are counted as contributing to the signal (i.e. the absolute change is important). If this definition is changed such that, for example only potentiation events are viewed as contributing to the memory signal, then the SNR would be harmed by saturation of LTP. In this case however, the SNR would be augmented by saturation of LTD. Thus if only potentiation events were important to the initial signal this would predict that saturating LTD should improve memory.

The state based models add the theoretical development, outlined above, that the timescale of decay of induced LTP could affect the dynamics of retrograde amnesia. This predicts that if weights could be saturated in vivo, but in a manner where two distinct timescales of decay of that saturation are visible, early LTP and late LTP like timescales, then this should have an impact on the behaving animal. In such circumstances we would expect that saturation of early LTP should allow the recovery of previously stored hippocampal memory traces.

An extension to the above point follows from the result obtained in the perceptron model. As described above, the 4 and 8 state models predict that multi-timescale synaptic dynamics should permit synaptic overloading. If this were the case in-vivo, it should be possible to cause reversible retrograde amnesia of a previous hippocampal

memory trace that was formed with protein synthesis dependent synaptic modification, with some sudden intense learning. Observation of this new effect would hinge on two factors: 1) It should somehow be ensured that the overlap in the population of synapses storing both the old and new memory is high, 2) it must be ensured that the new trace from the new learning is not itself composed of protein synthesis dependent synaptic modifications. Point 1. might be challenging because there is evidence to suggest that the hippocampus sparsifies and orthogonalises memory traces (Guzowski and Knierim, 2004; Bakker et al., 2008). Point 2. might be ensured by blocking protein synthesis pharmacologically.

The state based modeling approach suggests that there might be benefits to multi-timescale synaptic dynamics within the hippocampus in alleviating the plasticity stability dilemma. The cascade model is another state based model that also suggests benefits of multi-timescale dynamics from the point of view of memory trace strength and stability (Fusi, Drew, and Abbott, 2005). However, the state based models described in this thesis and the cascade model are very different. The cascade model is aimed at explaining the forgetting dynamics of familiarity of visual scenes. In this phenomenology, human beings demonstrate a remarkable capacity to recall having seen a picture after a long time elapse even if the picture was only seen on one occasion (Standing, Conezio, and Haber, 1970; Standing, 1973), and was one of an ensemble of many thousands. As discussed in chapter 2, the cascade model suggests that by optimising the steady state distribution of the synapses, such that the maximum memory lifetime is no longer a logarithmic function of the number of synapses, the stationary stability plasticity dilemma can be over come in large enough ensembles of synapses. It is thus suggested that synapses in the visual system might be able to learn and retain very long term one-shot familiarity traces by using an optimisation of their steady state.

The approach in this chapter largely explored the NPS and differs from the approach taken by the models encountered in chapter 2. Firstly there has been no attempt to optimise the steady state in terms of memory lifetime and initial signal. Here the transition rates have been set so as to be compatible with experimental data and the maximum memory lifetime still scales as a logarithm in the number of synapses[9]. Secondly the target system is different: The argument made here is that early LTP/D and late LTP/D allow synapses to superimpose memory traces at differing timescales thus

[9]It is possible that logarithmic scaling is less of a problem in the hippocampus where memory traces can be potentially off loaded to a slow learning (and hence relatively fluctuation resistant) neocortex.

affording the hippocampus with the ability to store episodic memory traces in an ongoing manner, while preserving important memory traces for a sufficient amount of time so as to allow them to be consolidated at the systems level. In chapter 8 it shall be shown that eLTP , lLTP and synaptic tagging could combine to allow the rescue of memory traces that would otherwise decay.

Chapter 8

A state based model of synaptic tagging

Throughout this thesis, the difference between early and late phase LTP has been emphasised. It has been suggested that memory traces might be stored more effectively in synapses with more than one timescale of synaptic plasticity and that this is the reason for the existence of more than one phase of LTP/D. If there is to be any such utility in the distinction between early and late LTP however, then the cell should have some mechanism for correctly targeting synapses that are to be consolidated. If late LTP is protein synthesis dependent and occurs at specific synapses, then how do the required proteins 'know' which synapse to go to? As was described in chapter 1 experiments have shown that plasticity related proteins (PRPs) are manufactured locally to the dendritic branch upon which the synapse is located. The production of PRPs is triggered by the induction of late LTP. The theory of synaptic tagging predicts that synapses that have recently undergone plasticity such as early LTP or late LTP are 'tagged'. The tag allows the PRPs to stabilise that synapse. Since synapses that have undergone early LTP are tagged (but do not trigger the production of PRPs themselves) they can nevertheless be stabilised by PRPs triggered by induction of late LTP in another synapse. Thus early LTP can be converted into late LTP if late LTP is induced in a nearby synapse within some time period. This is the process of synaptic tagging.

It has been suggested that synaptic tagging might have a cognitive corollary (Frey and Morris, 1997; Frey and Morris, 1998): Association between weak stimuli, that only elicit early LTP and strong stimuli that cause late LTP might mediate associations between emotionally 'weak' and emotionally 'strong' experiences. For example one might remember the colour of the shirt worn on the day of PhD thesis hand-in. This

process requires that memory traces that might otherwise have decayed are somehow rescued when a stronger long lasting memory trace is created. In the last chapter we saw that the state based models of LTP allow the superposition of rapidly decaying eLTP memory traces on top of slowly decaying lLTP memory traces. In this chapter it is demonstrated that synaptic tagging could in principle allow the on-line conversion of weak eLTP memory traces into strong lLTP traces.

In this chapter a state based model of synaptic tagging is proposed. The model is used to reproduce the electrophysiological manifestation of synaptic tagging. Next the model is used in conjunction with the perceptron model to show that synaptic tagging can lead the transformation of a fast decaying memory trace signal in to a slower decaying signal. This supports the idea that in principle the combination of early LTP, late LTP and synaptic tagging could account for association of weak stimuli with strong stimuli in episodic memories (flashbulb memories), if we assume that these processes apply to hippocampal memory traces mediating episodic memories.

8.1 The model

The experiments demonstrating synaptic tagging suggest that when a stimulation protocol is applied, there is an ordered sequence of events in time (chapter 1). Firstly the synapse undergoes early LTP and a tag is set, perhaps by virtue of phosphorylation of some synaptic molecule. Next, as the stimulation is continued, late LTP is induced and protein synthesis is engaged. A course grained description can be applied to this sequence by assuming that there are 6 basic stability states: stable depressed, early LTP tagged, late LTP protein synthesis, stable potentiated, early LTD tagged and late LTD protein synthesis, Fig. 8.1. It is again assumed that the synapses are binary. These states are arranged in a ring and application of a late LTP protocol can be considered as driving synapses around this ring from the stable depressed to the stable potentiated state. In this scheme one synapse (during the induction of late phase LTP/D) passes through the PRP state and if the other synapse happens to be in a tagged state (after induction of early LTP/D), this causes it to collapse into the stable potentiated (depressed) state if the synapse is in the potentiated (depressed) tagged state. However if this synapse is not in a tagged state having had no recent early LTP/D then it is unaffected.

It could be argued that a simpler arrangement of states is linear, progressing from depressed $\rightarrow$ tagged $\rightarrow$ protein synthesis $\rightarrow$ potentiated. However the nature of the tag

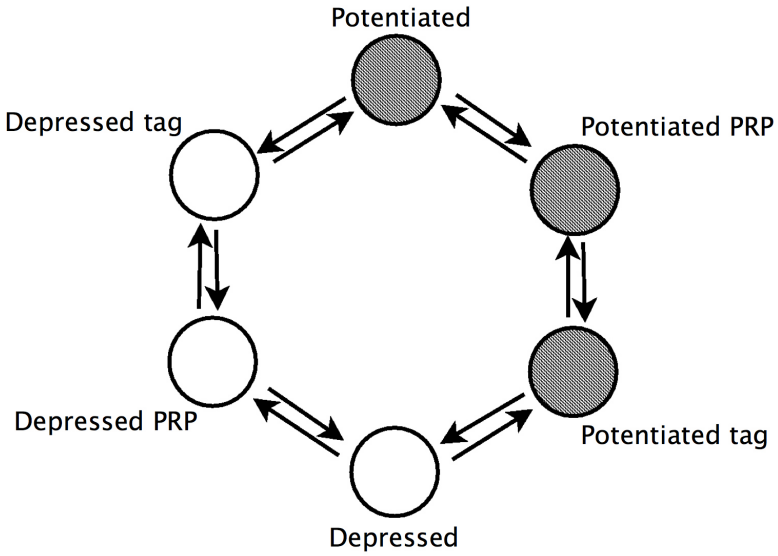

Figure 8.1: Schematic of the arrangement of the stability states in the synaptic tagging model. The filled circles represent states that are potentiated. This structure is motivated by the observation that in going from a depressed state to a potentiated state, a synapse must transition through states of being tagged and activation of protein synthesis.

state determines whether the synapse expresses late LTP or late LTD when plasticity related proteins are encountered. This choice between depression and potentiation in the tagged synapse is independent of whether the late LTP/D synapse is potentiated or depressed (Sajikumar et al., 2005). Hence the potentiated tagged and depression tagged states are distinct.

In the simplest model based on Fig. 8.1, the synaptic weight would depend directly on each of the 6 states. Experiments in which a synaptic tag was set by induction of early LTP but followed by low frequency stimulation, depotentiating the synapse, showed that given sufficient time elapse the tag remained, even though the weight had changed (Sajaykumar and Frey, 2004). For this reason each of the stability states is permitted to be in either a $w = 0$ or $w = 1$ weight state, but the baseline transition rates are arranged such that potentiated synapses (grey circles in Fig. 8.1) spend nearly all of their time in the $w = 1$ state (and visa versa). Thus, for every stability state there are two weight states $\{0, 1\}$, forming a bistable pair, Fig. 8.2A. The stability state in Fig. 8.1 determines the equilibrium probability density over these binary weight states. Therefore the states are arranged as follows: (odd numbers are the $w = 0$ states and even numbers are $w = 1$ states)

- Stable depressed (states $\{1,2\}$): The synapse tends to occupy weight state, $w = 0$. Thus the overall flow of states is from state 2 to state 1, Fig. 8.2A when at equilibrium.

- Potentiated, tagged (states $\{3,4\}$): The synapse has undergone potentiation and is tagged.

- Potentiated, PRP (states $\{5,6\}$): The synapse has undergone potentiation and has been tagged. Protein synthesis has also been activated and PRPs are present.

- Stable potentiated (states $\{7,8\}$): The synapse tends to occupy weight state, 8. Thus the overall flow is from 7 to 8.

- Depressed, tagged (states $\{9,10\}$): The synapse has undergone depression and is tagged.

- Depressed, PRP (states $\{11,12\}$): The synapse has undergone depression, has been tagged and PRPs are present.

At equilibrium the transition rates are configured such that about half of the synapses congregate in the stable depressed state (state 1) and about half congregate in the stable potentiated state (state 8). As was the case in the 8 state model in chapter 5, transition rates R_{ij} from state i to state j, are once again chosen by means of a powerlaw,

$$R_{ij} = \alpha x^{-\gamma} \tag{8.1}$$

for $x \in \{1,2,3,4\}$. The powerlaw is adjusted such that the decay of early LTP and late LTP in the tagging model, match the experimental data, Fig. 8.5 giving $\gamma = 6.1$ and $\alpha = 0.1s^{-1}$. The transition rates thus calculated are inserted into the baseline transition matrix, Fig. 8.3. The highest transition rate thus chosen ($x = 1$) is deemed to be the rate of transition from tagged states to PRP states when a tagged synapse is consolidated via the tagging interaction. This rate must be high in order that the tagging interaction has a significant effect.

Early LTP/D and late LTP/D can be induced in the model. Early LTP is induced by increasing the flow of states in to the potentiated state 2, and by increasing the flow of states toward the tagged states, 3 and 4, Fig. 8.2B. Late LTP is induced in an identical fashion to early LTP, but now states flow in to the tagged states and then to the PRP states, 5 and 6, and from there into the stable potentiated weight states, 7 and 8, Fig.

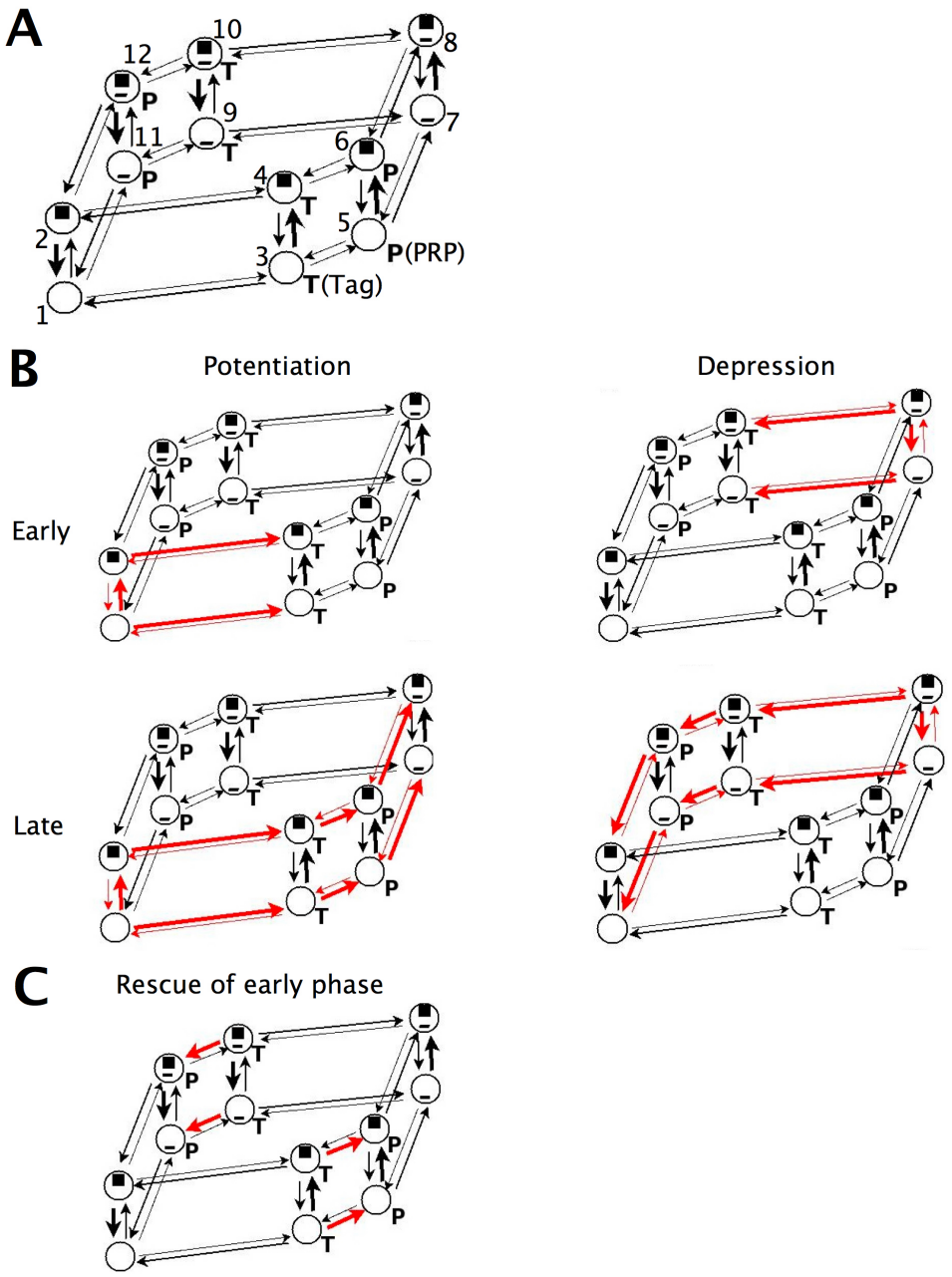

Figure 8.2: Schematic of the 12 state tagging model. A: The model at equilibrium. The magnitude of the transition rate is indicated by the thickness of the arrow. Greater thickness represents larger rates. B: The plasticity induction protocols for early phase and late phase potentiation and depression. Red arrows indicate transition rates that are altered by the induction protocol. C: During a tagging protocol involving two synapses, the transition rate from tagged states to PRP states can be elevated in one synapse if the other synapse is in a PRP state (see text). This diagram illustrates the effect of the tagging interaction upon the transition rates of the tagged synapse.

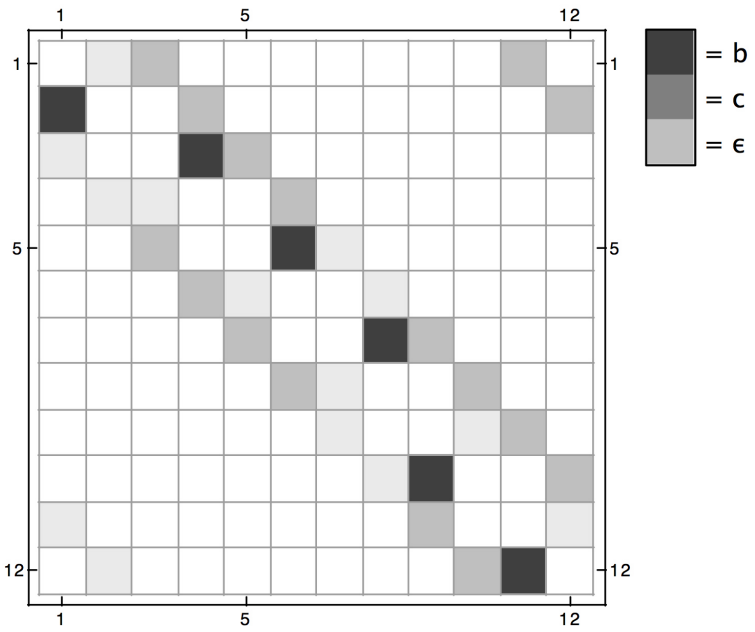

Figure 8.3: The baseline transition matrix of the 12 state model. The grey values represent values chosen from the power law. White indicates that the matrix element is zero. Note that the largest transition rate a chosen from the powerlaw Eq.(8.1) when $x = 1$ does not appear in this matrix because it mediates the tagging interaction and there is no tagging interaction in the single synapse model. It is present in the transition matrix for the full model however, Fig. 8.4.

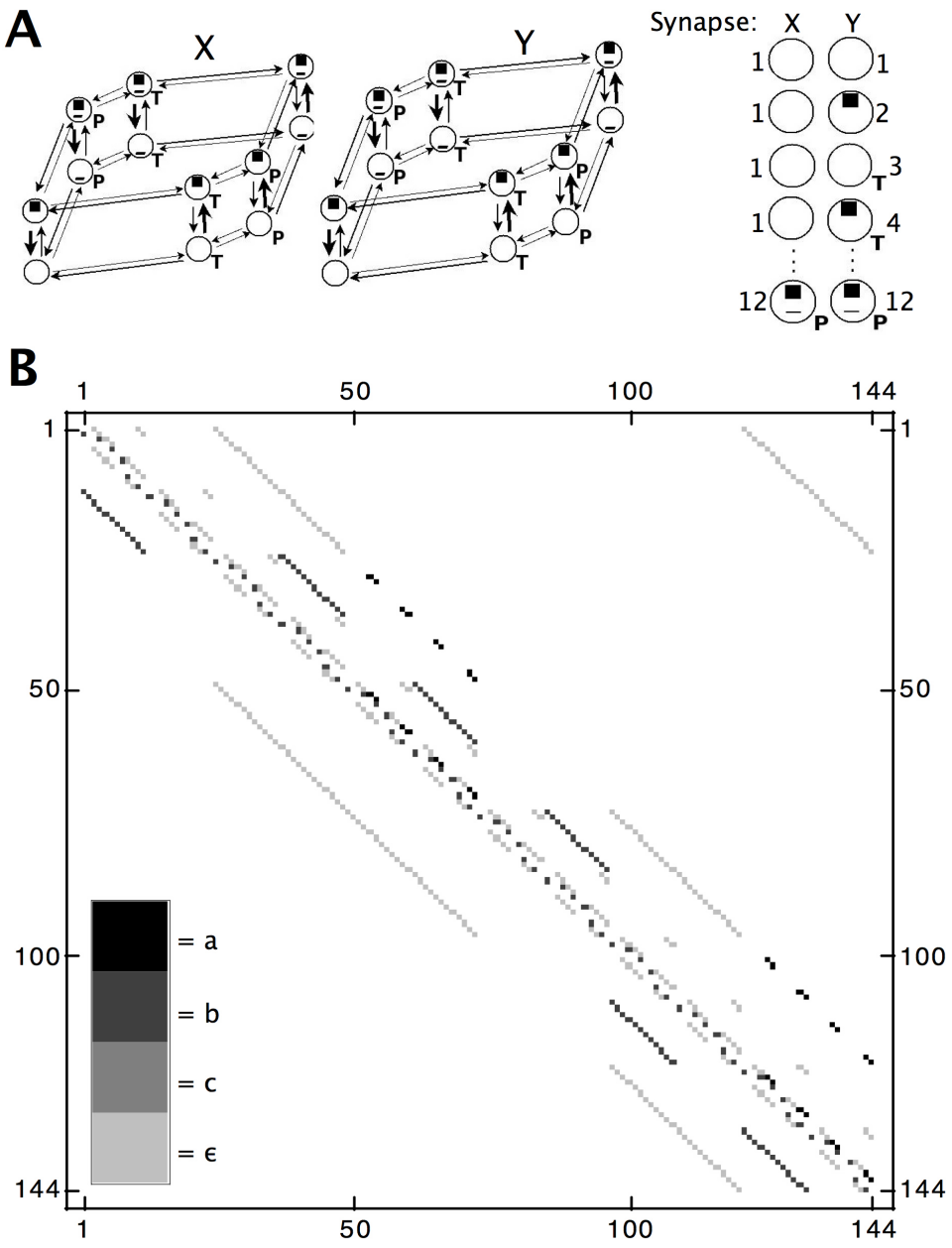

Figure 8.4: The state space is expanded when tagging is taken in to account. A: The possible states are now made up of all possible combinations of the 12 states allowed for synapse X and synapse Y. These combine to give 144 individual states of the synaptic pair. B: A transition matrix can be constructed for the expanded 2 synapse system. This matrix governs the evolution of the tagging model.

8.2B. The early LTD and late LTD protocols operate in precise antagonism to the LTP protocols, Fig. 8.2B.

In this model synaptic tagging occurs between two synapses. In the state based framework this requires that at any point in time the state of the system is defined not only in terms of the state of a single synapse (a 12 dimensional state space) but in terms of the combination of the states of two synapses (a 144 dimensional state space). This can be understood by imagining that for every single state of synapse X, there are twelve possible states of synapse Y, i.e. synapse X might be in state 1, while synapse Y can be in any one of its 12 allowed states. The 144 state tagging transition matrix, Fig. 8.4, thus contains the 12 state single synapse transition matrix, Fig. 8.3, in 12 blocks. Each of these blocks describes the transitions of synapse X when the other synapse is in one of its 12 states.

The tagging interaction is mediated as follows: If synapse Y is in a PRP state then the 12 state sub-block that specifies the behavior of synapse X, is selected such that the probability of making a transition from the tag state into the PRP state (and then on into a consolidated potentiated state) is elevated. The subblock for each PRP state has an elevated transition rate of tagged $\rightarrow$ PRP, Fig. 8.2C. The result of this is that if synapse X happens to be in the tagged state when synapse Y is in a PRP state, then it is more likely to be consolidated. If synapse X is not in a tagged state however, then this does not occur, regardless of the state of synapse Y.

Finally, the tagging model does not obey detailed balance. Thus the expansion in eigenfunctions method was not used to solve the dynamics of the model. Instead the model was integrated using the direct matrix method introduced in chapter 3.

8.2 Synaptic tagging

Synaptic tagging can be performed in the model in an analogous way to the experiments reviewed in chapter 1 (Frey and Morris, 1997; Sajaykumar and Frey, 2004; Sajikumar and Frey, 2004). The model was matched to experimental data gathered by Roger Redondo. In these experiments, late LTP was initially induced with 150s TBS. After 30 mins, early LTP was induced with HFS. This is emulated in the model by first matching late LTP and early LTP alone to the data, Fig. 8.5A+B. Next, late LTP is applied to synapse X. After 30mins early LTP is applied to synapse Y, Fig. 8.5C. The early LTP decay path is also calculated in the case that no LTP was applied previously. Early LTP that was preceded by late LTP has a more slowly decaying component than

early LTP that was not preceded by late LTP. In the terminology used in the literature, the protocol performed in Fig. 8.5C is termed 'strong before weak'. Alternatively, the protocol can be 'weak before strong' in which case the induction of early LTP precedes the induction of late LTP, in this case synaptic tagging still occurs in the same way, Fig. 8.5D.

8.3 Synaptic tagging rescues a weak pattern in the Perceptron model

Flashbulb memories are formed when an entire snapshot of information is fixed in memory by a single event, much as an old fashioned flashbulb fixes an entire visual scene in an instant when the photographer wishes to capture one salient but fleeting object. In human memory, flashbulb memories are caused when an important event leads to emotional arousal causing memory of both that important event and a number of peripheral, seemingly inconsequential events (Brown and Kulik, 1977). In some ways the analogy to a flashbulb is an unfortunate one, because flashbulb memories can lead to the retention of information that preceded or followed the event leading to the 'flash'. The classic example of this is to say that everyone remembers where they were when JFK was shot[1].

It is not clear what the behavioral role of flashbulb memories is. They might be a mechanism for ensuring retention of information about the circumstances leading up to and directly following an important event. It has been suggested that synaptic tagging might be a mechanism that mediates flashbulb memories at the low level (Frey and Morris, 1997; Frey and Morris, 1998). The temporal non-locality of synaptic tagging, meaning that synapses can alter their course of plasticity depending upon events that do not happen in strict synchrony, is the key feature that might enable this. The slowing of the decay of the signal of a memory stored with early phase transitions due to the storage of a late phase memory can be demonstrated using the model of synaptic tagging and the perceptron.

In the expanded state space of the tagging model, plasticity protocols can be applied to one or other of the synapses separately, by adjusting the transition rates for that synapse without altering the transition rates governing the other synapse. Specifically, alterations can be made to the transition rates in any subblock of the full 144 state

[1] Clearly, this does not apply to me because I was not alive when this Earth-shaking event occurred. The events of September 11th 2001 provide a tragic contemporary example.

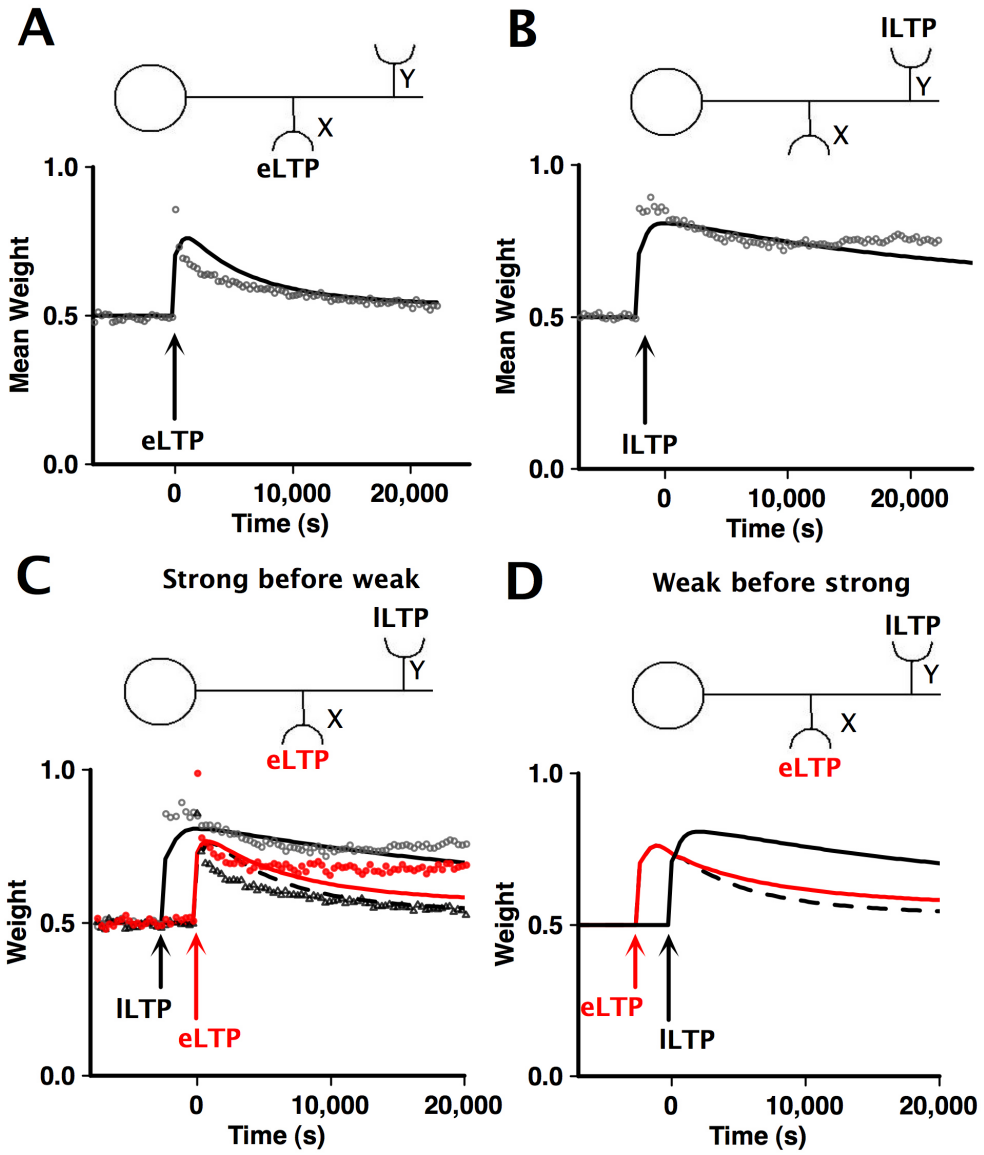

Figure 8.5: Synaptic tagging in the model, compared to experimental data. A: Early LTP alone is elicited in synapse X. The experimental data is indicated by the circles, while the behavior of the model is indicated by the solid black line. B: late LTP alone is elicited in synapse Y. C: Result of applying a tagging protocol. Late LTP is induced in synapse Y first (black solid line), then followed after 30mins with early LTP in synapse X (red solid line). The dashed black line is the decay of early LTP in synapse X in the case that the tagging interaction is removed from the model. Experimental data for the initial induction of late LTP (open circles), the subsequent rescue of early LTP via tagging (filled red circles) and control early LTP with no tagging (open triangles) is plotted for comparison. The decay of early LTP in synapse X is slowed by the initial induction of late LTP. D: As C, but early LTP is first elicited in synapse X and is followed after 30mins by late LTP in synapse Y. This leads to an identical rescue of early LTP.

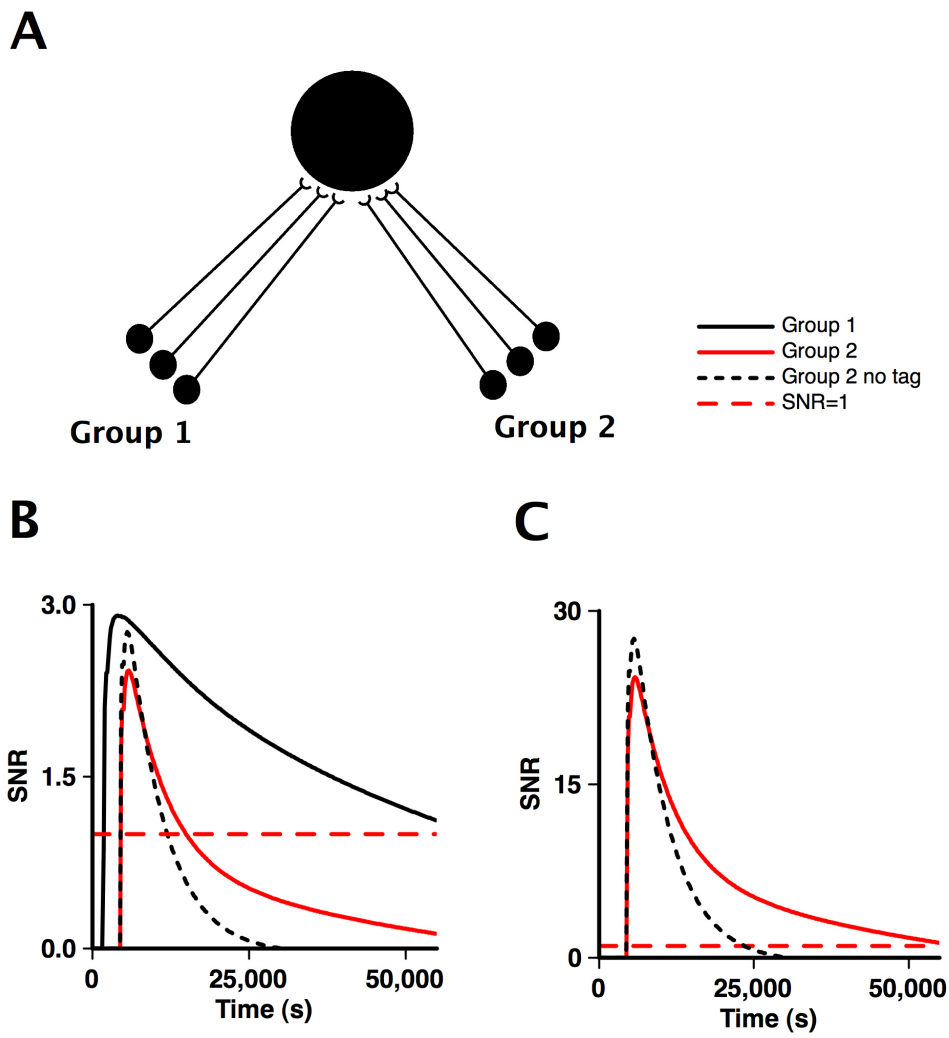

Figure 8.6: Pattern rescue in the synaptic tagging model. A: The inputs to the perceptron are divided into two groups of synapses, each storing a separate random input pattern. Although the patterns stored are independent, the two synaptic groups share a common state space such that their plasticity transitions can interact. B: The first pattern is stored with late LTP/D in group 1 leading to a slowly decaying SNR (black line). The red dashed line is SNR=1. In the second group of synapses the second pattern is stored 2600s later with early LTP/D (red line). The decay of the SNR of the second pattern is slower than the case where everything is identical, but the tagging transitions are turned off (black dashed line). Both groups contain 100 synapses. C: The second pattern is stored after late LTP in an identical fashion to B, but now 10000 synapses are in each group. The initial signal is higher and the difference in time for the SNR to reach 1 of the pattern stored with tagging and the pattern stored without tagging increases to 30,000s (8.3 hrs).

matrix without altering any other transition rates. Thus plasticity protocols can be applied independently to the two synapses. Here the inputs to the perceptron model are divided in to two groups each composed of 100 synapses, Fig. 8.6A, where each group is independently subjected to either early LTP/D or late LTP/D. After 2000s a random pattern is stored in group 1 with late phase transitions, in precisely the same way as patterns were stored in chapter 7. After a further 2600s another pattern, random with respect to the first, is stored with early phase transitions in the synapses of group 2, Fig. 8.6B. We see that the timecourse of decay of the SNR of the second pattern, stored with early phase transitions, is extended in comparison to the control case where the tagging interaction was removed from the transition matrix. Although the SNR decay timecourse is slowed by the tagging interaction, the absolute difference in time for the SNR to reach 1 is increased by only 3200s (53 mins).

The time-gap between the time for the SNR to reach 1 in the case with tagging as against the case where there is no tagging, can be improved by increasing the number of synapses within each of the subgroups. This increases the initial signal and raises the SNR decay curve. This allows the more slowly decaying tail of the curve to determine the time taken for the SNR to reach 1. With 10,000 synapses (around 1/3 of the synapses available to pyramidal neuron) the time-gap is increased to around 30,000s (8.3 hrs). Clearly this time difference is not long enough to allow the rescued memory (group 2) to persist for a lifetime, as is the case in typical examples of flashbulb memories. However, lasting for an extra 8 hours might be long enough to enable a memory trace to be consolidated at the systems level.

8.4 Discussion

In this chapter the state based modeling approach was extended to synaptic tagging. Using the model the electrophysiological hallmark of synaptic tagging - the rescue of rapidly decaying early LTP by the induction of late LTP in a separate input - was qualitatively reproduced. It was then shown that the synaptic tagging model can lead to the heterosynaptic interaction of the persistence of memory traces, where a single trace is stored within a subgroup of synapses on one neuron. The degree of rescue of the weak memory trace depends upon the number of synapses in the subgroups. In this model, and when the number of synapses in each subgroup is a significant proportion of the total synapses on a pyramidal neuron, the lifetime of the weak memory can be extended by around 8 hours when synaptic tagging is present. This is only a factor

2 extension to the memory lifetime. Thus if this model were to mediate flashbulb memories, it would have to be the case that a modest increase in memory lifetime could make the difference between a memory trace being consolidated or not, at the systems level.

There are many biological details of synaptic tagging that are not yet understood, for example the biochemical identity of the tag and plasticity related proteins. This means that in formulating a state based model, there were many uncertainties and so many assumptions were made. In particular, there is not much to constrain the topology of the state diagram. The state diagram chosen was the simplest that could be found that allowed the basic reproduction of electrophysiological synaptic tagging.

In this model and all other models that were explored in arriving at this one, it was found that the extent to which early LTP could be rescued by late LTP fell far short of the level that has been observed in experiment (Sajaykumar and Frey, 2004; Sajikumar and Frey, 2004). However, state based models that relax one of the key assumptions of this thesis, that plasticity is modelled by the instantaneous piecewise substitution of the transition matrix, are able to reproduce larger magnitudes of rescue of early LTP (Barrett et al., 2007). In models of this type, the transition probabilities of the Markov process underlying synaptic plasticity are free to vary as some continuous function of time. It makes some sense that this should be the case: The timescales of the dynamics of the biochemical processes underlying synaptic tagging are probably not small compared to the decay timescale of the plasticity itself. We can see this from experiments that measured the time interval that must elapse before synaptic tagging cannot be abolished. It was found that the time interval over which the tag is established is of the order of 10mins (Sajaykumar and Frey, 2004). Alternatively to introducing continuously varying transition rates, the Markovian property can be restored, but only at the expense of increasing the size of the state space such that state transitions are modeled that again happen on a timescale that is short as compared to the decay timescale of LTP itself. For this reason this model of synaptic tagging would need to be extended such that transitions between tag and PRP states consist of cascades of state transitions, rather than a single transition.

Chapter 9

Conclusion

In this chapter results are summarised and placed within the context of previous work. Finally, questions raised by the thesis that are open to further research are stated. This thesis assessed the plasticity stability dilemma in several biologically plausible models of synaptic plasticity, including novel models of long term potentiation and depression and synaptic tagging. This investigation led to the prediction of several ways that the brain might overcome the plasticity stability dilemma in principle.

9.1 Conclusions and the contribution made by this work

The key motivating influences behind the approach taken in this thesis were the ideas that synaptic plasticity can be abstracted into a stochastic process (Amit and Fusi, 1992; Amit and Fusi, 1994; Fusi, Drew, and Abbott, 2005; Senn and Fusi, 2005; Fusi and Abbott, 2007) and that statistical models can be combined with experimental data to model those processes (van Rossum, Bi, and Turrigiano, 2000). The tools of this approach were extended such that in addition to existing calculations of the steady state, the autocorrelation can be also calculated, allowing a characterisation of the 'stability' of correlation between the weights and a previously stored pattern.

9.1.1 Spike Timing Dependent Plasticity

Two existing models of STDP, non-weight dependent STDP (nSTDP) (Song, Miller, and Abbott, 2000) and weight dependent STDP (wSTDP) were cast in the stochastic processes framework. This explicitly revealed the long autocorrelation timescale present in nSTDP. It was found that nSTDP has an autocorrelation timescale that is

several orders of magnitude greater than the autocorrelation timescale of wSTDP for single units under comparable conditions. The origin of this striking difference is that the synaptic weight dynamics of nSTDP are bistable, whereas there is no bistability inherent in wSTDP. This strongly segregates weights into to two groups, allowing only slow diffusion between the two groups and hence endowing the autocorrelation with a slow timescale. It was demonstrated that this can be regarded as analogous to the stochastic escape problem, a process that is well known for producing long correlation times. Therefore this result applies more generally to any stochastic synapse having a linear bounded state space: In this case bistable dynamics provide resistance of the memory trace to fluctuations. This element of the thesis therefore provides an increase in our understanding about the chosen models of STDP, but also slightly extends the comments of Fusi and Abbott (Fusi and Abbott, 2007), who did not explore linear bounded models having bistable dynamics.

Another important factor in the memory trace survival time in the models of STDP was revealed by exploring the behavior of STDP in simple networks of neurons. It was found that while wSTDP in networks with lateral inhibition still gives rise to more unstable receptive fields than nSTDP, that those receptive fields can nevertheless remain correlated with their previous locations for a significant amount of time, of the order of hours, if the lateral inhibition is sufficiently strong. Thus the strength of inhibition in the network modifies the behavior of the plasticity rule as studied in the single unit case. The modulation of inhibition in the wSTDP model allowed the stability of the receptive fields to be varied. When inhibition was removed entirely, the receptive fields were rapidly destroyed by ongoing activity. Thus inhibition might offer a mechanism by which cortical processing can alter its stability depending upon the demands of the environment. This is a novel property of wSTDP in networks that has not been documented previously.

9.1.2 State based models of LTP/D

Previous authors had demonstrated that state based modeling can be applied to synaptic plasticity and that this allows calculation of steady state synaptic weight distributions (van Rossum, Bi, and Turrigiano, 2000) and the signal to noise ratio of the memory trace (Fusi, Drew, and Abbott, 2005; Fusi and Abbott, 2007). In this thesis the same approach was applied in order to directly model long term potentiation and depression. Experimental data was used to constrain the decay timescales of early and late LTP/D

in these models. These models are the first state based models that have been directly linked to experimental data in this way. This is an initial attempt to calculate the steady state memory trace survival time from the decay timescales of early and late LTP/D observed in experiment.

In the context of the stationary plasticity stability dilemma (SPS), memory traces are synaptic fluctuations within the steady state of the synaptic ensemble. In the SPS, it was found that the addition of the late phase timescale to the LTP models increased the memory trace lifetime to around 200 hours as compared to only 7 hours for the simple 2 state binary model at the same initial signal.

The memory trace was also studied in the non stationary plasticity stability dilemma (NPS). Two scenarios were explored: Firstly, the disruption to memory traces stored with synaptic fluctuations, when an LTP protocol is applied to the synapses. This mimicked amnesia experiments that have been performed on rats using in-vivo induction of LTP. It was found that the induction of amnesia, as measured by the degree to which the autocorrelation and signal to noise ratio of the memory trace was harmed depended upon the degree of saturation of the weights, and whether early or late LTP was induced.

The degree of retrograde amnesia was found to depend upon whether late LTP or early LTP was induced. Saturation of late LTP leads to the complete destruction of the memory trace. However saturation of early LTP leads only to a transient disruption of the trace, with the memory signal returning to the decay trajectory that it would have followed, had there been no intervention.

Anterograde amnesia was more difficult to induce robustly in the model. While the autocorrelation of the memory trace was harmed both by induction of early and late LTP, the signal to noise ratio was not necessarily disrupted. This is as a result of the fact that although correlations of a memory trace stored in a non-equilibrium state of the synaptic ensemble die away more rapidly than at the steady state, the initial signal can be larger due to increased fluctuations. This result does not carry directly over into the biological case since the encoding of memories in the brain is likely far more complex than the inner product scheme employed in chapter 7. However, it does indicate that in the NPS, even in this simple case and with a simple encoding scheme, disruption of the weights does not automatically imply an analogous disruption to memory. Indeed as was mentioned in chapter 7, the experimental results regarding the induction of anterograde amnesia have been mixed.

The second NPS scenario that was explored was the situation in which memory

traces are not stored in the steady state of the baseline synapses, but rather are stored by large LTP/D events. This has the advantage that the initial signal of the memory trace can now be very large, potentially scaling steeply with the square root of the number of synapses. However this is at the cost that the memories tend to be short lived, for two reasons: Firstly, the signal to noise can decay more rapidly than at the steady state. Secondly incoming memories overwrite existing memories instantly, a scenario that was demonstrated using the simple 2 state model. However, these problems can be alleviated by two factors: Firstly synaptic overload allows the superposition of memory traces formed using early phase plastic transitions on top of memory traces stored with late phase synaptic transitions. This means that information can be stored in an ongoing manner for a temporary time while the initial memory trace is retained for a long period. This would be suited to a function such as the hippocampus is thought to perform: The automatic recording of experience, where only a fraction of the information need be retained for a long period, but where it must all be initially stored. Secondly, even with synaptic overload, the memory lifetime of the state based models in the NPS was severely limited by the rate of storage of the late phase memory traces. It was demonstrated that sparse coding could in principle alleviate this situation.

9.1.3 Synaptic tagging

In chapter 8, a state based model of synaptic tagging was proposed and matched to experimental data. While the model was not in precise numerical agreement with the experimental data, it was able to reproduce the phenomenology of synaptic tagging. Using the same non-steady state storage scheme as defined in chapter 7, the model was used to demonstrate the conversion of a rapidly decaying memory trace stored with early LTP/D, in to a more slowly decaying trace via the tagging interaction. It was found that with a realistic number of synapses in each group, the difference in memory trace decay time between an early phase memory trace with no tagging, and an identical trace with tagging was 8.3 hours. It would be necessary for this modest time increase to allow systems level consolidation of the memory trace.

Recall that synaptic overload allows memory traces to be stored on a temporary basis initially, without disrupting a previous trace stored with late phase transitions. When combined with synaptic overload, synaptic tagging might allow the conversion of superimposed, rapidly decaying, memory traces into more slowly decaying memory traces. Importantly, this conversion could take place before or after the storage of

the short term memory trace and depend upon whether or not a separate population of synapses has caused the production of PRPs. Since conversion of an early phase memory trace into a late phase memory trace causes late LTP, this process would partially harm any pre existing memory trace relying on late LTP. However sparse coding could alleviate this problem for identical reasons to its usefulness in protecting long term memory traces during memory storage with synaptic overload.

9.1.4 Summary of conclusions

The principle contribution of this thesis is the demonstration that biological synapses might combat the plasticity stability dilemma in the following ways:

- The dynamics of the learning rule, or rather the drift term in the Fokker Planck formalism, greatly influences the stability of correlations stored within the synapses of a single unit. It was found that a bistable linear bounded learning rule conferred resistance to steady state fluctuations.

- Lateral inhibition might have a large influence on the stability of plastic feedforward weights and the stability of processing in networks. Here it was found that in the case of the weight dependent STDP learning rule the receptive fields of the output units could be stabilised against steady state fluctuations by increasing the magnitude of lateral inhibition. Furthermore, the inhibition could also modulate the presence of receptive fields in the network. Thus, removal of inhibition leads to the rapid obliteration of receptive fields in the wSTDP network. Replacing inhibition would allow a new set of receptive fields to be learned.

- In a model of early LTP and late LTP it was found that the presence of more than one timescale in the synaptic dynamics conferred considerable benefits upon memory trace retention. Thus it is reasonable that synaptic components such as the PSD or phosphorylating switches might increase memory retention time by providing hidden synaptic variables that are not modified by all regimes of activity. Crucially, we saw in chapter 6 that this can be achieved while maintaining an initial signal that is as high as a 2 state model having no additional variables.

- In principle the state based models of LTP/D allow memory traces to be stored away from the steady state of individual synapses. It was shown that this NPS scenario gives a very large initial signal, although at the cost of memory longevity.

Memory longevity can be partially restored by employing synaptic overload and sparse coding.

- Synaptic overload provides a function supporting synaptic tagging, which allows short term memory traces to have their decay timescale lengthened in an online manner. Synaptic overload removes the necessity for all memory traces to be stored on the longest timescale of plasticity and hence might aid the automatic recording of experience.

9.2 Predictions

Processes similar to those stated in §9.1.4 predict:

- Receptive field stability can be directly related to the strength of inhibition in the network. If receptive fields display stability characteristics that are heavily dependent upon inhibition then it suggests that the underlying synaptic dynamics is not intrinsically stable, in the sense that correlations are rapidly lost. This situation would be compatible with a learning rule that is not bistable such as wSTDP. On the other hand, synapses that are each plastic but that operate according to a learning rule having a high degree of intrinsic stability, should lead to a network whose processing can also be very stable (again in the sense of the survival of correlations). This behavior would be compatible with the bistable nSTDP learning rule studied here.

- It should be possible to cause *reversible* retrograde amnesia of recently formed memories by inducing early LTP in the hippocampus, and this effect should be more robust than anterograde amnesia induction.

- If a process such as synaptic overload occurs, whereby memory traces can be stored on more than one timescale, then it might be possible to *reversibly* disrupt prior learning with intense novel learning under the conditions stated in chapter 7.

9.3 Further Work

The matching of the state based models to experimental data here was somewhat informal. This was primarily due to time constraints, but also as a result of a lack of

a sufficient quantity of raw experimental data. It would be useful to devise a scheme whereby models could be automatically selected as a function of an electrophysiological data set, i.e. the likelihood of the data could be maximised with respect to the number of states, the transition rates and possibly even the topology of the model.

In synaptic overload, the direct superposition of early LTP/D and late LTP/D occurs. This superposition means that the synaptic weight can be altered on the short term without altering the long term weight. Intuition gained from working with various state based models of LTP/D suggests that this superposition of timescales is made possible by the ring topology of the state diagram (for piecewise homogeneous models). Is this true? This raises the question of how the topology of state diagrams relates to the dynamic characteristics of the synaptic weight: Are there more general classes of state based model categorised according to topology? This amounts to understanding how the structure of the graph in the state diagram relates to the superposition of eigenvectors in the solutions (at least in the case that detailed balance is obeyed).

The implementation of the state based models in this thesis was extremely simple. It would be very interesting to embed state based models within neural networks performing processing. In this more complex case do the conclusions in §9.1.4 still apply? Can synaptic overload operate in more sophisticated networks?

There is still much work to be undertaken to understand how the interaction between learning rules and network dynamics operates. This is a difficult problem because the dynamics of the weights determine the output activity, which in turn affects the dynamics of the weights. This applies most directly to chapter 5 of this thesis, but is an issue that is relevant to any plasticity process be it LTP or STDP.

Finally, to better match data, the synaptic tagging model in this thesis suggests that either more states must be introduced in order to mediate the tagging interaction, or the underlying Markov process should be inhomogeneous. It would be interesting to attempt to model the biochemical pathways underlying tagging in more detail and hence to expand the state space of the model. This would have the advantage that the model could be directly mapped on to transition rates of putative biochemical transitions. An alternative approach is to derive the properties of the inhomogeneous model from the supposed biochemical cascade. If a system could be devised that allows this process to be carried out rapidly, then the consequences of many alternative predictions about the mechanism of synaptic tagging could be explored.

Bibliography

Abbott, L.F. and S.B. Nelson (2000). Synaptic plasticity: Taming the beast. *Nature Neuroscience* 3: 1178–1183.

Abraham, W.C. (2003). How Long will Long Term Potentiation Last. In Bliss, T, G.L. Collingridge, and R.G.M. Morris, editors, *LTP: Enhancing Neuroscience for 30 years*, chapter 18, pp. 211–228. Oxford, Clarenden St, Oxford.

Abraham, W.C. and A. Robins (2005). Memory retention: The synaptic stability versus plasticity dilemma. *Trends Neurosci* 28: 73–78.

Abraham, W.C. and J.M. Williams (2003). Properties and Mechanisms of LTP Maintenance. *The Neuroscientist* 9: 463–474.

Aihara, T., Y. Abiru, Y. Yamazaki, H. Watanabe, Y. Fukushima, and M. Tsukada (2007). The relation between spike timing dependent plasticity and calcium dynamics in the hippocampal CA1 network. *Neuroscience* 145: 80–87.

Alpermann, C.R., R.G.M. Morris, M. Korte, and T. Bonhoeffer (2006). Homeostatic shutdown of long-term potentiation in the adult hippocampus. *PNAS* 103: 11039–11044.

Alvarez, P., S. Zola-Morgan, and L.R. Squire (1995). Damage limited to the hippocampal region produces long lasting memory impairment in monkeys. *Journal of Neuroscience* 15: 3796–3807.

Amit, D. J. and M. V. Tsodyks (1992). Effective neurons and attractor neural networks in cortical environment. *Network* 3: 121–137.

Amit, D.J. and S. Fusi (1992). Constraints on learning in dynamic synapses. *Network* 3: 443–464.

Amit, D.J. and S. Fusi (1994). Dynamic learning in neural networks with material synapses. *Neural Computation* 6: 957–982.

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

Appleby, P.A. and T. Elliot (2005). Synaptic and temporal ensemble interpretation of spike timing dependent plasticity. *Neural Computation* pp. 2316–2336.

Appleby, P.A. and T. Elliott (2006). Stable competitive dynamics emerge from multispike interactions in a stochastic model of spike timing dependent plasticity. *Neural Computation* 18: 2414–2464.

Bakker, A., C.B. Kirwan, M. Miller, and C.E.L Stark (2008). Pattern separation in the human Hippocampal CA3 and Dentate Gyrus. *Science* 319: 1640–1642.

Baldassi, C., A. Braunstein, N. Brunel, and R. Zecchina (2007). Efficient supervised learning in networks with binary synapses. *PNAS* 104: 11079–11084.

Barlow, H.B. (1989). Unsupervised learning. *Neural Computation* 1: 295–311.

Barnes, C.A., M.W. Jung, B.L. McNaughton, D.L. Korol, K. Andreasson, and P.F. Worley (1994). LTP saturation and spatial learning disruption: Effects of task variables and saturation levels. *The Journal of Neuroscience* 14: 5793–5806.

Barrett, A., G.O. Billings, R.G.M. Morris, and M.C.W Rossum (2007). A biophysical model of long-term potentiation and synaptic tagging. *Soc. Neurosc. Abstr* 33: 583.1.

Bashir, Z.I. and G.L. Collingridge (1994). An investigation of depotentiation of long-term potentiation in the CA1 region of the hippocampus. *Exp. Brain Res.* 100: 437–443.

Beattie, E.C., R.C. Carroll, R.C. Malenka, W Morishita, M Zastrow, H Yasuda, and X Yu (2000). Regulation of AMPA receptor endocytosis by a signaling mechanism shared with LTD. *Nature Neuroscience* 3: 1291–1300.

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

Ben Dayan Rubin, D.D. and Stefano Fusi (2007). Long memory lifetimes requires complex synapses and limited sparseness. *Frontiers in computational neuroscience* 1(Article 7): 1–13.

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

Berman, D.H. (1975). The fluctuation dissipation theorem for contracted descriptions of Markov processes. *Journal of Statistical Physics* 20: 57–81.

Bi, G.Q. and M.M. Poo (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience* 18: 10464–10472.

Bi, G.Q. and M.M. Poo (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.* 24: 139–166.

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

Billings, G.O. and M.C.W Rossum (2006). Stability and plasticity in network and single unit models. *Soc. Neurosc. Abstr* 32: 133.4.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, New York.

Bliss, T.V. and A.R. Gardner-Medwin (1973). Long-lasting increases of synaptic influence in the unanesthetized hippocampus. *Journal of Physiology* 216: 32–33.

Bliss, T.V. and T. Lomo (1973). Long-lasting potentiation of synaptic transmission in the dentate gyrus of the anaesthetized rabbit. *Journal of Physiology* 232: 331–356.

Borgdoff, A.J. and D Choquet (2002). Regulation of AMPA receptor lateral movements. *Nature* 417: 649–653.

Bredt, D.S., M. Fukata, R.A. Nicoll, and S. Tomita (2004). Dynamic interaction of Stargazin-like TARPs with cycling AMPA receptors at synapses. *Science* 303: 1508–1511.

Brown, R. and J. Kulik (1977). Flashbulb memories. *Cognition* 5: 73–99.

Burkitt, A.N., H. Meffin, and D.B. Grayden (2004). Spike-timing-dependent plasticity: The relationship to rate-based learning for models with weight dynamics determined by a stable fixed point. *Neural Computation* 16: 885–940.

Cain, D.P., E.L. Hargreaves, F. Boon, and Z. Dennison (1993). An examination of the relations between hippocampal long term potentiation, kindling, afterdischarge and place learning in the water maze. *Hippocampus* 3: 153–164.

Carpenter, G.A. and S. Grossberg (2003). Adaptive resonance theory. In Arbib, M. A., editor, *The handbook of brain theory and neural networks*. MIT press, 2nd edition.

Carroll, R.C., E.C. Beattie, M Zastrow, and R.C. Malenka (2001). Role of AMPA receptor endocytosis in synaptic plasticity. *Nature Reviews Neuroscience* 2: 315–323.

Castro, C.A., L.H. Silbert, B.L. McNaughton, and C.A. Barnes (1989). Recovery of spatial learning deficits after decay of electrically induced synaptic enhancement in the hippocampus. *Nature* 342: 545–548.

Choquet, D. and A. Triller (2003). The role of receptor diffusion in the organization of the postsynaptic membrane. *Nature Reviews Neuroscience* 4: 251–265.

Chung, H.J., J. Xia, R.H. Scannevin, X. Zhang, and R.L. Huganir (2000). Phosphorylation of the AMPA receptor subunit GluR2 differentially regulates its interaction with PDZ domain-containing protiens. *The Journal of Neuroscience* 20: 7258–7267.

Colquhoun, D., K.A. Dowsland, M. Beato, and A.J.R. Plested (2004). How to impose microscopic reversibility in complex reaction mechanisms. *Biophysical Journal* 86: 3510–3518.

Corkin, S. (2002). What's new with the amnesiac patient H.M.? *Nature Reviews Neuroscience* 3: 153–160.

Dan, Y. and M.M. Poo (2006). Spike timing-dependent plasticity: From synapse to perception. *Physiol Rev* 86(3): 1033–1048.

Dayan, P. and D.J. Willshaw (1991). Optimising synaptic learning rules in linear associative memories. *Biological Cybernetics* 65: 253–265.

Debanne, D., B. H. Gähwiler, and S. M. Thompson (1996). Cooperative interactions in the induction of long-term potentiation and depression of synaptic excitation between hippocampal CA3-CA1 cell pairs in vitro. *Proc Natl Acad Sci USA* 93: 11225–11230.

Debanne, D., B. H. Gähwiler, and S. M. Thompson (1999). Heterogeneity of synaptic plasticity at unitary CA1-CA3 and CA3-CA3 connections in rat hippocampal slice cultures. *Journal of Neuroscience* 19: 10664–10671.

Delorme, A., L. Perrinet, M. Samuelides, and S.J. Thorpe (2001). Network of integrate-and-fire neurons using rank order coding B: Spike timing dependant plasticity and emergence of orientation selectivity. *Neurocomputing* 38–40(1–4): 539–45.

Dudek, Serena M. and M.F. Bear (1992). Homosynaptic long-term depression in area CA1 of hippocampus and effects of N-methyl-D-aspartate receptor blockade. *PNAS* 89: 4363–4367.

Duprat, F, G.L. Collingridge, M Daw, W Lim, and J.T.R. Isaac (2003). GluR2 protien-protien interactions and the regulation of AMPA receptors during synaptic plasticity. In Bliss, T, G.L. Collingridge, and R.G.M. Morris, editors, *LTP: Enhancing Neuroscience for 30 years*, chapter 15, pp. 175–181. Oxford, Clarenden St, Oxford.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition letters* 27: 861–874.

Felderhoff, B.U. (1978). On the derivation of the fluctuation-dissipation theorem. *J. Phy. A: Math. Gen.* 11: 921–927.

Földiák, P. (2002). Sparse coding in the primate cortex. In Arbib, M.A., editor, *The handbook of brain theory and neural networks*. MIT press.

Földiák, P. and D. Endres (2008). Sparse coding. *Scholarpedia* 3: 2984.

French, R.M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Neurosciences* 3: 128–135.

Frey, U. and R.G.M. Morris (1997). Synaptic tagging and long term potentiation. *Nature* 385: 533–536.

Frey, U and R.G.M. Morris (1998). Synaptic tagging: Implications for late maintenance of hippocampal long-term potentiationmplications for late maintenance of hippocampal long-term potentiation. *Trends in Neurosciences* 21: 181–188.

Froemke, R.C. and Y. Dan (2002). Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature* 416: 433–438.

Froemke, R.C., M.M. Merzenich, and C.E. Schreiner (2007). A synaptic memory trace for cortical receptive field plasticity. *Nature* 450: 425–429.

Froemke, R.C., I.A. Tsay, M. Raad, J.D. Long, and Y. Dan (2006). Contribution of individual spikes in burst induced long term synaptic modification. *Journal of Neurophysiology* 95: 1620–1629.

Fusi, S. and L. F. Abbott (2007). Limits on the memory storage capacity of bounded synapses. *Nat Neurosci* 10: 485–493.

Fusi, S., P. J. Drew, and L.F. Abbott (2005). Cascade models of synaptically stored memories. *Neuron* 45: 599–611.

Fusi, S. and W. Senn (2006). Eluding oblivion with smart stochastic selection of synaptic updates. *Chaos* 16: 026112; 1–11.

Gardner-Medwin, A.R. (1989). Doubly modifiable synapses: A model of short and long term potentiation. *Proc. Roc. Soc. B* 238: 137–154.

Gerstner, W., R. Kempter, J. L. van Hemmen, and H. Wagner (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature* 383: 76–78.

Golomb, D., N. Rubin, and H. Sompolinsky (1990). Willshaw model: Associative memory with sparse coding and low firing rates. *Phys. Rev. A* 41: 1843–1854.

Groc, L, M Heine, L Cognet, K Brickley, F.A. Stephenson, B Lounis, and D Choquet (2004). Differential activity-dependent regulation of the lateral mobilities of AMPA and NMDA receptors. *Nature Neuroscience* 7: 695–696.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science* 11: 23–63.

Gutig, R., R. Aharonov, S. Rotter, and H. Sompolinsky (2003). Learning input correlations through nonlinear temporally asymmetric hebbian plasticity. *The Journal of Neuroscience* 23: 3697–3714.

Guzowski, J.F. and J.J. Knierim (2004). Ensemble dynamics of hippocampal regions CA3 and CA1. *Neuron* 44: 581–584.

Harris, K.M., J.C. Fiala, and L Ostroff (2003). Structural changes at dendritic spine synapses during long-term potentiation. In Bliss, T, G.L. Collingridge, and R.G.M.

Morris, editors, *LTP: Enhancing Neuroscience for 30 years*, chapter 19, pp. 230–234. Oxford, Clarenden St, Oxford.

Hayer, A. and U.S. Bhalla (2005). Molecular switches at the synapse emerge from receptor and kinase traffic. *PLoS Computional Biology* 1: 137–154.

Hertz, J., A. Krogh, and R. G. Palmer (1991). *Introduction to the theory of neural computation*. Perseus, Reading, MA.

Hille, B. (2001). *Ionic Channels of excitable membranes*. Sinauer, Sunderland, MA.

Izhikevich, E.M. and N.S. Desai (2003). Relating STDP to BCM. *Neural Computation* 15: 1511–1523.

Jeffery, K.J. and R.G.M. Morris (1993). Cumulative long-term potentiation in the rat dentate gyrus correlates with, but does not modify performance in the water maze. *Hippocampus* 3: 133–140.

Kauer, J.A., R.C. Malenka, and R.A. Nicoll (1988). NMDA application potentiates synaptic transmission in the hippocampus. *Nature* 334: 250–252.

Kelleher, R.J., A. Govindarajan, H-Y. Jung, H. Kang, and S. Tonegawa (2004). Translation control by MAPK signaling in long-term synaptic plasticity and memory. *Cell* 116: 467–479.

Kepecs, A., M.C.W. van Rossum, S. Song, and J. Tegner (2002). Spike timing dependent plasticity: Common themes and divergent vistas. *Biological Cybernetics* 87: 446–458.

Kim, M.J., K. Futai, Y. Hayashi, K. Cho, and M. Sheng (2007). Synaptic accumulation of PSD-95 and synaptic function regulated by phosphorylation of serine-295 of PSD-95. *Neuron* 56: 488–502.

Kistler, W.M. (2002). Spike-timing dependent synaptic plasticity: A phenomenological framework. *Biological Cybernetics* 87: 416–427.

Kubo, R. (1966). The fluctuation dissipation theorem. *Rep. Prog. Phys.* 29: 255–284.

Kubo, R., M. Toda, and N. Hashitsume (1998). *Statistical physics II: Nonequilibrium statistical mechanics*. Springer.

Larson, J. and G. Lynch (1986). Induction of synaptic potentiation in hippocampus by patterned stimulation involves two events. *Science* 232: 985–988.

Laurent, G. (2002). Olfactory network dynamics and the coding of multidimensional signals. *Nature Reviews Neuroscience* 3: 884–895.

Lee, H, M Barbarosie, K Kameyama, M.F. Bear, and R.L. Huganir (2000). Regulation of distinct AMPA receptor phosphorylation sites during bidirectional synaptic plastcity. *Nature* 405: 955–959.

Leibold, C. and R. Kempter (2008). Sparseness constrains the prolongation of memory lifetime via synaptic metaplasticity. *Cerebral Cortex* 18: 67–77.

Levy, W.B. and O. Steward (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience* 8: 791–797.

Li, S., W.K. Cullen, R. Anwyl, and M.J. Rowan (2003). Dopamine dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. *Nature Neuroscience* 6: 526–531.

Lin, J.W., W Ju, K Foster, S Hyoung Lee, G Ahmadian, M Wyszynski, Y Tian Wang, and M Sheng (2000). Distinct molecular mechanisms and divergent endocytotic pathways of AMPA receptor internalization. *Nature Neuroscience* 3: 1282–1290.

Lin, L., G. Chen, H. Kuang, D. Wang, and J.Z Tsien (2007). Neural encoding of the concept of nest in the mouse brain. *Proc Natl Acad Sci USA* 104: 6066–6071.

Lisman, J. and S. Raghavachari (2006). A unified model of the presynaptic and postsynaptic changes during LTP at CA1 synapses. *Science STKE* 356: re 11;1–15.

Malenka, R.C. (1994). Synaptic plasticity in the hippocampus: LTP and LTD. *Cell* 76: 535–538.

Malenka, R.C. and M.F. Bear (2004). LTP and LTD an embarrassment of riches. *Neuron* 44: 5–21.

Malenka, R.C. and S.A. Siegelbaum (2001). Synaptic plasticity: Diverse targets and mechanisms for regulating synaptic efficacy. In Cowan, W.M., C.F. Stevens, and T.C. Sudhof, editors, *Synapses*, chapter 9, pp. 393–453. Johns Hopkins, North Charles St, Baltimore, Maryland.

Man, H, J.W. Lin, W Ju, G Ahmadian, L Liu, L.E. Becker, M Sheng, and Y.T. Wang (2000). Regulation of AMPA receptor-mediated synaptic transmission by Clatherin-dependent receptor internalization. *Neuron* 25: 649–662.

Markram, H., J. Lübke, M. Frotscher, and B. Sakmann (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275: 213–215.

Marr, D. (1971). Simple memory: A theory for archicortex. *Phil. Trans. R. Soc. B* 841: 23–81.

Martin, S. J. and R. G. Morris (2002). New life in an old idea: The synaptic plasticity and memory hypothesis revisited. *Hippocampus* 12: 609–636.

McNaughton, B.L., C.A. Barnes, G. Rao, J. Baldwin, and M. Rasmussen (1986). Long-term enhancement of Hippocampal synaptic transmission and the aquisition of spatial information. *The journal of neuroscience* 6: 563–571.

McNaughton, B.L. and R.G.M. Morris (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences* 10: 408–415.

Meffin, H., J. Besson, A.N. Burkitt, and D.B. Grayden (2006). Learning the structure of correlated synaptic subgroups using stable and competitive spike-timing-dependent plasticity. *Phys. Rev E* 73: 041911.

Megias, M., Z.S. Emri, T.F. Freund, and A.I. Gulyas (2001). Total number and distribution of inhibitory and excitatory synapses on hippocampal CA1 pyramidal cells. *Neuroscience* 102: 527–540.

Miguel, Maxi San and Raul Toral (1997). Stochastic effects in physical systems. In *Instabilities and Nonequilibrium structures*. Kluwer, arXiv:cond-mat/9707147.

Montgomery, J. M., P. Pavlidis, and D. V. Madison (2001). Pair recordings reveal all-silent synaptic connections and the postsynaptic expression of long-term potentiation. *Neuron* 29: 691–701.

Morris, R.G.M. and U. Frey (1997). Hippocampal synaptic plasticity: Role in spatial learning or the automatic recording of attended experience? *Phil. Trans. R. Soc. B* 352: 1489–1503.

Morris, R.G.M., E.I. Moser, G.Riedel, S.J. Martin, J. Sandlin, M. Day, and C.O. O'Carroll (2003). Elements of a neurobiological theory of the hippocampus: The role of activity-dependent synaptic plasticity in memory. *Phil. Trans. R. Soc. B* 358(773-786).

Morrison, A., A. Aertsen, and M. Diesmann (2007). Spike-timing-dependent plasticity in balanced random networks. *Neural Computation* 19: 1437–1467.

Moser, E.I. and R.G.M. Morris (1998). Impaired spatial learning after saturation of long term potentiation. *Science* 281: 2038–2042.

Moser, E.I. and M-B. Moser (1999). Is learning blocked by saturation of synaptic weights in the hippocampus? *Neuroscience and behavioral reviews* 23: 661–672.

Moser, M.-B., E.I. Moser, E.Forrest, P. Andersen, and R.G.M. Morris (1995). Spatial learning with a minislab in the dorsal hippocampus. *PNAS* 92: 9697–701.

Mu, Y. and M.M. Poo (2006). Spike timing-dependent LTP/LTD mediates visual experience-dependent plasticity in a developing retinotectal system. *Neuron* 50: 115–125.

Nadal, J.P., G. Toulouse, J.P. Changeux, and S. Dehaene (1986). Networks of formal neurons and memory palimpsests. *Europhys. Lett.* 1: 535–542.

O'Brien, R. J., S. Kamboj, M. D. Ehlers, K. R. Rosen, G. D. Fischbach, M. P. Kavanaugh, and R. L. Huganir (1998). Activity-dependent modulation of synaptic AMPA receptor accumulation. *Neuron* 21: 1067–1078.

O'Connor, D.H., G.M. Wittenberg, and S.S.H. Wang (2005). Dissection of bidirectional synaptic plasticity into saturable unidirectional processes. *Journal of Neurophysiology* 94: 1565–1573.

O'Keefe, J. and N. Burgess (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature* 381: 425–428.

O'Keefe, J. and D.H. Conway (1978). Hippocampal place units in the freely moving rat: Why they fire when they fire. *Exp. Brain Res.* 31: 573–590.

O'Keefe, J. and J. Dostrovsky (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research* 34: 171–175.

Olshausen, B. A. and D. J. Field (1996). Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature* 381: 607–609.

Olshausen, B.A. and D.J. Field (1997). Sparse coding with and overcomplete basis set: A strategy employed by v1? *Vision Research* 37: 3311–3325.

O'Reilly, R.C. and J.L. McClelland (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus* 4: 661–682.

Otani, S., C.J. Marshall, W.P. Tate, G.V. Goddard, and W.C. Abraham (1989). Maintenance of long-term potentiation in rat dentate gyrus requires protein synthesis but not messenger RNA synthesis immediately post-tetanization. *Neuroscience* 3: 519–26.

Park, M, E.C. Penick, J.G. Edwards, J.A. Kauer, and M.D. Ehlers (2004). Recyling endosomes supply AMPA receptors for LTP. *Science* 305: 1972–1975.

Pastalkova, E, P Serrano P, Pinkhasova, E Wallace, AA Fenton, and TC Sacktor (2006). Storage of spatial information by the maintenance mechanism of LTP. *Science* 25: 1141–4.

Petersen, C. C. H., R. C. Malenka, R. A. Nicoll, and J. J. Hopfield (1998). All-or-none potentiation at CA3-CA1 synapses. *Proceedings National Academy of Sciences* 95: 4732–4737.

Pfister, J.P. and W. Gerstner (2006). Triplets of spikes in a model of spike timing-dependent plasticity. *Journal of Neuroscience* 26: 9673–9682.

Reymann, K.G. and J.U. Frey (2007). The late maintenance of hippocampal LTP: Requirements, phases, 'synaptic tagging', 'late-associativity' and implications. *Neuropharmacology* 52: 24–40.

Risken, H. (1996). *The Fokker-Planck Equation*. Springer, New York, 2nd edition.

Robbins, A. (2004). Sequential learning in neural networks: A review and a discussion of pseudorehearsal based methods. *Intelligent Data Analysis* 8: 301–322.

Roche, K.W., R.J. O'Brien, A.L. Mammen, J. Bernhardt, and R.L. Huganir (1996). Characterization of multiple phosphorylation sites on the AMPA receptor GluR1 subunit. *Neuron* pp. 1179–88.

Rolls, E. and A. Treves (1998). *Neural networks and brain function*. Oxford University Press, Clarenden St, Oxford, 1st edition.

Rolls, E.T. (1996). A theory of Hippocampal function in memory. *Hippocampus* 6: 601–620.

Rubin, J., D. D. Lee, and H. Sompolinsky (2001). Equilibrium properties of temporally asymmetric Hebbian plasticity. *Phys. Rev. Lett* 86: 364–367.

Rubin, J.E., R.C. Gerkin, G.Q. Bi, and C.C. Chow (2005). Calcium time course as a signal for spike-timing-dependent plasticity. *J Neurophysiol* 93: 2600–2613.

Sajaykumar, S. and J.U. Frey (2004). Resetting of 'synaptic tags' is time and activity dependent in rat hippocampal CA1 in vitro. *Neuroscience* 129: 503–507.

Sajikumar, S and J.U. Frey (2003). Anisomycin inhibits the late maintenance of long-term depression in rat hippocampal slices in vitro. *Neuroscience Letters* 338: 147–150.

Sajikumar, S and J.U. Frey (2004). Late-associativity, synaptic tagging, and the role of dopamine during LTP and LTD. *Neurobiology of Learning and Memory* 82: 12–25.

Sajikumar, S, S Navakkode, TC Sacktor, and JU Frey (2005). Synaptic tagging and cross-tagging: The role of protein kinase Mzeta in maintaining long-term potentiation but not long-term depression. *Journal of Neuroscience* 25(24): 5750–6.

Schoville, W.B. and B. Milner (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurorolgy, Neurosurgery and Psychiatry* 20: 11–21.

Senn, W. and S. Fusi (2005). Learning only when necessary: Better memories of correlated patterns in networks with bounded synapses. *Neural Computation* 17: 2106–2138.

Shapley, R., M. Hawken, and D.L. Ringach (2003). Dynamics of orientation selectivity in the primary visual cortex and the importance of cortical inhibition. *Neuron* 38: 689–699.

Sheng, M and S Hyoung Lee (2003). AMPA receptor trafficking and synaptic plasticity: Major unanswered questions. *Neuroscience Research* 46: 127–134.

Shigemoto, R. (2006). Memory traces in short and long term cerebellar motor learning. *Soc. Neurosc. Abstr* 32: 495.3.

Shouval, H.Z., M.F. Bear, and L.N. Cooper (2002). A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proc Natl Acad Sci USA* 99: 10831–10836.

Sjöström, P.J., G.G. Turrigiano, and S.B. Nelson (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron* 32: 1149–1164.

Smolen, P. (2007). A model of long term potentiation simulates aspects of memory maintenance. *PloS One* is 5: e445.

Song, S. and L.F. Abbot (2001). Cortical development and remapping through spike timing-dependent plasticity. *Neuron* 32: 339–350.

Song, S., K. D. Miller, and L. F. Abbott (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nat Neuroscience* 3: 919–926.

Song, S., P.J. Sjostrom, M. Reigl, S. Nelson, and D.B. Chklovski (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biology* 3: 507–519.

Squire, L.R. and S. Zola-Morgan (1991). The medial temporal lobe memory system. *Science* 253: 1380–1386.

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology* 25: 207–222.

Standing, L., J. Conezio, and R.N. Haber (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science* 19: 73–74.

Staubli, U. and D. Chun (1996). Factors regulating the reversibility of long-term potentiaton. *Journal of Neuroscience* 16: 853–860.

Staubli, U. and G. Lynch (1987). Stable hippocampal LTP elicited by ”theta” pattern stimulation. *Brain research* 435: 227–234.

Steele, R.J. and R.G.M. Morris (1999). Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. *Hippocampus* 9: 118–136.

Sterratt, D.C. and D. Willshaw (2008). Inhomogeneities in heteroassociative memories with linear learning rules. *Neural Computation* 20: 311–244.

Stevens, C.F. and Y. Wang (1995). Facilitation and depression at single central synapses. *Neuron* 14: 795–802.

Storkey, A. and R. Valabregue (1997). Hopfield learning rule with high capactiy storage of time-correlated patterns. *Electronics Letters* 33: 1803–1804.

Sutherland, R.J., H.C. Dringenberg, and J.M. Hoesing (1993). Induction of long-term potentiation at perforant path dentate synapses does not affect place learning or memory. *Hippocampus* 3: 141–148.

Tardin, C, L Cognet, C Bats, B Lounis, and D Choquet (2003). Direct imaging of lateral movements of AMPA receptors inside synapses. *The EMBO Journal* 22: 4656–4665.

Teng, E. and L.R. Squire (1999). Memory for places learned long ago is intact after hippocampal damage. *Nature* 400: 675–677.

Toyoizumi, T., J.P. Pfister, K. Aihara, and W. Gerstner (2007). Optimality model of unsupervised spike-timing-dependent plasticity: Synaptic memory and weight distribution. *Neural Computation* 19: 639–671.

Triller, A. and D. Choquet (2003). Synaptic structure and diffusion dynamics of synaptic receptors. *Biology of the Cell* 95: 465–476.

Triller, A. and D. Choquet (2005). Surface trafficking of receptors between synaptic and extrasynaptic membranes: And yet they do move! *Trends in Neurosciences* 28: 133–139.

Tsukada, M., T. Aihara, Y. Kobayashi, and H. Shimazaki (2005). Spatial analysis of spike timing dependent LTP and LTD in the CA1 area of hippocampal slices using optical imaging. *Hippocampus* 15: 104–109.

Turin, G.L. (1960). An introduction to matched filters. *IRE Transactions on information theory* 6: 311– 329.

Turrigiano, G.G. (2000). AMPA receptors ubound: Membrane cycling and synaptic plasticity. *Neuron* 26: 5–8.

Turrigiano, G.G., K.R. Leslie, N.S. Desai, L.C. Rutherford, and S.B. Nelson (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* 391: 845–846.

van Kampen, N. G. (1981). Itô versus Stratonovich. *Journal of Statistical Physics* 24: 175–187.

van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*. North-Holland, Amsterdam, second edition.

van Rossum, M.C.W, G.Q. Bi, and G.G. Turrigiano (2000). Stable Hebbian learning from spike timing-dependent plasticity. *Journal of Neuroscience* 20: 8812–8821.

Wang, H.X., R.C. Gerkin, D.W. Nauen, and G.Q. Bi (2005). Coactivation and timing-dependent integration of synaptic potentiation and depression. *Nat Neurosci* 8: 187–193.

Weber, J. (1956). Fluctuation dissipation theorem. *Physical Review* 101: 1620–1626.

Wenisch, O.G., J. Noll, and J. L. van Hemmen (2005). Spontaneously emerging direction selectivity maps in visual cortex through STDP. *Biological Cybernetics* 93: 239–247.

Whitlock, J.R., A.J. Heynen, M.G. Shuler, and M.F. Bear (2006). Learning induces long-term potentiation in the Hippocampus. *Science* 313: 1093–1097.

Willshaw, D.J. and T.J. Buckingham (1990). An assesment of Marr's theory of the hippocampus as a temporary memory store. *Phil. Trans. R. Soc. B* 329: 205–215.

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

Woo, N.H. and P.V. Nguyen (2002). "Silent" metaplasticity of the late phase of long term potentiation requires protien phosphatases. *Learning and Memory* 9: 202–213.

Yao, H. and Y. Dan (2001). Stimulus timing-dependent plasticity in cortical processing of orientation. *Neuron* 32: 315–323.

Yao, H., Y. Shen, and Y. Dan (2004). Intracortical mechanism of stimulus-timing-dependent plasticity in visual cortical orientation tuning. *Proc Natl Acad Sci USA* 101: 5081–5086.

Young, J.M., W.J. Waleszczyk, C. Wang, M.B. Calford, B. Dreher, and K. Obermayer (2007). Cortical reorganization consistent with spike timing but not correlation-dependent plasticity. *Nature Neuroscience* 10: 887–895.