Internationl School for Advanced Studies (SISSA/Trieste)

Attractors, memory and perception

by

Athena Akrami

Supervisor

Alessandro Treves

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Cognitive Neuroscience Sector
Scuola Internazionale Superiore di Studi Avanzati

January 2010

Summary of the work

In this Thesis, the first three introductory chapters are devoted to the review of literature on contextual perception, its neural basis and network modeling of memory. In chapter 4, the first two sections give the definition of our model; and the next two sections, 4.3 and 4.4, report the original work of mine on retrieval properties of different network structures and network dynamics underlying the response to ambiguous patterns, respectively. The reported work in chapter 5 has been done in collaboration with Prof Bharathi Jagadeesh in University of Washington, and is already published in the journal "Cerebral Cortex". In this collaboration, Yan Liu, from the group in Seattle, carried out the recording experiments and I did the data analysis and network simulations. Chapter 6, which represents a network model for "priming" and "adaptation aftereffect" is done by me. The works reported in 4.3, 4.5, and the whole chapter 6 are in preparation for publication.

Acknowledgements

I thank Alessandro who taught me how to face ambiguity in Limbo.
I thank him again for his (failed) attempts to foster a minimalistic, concise and essential style in my way of thinking.
I thank Bharathi and her family with whom I had memorable moments in Seattle, while exploring the experimental world of Neuroscience.
I thank my family whose omnipresent memories blessed every moment of my life at migrate.
I thank SISSA for ”considerate la vostra semenza: fatti non foste a viver come bruti, ma per seguir virtute e canoscenza”.
I thank my unique friends, Arash, Eleonora, Valentina and Erika who often had to bear the brunt of my frustration, rages and recalcitrance against the world and offered me their open hearts, in lieu, to regain my equanimity.
I thank the gang of my fellow friends with whom I practiced life and refreshed identity.
I thank my homeland whose fervent resistance and bravery inspired me with determination and hope.
and my special thanks to the one whose name may not be revealed explicitly, for the countless support in any sense of the word.

Contents

From the tablet of my heart and soul, Thy image, ever– goeth not. From my recollection, that proudly moving cypress ever– goeth not.

O true (beloved!) from my distraught brain, the image of thy cheek, By the sky's violence and time's wrath,– goeth not.

In eternity without beginning, covenant with thy tress-tip, my heart established. Till eternity without end, it draweth not forth its head; and, from the head of the covenant,– goeth not.

In my heart and soul, my love for thee hath taken a place, such That (even) if my head (life) goeth,-from my soul, my love for thee– goeth not.

Save the load of grief for thee, whatever is in my wretched heart, Goeth from my heart; but from my heart that (grief's load)– goeth not.

If for the pursuit of lovely ones, my heart goeth, 'tis excusable It hath (love's) pain. What may it do if, for remedy-sake, it– goeth not?

Whoever head-bewildered like Hafiz, wisheth not to become Giveth not his heart to lovely ones; and, in pursuit of them,– goeth not.

Chapter 1

Introduction

Footfalls echo in the memory
Down the passage which we did not take
Towards the door we never opened
T.S. Eliot

1.1 General issues

1.1.1 Adaptation

Adaptation is a characteristic of systems as diverse as economics, plant and ant colonies, sense organs and, of particular interest to us here, the brain. It is a fundamental property of neurons in our brain, with moment to moment impact on our perception of the world. Spatial and temporal context constantly shape our perception and the underlying neural responses in a way that can be related to the statistical structure of the inputs (Schwartz et al. 2007). Perceptual adaptation happens when a perceptual system changes its operating properties in response to changes in its inputs. This adjustment can occur over a wide range of time scales. Some forms of visual adaptation, such as the accommodation of lenses to viewing distance, or adjustments of the eye to variations in intensity of light, are so rapid as to escape our notice. Other forms of adaptation occur over seconds or minutes, and can produce dramatic aftereffects, such as the waterfall illusion, tilt aftereffect, and distortion in appearance of shapes, faces and objects. The perceptual learning of a fine discrimination, on the other hand, tends to occur over days, weeks, or years. The development of an individual human visual systems takes years. Finally, evolution has been refining our vision for millions of years, through a process of adaptation through natural selection. Despite their very different time frames, these changes can all be thought of as different forms of adaptation which serve a common function: to harmonize the mind to the world. At a computational level, of course, the mechanisms that produce this fit may be quite different in each case. This thesis focuses on adaptation which produces reversible changes over the relatively short time frames of seconds.

1.1.2 Ambiguous perception

Our brain receives only a restricted amount of information from the visual world mostly because the 3D world is projected onto 2D retina. Each percept is thus the result of a constructive process, starting already at the earliest steps of visual analysis, as demonstrated impressively by many optical illusion studies (Kornmeier and Bach 2005). Although our eyes receive incomplete and ambiguous information, our perceptual system is usually able to successfully construct a stable and faithful representation of the world. However, sometimes our perceptual system fails to produce a stable unambiguous percept, especially if the visual information is equally compatible with different perceptual interpretations, as in the case of ambiguous figures.

Intrinsically ambiguous sensory information thus necessitates prior experience as a constraint on perception to parse stimuli into well defined categories. Priming paradigms have been used extensively in order to study the detailed features of such perceptual influences on relatively short time scales. Priming is the phenomenon where the perception of a given stimulus, or the prime, affects the perception of a succeeding stimulus, or the target, even when the target is presented after a long delay or the prime is not explicitly perceived.

1.1.3 Memory-based prediction

Understanding how the brain translates sensory inputs into useful information about objects is a difficult problem already on theoretical grounds alone. The difficulty of visual perception arises because natural images are both complex and objectively ambiguous.

In the book ”On Intelligence”, Hawkins offers us a picture of how an intelligent agent does perceive and with that survives in an ambiguous world. Hawkins' basic idea is that the brain is a machine to predict the future (Hawkins and Blakeslee 2005). Specifically, hierarchically organized regions of cortex are designed to predict their future input sequences. Perhaps not always far in the future, but far enough to be of real use to an organism. The memory-prediction framework concerns the role of the mammalian neocortex and its associations with the hippocampus and the thalamus in matching sensory inputs to stored memory patterns, and how this process leads to predictions of what will happen in the future. Memory seems to have a direct top-down effect that continuously modulates the incoming sensory data and changes basic sensations (Henson et al 2006).

Linking memory and perception Theoretical models of cognition usually address processes that are said to occur after the perceptual system has completed its work. According to such views, output from a putative perceptual system merely provides the initial information that is fodder for subsequent cognitive operations. This approach is common in the human cognitive psychology literature, in which a set of primitive elements usually taken to be the building blocks of cognition (Biederman 1987; Julesz 1981; Treisman and Gelade 1980). Many researchers in animal associative learning theory have adopted a similar paradigm by assuming that any event that can be learned will correspond to a static perceptual representation that is always fully activated in the presence of the stimulus. This division of behaviour into serial stages reflects the sense-think-act paradigm that is also common in the fields of artificial intelligence and robotics. The strength of this approach is that it separates the processes contributing to behaviour, thereby parsing a complex system into a form that may be more easily understood.

The assumption upon which the above approach is based holds if perceptual representations are static. If perceptual organization changes as a result of experience, however, this should have a direct impact on cognitive processes. For example, the discriminability of a stimulus could increase or decrease over time, leading to changes in the learning rate as learning proceeds.

Many alternatives to the sense-think-act theory have been suggested; one approach is to assume that perceptual representations are not fixed or finite but adapt to the requirements of the task for which they are employed. Gibson, an early proponent of this view, suggested that the perceptual interpretation of an event depends on the observer's history, training, and acculturation (Gibson 1969). No set of primitives exists because the perceptual building blocks themselves are adaptable. Rather than providing an immutable substrate for cognition, representations might instead adapt flexibly to the requirements of cognitive tasks. So, although cognitive processes involved in, for example, learning, might alter perception, the alteration is beneficial because, as a result, perception becomes better tuned to the task at hand.

1.1.4 Neocortex as an autoassociative memory

Any adaptive, memory guided perception should be implemented in (sub)cortical structures. It is a widespread assumption that the substantial anatomical self-similarity of neocortical structure underlies a set of elementary operations that are carried out, on different incoming inputs, by different patches of cortex (Marr 1970; Braitenberg 1978; Abeles 1991). In other words, the local neocortical circuits are characterized by stereotypical physiological and structural features that subserve generic computational operations. A combination of neuronal connectivity, intrinsic cellular and synaptic properties give rise to these basic computations of the cortical microcircuit. It is still a question how to translate such hypothetical universal processing in simple conceptual terms (O'Kane and Treves 1992; Puccini et al. 2007; Rolls et al. 2001). In this context, considerable theoretical and experimental evidence supports the notion that cortical networks have been specialized in evolution to serve a memory function. The system of local, or intrinsic, connections among pyramidal cells has often been thought, for example by David Marr, to implement an autoassociative memory function. A (presumed) Hebb-like synaptic plasticity of these connections may enable them to store a set of local activity patterns. Later, afferent activity containing a fraction of the information associated with one such pattern-which would, by itself, elicit a distorted or partial version of the pattern may trigger recurrent interactions through the local connections, resulting in the original activity pattern, or a very close version of it, with most of its information content, being relayed on for the next stage of processing.

1.2 This Thesis: specific questions

In this thesis, we aim to better understand the dynamical process of memory storage and retrieval in the cortex, and how it affects the perception of ambiguous stimuli, via realistic neural modeling. We would like to assess verious computational abilities that can emerge from a generic network model, considering different structural, temporal or biophysical options. There are three questions that we would like to address here, namely (*i*) the storage capacity of different network structures, (*ii*) network dynamics underlying response to morphed patterns, and (*iii*) the possible mechanism responsible for complex behavioral phenomena like *priming* and *adaptation aftereffects*, which may emerge from attractor dynamics in cortical networks. The next two chapters are devoted to a literature review of some experimental and theoretical studies, concerning visual memory in the cortex and in the model. The rest of the thesis is organized in 3 more chapters, the contexts of which summarize here:

1.2.1 The storage capacity in different network architectures

Once the two neural activity configurations that are stored in a pattern associative memory, are not orthogonal, interference may occur. The influence of such crosstalk can be controlled in part with sparse encoding of the stored patterns, effectively by setting high threshold for the firing of output units. In other words, the original vectors need not to be orthogonal, but if they are too similar, some interference will still be likely to occur. In general, the fact that interference is a property of neural network pattern associator memories is of interest, because interference is a major property of human memory. It is one of the fundamental cognitive neuroscience questions, how the brain copes with interference (Rolls and Deco 2002). If patterns are too similar to be stored in associative memories, then one solution that the brain seems to use repeatedly is to expand the encoding to a form in which the different stimuli are less correlated, that is, more orthogonal, before they are presented as final patterns to a memory storehouse to be stored. In autoassociative networks, the problem raised by interference is related to the correlation between the attractors that formed by changing the synaptic weights in order to store patterns. Later on, during the retrieval phase, this interference affects the reading out process. Quantitatively, in chapter 4, we aim to compare the retrieval properties, in particular the storage capacity, of different network structures which differ in *architecture*, *pattern statistics*, and *pattern generation*. We will explore the effect of diluted connectivity, fast-noise, and the number of stored patterns on the retrieval process of different types of patterns in different network structures.

1.2.2 The response to morphed patterns in analog networks

Can we see retrieval dynamics by probing a local cortical network? In network models of associative memory, in which information is supposedly carried with temporally coarse resolution by rates of emission of action potentials, units have been chosen in which a single output variable (often even binary) represented some short-time average of neuronal spiking rate. The focus then has been on studying the steady-state behaviour, i.e. the attractor state, with the dynamics either neglected altogether or artificially defined (for example, in terms of ”updatings”) merely in order to fully specify the model (Amit et al. 1987).

There have been few studies trying to understand the collective properties of large networks, with more realistic description of the relevant biophysical processes, to address questions like: over what time scales can a local firing pattem -elicited by an incoming stimulus- reach a steady-state distribution? Which of the biophysical parameters at the single unit level contributes to set those time scales? (Treves 1993; Matsumoto et al. 2005) Yet the need to obtain an analytical grasp of the dynamics of large networks, endowed with specific synaptic weights, coding for different memory patterns, is apparent, as new experimental data is beginning to throw light on the temporal course of information processing, e.g. in neocortical association areas of primates (Tovee et al. 1993; Rolls and Tovee 1994; Sugase et al. 1999; Rainer and Miller 2002; Akrami et al. 2009; Woloszyn and Sheinberg 2009).

In chapter 4, after the discussion about storage capacity of different network structures, we extend an analytical approach and simulation to study the transient dynamics of networks of threshold-linear model neurons that include, as a necessary ingredient of the relevant computational mechanism, a simple feature of pyramidal cell biophysics: firing rate adaptation. Such analysis yields the attractor states of the network and the full spectrum of time constants of the transients associated with different steady states (Treves 1993). We provide the analytic solution for the dynamics with which the network model responds to an ambiguous pattern, that is ”equally” correlated with two stored patterns. We then generalize the model to probe the response to the other morph levels with numerical simulations. In general, studying these transients, in response to external inputs that are morphed between two stored patterns, and are influenced by previous activity of the network, may shed light on the possible contribution of attractor dynamics to perceptual boundary shifts. We leave this issue to be further discussed in chapter 6.

1.2.2.1 The response to morphed patterns in IT Cortex, relevance of attractor dynamics

Electrophysiological evidence for attractor dynamics? Later, in Chapter 5, we present a piece of experimental data, collected in a paradigm in which the animal has to judge about an ambiguous stimulus. We recorded from individual neurons in IT cortex while monkeys performed a classification task on morphed visual images. We report here a population of IT neurons whose responses evolve gradually over the course of a trial, first representing parametrically the morphed image and later converging to represent one of the 2 ”categories”, corresponding to the 2 images. Then we discuss the results in the context of the attractor network simulations, which highlight key features of IT dynamics and provide insights into the local network properties that might underlie them.

The convergence of IT activity from a stimulus-based representation to a category-based representation was asymmetric, in that only responses to the morphed images that resemble the effective stimulus for an individual cell converge, whereas responses to morphed images that resemble the ineffective stimulus remain segregated by morph level. An asymmetric convergence may result from multiple mechanisms, of course. We have tried to assess 2 possible underlying mechanisms, a gradual decay of the response over time, due to the firing rate adaptation, and attractor dynamics in the local recurrent networks. With our first simulation, we could rule out the possibility that the convergence was the result of simple linear decay of neural responses. A linear decay of responses had an equivalent effect at each individual morph level, and did not produce a change from the linear dependence visible at stimulus onset, whereas the operation of a simple attractor network produced qualitatively similar convergence to that observed in the neural data, allowing a more detailed interpretation of the observations.

1.2.3 Experience-dependent dynamics of attractor networks in response to morphed patterns

Can local attractor dynamics help us understand the effect of recent experience?

In chapter 6, we aim to see whether a generic model that in chapter 5 could replicate the neurophysiologycal observations, can be used to model more complicated neural and behavioural phenomena, including ”priming” and ”adaptation aftereffects”. Puzzled by the complexity, even controversy, in experimental results of priming paradigms, we hypothesized that this complexity may stem from the unfolding in time of the same basic mechanisms, and we asked what determines the direction and magnitude of the effects in priming experiments simulated in an attractor network model.

We postulated that the notion of local attractor networks, subject to adaptation and distributed over multiple cortical patches (O'Kane and Treves, 1992; Treves 2005), may provide a conceptual framework, accessible to neurophysiological investigation, with behavioral predictions, for analyzing the expression of the effects of recent experience on a particular class of visual stimuli.

In chapter 6, the behavior of the system beyond the various bifurcation lines is studied both through numerical integration of coupled nonlinear differential equations and through numerical simulations of the model. This allows us to characterize the phase diagrams of networks of sparsely connected excitatory units endowed with global inhibition and firing rate adaptation.

They show a very rich behavior. Depending on the values and temporal characteristics of adaptation, the strength of external input, the time duration of the presentation of the prime or target and the time lag between these two, we found different emergent behaviors of the network - a complexity which might correspond to the complexity of the psychophysical findings.

Chapter 2

Visual Adaptation

2.1 Repetition: perceptual consequences

When stimuli are repeated, neural activity is usually reduced. This neural repetition effect has been reported at multiple spatial scales, from the level of individual cortical neurons in monkeys (Li et al. 1993;Sobotka and Ringo 1996; Liu et al. 2009) to the level of hemodynamic changes (measuring the pooled activation of millions of neurons) in humans using functional magnetic resonance imaging (e.g. fMRI (Demb et al. 1995; Buckner et al. 1995)). Repetition-related reductions also occur at multiple temporal scales, both in their longevity from milliseconds (Sobotka and Ringo 1996) to minutes (Henson et al. 2000) and days (van Turennout 2000) and in the latency of their expression (Henson et al. 2004). The phenomenon also occurs in multiple brain regions, and across an impressively large number of experimental conditions. All of these changes in neural activities are normally accompanied with behavioral manifestations.

In this chapter, we review experimental data describing how single cells, BOLD signals, and perceptions are affected by recent experience and repetition, and describe models that link these two datasets. We end by discussing some of the many gaps in the data and in our theoretical understanding of contextual effects.

2.1.1 Perceptual priming

The classical definition of priming stands for the behavioral phenomenon of improved processing of a stimulus following prior exposure to that stimulus (Schacter and Buckner 1998; Tulving and Schacter 1990). Behaviorally, priming typically manifests as increased accuracy and/or faster speed in making judgments on a stimulus that has been previously encountered. It is thought to reflect an implicit form of memory and learning, as it does not involve explicit memory of the prior experience (Grill-Spector 2008; Schacter and Buckner 1998; Tulving and Schacter 1990).

In a typical priming experiment, subjects are shown an initial stimulus (prime) and are required to make a decision (e.g., categorize the stimulus) or produce a response (generate a word) on a subsequent stimulus (test) that is identical or related to the initial stimulus (e.g., the same object in different views, or a new object that is related perceptually, conceptually, or semantically to the prime). The priming effect (i.e., improvement in performance) is largest when the repeated stimulus is identical to the initial stimulus (prime). In some behavioral paradigms of priming, many intervening stimuli occur between the test and the prime. However, in other paradigms, the test immediately follows the prime. Priming, could be evident also in producing a shift in categorical boundary, in classification tasks, in which the subject is asked to judge about the similarity of a morphed stimulus with two target stimuli (Van Rijsbergen et al. 2008 ; Furl et al. 2007; Daelli and Treves, accepted)

One particular striking aspect of priming is that it can be manifested after a single exposure to an object and is preserved in timescales ranging from seconds to even an year (Cave 1997).

The level of priming is modulated by several factors such as the number of stimulus repetitions, the number of intervening stimuli, and the time between repeats. The magnitude of response time (RT) priming increases with the number of stimulus repetitions both in short timescales (seconds/minutes) and in longer timescales (days and weeks), and this advantage remains over weeklong delays compared to single exposures of stimuli (Brown et al. 1996). Similarly, RT priming is largest when there are no intervening stimuli between the prime and the test stimulus and when the temporal interval between them is shortest. Thus, immediate repetitions produce a larger priming effect compared to when repetitions occur after several minutes or days (van Turennout et al. 2000 ; Sayres and Grill-Spector 2006).

2.1.1.1 Priming as a memory phenomenon

Foremost is the interest in priming as an example of ”implicit memory”. Implicit memory represents an effect of prior experience on behaviour, in the absence of conscious awareness of the past (Graf et al. 1984). This term arose from studies of amnesiac patients with damage to medial temporal lobe structures, who can show priming even though they appear unaware of any prior exposure to the primed stimulus (i.e. lack ”explicit” memory). Warrington and Weiskrantz (1974), for example, showed that amnesiacs were impaired relative to controls on the ”direct” memory tests of recall and recognition for previously studied words. On indirect tests however, in which the participants simply tried to identify degraded versions of words, amnesiacs showed an advantage for studied words that was equivalent to that in controls.

Findings like these led to the proposal that priming reflects the operation of ”non-declarative” memory systems supported by regions outside the medial temporal lobe, as distinct from the ”declarative” memory system that is impaired following medial temporal damage (Squire and Cohen 1984. This proposal has been bolstered by reports of a few patients with more posterior cortical lesions, who show intact performance on direct memory tests but impaired performance on certain indirect tests (Gabrieli et al. 1995; Keane et al. 1995). Priming is usually distinguished from other types of implicit memory that are intact in amnesia, such as skill-learning (Milner et al. 1968), because it can occur after a single stimulus presentation, rather than requiring repeated trials (Hauptmann and Karni 2002), and because it is normally specific to a particular stimulus or process, unlike a generalised skill (see Gabrieli 1998; Schacter and Tulving 1994 for reviews). The association of priming with implicit memory is further supported by data from healthy individuals, such as functional dissociations between direct and indirect tests as a function of study task (Jacoby and Dallas 1981) or retention interval (Tulving et al. 1982), and reports of intact priming when the prime is subliminal (Forster and Davis, 1984) or shows no evidence of explicit memory (Stark and McClelland, 2000).

Nonetheless, demonstrations that priming can occur in the absence of explicit memory do not imply that priming effects measured under normal conditions are a pure reflection of implicit memory. Even though an indirect memory test does not refer participants to previous encounters with stimuli, participants may voluntarily, or involuntarily, recollect such encounters. As a consequence, considerable effort has been devoted to developing methods that dissociate implicit and explicit contributions to memory tasks (e.g. Hayman and Tulving 1989; Jacoby et al. 1993; Richardson-Klavehn and Gardiner 1995; Schacter et al. 1989).

2.1.2 Adaptation aftereffect

In some conditions, experiencing the adapter subsequently distorts perception of particular stimulus attributes, typically biasing perception towards the opposite of the adapting stimulus. This phenomenon, termed an adaptational aftereffect can affect the perception of various stimulus attributes and is often linked directly to diminished responses of feature-selective neurons in the visual cortex (Blakemore and Campbell 1969; Coltheart 1971; Tolhurst and Thompson 1975; Barlow 1990; Bednar and Mukkulainen 2000). Historically, stimuli demonstrating such aftereffects are impoverished, containing little information other than the relevant adapting feature, such as a particular colour or direction of motion.

Many diverse stimuli are effective in adaptation, and the resulting aftereffects, though they impact on different aspects of perception, have much in common. For example, most aftereffects display interocular transfer if the adapting and test eye are different (Gibson 1937; Wade et al. 1993), exhibit storage across blank periods (Spigel 1960; Thompson and Movshon 1978), and are restricted to confined spatial zones in the visual field (Gibson 1937; Anstis and Gregory 1965). Also, aftereffects have a finite duration that depends upon adaptor strength and exposure time (Gibson and Radner 1937; Wolfe 1984; Magnussen and Greenlee 1987; Hershenson 1989).

But how does adaptation to a stimulus predispose the brain towards perceiving its opposite It is alluring to ascribe aftereffects to the aberrant functioning of sensory neurons that have been previously overstimulated. Orientation aftereffects would then derive from an imbalance among orientation-selective neurons in the primary cortical area V1, and motion aftereffects might similarly arise from direction-selective neurons in the motion-selective middle temporal (MT) cortex, or adapting to a tilted line causes a subsequently presented vertical line to appear tilted in the direction opposite of the adaptation line. These observations are thought to tap into competing populations of orientation tuned neurons in the visual cortex (Coltheart 1971; Clifford et al 2000; Jin et al 2005; Liu et al. 2007). But while there is little doubt that such feature-selective neurons contribute to the expression of aftereffects, their precise role is difficult to pinpoint. The fact that very different stimuli cause similar aftereffects poses a challenge for any theory of adaptation tied to a particular functional architecture (van der Zwan and Wenderoth 1995; Clifford 2002; for a review see Leopold and Bondar 2005). Within the domain of orientation, for example, adaptation to luminance-defined bars, known to activate V1 neurons, or to more complicated stimuli unlikely to activate V1 neurons, causes aftereffects with very similar properties (van der Zwan and Wenderoth 1995; Paradiso et al. 1989; Joung et al. 2000).

2.1.3 Stereotypical dynamic properties for different adaptation aftereffect

Face-identity aftereffect has stereotypic dynamic properties that resemble those described previously for traditional aftereffects. In testing the effects of adapting and test duration using the method of constant stimuli, Leopold et al (2005) found that, as with traditional aftereffects (tilt aftereffect Magnussen and Johnsen 1986 or motion aftereffect), the face identity aftereffect grew stronger as a function of adaptation time, and weaker as a function of test duration. Clearly, some aftereffects bear a signature of early processing (e.g. retinotopically restricted adaptation fields) and others of late processing (e.g. invariance to scale and position). But it may be that the other shared aspects of their phenomenology, such as their temporal dynamics, can be attributed to a stereotypical activation of the entire visual cortex that is independent of the specific stimulus.
These results underscore the difficulty in pinpointing the neural locus of aftereffects in general. Given their positional invariance, face aftereffects are unlikely to derive exclusively from circuits in the primary visual cortex. At the same time, their dynamics suggest a high degree of mechanistic overlap with simpler aftereffects, which are often considered to have their origins in early retinotopic processing. The question therefore arises, how is it possible that such diverse aftereffects, apparently resulting from adaptation of distinct populations of neurons in the visual cortex, share temporal properties to such a degree?

One possibility is that the wiring among visual neurons throughout the brain is sufficiently stereotyped that the same dynamics arise wherever competing groups of neurons become differentially adapted. This notion might allow for the present results to fit into a scheme not so different from traditional accounts of simple aftereffects (Coltheart 1971). Such notions normally invoke antagonistic connectivity between a pair of alternative stimulus representations, which could be orientation-selective neurons in V1 or face-selective neurons in the inferotemporal cortex, with the circuit dynamics generated locally in each case. While this may be a parsimonious explanation for the observed results, there is little evidence for it.

A second possibility might be that aftereffects are, by nature, a product of interactions between different processing stages in the brain. With respect to the cortical hierarchy, a perceptual aftereffect might never be accurately described as purely ”low-level” or ”high-level”, since very different visual stimuli adapt the same, multiple processing stages, albeit in different ways. While the evidence for this hypothesis is also minimal, recent neuroimaging studies do verify that large-scale networks in the brain, at different cortical processing stages, are affected by periods of prolonged stimulus adaptation (Taylor et al. 2000; Tolias et al. 2001).

2.1.4 Switch between adaptation aftereffect and priming

Recent experience with clear, prototypical stimuli may however induce complex effects on the subsequent perception of ambiguous ones, ranging from attraction (priming) to repulsion (adaptation aftereffects).

In experiments with binocular rivalry, Pearson and Clifford 2005 argued that unambiguous stimuli produce aftereffects, whereas ambiguous stimuli produce priming. Although in this study they mainly used a different type of bistable stimuli (i.e., ambiguous motion), the results suggest that the difference between ambiguous and unambiguous primes may not be qualitative, but could be a quantitative difference in the temporal pattern of the priming effect following an exposure to ambiguous and unambiguous stimuli. Kanai and Verstraten (2005) found that adaptation to unambiguous motion leads to both aftereffect and priming depending on the adaptation-test interval. With short ISIs, unambiguous motion indeed produces an aftereffect, but later the effect of adaptation switches to priming. A number of studies in the past consistently revealed that a brief exposure produces a priming effect, whereas prolonged viewing of the same stimulus produces an opposite effect (Huber and O'Reilly 2003; Kanai and Verstraten 2005; Long et al. 1992 and Pinkus and Pantles 1997). In the study by Pearson and Clifford, the blank interval was relatively short (1 s), where a predominant negative bias is expected to occur after adapting to an unambiguous prime. Thus, it is plausible that an unambiguous prime turns to a positive bias with a longer inter-stimulus interval. Parsimoniously, this would obviate the need to assert the existence of two categorically distinct priming or adapting mechanisms depending on the ambiguousness of the priming stimulus.

In a very recent study in our group, Valentina Daelli, applied a well-established aftereffects paradigm to test how recent experience influences perception of complex, everyday, non-face objects. She used morphing software to create continua between paired images of animals, plants or artifacts, and then tested how the discrimination of an ambiguous image is affected by previous adaptation to one of the end-points of the morphing continuum (Fig. 2.1). Based on the simple hypothesis of common but distributed local mechanisms, which just summate their effects, she intended to experimentally isolate distinct factors: (a) the contribution of early, low-level processes; (b) the time scale and direction of perceptual shift; (c) the role of semantic, post-perceptual processes; and (d) the influence of intervening distractors.

Interestingly, she found a basic shift from aftereffects to priming effects as the delay lengthens between experiencing a prototype and seeing the ambiguous stimulus(2.2). On the other hand,

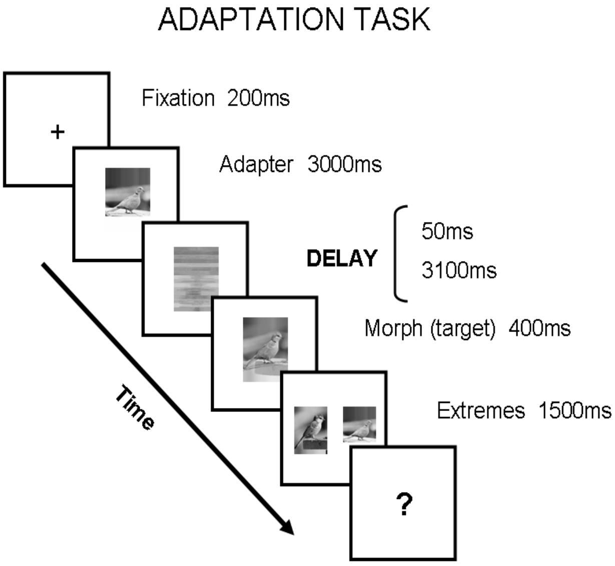


Figure 2.1: **Schematic view of the behavioral paradigm, adapted in our group**

she observed the replacement of negative (aftereffect) with positive (priming) bias, given the adapter is chosen from one of the intermediate morphed patterns that has high similarity with both of the two end points. On top of this main finding, she describes a pattern of aftereffect modulation by the perceptual and semantic similarity to the prototype that does not appear to match any simplistic box model. These results provide a challenge to the development of mechanistic neuronal models of ecologically relevant visual perception.

Overall, this evidence suggests the hypothesis that the same fundamental mechanisms may underlie aftereffects across low- and high-level stimulus dimensions, although expressed by different neural populations and subject to different influences. Indeed, within a simplified network model of a single (local) neural population, ”aftereffects” result simply from the presence of firing rate adaptation (Menghini et al 2007), a ubiquitous property of cortical neurons. Their temporal dynamics, moreover, and whether they can turn into priming effects, may strongly influenced, in the model, by the structure of attractors, whether the patterns of activity corresponding to unambiguous stimuli have been previously stored as local attractor states.

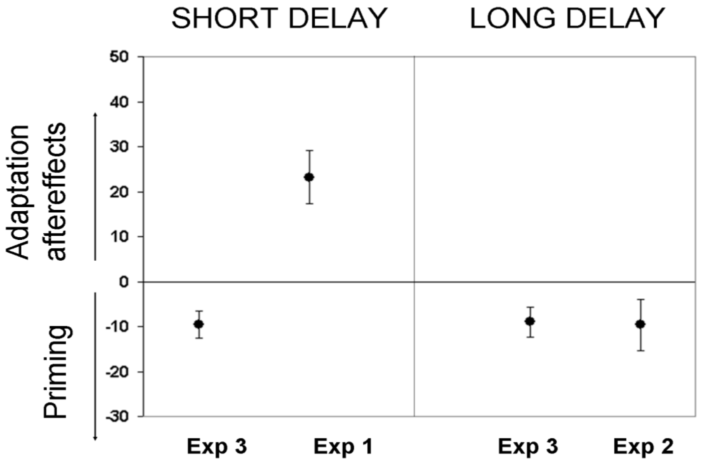


Figure 2.2: **Behavioral result, summarized in terms of "category shift".** Exp1: The adapter is one of the end points (Full adapter), long delay. Exp2: Full adapter, short delay. Exp3: The adapter is one of the ambiguous morphs.

2.2 Masking

Firing of neurons within a cortical area normally continues to a visual stimulus for several hundred milliseconds. Over this long period of time different factors will influence the firing of neurons in a given cortical area. Initially the firing will be based mainly on incoming information from the preceding cortical area (feed-forward information), but at varying temporal intervals different feed-back mechanisms will play a modulating role.

In masking paradigms, a stimulus is rendered invisible through the presentation of a second stimulus shortly after the first. Over the years, authors have typically explained masking by postulating some early disruption process. In these feedforward type explanations, the mask somehow "catches up" with the target stimulus, disrupting its processing either through lateral or interchannel inhibition. However, studies from recent years indicate that visual perception and most notably visual awareness itself may depend strongly on cortico-cortical feedback connections from higher to lower visual areas (Fenske et al. 2006). This has led some researchers to propose that masking derives its effectiveness from selectively interrupting these reentrant processes. In an EEG experiment, Fahrenfort et al used electroencephalogram measurements to determine what happens in the human visual cortex during detection of a texture defined square under non-masked (seen) and masked (unseen) conditions. Electroencephalogram derivatives that are typically associated with reentrant processing turn out to be absent in the masked condition. Moreover, extrastriate visual areas are still activated early on by both seen and unseen stimuli, as shown by scalp surface Laplacian current source-density maps. This conclusively shows that feedforward processing is preserved, even when subject performance is at chance as determined by objective measures. From these results, one can conclude that masking derives its effectiveness, at least partly, from disrupting reentrant processing, thereby interfering with the neural mechanisms of figure-ground segmentation and visual awareness itself.

Rolls and Tovee 1995 run an experiment to investigate the duration of the time for which cortical neurons respond when the identification of a visual stimulus is just possible. In this study, authors presented a test face stimulus for 16 ms, and followed it at different intervals by a masking stimulus (either an N-O pattern or a face) while recording from single neurons in the temporal visual cortex of macaques. When there was no mask the cells responded to the 16 ms of the test stimulus for 200-300 ms, far longer than the presentation time. The authors suggest that this reflects the operation of a short-term memory system implemented in cortical circuitry. If the mask was a stimulus which did not stimulate the cells (either a non-face pattern or a face which was a non-effective stimulus for that cell), then, as the interval between the onset of the test stimulus and the onset of the mask stimulus (the stimulus onset asynchrony) was reduced, the length of time for which the cells fired in response to the test stimulus was reduced. It is suggested that this is due to the mask stimulating adjacent cells in the cortex which by lateral inhibition reduce the responses of the cells activated by the test stimulus. When the stimulus onset asynchrony was 20 ms, face-selective neurons in the inferior temporal cortex of macaques responded for a period of 20-30 ms before their firing was interrupted by the mask. With the same test-mask stimulus onset asynchrony of 20 ms, humans could just identify which of six faces was shown.

In conclusion, the effect of the mask is to reduce the total information available for the cell, the size of the information peak, and the length of time it is signaling information. It is notable that the information in the no-mask condition did outlast until the end of the stimulus by as much as 200 to 300 msec, indicating some short-term memory trace property of the neuronal circuitry which can be interpreted as visual memory, perhaps implemented by the recurrent collateral connections, made between nearby pyramidal cells in the cerebral cortex.

2.3 Neural substrates for perceptual priming

As with other cognitive tasks, priming is likely to involve a sequence of more or less distinct processes, such as stimulus identification, memory retrieval and response preparation and execution (for a review, see Meyer et al. 1988). Although reaction time measures of accelerated identification have provided important insights into cognitive variables that affect priming, it is long recognized that the neural architecture of the underlying brain processes are not deductible solely from the timing of overt responses.

It has become evident, that at least two mechanisms can be characterized, one is improved stimulus identification through implicit learning in perceptual and conceptual representational brain regions (Tulving and Schacter 1990; Wiggs and Martin 1998), and the other is facilitated expression of the overt response as a result of learning in fronto-striatal networks (Bayley et al. 2005; Dobbins et al. 2004; Maccotta and Buckner 2004; Poldrack and Gabrieli 2001; Wig et al. 2005). Psychophysical evidence for contextual effects is particularly widespread in vision (Clifford and Rhodes 2006), including motion (Wohlgemuth 1911; Kanai and Verstraten 2005), brightness(Adelson 1999; Eagleman et al. 2004), orientation (Gibson 1937), blur(Webster et al. 2002), faces(Webster et al. 2004; Leopold et al. 2005; Furl et al. 2007; Van Rijsbergen et al. 2008), and objects (Daelli and Treves, accepted). Contextual influences also extend to other modalities, such as audition (Oxenham 2001) and somatosensory processing(Wallace 2004). However, Neurophysiological evidence for contextual influences is most extensive in the early visual processing of orientation and motion, and in the whisking activity in the rodent somatosensory cortex; but context is also likely to influence the neural processing of many other attributes, including color and border ownership. In the following section we bring, at first, a general review of temporal dynamics of neural responses manipulated by the context, and then later, we continue with reported neural processes correlated with perceptual phenomena.

2.3.1 Timecourse of information processing in visual cortex

Experimental studies on humans provide a rich phenomenology of behavioral effects, but it is hard to infer from these data the mechanisms underlying such phenomena at the neuronal or network levels. There have been several experimental studies trying to shed light on the temporal course of information processing in different brain areas, but mostly those that are involved in more abstract representation of external world, e.g. prefrontal or inferotemporal cortex (Tovee et al. 1993; Sugase et al. 1999). Recent electrophysiological experiments on behaving monkeys provide invaluable information on the dynamics at the neuronal level underlying priming-like effects. Pair associate tasks used to probe neuronal correlates of learning of associations between stimuli present striking similarities with human priming protocols. In such tasks, the monkey learns associations between arbitrary visual stimuli (Erickson and Desimone 1999; Rainer et al. 1999; Liu et Jagadeesh 2008; McMahon and Olson 2007).

Rainer and Miller studied the timecourse of neural activity in the primate prefrontal (PF) cortex during an object delayed-matching-to-sample (DMS) task, to assess the effects of experience on this timecourse. They have conducted the task using both novel and highly familiar objects(Rainer and Miller 2002. In addition, noise patterns containing no task-relevant information were used as samples on some trials. Comparison of average prefrontal ensemble activity relative to baseline activity generated by objects and noise patterns revealed three distinct activity periods. (i) Sample onset elicited a transient sensory visual response. In this sensory period, novel objects elicited stronger average ensemble activity than both familiar objects and noise patterns. (ii) An intermediate period of elevated activity followed, which began before sample offset, and continued well into the delay period. In the intermediate period, activity was elevated for noise patterns and novel objects, but near baseline for familiar objects. (iii) Finally, after average ensemble activity reached baseline activity at the end of the intermediate period, a reactivation period occurred late in the delay. Experience had little effect during reactivation, where activity was elevated for both novel and familiar objects compared to noise patterns. They show that the ensemble average resembles the activity timecourse of many single prefrontal neurons. These results suggest that PF delay activity does not merely maintain recent sensory input, but is subject to more complex experience-dependent dynamics. This has implications for how delay activity is generated and maintained.

Sugase et al. (1999) found that global information is represented at the initial transient firing of a single face-responsive neuron in inferior-temporal (IT) cortex, and that finer information is represented at the subsequent sustained firing. A feed-forward model and an attractor network was suggested by Matsumoto et al. (2005) to reproduce this dynamics. The attractor network, specifically an associative memory model, is employed to elucidate the neuronal mechanisms producing the dynamics. The results obtained by computer simulations show that a state of neuronal population initially approaches to a mean state of similar memory patterns, and that it finally converges to a memory pattern. This dynamics qualitatively coincides with that of face-responsive neurons. The dynamics of a single neuron in the model also coincides with that of a single face-responsive neuron.

2.3.2 Repetition: average network effects

Using the haemodynamic techniques of functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), priming-related effects have been observed in numerous regions of the human brain, with the specific regions depending on the type of stimulus and the manner in which it is processed. The most common finding is a decreased haemodynamic response for primed versus unprimed stimuli, though priming-related response increases have been observed.

2.3.2.1 Repetition adaptation in single neurons

Neural responses throughout the sensory system are affected by stimulus history. In the inferotemporal cortex (IT) an area important for processing information about object shapes, there is a substantially reduced response to the second presentation of an image. Understanding the mechanisms underlying repetition suppression may provide important insights into the circuitry that generates responses in IT. In addition, repetition suppression may have important perceptual consequences. The characteristics of repetition suppression in IT are poorly understood, and the details, including the interaction between the content of the first and second stimulus and the time course of suppression, are not clear. There have been several studies, looking at different aspects of the effects produced by repetition in IT (Liu et al. 2009;Sawamura et al. 2006;; Woloszyn and Sheinberg 2009; McMohan and Olson 2007;).
Stimulus-specific repetition-related reductions in firing rates have been found in physiological recordings of neurons in macaque inferior temporal (IT) cortex (Miller et al. 1993; Sobotka and Ringo 1995; Sobotka and Ringo 1994. These repetition effects have been reported for awake behaving animals performing various visual tasks (e.g. match to sample (Miller et al 1993, recognition memory (Brown, and Xiang 1998)), as well as in anesthetized animals (Miller and Desimone 1993), and occur for both behaviorally relevant and irrelevant stimuli. RS effects are stimulus specific in the sense that they do not appear to reflect a global reduction in the firing of a population of neurons to all subsequent stimuli (Miller et al 1993). Nonetheless, the precise definition of a ”stimulus” is important, because neural RS can exhibit invariance to some changes in stimulus dimensions (such as the size or position of an object (Lueschow et al. 1994). The degree of RS depends on several factors, although the precise effect of each factor can vary with brain region. RS persists despite many intervening stimuli, particularly in more anterior regions of IT cortex (Miller et al. 1993; Xiang and Brown 1998). RS increases with more repetitions of the same stimulus, such that firing rates resemble an exponentially decreasing function of presentation number

(Li et al 1993). This reduction in firing rates occurs primarily for visually excited neurons, and the greatest reduction tends to occur for neurons that were most active on the first presentation (Li et al. 199). Importantly, several studies indicate that RS onsets rapidly, as fast as 70-80 ms in some perirhinal neurons (Xiang and Brown 1998) and with a mean population latency of around 150 ms in IT (Ringo 1996). Indeed, Xiang and Brown (Xiang and Brown 1998) suggested that these effects are too fast for "top-down" influences.

Liu et al. (2009) examined the time course of suppression in IT by varying both the duration and stimulus content of two stimuli presented in sequence. The data show that the degree of suppression does not depend directly on the response evoked by the first stimulus in the recorded neuron. Repetition suppression was also limited in duration, peaking at approx 200 ms after the onset of the second (test) image and disappearing before the end of the response. Neural selectivity to a continuum of related images was enhanced if the first stimulus produced a weak response in the cell. The dynamics of the response suggests that different parts of the input and recurrent circuitry that gives rise to neural responses in IT are differentially modulated by repetition suppression. The selectivity of the sustained response was preserved in spite of substantial suppression of the early part of the response. The data suggest that suppression in IT is a property of the input and recurrent circuitry in IT and is not directly related to the degree of response in the recorded neuron itself.

2.3.2.2 Match enhancement

One should discuss the difference between "match enhancement" and "match suppression" (Baylis and Rolls 1987; Eskandar et al. 1992; Miller et al. 1993; Miller and Desimone 1994). Even though many single cells show suppression of responses to repetition, some show enhancement. In a recent study, for instance, responses to repetition were suppressed, enhanced or unaffected in approximately equal proportions in macaque IT (Eifuku et al. 2004). Functional imaging studies show a similarly mixed range of effects across areas. Zago et al. 2005 demonstrated repetition suppression for everyday objects in most ventral areas, but several parietal and occipital areas showed significant repetition enhancement.

Previous single-cell studies of PFC reported a variety of repetition effects ranging from suppression (Kubota et al. 1980; Rainer et al. 1999; Zaksas and Pasternak 2006) to enhancement (Rainer et al., 1999) of the responses to a repeated stimulus.

Woloszyn and Sheinberg 2009 explored neural dynamics in IT cortex during a visual working memory task. From the collection of match analyses, they conclude that the population of ITC cells retained enough information about the encoded sample to initially respond more strongly/quickly and to later dampen its response to a matching comparison stimulus. Crucially, this memory trace persisted through the presentation of intervening sensory information.

In fact, the emergence of match suppression lagged visual selectivity by, on average, 35.37 ms, consistent with recent ITC studies exploring the dynamics of repetition suppression (Sawamura et al. 2006; McMahon and Olson 2007; Liu et al. 2009). This suggests to us that match suppression could be a marker of finished cortical processing, which occurs faster for the matching compared with the nonmatching stimuli. The earliest significant differences (in this case match enhancement) always arose in the single-unit activity. This finding supports the conclusion that the observed match enhancement is the result of neural processes local to ITC.

Match enhancement has been hypothesized previously to reflect the augmentation of task-relevant visual representations (Miller and Desimone 1994; Miller et al. 1996. In agreement with the original proposal (Desimone and Duncan 1995; Miller and Cohen 2001), can be postulated that, by biasing information flow in posterior visual areas, feedback from PFC and/or PRh, as well as intrinsic ITC circuitry, allows ITC neurons to respond more strongly and/or more quickly to the remembered stimulus. The work by Woloszyn and Sheinberg 2009 has clarified the nature of this augmentation by showing that it specifically targets those neurons essential for the representation of the stimulus, because nonpreferred stimuli did not elicit an enhanced response. Furthermore, it has been shown it to be more evident in ITC than previously thought, likely a result of the selectivity of the sample of neurons.

2.3.2.3 Repetition suppression in fMRI signals

Repetition suppression (RS) is a reduction of neural response that is often observed when stimuli are presented more than once. Many functional magnetic resonance imaging (fMRI) studies have exploited RS to probe the sensitivity of cortical regions to variations in different stimulus dimensions; however, the neural mechanisms underlying fMRI-RS are not fully understood.

Epstein et al. 2008 tested the hypothesis that long-interval (between-trial) and short-interval (within-trial) repetition suppression effects are caused by distinct and independent neural mechanisms. Subjects were scanned while viewing visual scenes that were repeated over both long and short intervals. Within the parahippocampal place area (PPA) and other brain regions, suppression effects relating to both long- and short-interval repetition were observed. Critically, two sources of evidence indicated that these effects were engendered by different underlying mechanisms. First, long- and short-interval repetition suppression effects were entirely noninteractive even although they were measured within the same set of trials during which subjects performed a constant behavioral task, thus fulfilling the formal requirements for a process dissociation. Second, long- and short-interval RS were differentially sensitive to viewpoint: short-interval repetition suppression was only significant when scenes were repeated from the same viewpoint while long-interval RS less viewpoint-dependent. Taken together, these results indicate that long- and short-interval fMRI-RS are mediated by different neural mechanisms that independently modulate the overall fMRI signal. These findings have important implications for understanding the results of studies that use fMRI-RS to explore representational spaces.

Repetition lag In (Henson et al. 2003), the modulation of repetition effects by the lag between first and second presentations of a visual object during a speeded semantic judgment task was examined using both scalp event-related potentials (ERPs) and event-related functional magnetic resonance imaging (efMRI). Four levels of lag were used within a single session, from zero to one, to tens of intervening stimuli, and which allowed partial separation of the effects of interference from the effects of time. Reaction times (RTs) showed that the magnitude of repetition priming decreased as lag increased. The ERP data showed two distinct effects of repetition, one between 150 and 300 ms post stimulus and another between 400 and 600 ms. The magnitude of both effects, particularly the earlier one, decreased as lag increased. The fMRI data showed a decrease in the haemodynamic response associated with repetition in several inferior occipitotemporal regions, the magnitude of which also typically decreased as lag increased. In general, and contrary to expectations, lag appeared to have mainly quantitative effects on the three types of dependent variable: there was little evidence for qualitative differences in the neural correlates of repetition effects at different lags.

2.3.3 Is repetition suppression related to behaviour?

The priming of a stimulus by another has become an important tool for exploring the neural underpinnings of conceptual representations. However, priming effects can derive from many different types of relationships and it is important to distinguish between them in order to be able to develop theoretical accounts of the representation of conceptual knowledge.

One way to examine the correlation between fMRI-adaptation and visual repetition priming is to dissociate behavioral performance from neural activity. For example, Henson et al. 2000 used familiar and unfamiliar stimuli to manipulate stimulus familiarity. They found attenuated responses to the repetition of familiar stimuli but enhanced responses to the repetition of unfamiliar stimuli, which excludes a simple one-to-one correspondence between adaptation and repetition priming. Xu et al. (2007) reports a full dissociation between adaptation and repetition priming in the scene-specific region in the ventral visual cortex, the parahippocampal place area (PPA). Observers viewed pairs of very similar and less similar scene photographs. Two tasks were used to induce opposite behavioral patterns to identical stimuli. In the scene task, observers judged whether two photographs originated from the same scene and, thus, needed to attend to the photos as a whole. As such, behavioral responses were faster and more accurate when the two photographs were very similar than when they were less similar. In the image task, observers judged whether the two photographs were identical pixel by pixel and, thus, needed to focus on feature analysis. This, however, resulted in faster and more accurate behavioral responses when the two photographs were less similar than when they were very similar. Because overall reaction time and accuracy showed no difference between the two tasks, processing load was matched. This was further supported by the lack of difference in peak amplitude and latency of the PPA responses between the two tasks. Although observers might use different strategies and displayed different eye movement patterns, these parameter did not seem to affect neural and behavioral responses. Most interestingly, the significant interaction between stimulus similarity and task observed in the behavioral performance was not reflected in the responses in PPA. Rather, PPA responses were greater for the less similar image pairs than for the very similar image pairs, independent of stimulus processing time. This suggests that attenuation in ventral visual areas is stimulus-specific rather than processing load-specific. Using a whole-brain random-effects analysis, rather than region-of-interest (ROI) analysis to localize the PPA, the authors identified two regions that mirrored the behavioral performance, the anterior cingulate cortex (ACC) and left insula. Both regions are involved in decision making (Wig et al., 2005). As such, both task-independent and task dependent fMRI responses were identified in ventral visual cortex and prefrontal cortex, respectively.

The strength of this study is the manipulation of task difficulty while keeping the stimuli identical, thus providing evidence that fMRI-adaptation is stimulus specific. However, given the limitations of fMRI, single-cell level evidence is necessary to fully demonstrate the stimulus specificity of fMRI-adaptation. For example, mon-keys exhibit repetition suppression and repetition priming, but such correlation disappeared with trial-by-trial analysis (McMahon and Olson 2007). A combination of imaging and single-cell recordings with closely matched paradigms would help resolve whether dissociation between fMRI-adaptation and behavioral performance can be observed in monkey single-unit responses across trials.

Semantic priming While it is well known that repetition priming (the repeated presentation of the same stimulus) is associated with a reduced neural response, called repetition suppression (RS), the neural correlates of semantic priming (when two stimuli are related in meaning but not identical) are not so well established. Raposo et al (2006) compared the neural correlates of repetition and semantic priming using written words, independently manipulating form and meaning. In an fMRI study, subjects saw single words and made a concrete abstract decision. Two consecutive words were identical (town-town) or varied along a continuum of semantic relatedness, from highly related (cord-string) to unrelated (face-sail). The authors found distinct patterns of activation for repetition and semantic priming. Repetition priming was associated with repetition suppression in left inferior frontal gyrus (LIFG), bilateral parahippocampal gyrus and right fusiform gyrus. The authors also observed increased activation for word repetition in the right middle frontal gyrus (RMFG) and ight middle temporal gyrus (RMTG), which may reflect recognition of item's earlier presentation. There was no evidence of suppression for semantic relatedness. Semantic priming was associated with enhanced activation in multiple bilateral fronto-temporal areas, i.e. semantic enhancement. The results suggest that repetition and semantic priming in visual word recognition depend on distinct cognitive processes and neural substrates.

2.4 Review of the current models

All the experiments described in the previous sections have revealed a rich phenomenology on how reaction times, or perceptual boundaries, depend on various factors such as strength and nature of associations between memory items, time intervals between stimulus presentations, and so forth. Interestingly, experimental protocols on humans present striking similarities with pair association task experiments in monkeys. In several animal experiments, electrophysiological recordings of cortical neurons in such behavioral tasks have found two types of task-related activity, ”retrospective” (related to a previously shown stimulus), and ”prospective” (related to a stimulus that the monkey expects to appear, due to learned association between both stimuli) (Sakai and Miyashita 1991). Mathematical models of cortical networks allow theorists to understand the link between the physiology of single neurons and synapses, and network behavior giving rise to retrospective and/or prospective activity. Equally common, are mathematical/connectionist models, more abstract in their level of representation, that have been used to account for the variety of priming effects reported in humans.

In the following, we try to review some of the existing models on contexual perception.

- **Models of semantic priming** In modeling contextual perception, there have been several lines of studies. Some works were focused on semantic priming. According to the association-based view of priming, semantic memory is assumed to be organized as a semantic network where concepts and features are encoded in a localist way by single nodes (Anderson 1983; Collins and Loftus 1975). In localist networks, direct and indirect strengths would lead to priming effects through automatic spreading of activation from node to node. According to the feature-based view of priming, semantically related concepts would share common semantic features (Cree et al. 2003)in distributed networks (McRae 2004; Randall et al. 2004; Becker et al. 1997 Bullinaria 1995; Masson 1995; Plaut 1995; Moss et al. 1994) based on attractor network architectures (Hopfield, 1982). In distributed networks, recall of a given concept in memory corresponds to convergence of the network to an attractor state, that is, a stable distributed pattern of activation or inhibition of units coding for features in a localist way (Akrami et al. 2009). The level of overlap between prime and target features then leads to priming effects through activation/inhibition of features units, without direct associations between concepts but with direct associations between features. Another group of theorists has used more biologically realistic models of cerebral cortex to account for the monkey neurophysiological data (Brunel 2009; Amit 1995). These models account both for retrospective activity in working memory (Amit et al. 2003; Haarmann and Usher 2001; Renart et al. 2001; Wang 2001; Amit and Brunel 1997) and prospective activity in paired associate tasks (Lavigne 2004; Mongillo et al. 2001). However, the wide variety of semantic priming effects in human challenges cortical network models of biophysically realistic neurons.

  There have been also several studies on semantic priming in which the authors investigated how priming effects depend on the type of relationships between prime and target (direct vs. indirect), and on the association strength (estimated in production norms as the percentage of production of targets associated to a given prime word among several subjects; McRae et al. 2005; Cree and McRae 2003; McRae et al. 1997; Battig and Montague 1969; Shapiro and Palermo 1968). Below, we discuss further two of such models, ”Long-term repetition priming” and ”similarity-based priming.

  - **Long-term repetition priming** One of the earliest accounts of long-term repetition priming was based on Morton’s (1969) logogen model of word recognition. Morton proposed that long-term priming is the result of a word-detector unit’s threshold being lowered as a result of previous suprathreshold activation of that word unit. Bavelier and Jordan (1992) proposed a similar account of long-term repetition priming in a multilayered connectionist model. McClelland and Rumelhart (1986) proposed an alternative mechanism for long-term repetition priming in neural networks with recurrent connections. They postulated that exposure to each pattern involves some incremental learning. Later, Becker et al (1997) adapted the model proposed by McClelland and Rumelhart and refer to this postulate as the "incremental learning hypothesis." It predicts that all of the connections in the network that are involved in processing a pattern, not just the thresholds of units, should undergo some incremental learning as a result of priming. In McClelland and Rumelhart's model, each learning step involves a large initial change in each weight, which rapidly decays down to a permanent or very slowly decaying smaller change. Thus, priming is thought to reflect the normal course of learning. When a network is exposed to a previously primed pattern, it should settle to a stable response more quickly because the connections involved in producing the response have been reinforced.

- **Similarity-based priming** The threshold lowering and incremental learning hypotheses both can account for long-term repetition priming, in which responses to the same input pattern are faster or more accurate on repeated presentations. However, when the prime and target stimuli are not identical but related, the threshold account predicts no long-term effect. The incremental learning account makes no specific prediction about how long-term priming would depend on the level of similarity between primes and targets, although McCleUand and Rumelhart (1986) did make the general argument that long-term priming should generalize from primes to similar targets on the basis of the amount of overlap in their representations. For example, in form-based priming (Forster 1987), the prime and the target are similar in that they share perceptual features but are otherwise unrelated, whereas in semantic priming the prime and the target are semantically similar or are semantic or contextual associates. Any theory capable of making specific predictions about these cases must make stronger assumptions about the nature of the representation, the processing of stimuli, or both.

In a recent memory-based model (Brunel and Lavigne 2008), different types of semantic priming (restricted to positive priming) were studied. However, this model lacks enough ingredients possibly underlying the dynamics producing adaptation aftereffect, namely firing rate adaptation.

- **Source confusion account** Huber and O'Reilly (2003) hypothesized that the constant integration of perceptual information over time implies the possibility that we may incorrectly blend together one face and the next as our eyes scan across positions, or as the crowd moves. Yet we do not suffer greatly from such source confusion between successive faces. To explain this and other results that involve the immediate effect of successively presented stimuli, they suggest that the perceptual system includes a discounting mechanism that appropriately reduces the response to a previously presented face, and this reduction serves to offset the effect of source confusion. They proposed that neural habituation is the basic mechanism behind temporal discounting, which automatically parses the stream of perceptual events. This theory initially was developed to explain priming effects with words, but recently was generalized to examine immediate face repetition priming in a threshold identification task as a function of prime duration and face inversion (Huber 2008). In this study, the behavioral priming effect, which is faster or slower reaction time to the target after presenting with a short or long duration prime, respectively, is modeled by means of the conflicting acts of lingering activation from the prime and accumulating depletion of synaptic resources. This study is limited to the role of prime duration, and how its prolongation makes any positive priming disappear. They have not discussed negative priming (adaptation aftereffect in their model).

- **Models of adaptation aftereffect in face perception** Some other studies looked at the interaction between local shunting adaptation and a near-threshold neural baseline (Noest et al. 2007). This neural model explains the observed behavior, without invoking any "high-level" decision making or memory process. However, it has been shown that also high-level, complex processes such as face perception are subject to aftereffects, which cannot be explained based on a combination of adaptation to low-level features (Webster and MacLin 1999).

  Several recent demonstrations using visual adaptation have revealed high-level aftereffects for complex patterns including faces. While traditional aftereffects involve perceptual distortion of simple attributes such as orientation or color that are processed early in the visual cortical hierarchy, face adaptation affects perceived identity and expression, which are thought to be products of higher-order processing. And, unlike most simple aftereffects, those involving faces are robust to changes in scale, position and orientation between the adapting and test stimuli.

- **Models of bi-stable perception** Settling down in one or the other competing percept, can be viewed also in a general framework of ambiguous perception: Certain visual displays are not perceived in a stable way but, from time to time and seemingly spontaneously, their phenomenal appearance wavers and settles in a distinctly different form. This phenomenon is called bistable perception and occurs with a variety of ambiguous visual displays. The most extensively studied instance is binocular rivalry (Wheatstone 1838; Blake and Logothetis 2002), where the phenomenal experience of an observer alternates between two images that are continuously presented to the left and right eye, respectively. In spite of the somewhat "unnatural" method of stimulus delivery, there is good evidence that binocular rivalry shares the typical properties of other instances of bistable perception (Andrews and Purves 1997). Another classical example is the phenomenan in which during prolonged observation of an ambiguous figure, sudden perceptual reversals occur, while the stimulus itself stays unchanged.

Almost two centuries of research on the phenomena of perceptual instability culminated in two explanatory approaches: The bottom-up approach assumes passive, automatic and locally adaptable mechanisms during early visual processing as underlying perceptual reversals (e.g. Toppino and Long 1987). In apparent contradiction, the top-down approach assumes that active, volitional processes near perceptual awareness cause reversals (e.g. Horlitz and O'Leary 1993; Leopold Logothetis 1999; Rock et al. 1994). Both approaches are based on numerous experimental evidence (for a review see Long and Toppino, 2004) and although several authors suggest that both bottom-up factors and top-down factors play an important role for perceptual reversals (Blake and Logothetis 2002; Kornmeier and Bach 2005, Kornmeier and Bach 2006; Long and Toppino 2004;), a theory integrating all empirical findings is still missing to date.

There have been several proposed models focusing on neural mechanisms underlying spontaneous percept switching under steady viewing of an ambiguous stimulus (Noest et al. 2005) most of them mainly relying on the rule of adaptation and/or noise in causing a form of instability in response of the neural assembly which codes for one percept, followed by a transition to the other competing percept. Bistable perception, and binocular rivalry in particular, are often modeled mechanistically using reciprocal inhibition architecture, as shown schematically in Fig. 2.3 (left). There are two neuronal populations whose activities represent the two competing interpretations of the stimulus. The dominant population exerts a strong inhibitory influence on the competing one, so that the latter is suppressed, and only one stimulus is being perceived at a time, a scenario known as mutual exclusivity in most models. The switching in dominance between the two populations is realized by an adaptation mechanism, such as spike frequency adaptation and/or synaptic depression. The adaptation process weakens the inhibition either by decreasing the activity of the dominant population or by decreasing the strength of inhibitory connection between the populations, thus allowing the suppressed population to become active. The resulting activity of such system are deterministic continuous anti-phase oscillations of the firing rates of the two populations, and corresponding switches of two percepts. These general principles have been

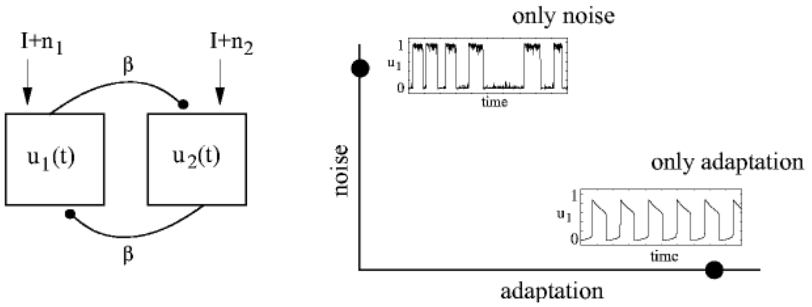


Figure 2.3: **Bistable perception**, Binocular rivalry and bistable perception are often modeled using reciprocal inhibition mechanism, as shown schematically on left. There are two neuronal populations whose activities represent the two competing interpretations of the stimulus (u_1 and u_2). Each population receives a deterministic input of equal strength I_i and independent noise n_i. Lines with circles represent inhibitory connections of the strength β between the two competing populations. The dominant population exerts a strong inhibitory influence on the competing one, so that the latter is suppressed, and only one stimulus is being perceived at a time. (Right) Schematic representation of the extreme versions of the model, either noise only or adaptation only. This two figures are adapted from Shpiro et al. 2009

incorporated in many mathematical models of binocular rivalry (Lehky 1988; Kalarickal and Marshall 2000; Lago-Fernandez and Deco 2002; Laing and Chow 2002; Stollenwerk and Bode 2003; Shpiro et al. 2007). These models can be termed oscillator models and they do not rely on noise.

On the other hand, there is another set of neural competition models, in which, different percepts are represented by multiple stable states (attractors) of the system (Hertz et al. 1991; Haken 1994; Salinas 2003; Riani and simonotto 1994; Kim et al. 2006; Moreno-Bote et al. 2007;freeman 2005). Noise is responsible for the switches between the states, so that in the absence of noise no alternations are possible. These are the noise-driven attractor models (Moreno-Bote et al. 2007).

Recently, Shpiro et al have shown that these two types of models, can be realized within a single theoretical framework (Shpiro et al. 2009). They show that oscillator and attractor behavior can be the two regimes of a single neuronal competition model that includes both noise and adaptation processes (Figure 2.3). In the oscillator regime, adaptation causes the populations to alternate in dominance, while noise is only the source of the irregularity in the switching times. In the attractor regime, noise is the primary cause of the switches. While the adaptation process is still present, it is not strong enough to cause alternations on its own. In the absence of noise, the model in the attractor regime would not switch. However, still in this model, the two percepts are reduced to two homogenous neural populations, reciprocally inhibiting each other. Although the attractor dynamics emerges from this reduced model, its relevance to ”memory” and all phenomena attached to it, is not clear.

Is bistable perception a memory-less phenomena? Whereas bistable perception was long considered a memoryless process (Fox and Hermann 1967; Borsellino et al. 1972), it has become clear that phenomenal appearance can be influenced by past perceptual states. For example, when the presentation of an ambiguous display is interrupted and later resumed, the dominant appearance often remains the same (Maier et al. 2003. This persistence of the dominant appearance stabilizes perception considerably, slowing or even arresting perceptual reversals for intermittently presented displays. The ”memory” in question reflects a longer history of dominance periods, not merely the last dominance period before the stimulus interruption (Brascamp et al. 2008; Pastukhov and Braun 2008).

2.4.1 Concluding remarks

We have reviewed some of the studies, from behavioral and neuronal point of view, related to contextual recall of different sensory inputs. These studies reveal a rich phenomenology on how in terms of reaction times or category boundary, the performance may change. In parallel, there have been several theoretical themes that modeled various pieces of such phenomenology. However, it remained a challenge whether a generic cortical model is able to unify together different aspects of contextual recall.

Chapter 3

Associative Memory Networks

In this chapter, we start with a review of memory representations in the cortex and of the possible neural substrates of the storage of visual memories and of their retrieval later on, conditioned on different sensory input.
Then, in section 2, we discuss the contribution of different modeling approaches to study memory related phenomena in the brain.

3.1 Memory in the cortex

Memory is the record of experience represented in the brain. There are multiple forms of memory supported by distinct brain systems. Specific forms of memory are characterized by whether they last a short or long period(Dobbin et al. 2002), by whether they involve unique experiences or accumulated knowledge(Squire et al. 2007; Eichenbaum et al. 2007), and by whether memory is expressed explicitly by conscious remembering or implicitly through changes in the speed or bias of performance in particular tasks(Schacter 1992). All forms of memory are based on changes in synaptic connections within neural circuits of each memory system. The strength of memory is modulated by emotional arousal and declines in aging.
Memory about episodes in life (episodic memory), about facts of common knowledge (semantic memory) and acquired skills are best examples of long-term memory, the first two of which are usually called declarative memory, whereas the last one can be considered as non-declarative.

Memory about the immediate past, for instance keeping track of the previous words in this sentence, which is necessary in order to understand the meaning of the whole sentence, is an example of working/short-term memory. Long-term memory is commonly hypothesised to be implemented via changes in synaptic efficacies and working memory in many cases is considered to be on-going neuronal activity, that can be related to a previously learned long-term memory pattern.

According to most neuroscientists, the cortex and the structures inside the medial temporal lobe form the neural system of declarative memory. The hippocampus is important during the formation of long-term memories and the neo-cortex serves as thr storehouse where memories are consolidated after being temporarily maintained in the hippocampus.

A major breakthrough in understanding memory systems and their underlying brain mechanisms began with the study of a patient known by his initials H.M. (HM Patient; Scoville and Milner 1957). This case involved an experimental surgical treatment for epilepsy in which the medial temporal lobe was removed. The surgery largely ameliorated the seizures but unexpectedly left H.M. with a severe amnesia which allows him to remember a limited amount of information for a short time (up to a few minutes). Despite his inability to remember new information, H.M. has considerable intact memories from his childhood and information obtained up to a few years before his surgery. From these observations researchers concluded that the parts of the medial temporal lobe that were removed in H.M, including the hippocampus and adjacent parahippocampal region, play a critical role in converting a short-term memory to a long-term, permanent memory store. Furthermore, the fact that H.M. retains memories for events that occurred remotely prior to his surgery indicated that these brain areas are not the site of permanent storage but instead play a role in the organization and permanent storage of memories elsewhere in the brain through the process known as memory consolidation.

The evidence for the involvement of the cortex in long-term memory also primarily came from neuropsychology. Patients with anterior temporal lobectomy were impaired in task, which involved recognition of visual stimuli regardless of the level of the hippocampal lesion (Milner 2003). Animals with ablation of the prehinal cortex showed deficit in delayed matching to sample tasks (Buckley and Gaffan 1998). Neurophysiological evidence for the role of cortex in memory, from monkey IT cortex and prefrontal cortices support this position(Funahashi et al. 1989; Romo et al. 1999; Akrami et al. 2009. In what follows we review some of these evidence from electrophysiology of monkey temporal cortex.

3.1.1 Synaptic plasticity and memory

The cellular basis of memory involves activity dependent plasticity in synaptic connections. An important model in the study of the cellular basis of memory is the phenomenon of long-term potentiation (LTP), a long-lasting increase in the strength of a synaptic response following stimulation (Bliss and Collingridge, 2007). LTP is prominent in the hippocampus, as well as in the cerebral cortex and other brain areas that are involved in different forms of memory. LTP is typically induced by the co-occurrence of excitatory input and intracellular depolarization at the so-called Hebbian synapse, involving N-methyl-d-aspartate (NMDA) receptors that allow the entry of Ca++ into the synapse, which activates cyclic adenosine monophosphate (cAMP). Subsequently, cAMP activates several kinases, some of which increase the number of synaptic receptors. In addition, cAMP activates cAMP-response element binding protein (CREB), which operates within the nucleus to activate a class of genes called immediate early genes, which, in turn, activate other genes that direct protein synthesis. Among the proteins produced is neurotrophin, which activates growth of the synapse. Thus, a series of molecular reactions plays a vital role in fixating the changes in synaptic function that occur in LTP.

Evidence that the permanent fixation of memories depends on this molecular and cellular cascade of events comes from studies showing that memory fixation can be halted by interference with the molecules in this cascade. Many studies have shown that drugs that block NMDA receptors, cAMP, CREB, or other molecules involved in protein synthesis block memory. These treatments are effective when given before or within minutes after learning, and but are not effective if they are delayed, indicating that the molecular cascade leading to protein synthesis is not essential to initial learning or to maintaining short-term memory, but is essential for permanent memory fixation. In addition, studies using genetically modified mice have shown that alterations in specific genes for these molecules can dramatically affect the capacities for LTP and memory fixation.

In addition to LTP, there is also mechanism that diminishes the strength of connections at infrequently used synapses called long term depression (LTD). LTD involves the same molecular substrates as LTP but occurs with different timing rules of activity at synapses. The combination LTP and LTD allow for a sophisticated reorganization of circuits that create neural representations of information. LTP and LTD occur among all brain structures that are known to participate in different kinds of memory. These cellular and molecular events occur on a time scale of seconds and minutes, are essential for the transition from short-term storage to long-term memory, and occur in every brain structure that participates in memory.

3.2 Visual memory and object representation in cortex

Object recognition can be thought of as the process of matching the image of an object to its representation stored in memory. Because different viewing, illumination and context conditions generate different retinal images, understanding the nature of the stored representation and the process by which sensory input is normalized is one of the greatest challenges in research on visual object recognition. It is well known that familiar objects are recognized regardless of viewing angle, scale or position in the visual field. How is such perceptual object constancy accomplished? Does the brain transform the sensory or stored representation to discard the image variability resulting from different viewing conditions, or does generalization occur as a consequence of perceptual learning, that is, of being acquainted with different instances of any given object?

Most theories which postulate that transformations of an image representation precede matching assume either a complete three-dimensional description of an object (Ullman 1989), or a structural description of the image that specifies the relationships among viewpoint-invariant volumetric primitives (Marr 1982; Biederman 1987). In such theories, the locations are specified in a coordinate system defined by the viewed object. In contrast, theories assuming perceptual learning are viewer-centered, postulating that three-dimensional objects are modeled as a set of familiar two-dimensional views, or aspects, and that recognition consists of matching image features against the views held in this set.

An important question for understanding brain function is whether a particular object (or face) is represented in the brain by the firing of one or a few (gnostic or "grandmother") cells (Barlow 1972), or whether instead the firing of a group or ensemble of cells each with different profiles of responsiveness to the stimuli provides the representation. A grandmother cell representation is a code which is very sparse, in that each neuron responds to only one object or stimulus. A very large number of neurons would be required, since each neuron responds to only one stimulus. This encoding is described as local, in that all the information that a particular object is present is carried by one neuron. In contrast, ensemble encoding is described as distributed, in that the information that a particular stimulus was shown is distributed across a population of neurons. Many more stimuli can potentially be represented by a distributed code, as each object is represented by a combination of different neurons firing, and this type of code can have many other advantages, as described below. The actual representation found is distributed. Baylis et al. 1985 showed this with the responses of temporal cortical neurons that typically responded to several members of a set of 5 faces, with each neuron having a different profile of responses to each face (Baylis et al. 1985). In a more recent study using 23 faces and 45 nonface natural images, a distributed representation was found again (Rolls and Tovee, 1995), with the average sparseness being 0.65.

Edmund Rolls has discussed the advantages of the distributed encoding actually found are as follows (Rolls 2000).

Exponentially high coding capacity This property arises from two factors: (1) the encoding is sufficiently close to independent by the different neurons (i.e., factorial), and (2) the encoding is sufficiently distributed. Part of the biological significance of the exponential encoding capacity is that a receiving neuron or neurons can obtain information about which one of a very large number of stimuli is present by receiving the activity of relatively small numbers of inputs (in the order of hundreds) from each of the neuronal populations from which it receives.

Ease with which the code can be read by receiving neurons For a code to be plausible, it is a requirement that neurons should be able to read the code. This is why when we have estimated the information from populations of neurons, we have used in addition to a probability estimating measure (PE, optimal, in the Bayesian sense) also a dot product (DP) measure, which is a way of specifying that all that is required of decoding neurons would be the property of adding up postsynaptic potentials produced through each synapse as a result of the activity of each incoming axon (Abbott et al. 1996; Rolls et al. 1997). It was found that with such a neuronally plausible decoding algorithm (the DP algorithm), the same generic results were obtained, with only a 40% reduction of information compared to the more efficient (PE) algorithm.

Generalization, completion, graceful degradation, and higher resistance to noise Because the decoding of a distributed representation involves assessing the activity of a whole population of neurons, and computing a DP or correlation between the set (or vector) of inputs and the synaptic weights, a distributed representation provides more resistance to variation in individual components than does a local encoding scheme, and this provides for higher resistance to noise (Panzeri et al. 1996) and for graceful (in that it is gradual) degradation of performance when synapses or input axons are lost.

Speed of readout of the information The information available in a distributed representation can be decoded by an analyzer more quickly than can the information from a local representation, given comparable firing rates. Within a fraction of an interspike interval, with a distributed representation, much information can be extracted (Treves 1993; Rolls et al. 1997; Treves et al. 1997). In effect, spikes from many different neurons can contribute to calculating the angle between a neuronal population and a synaptic weight vector within an interspike interval. With local encoding, the speed of information readout depends on the exact model considered, but if the rate of firing needs to be taken into account, this will necessarily take time, because of the time needed for several spikes to accumulate in order to estimate the firing rate.

Invariance in the neuronal representation of stimuli One of the major problems that must be solved by a visual system is the building of a representation of visual information that allows recognition to occur relatively independently of size, contrast, spatial frequency, position on the retina, angle of view, etc. This is required so that if the receiving regions such as the amygdala, orbitofrontal cortex, and hippocampus learn about one view, position, or size of the object, the animal generalizes correctly to other positions, views, and sizes of the object. The majority of face-selective neurons in the inferior temporal cortex have responses that are relatively invariant with respect to the size of the stimulus (Rolls and Baylis 1986).

3.2.1 Role of IT

Indeed, early lesion studies had suggested that IT contributes to the retention of visual information (Mishkin, 1982) and single-cell recordings in IT in subjects performing delayed response tasks showed visually selective delay period activity (Fuster and Jervey 1982; Miyashita and Chang 1988. Despite these initial reports, additional work revealed the selective delay period activity in IT to be susceptible to the presence of intervening sensory information (Miller et al. 1993); concomitantly, it was shown that PF neurons were resistant to such interference (Miller et al. 1996). Nonetheless, the acknowledged involvement of IT in visual working memory extends beyond mere encoding and into the memory epoch as it has been shown recently that a population of IT neurons does regain category selectivity toward the latter stages of the delay interval (Meyers et al. 2008; Akrami et al. 2009). In addition, IT is involved in the matching phase of visual working memory because some IT neurons respond more strongly to a stimulus if it matches the on actively held in mind or more weakly if a stimulus simply repeats itself, effects commonly referred to as match enhancement and match suppression, respectively (Baylis and Rolls 1987; Eskandar et al. 1992; Miller et al. 1993; Miller and Desimone 1994). In fact, match enhancement is hypothesized to reflect the biasing of visual activity in IT by PF feedback (Miller and Desimone 1994; Miller et al. 1996). The precise cortical origins and circuitry underlying match effects, however, remain elusive.

Woloszyn and Sheinberg (2009) examined the temporal evolution of the single-cell responses, individually and jointly as a population, during the various phases of visual working memory. They conclude that IT is an integral component of the working memory system for visual objects. Importantly, by comparing single-cell activity with local field potential (LFP) dynamics, they show that match enhancement is likely a consequence of neural processing occurring within IT, supporting the view that it plays a central role not only in object perception but also in the matching component of visual working memory. Their results support the view that IT activity depends on both the physical visual world and memory demands for objects within that world.

3.2.1.1 Representation of familiar and novel objects in IT

Long-term familiarity facilitates recognition of visual stimuli. IT neurons show modulations of their responses to novel and familiar images. Li et al. (1993) screened for visual activity with familiar visual stimuli and then tested monkeys' abilities to report repeated presentations of novel stimuli when there were variable numbers of intervening familiar image presentations. In general, there was a declining neuronal response to novel images with repetition. Using a serial recognition task, Xiang and Brown (1998) found that 40 percent of the visually responsive neurons in macaque temporal cortex showed a decrease in spike magnitude with stimulus familiarity. They also reported "novelty" cells that responded to first presentations of new stimuli, but not familiar stimuli. Kobatake et al. (1998) measured the response of IT neurons to complex shapes in 5 monkeys with different levels of familiarity with those shapes. Two monkeys trained to discriminate the shapes showed a higher proportion of neurons with robust responses to shapes within the training set compared with the 3 naive animals. More recently, Freedman et al. (2006) studied the response of IT neurons to rotated versions of familiar images and between novel and familiar images. In general, they found that neurons showed greater stimulus selectivity for familiar items. For rotated versions of familiar items, stimulus selectivity was typically greatest at the familiar, learned orientation. In addition, however, the average spiking response was less to familiar stimuli than novel stimuli due to a greater sustained firing of IT neurons to novel stimuli in the phasic period after the initial transient response (approx. 150-600 ms). In the perirhinal portion of IT, Holscher et al. (2003) found that the response of single neurons increased gradually over a few weeks of experience training, and that this effect could combine with short term response reductions observed within a session.

Anderson et al. (2008) measured the local field potential (LFP) and multiunit spiking activity (MUA) from the inferior temporal (IT) lobe of behaving monkeys in response to novel and familiar images. In general, familiar images evoked larger amplitude LFPs whereas MUA responses were greater for novel images. Familiarity effects were attenuated by image rotations in the picture plane of 45 degrees. Decreasing image contrast led to more pronounced decreases in LFP response magnitude for novel, compared with familiar images, and resulted in more selective MUA response profiles for familiar images. The shape of individual LFP traces could be used for stimulus classification, and classification performance was better for the familiar image category. Recording the visual and auditory evoked LFP at multiple depths showed significant alterations in LFP morphology with distance changes of 2 mm. In summary, IT cortex shows local processing differences for familiar and novel images at a time scale and in a manner consistent with the observed behavioral advantage for classifying familiar images and rapidly detecting novel stimuli.

3.2.1.2 Delay period and contribution of IT, prefrontal cortex and prihinal coretx

For more than three decades, it has been known that neurons in prefrontal cortex exhibit persistent activity during the delay periods of working memory tasks (Fuster and Alexander, 1971). The view that this activity is the neural correlate of short-term storage has come under close scrutiny recently (Rushworth et al. 1997; Rowe et al. 2000; Lebedev et al. 2004), with some proposing that the major role of PF in visual working memory maintenance is not storage per se but rather the reactivation of the appropriate visual representations (Ranganath and D'Esposito 2005). In this latter framework, activity in both the PF and IT is necessary to accurately recall visual memories, but the actual memory reinstatement occurs within IT. Woloszyn and Sheinberg (2009) showed that the reliability with which individual IT neurons signal stimulus identity increased toward the latter stages of the delay period, suggesting that the relevant memory was reactivated just before the subject was to be faced with a same/different decision. This rise in stimulus selectivity was also apparent when the activities of multiple neurons were combined, consistent with a previous population analysis showing the same phenomenon but with regard to category instead of stimulus identity information (Meyers et al. 2008). It is appealing to speculate that feedback projections from PF, known from anatomical studies to be abundant (Pandya and Yeterian 1990), were responsible for reinstating stimulus-specific neural activity within IT. This conjecture is further supported by studies showing anticipatory delay period activity within PF itself (Rainer et al. 1999; Rainer and Miller 2002). This would be consistent with the notion that, when bottom-up input and short-term memory meet in IT, bottom-up input dominates, allowing for a more veridical representation of the sensory environment. Another possible source of feedback to IT is the medially adjacent perirhinal cortex (PRh). Although both IT and PRh have the ability to respond robustly to a non-optimal stimulus if it signals the presentation of the preferred stimulus, this retrieval signal appears first in the PRh and only later in IT, suggesting that it flows backward from PRh to IT (Sakai and Miyashita 1991; Naya et al. 1996 and 2001). Most germane to this discussion is the observation that delay period selectivity in both PRh and IT persists through distractors only for the sought after target and not for the stimulus that cued the retrieval of that target (Takeda et al. 2005). Assuming that the rise in delay period selectivity is target related and PRh neurons have the ability to retrieve the same item that served as the cue, then it is possible that PRh neurons participate in reinstating working memory contents in IT Other medial temporal lobe structures, for example, the entorhinal cortex (Suzuki et al., 1997), could also contribute to this process, but their relevance to visual object retrieval specifically is less well understood.

3.2.1.3 Silent delay activity

During the recording from cells in extrastriat cortex, the lack of firing by a neuron does not mean that the cell does not contribute to storing the contents of working memory. For example, recent computational work by Mongillo et al. (2008) has shown that synaptic calcium kinetics in a recurrent network of neurons can be used to store traces of past spiking activity. Additional modeling work by Sugase-Miyamoto et al. (2008) has focused on the ability of IT neurons to store past visual inputs by "remembering" specific patterns of synaptic activity that occur at the time of encoding. Both of these models hypothesize that memory traces are implicitly stored in synapses and only explicitly read out in spiking form at the time of a memory recall signal and/or new visual input. Such quiescent storage models are compatible with the evidence presented in a recent paper by Woloszyn and Sheinberg (2009), wherein individual neurons during the complex occluder condition do not show widespread delay period selectivity but do exhibit match enhancement that occurs essentially simultaneously with the arrival of new visual information. In other words, it could be the case that, if IT spiking delay selectivity is washed out by interfering visual input, an implicit synaptic memory trace persists that manifests itself in the form of match enhancement at the time of recall. In fact, delay period selectivity itself could be reflecting synaptically stored memories, which would suggest a tight link between the neural processes underlying delay period selectivity and match enhancement.

3.3 Simple models of memory storage and memory retrieval

All regional anatomical explorations implicate this postulate: a common functional identity [is determined by] the same type of structure and connections, whatever the mammal examined. Cajal 1922

A century and a decade after Cajal wrote these words, many of the enduring questions of cortical neurobiology that he helped identify remain unanswered. One such question is the degree to which computations in different cortical regions of different species can be encapsulated in a single canonical microcircuit: a kind of basic wiring diagram which, although embellished, remains fundamentally unaltered from mouse to man and across all cortical regions.

3.3.1 A canonical microcircuit for neocortex

The uniformity of the mammalian neocortex (Hubel and Wiesel 1974; Rockel et al. 1980) has given rise to the proposition that there is a fundamental neuronal circuit (Creutzfeldt 1977; Braitenberg and Schuz 1998) repeated many times in each cortical area. The anatomical organization that has emerged from studies (Gilbert and Wiesel l979; Douglas and Martin 1991) of neuronal morphology and immunochemistry is one of stereotyped connections between different cell types: pyramidal cells connect principally to other pyramidal cells, and the smooth cells connect principally to pyramidal cells. Pyramidal cells are excitatory; smooth cells are GABAergic and thought to be inhibitory. Some neurons of both types are driven directly by thalamic input and others indirectly (3.1).

The implication of this anatomical uniformity is that the computations performed by different areas of neocortex may also be uniform, or at the very least, highly similar. Thus although there are undoubtedly some anatomical variations across areas (e.g. Ding et al. 2009 for perihinal cortex, Lund 1988 for primary visual cortex), an argument can be made that functionally, the similarities outweigh these differences (Creutzfeldt 1977, Phillips et al. 1984). Given this computational uniformity, many have hypothesized that differences in the inputs to and the representations in a network may be sufficient to generate the broad functionality associated with the neocortex. In the following we briefly present the properties of *excitatory* and *inhibitory* neurons. Then, we continue with a review of exciting models of memory in cortex.

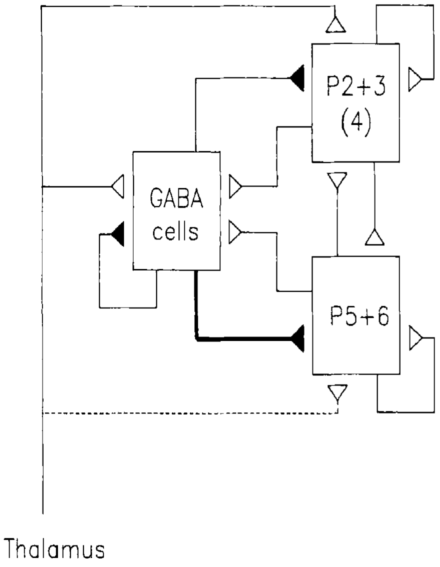


FIGURE 3.1: **The Canonical Cortical Microcircuit**, circa 1989

excitatory connections The middle layer IV, which in primary cortical areas appears as a thick layer of granule cells, is the main termination site for the thalamic afferents to the cortex. In secondary and association cortices, this layer contains pyramical cells which are smaller than those in superficial and deep layers. Layer IV neurons send glutamergic project to layer II and III and receive input also from deep layers. It is interesting that even though the main thalamic afferents to the cortex terminate on dendrites of layer IV neurons, they form no more than 20 % of the synapses in this layer (this number has been estimated to be 5 % in cat primary visual cortex(Ahmed et al. 1994;)).

The superficial layers II/III appears later in development than the deep layers and might have originated as result of an extension of the neurogenic period in mammals.

Inhibitory interneurons and their connections Inhibitory interneurons are GABAergic neurons and comprise 20 % of neurons in the cortex. Inhibitory neurons lack dendritic spines and hence are also called smooth neurons. Their major output is to spiny neurons, but they do not form than 15 % of the targets of the spiny neurons. There are various types of inhibitory neurons in the cortex, among which the *basket cells*are the most salient ones and amount to 20% of the GABAergic cells in the brain(Kisvarday 1992). Basket cells form connections with pyramidal cells in layers both below and above the payers where their cell bodies are located. However, in the middle layers they project mostly within the layer of their cell bodies. Their major targets are spines, dendritic shafts, and to lesser extent somata of pyramidal and spiny stellate cells.

Double banquet cells form another major class of inhibitory interneurons. Their somata are mainly located in layers II/III and their axons run in layers II-V. They receive inhibitory inputs from other interneurons and contact presumably the apical dendrites of pyramidal cells, thus forming a double inhibitory circuit.

Chandelier cells are found in the superficial layers and in layer IV, and form contacts with the initial parts of the axons of pyramidal cells (mainly in layers II/III).

Martinotti cells form another type of inhibitory neurons whose somata are located in deep layers, and whose axons arborize in layer I. GABAergic sparsely spiny non-pyramidal cells are also a source of feedback inhibitory currents to pyramidal cells.
The exact role of the above-mentioned interneurons is not clear and it is not known why there are so many morphologicallly different types of interneurons. Nevertheless, one may speculate about the dunctional significance of these cells, from their peculiarities. For instance, axoaxonal double bouquet cells may emphasize activity within a narrow cortical column.
It must be added that current view of the role of GABAergic interneurones in cortical-network function has shifted from one of merely dampening neuronal activity to that of an active role in information processing. Paulsen and Moser 1998 explored a potential role of hippocampal GABAergic interneurones in providing spatial and temporal conditions for modifications of synaptic weights during hippocampus dependent memory processes (Paulsen and Moser 1998).

3.3.2 Attractor neural networks: the Hopfield model and its variants

The variety of theoretical approaches to the modeling of memory is as wide as the variety of experimental approaches. Modelers explore many different levels, from molecular and single cell models to very high level models (Fuster 1995). In an intermediate family of models, can be built with more or less complex ”formal neurons”, from the binary neuron of the Hopfield model (Hopfield 1982) to spiking neurons like the integrate-and-fire neuron (Amit et al, 1994). These networks were classically developed to model working memory but in turn can be used as building

blocks for models with more complex functions (Dehaene and Changeux 1997).
Attractor neural networks belong to such category in which ”memory states” are formed during learning. A stimulus, when shown to the neural network (assembly), elicits a configuration of activity specific to that stimulus. This configuration of activity is then learned via Hebbian synaptic modifications. These synaptic modifications in turn enable the neural assembly to sustain an active representation of the stimulus (i.e. the ensemble of neural activities specific to that stimulus), in the absence of it: a ”memory state”. It was Hopfield who introduced the general concept of attractor neural network, in which this behaviour is generically observed. More specifically, in his paper of 1982 he defines an associative memory model based on formal neurons which represents the first full mathematical formalization of Hebb ideas and proposals on the neural assembly, the learning rule, the role of the connectivity in the assembly and the neural dynamics.
Most associative memory models are based on Hebbian learning rules. In fact, after the basic proposals of Stent and Hebb, remarkable progresses have been made in both the experimental studies of synaptic plasticity (see talks given at this symposium) and the theoretical analysis of Hebbian type learning rules. In many cases the so called covariance rule (Sejnowski 1977) and other closely related rules (such as the BCM rule) are compatible with experimental data, and on the theoretical side these rules can be shown to be efficient in models of formal neurons (in particular the Hopfield model uses a covariance rule).

3.3.2.1 The basic formalism of the Hopfield model

Donald Hebb proposed that if a neuron often contributes to the firing of another neuron the synapse from the first neuron to the second one becomes stronger. He proposed that this leads to the formation of cells assemblies which underlie memory storage in the brain (Hebb, 1949). Following Hebb, the theoretical analysis of the formation of attractors corresponding to memory patterns through recurrent connections has been put forward from the 1970s (Amari 1971; Anderson et al. 1977, Little and Shaw 1978, and Nakano 1972).
Hopfield (1982) presented an autoassociative memory model based on a network of highly interconnected two-state threshold units (’neurons’). He showed how to store randomly chosen patterns of OS and 1s in the network by using a Hebbian learning algorithm. He was also able to show that for symmetrical connections between the units, the dynamics of this network is governed by an energy function that is equivalent to the energy function of an Ising model spin glass (Kirkpatrick and Sherrington 1978).

Probably, the most influential work in this area was the work of Amit, Gutfreund and Sompolinsky (1985), that used the replica method to solve the Hopfield attractor network. They were able to calculate analytically the number of patterns that can be stored in a fully connected network of spin neurons. The model was defined as follows.

Consider a network of N neurons each of which represented by a spin variable $v_i = +1, -1, i = 1...N$. At any time each neuron receives an input from other neurons that is equal to:

$$h_i(t) = \sum_{j \neq i} J_{ij} v_j(t)$$

which J_{ij} is the weight of the synapse from neurons $j$to i. Furthermore, consider $p = \alpha N$ randomly generated patterns $\{\eta_i^\mu\}, \mu = 1...p$, which each η_i^μ is equal to 1 or -1 with equal probabilities and α is the storage load. Amit et al (1985) showed that if the synaptice weights take the form of

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \eta_i^\mu \eta_j^\mu \tag{3.1}$$

and at each time step one of the units (say unit k) is updated according to

$$Pr(v_k(t+1) = 1) = \frac{e^{\beta h_k(t)}}{e^{-\beta h_k(t)} + e^{\beta h_k(t)}}, \tag{3.2}$$

then there exists a critical storage load $\alpha_c(\beta)$ so that for $\alpha < \alpha_c(\beta)$the network would be able to retrieve the stored patterns after being provided with a partial cue. Here $\beta = \frac{1}{T}$is the inverse temperature. The main idea for solving this model was that the dynamics of the network can be described by the following Hamiltonian:

$$H = -\frac{1}{2} \sum_{i \neq j} J_{ij} v_i V_j \tag{3.3}$$

Using this Hamiltonian, one can calculate the free energy of the system and average it over the quenched noise using the replica trick (Mezard, Parisi, Virasoro 1987). This will lead to the following replica symmetric mean-field equations:

$$\begin{aligned} m &= \langle \eta \int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} tanh(\beta(m\eta + \sqrt{\alpha r} z)) \rangle && (3.4) \\ q &= \langle \int_{-\infty}^{+\infty} tanh^2(\beta(m\eta + \sqrt{\alpha r} z)) && (3.5) \\ r &= \frac{q}{(1 - \beta + \beta q)^2} && (3.6) \end{aligned}$$

Figure 3.2: **Storage capacity α versus temperature T.** In the inset the region of replica symmetry breaking (below T_R) is shown

where m is the order parameter which measures the overlap between the activity of the network and that of the stored pattern and $\langle\rangle$ represents the average over $Pr(\eta)$. Non-zero solution for m mean the existence of a retrieval state (of course one should check the stability of the solution too). The phase diagram of the system is shown in Fig 3.2. The Hopfield neural-network model is attractive for its simplicity and its ability to function as a massively parallel, autoassociative memory. Nevertheless, a number of limitations of the Hopfield model have been pointed out. First of all, the storage capacity (the number of memory patterns that can be stored in the network) is limited to a fraction of the number of processing elements (McEliece et al. 1986). Second, the standard Hopfield model is unsuccessful at storing temporal sequences of memory patterns (Hopfield, 1982). Third, as a model of the brain, it is unrealistic, due to the requirement of symmetrical connections between the units. Finally, it is quite limited in its ability to store sets of correlated patterns. After the work of Amit, Gutfreund and Sompolinsky (1985), many others worked on various modification of the same system.

3.3.2.2 Simple models meet psychology, not neurophysiology

Short term memory models (Nadal et al 1986), obtained as simple variants of the original Hopfield model, allow for a comparison with psychological data. These models reproduce basic properties of human working memory as studied by psychologists - forgetting of old memories which are erased by new ones, but also more elaborate phenomena, such as primacy effects and proactive interference (Nadal et al 1986). It is not clear, however, whether such simple models can account for more complex phenomena as described by Baddeley 1988. Another successful domain is the modeling of the effect of lesions in the cortex. In particular, a phenomenon similar to prosopagnosia is generically observed in neural networks (Virasoro 1988; Kropff and Treves 2007; Ruppin, Reggia 1998): information relative to individual memories is lost before the information which characterizes the class.

But what about insights that "models" could possibly offer us in electrophysiological experiments? From the very definition of the first Hopfield type models, it was obvious that no direct comparison with data at the neurophysiological level could be possible. Quite recently, it has been understood how to build a new generation of models (Amit et al 1994), as a compromise between the need for preserving simple systems, exhibiting the same collective properties as in the Hopfield model, and the need for incorporating more realistic details which would allow for a direct comparison with the phenomenology of recordings during memory tasks. Very encouraging results have been obtained: the self-sustaining selective neural activity exhibited in these models is in nice correspondence with the phenomenology of single unit recordings in monkeys during delayed response tasks, for example in inferior temporal (IT) cortex (Miyashita 1988) or in prefrontal (PF) cortex (Fuster 1995).

Towards a test of the attractor neural network paradigm An attractor network has several preferred activity states, such that relevant external inputs cause network activity to change dynamically and approach one of these preferred states, usually the one most closely correlated with some aspects of the inputs. An interesting development of such models concerns learning of temporal context. The hypothesis is that when two stimuli are often shown one after the other, synaptic modifications will occur in such a way as, when one of the stimuli is shown, neurons selective to the other also tend to be activated. Thus memory states corresponding to two stimuli which often appear one after the other become correlated. Models implementing such type of learning (Amit et al 1994) have been shown to reproduce quantitatively the results of the experiment of Miyashita, in which precisely these correlations were measured in IT cortex(Miyashita 1988). The availability of a detailed learning dynamics (Brunel 1996) enables to predict these correlations as a function of the temporal correlations existing in the sequence of stimuli. The ability of these models to analyse experimental data and to make new predictions has made possible a collaboration between physicists and neurobiologists, in order to set experiments which will test these predictions. As a result, many attractor networks have been proposed to account for numerous cerebral functions with discrete end values, including spatial orientation, sensory pattern recognition, categorical perceptual judgments, and execution of movement trajectories (Lukashin et al. 1996; Wyttenbach et al. 1996; Bartlett and Sejnowski 1998; Gala et al. 2004; Wills et al. 2005; Wong and Wang 2006).

Given the properties of a specific cortical region, Inferotemporal cortex that is believed to be the storehouse of the visual memories, we have hypothesized a possible detectable role, played by attractor dynamics which will be explained fully in Chapter 5.

3.4 Dynamics of large, more realistic associative networks

The focus of many analytical studies of the collective properties of neural networks, has been on studying the steady-state behaviour, i.e. the attractor structure. If any, the typical dynamics of symbolic neural units considered in earlier, statistical flavoured models, and in models form the Connectionist school, and even implemented in simulations is *Glauber dynamics*, or some more or less close relative. In this class of artificially defined dynamics, at each discrete time step, the activity value of each unit is modified according to some update rule. The algorithm is asynchronous in the sense that each unit is updated according to its own clock, as in distributed systems for parallel processing, or it is asynchronous in the sense that only one randomly chosen unit is updated at each step, as in Monte Carlo algorithms. The somewhat tricky point is now to identify to which time scale a discrete time step is correspond to, if sensible conclusions about the collective phenomena time scales are to be drown. The inter-spike interval, or the synaptic conductances and dendritic integration time constants seemed to be viable choices.

Starting with the assumption that Glauber-type, discrete time dynamics are not enough to realistically describe dynamics of neural assemblies, many researcher adapted a model of Integrate-and-Fire spiking neurons (Treves 1993; Abbott and Vreeswijk 1993; Battaglia and Treves 1998; Brunel and Hakim 1998; Brunel 2000. IF formal models are capable of representing with a degree of accuracy what happens at the level of single spikes; they are simple enough to give some hope for an analytical treatment; And last, single IF neurons have been shown in many cases to provide a good approximation to the dynamics of more complex model neurons (e.g.,Bernander et al., 1991).

Here, we briefly review some of these works and in the next session, will discuss the possible ways to define dynamics in a rate-model network, ignoring the very time fast time scales of membrane voltage dynamics.

3.4.1 Dynamics of sparsely connected spiking neurons

The simplest suitable model is the integrate-and-fire neuron with conductance-based synaptic transmission (Eccles, 1964) This latter feature amounts to assuming that an action potential event in the pre-synaptic cell causes a conductance change in the post-synaptic cell. The synaptic conductance successively follows its own dynamics, typically inactivating, or decaying, for example with an exponential time behavior. Treves (1993) analysed this family of models with analytical tools and yielded an analytical formula for the time constants of the exponentially decaying transients modes, through which firing activity in the network approaches the firing at steady state. We bring here a summarized review of Treves 1993.

The time evolution of the membrane V_i of cell i during an interspike interval follows the RC equation

$$C_i \frac{dV_i(t)}{dt} = g_i^0(V_i^0 - V_i^{(}t)) + g_i^K(t)(V_i^K - V_i^K) + \sum_\alpha g^\alpha(t)(V^\alpha - V_i(t)) + I_i(t) \tag{3.1}$$

where C_i is the capacity of the cell membrane, g_i^0 its passive conductance, V_i^o the resting potential, g_i^k an active potassium conductance producing firing rate adaptation, $V_i^K$the corresponding equilibrium potential, and g^α and V^α the conductance and equilibrium potential of each input synapse, α. $I_i(t)$ is an external current injected into the cell, condirered here solely in order to illustrate, the way the model would mimic *in vitro* data on the response characteristics of neorcortical cells.

The process of spike emission is faster than the time scales of interest, and hence it is not modeled in detail; rather, as the potential reaches a threshold value, a spike is added to the spike count, and the potential itself is reset to a hyperpolarization potential, from which the evolution resumes as in Eq. (3.1). 3 In Eq. (3.1) is introduced an intrinsic potassium conductance in order to produce a simplified version of the phenomenon of rate adaptation. Its dynamics is simply described by assuming it to decay exponentially in between spikes and to be incremented by a fixed amount during each spike emission

$$\frac{dg^K(t)}{dt} = -\frac{g^K(t)}{\tau^K} + \Delta g^K \sum_k \delta(t - t_{k,i}) \tag{3.2}$$

Synapses Synaptic conductances, both intrinsic and on afferents, are modeled by a conventional exponentially decaying behaviour. A spike emitted at time t_{k,j_α} by a presynaptic cell j_α activates, after a short interval Δt summarizing axonal and synaptic delays, a conductance on the postsynaptic membrane which then follows the equation

$$\frac{dg^\alpha(t)}{dt} = -\frac{g^\alpha(t)}{\tau^\alpha} + \Delta g^\alpha \sum_k \delta(t - \Delta t - t_{k,j_\alpha}) \tag{3.3}$$

The three above equation, together with the prescription concerning spikes, form a closed system, once the architecture of the network, the incoming afferents and all the various parameters are specified.

It is convenient to express all characteristic potentials in terms of the variable

$$x = \frac{V - V_F^{ahp}}{V_F^{thr} - V_F^{ahp}}$$

measuring the excursion of the membrane potential between spikes, for the cells in class F. The simple assumption concerning the connectivity and the synaptic efficacies are reflected in the fact that the inputs to any cell in the network are determined by globally defined quantities, namely the mean fields. These measure, as a function of time, the effective fraction of synaptic conductances (in units of Δ g) opened on the membrane of any cell of a given class (say, F) by the action of all presynaptic cells of another given class (G):

$$Z_F^G = \frac{1}{N_G} \sum_{\alpha \in G} \frac{g^\alpha(t)}{\Delta g_G^G} \tag{3.4}$$

Mean-field analysis A mean-field description is obtained by summing up the equations describing the dynamics of individual units to get (fewer) equations that describe collective behaviour. Thus grouping Eqs. 3.1 results in N_c functional equations describing the evolution in time of the fraction of cells of a particular class that at a given instant have a given membrane potential, while grouping Eqs. 3.4 results in $N_c in N_c$ equations describing the dynamics of the summed conductance opened on the membrane of a cell of a particular class by all the cells of another given class.

For the moment, we assume a dynamic in absence of firing rate adaptation. The mean-field equations governing the dynamic of the system are of the form

$$\begin{aligned}
\dot{x}_i(t) &= A_F(t) - x_i(t) B_F(t) && (3.5)\\
A_F(t) &= x_F^0 \omega_F^0 + x_F^E \omega_F^E z_F^E(t) + x_F^I \omega_F^I z_F^I(t) + x_S^E \omega_F^S S_F(t) && (3.6)\\
\rho_F(x,t) &= \frac{1}{N} \sum_i \delta(x - x_i(t)) && (3.7)\\
\dot{\rho}_F(x,t) &= B_F(t)\rho_F(x,t) - [A_F 9t) - x B_F(t)] \frac{\partial}{\partial x} \rho_F(x,t) && (3.8)\\
&+ \delta(x) A_F(t) [\rho_F(0^+,t) - \rho_F(0^-,t)] && (3.9)\\
&- \delta(x-1)[A_F(t) - B_F(t) W] \rho_F(1^-,t) && (3.10)\\
\dot{z}_G^F(t) &= -\frac{1}{\tau_G^F} z_G^F(t) + [A_F(t - \Delta t) - B_F(t - \Delta t)] \rho_F(1^-, t - \Delta t) && (3.11)
\end{aligned}$$

where, w_F^0 denotes the inverse of the time constant for passive membrane leakage in cells of class F, W_F^G is in fact $N_G \frac{\Delta_{g_F^G}}{C_F}$ that measures, in units of frequency, the total synaptic strengths of cells of class G (of which there are N_G) onto a cell of class F.

The density function $\rho(x,t)$ should be completed by setting the boundary condition as below:

$$[A_F(t) - B_F(t)]\rho_F(1^-,t) = A_F(t)9\rho_F(0^+,t) - \rho_F(0^+,t)] \tag{3.12}$$

This system of mean-field equations has stationary solutions characterized by a constant firing rate for each cell class. As the neuronal current-to-frequency transfer function under stationary conditions is rather similar to a threshold-linear function, while each synaptic conductance is constant in time, the stationary solutions are essentially the same as can be obtained using much simpler, non-dynamical methods with threshold-linear units. The advantage of the dynamical treatment, however, is that one can describe the approach to the asymptotic solutions, characterized by transient modes that decay with an exponential time dependence $e^{\lambda t}$. By linearizing the dynamical equations near a stable stationary solution and manipulating them algebraically, as explained in detail in, one obtains Eq. (3.13) yielding all the possible values of λ (they must all have a negative real part for the solution to be stable).

Different time scales For any possible stationary asymptotic firing behaviour of the network of formal neurons before, there is an infinite set of transient modes, each associated with a complex time constant λ. The full λ spectrum for a given stationary state is obtained as the set of solutions of the complex equation

$$|Q(\lambda) - e^{\lambda \Delta t} 1| = 0 \tag{3.13}$$

where Eq. 3.13 denotes the determinant of a matrix of rank N_c (the number of classes), composed of the matrix $Q(k)$ and the identity matrix 1. The form of $Q(\lambda)$ is determined by the description chosen to model single units, i.e. neurons and synapses. Once that description, and all the parameters that accompany it, are specified, one has to solve Eq. (3.13) to find the time constants that govern the collective behaviour of the network. This last step may involve a straightforward numerical procedure, if Nc is small, or further analytical manipulation followed by numerical evaluation, if N_c is large or, in particular, tends to infinity. In either case, the set of k values obtained reflects the parameters describing cells and synapses, and thus explicitly links the single unit dynamics (the level at which the model attempts to capture relevant biophysical features) with network dynamics (where contact with neurophysiological experiment can be made).

The dependence of the spectrum on the underlying parameters is in general very complicated, but may be characterized as follows. The spectrum, plotted on the complex plane $Re(\lambda), Im(\lambda)$, where both axes are measured in Hz, presents a gross, a fine and a hyperfine structure (see Fig. 2). The gross structure consists of those λ values which satisfy Eq. (3.13), whose real part is in

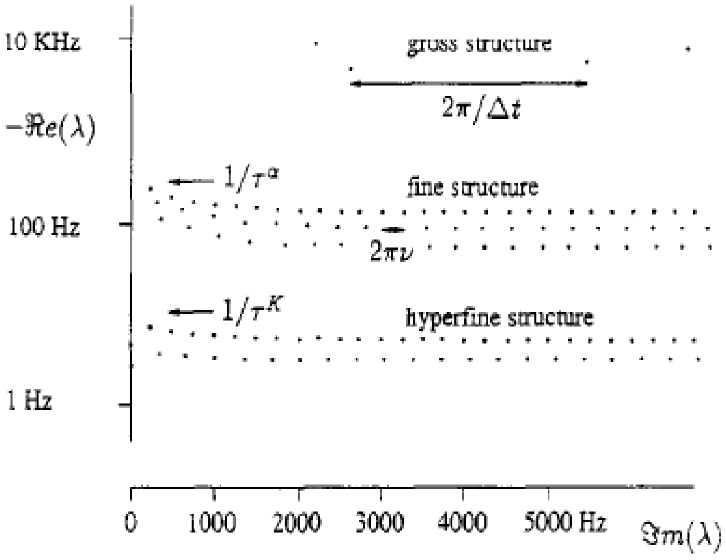


Figure 3.3: **Time constant spectrum**. Schematic representation of the spectrum of time constants characterizing collective dynamics, representing the dependence on the main model parameters, as found by numerically solving Eq. (3.13.

the kHz range and beyond. These values correspond thus to fast time scales, and are determined by fast time parameters (e.g. 1 ms) such as the delay At. They are also very sensitive to global conditions like the detailed balance of excitation and inhibition, and in fact an instability associated with an imbalance of this sort may show up as one or more of the λs of the gross structure acquiring a positive real part. There are two reasons, however, for not focusing one's attention on the gross structure of the spectrum. One is that, inasmuch as fast phenomena characterizing the real neural system have either been omitted altogether from the model (such as the Hodgkin-Huxley details of spike emission) or crudely simplified (such as unifying all axonal and synaptic delays into a unique delay parameter At), the model itself is not likely to be very informative about the fast dynamics of the real system.

Second, in the presence of a stable stationary solution and of transients (those associated with the fine and hyperfine structures) lasting longer than, say $5ms$, the faster transients are not very significant anyway. The fine and hyperfine structures both consist of orderly series of λ values satisfying Eq. (3.13), whose real parts are similar (within a series), while the imaginary parts take discrete values ranging from zero all the way to infinity, with approximate periodicity $2Im(\lambda) = 2\pi v$, where v is the firing rate of a particular group of cells. The (negative) real parts are complicated functions of all the parameters, but in broad terms they are determined by (and similar in value to) the inverse of the conductance time constants. The fine and hyperfine structures differ in the magnitude of their real parts, i.e. in their time constants, in that relatively fast inactivating conductances (e.g., those associated with excitatory transmission via AMPA receptors) produce the fine structure, while slower conductances (e.g. the intrinsic ones underlying firing-rate adaptation (Brown et al. 1983), or GABAB inhibitory ones (Connors et al. 1988)) produce the hyperfine structure. The fine structure, which therefore corresponds to time scales intermediate between those of the gross and hyperfine structures, is perhaps the most interesting, as it covers the range in which the formal model best represents dynamical features of real neuronal networks.

Summary In this analysis the phenomena relying on synchronicity can not survive. Instead, in discussing transients towards stationary asynchronous solutions, therefore in situations in which it is a good approximation to consider those solutions as both relevant and stable, and external inputs as fixed in time after onset, the method already yields a way of relating the transient time scales to the biophysical parameters introduced in the model. Fast time parameters, such as axonal conduction times and synaptic delays, may have an important effect on stability but, within a stable solution, only result in externally rapid transients (the gross structure of the spectrum). Conductance inactivation times have been shown to play the main role; those corresponding to synaptic conductances in determining the relatively fast transients of the fine structure, and that of the intrinsic (adapting) potassium conductance in determining the slow transients of the hyperfine structure.

This finding that the time of relaxation to stationary state can be of the order 5 ms, indicates that recurrent networks of spiking neurons can be very fast and is in fact relevant to the experimental data showing that about 60% of the information about visual stimuli in the IT is transmitted by the firing rate of neurons in the first $20 - 50ms$ after response onset.

3.4.1.1 Realistic time course of synaptic conductance

The type of synaptice conductance dynamics used in Treves 1993 was meant as a representation of the summed dynamics of many individual channels, and was an approximation of the more accurate α-function representation.

Abbott and Vreeswijk (1993) adapted the same mean-field approach to analyse the stability of the asynchronous state in a population of all-to-all, pulse-coupled, nonlinear neurons or oscillators. Moreover, they have considered an α-function type dynamics for the variable E characterising the input coming from one oscillator to the population (here we keep the original notations of). The state of an oscillator x_i is described

$$\frac{dx_i}{dt} = F(x_i) + G(x_i)E(t) \tag{3.14}$$

where F is an arbitrary positive-definite function that characterizes the behavior of the neuron in the absence of coupling; G determines the dependence og the coupling on x_i (where x_i runs between zero and one), and $E(t)$ is a dynamical variable modeling the inputs coming from the other units. If oscillator i reaches $x_i = 1$ at time t_0, $x_i$is reset to zero and $E(t)$ is incremented by an amount

$$E(t) \rightarrow E(t) + \frac{\alpha_1\alpha_2}{(\alpha_2 - \alpha_1)N}(e^{\alpha_1(t_0-t)} - e^{\alpha_2(t_0-t)} \tag{3.15}$$

with α_1 and α_2 arbitrary constants except that $\alpha_2 > \alpha_1$. In the limit $\alpha_2 \rightarrow \alpha_1 = \alpha$ this becomes the familiar α-function respone

$$E(t) \rightarrow E(t) + \frac{\alpha^2}{N}(t - t_0)e^{\alpha(t_0-t)} \tag{3.16}$$

If we assume that the oscillators pulse at a constant, single oscillator rate R (or, equivalently, a rate NR for the total population), then, form Eq. 3.16, we have an expression for E_0, the steady-state, single unit firing rate:

$$E(t) = E_0 = \int_{-\infty}^{t} dt' \frac{\alpha_1\alpha_2 R}{(\alpha_1 - \alpha_2)}(e^{\alpha_1(t-t')} - e^{\alpha_2(t-t')}) = R \tag{3.17}$$

To analyse the asynchronous state, it is convenient to make a change of variables, replacing x_i by y_i:

$$y_i = \int_0^{x_i} \frac{E_0 dx}{F(x) + E_0 G(x)}$$

Then, to describe the state of the full population, they use a density function ρ, and flux J (similar to Treves, 1993) defined by

$$\begin{aligned} \rho(y,t) &= \frac{1}{N}\sum_{i=1}^{N}\delta(y - y_i(t)) && (3.18) \\ J(y,t) &= [E_0 + \Gamma(y)\epsilon(t)]\rho(y,t) && (3.19) \\ & && (3.20) \end{aligned}$$

where

$$
\begin{aligned}
\frac{dy_i}{dt} &= E_0 + \Gamma(y_i)\epsilon(t) &(3.21)\\
\epsilon(t) &= E(t) - E_0 &(3.22)\\
\Gamma(y) &= \frac{E_0 G(x)}{F(x) + E_0 G(x)} &(3.23)
\end{aligned}
$$

In the range $0 < y < 1$, these satisfy the continuity equation $\frac{\partial \rho}{\partial t} = -\frac{\partial J}{\partial t}$, and at the end points, te boundary condition $J(0,t) = J(1,t)$. They have determined the conditions that must be satisfied by the time constants and phase dependence characterising the coupling between the oscillators in order for the asynchronous state to be stable.

The eigenvalues of such oscillators are highly dependent on the form of coupling Γ. For zero coupling, one can find the exact solution of both imaginary part and real part of the eigenvalues. Then, the eigenvalues obtained for zero coupling can be used to construct a weak-coupling perturbation expansion for the case of small $\Gamma(y)$. The zero coupling case is marginally stable with the imaginary eigenvalues $\lambda = 2\pi i n E_0$. Small coupling adds a real part to these eigenvalues determining the stability or instability of the asynchronous state. The real part of the eigenvalues for large n is proportional to $-1/n^2$.

3.4.2 Dynamics in rate-models

The most direct way to model neural systems is to gather data on the currents expressed by individual neurons (Connors et al 1982; McCormick et al. 1985) and to use this information to compute the voltages of each neuron in the network. However, the resulting model could be very difficult to analyze and understand. Instead, most work on modeling neural network uses an approach that describes neural activity solely in terms of the rate of action potential firing. The advantages of using a firing-rate model are simplicity and ease of analysis and simulation. Firing rate models have been developed and derived many times in the literature(Wilson and Cowan 1972, 1973;; Hopfield (1984). In classical neural network models, the input current is generally taken to be linear function of the total excitatory and inhibitory input firing rates. However, most of these rate models were dealing with solely stationary state-state of a network.

In the previous section, we discussed different approaches to describe the dynamics of a local network of spiking neurons. At this point, it begs the question, how to define the dynamics in a rate-based model in which the information is coded in coarse time averaged of spiking activities. It is interesting to see whether the asynchronous state, analyzed in (Treves 1993; Abbott and

Vreeswijk 1993) can be described by a firing rate model. One can use the result to compare the behaviour of transients about the asynchronous state in the full model using the density $\rho(y,t)$ to describe the population and in a simpler model where the population is described solely by the average firing rate $R(t) = J(1,t)$ and the detailed distribution over y is ignored. Under what conditions is this valid?

To construct a firing-rate model, once can first determine the single oscillator firing rate as function of the coupling variable $E(t)$ in the static case when $E(t)$ is a constant, say $R_0(E)$. In the simplest models, the dynamic firing rate $R(t)$ is determined by the equation

$$\frac{dR}{dt} = \alpha_0[R_0(E) - R] \tag{3.24}$$

where α_0 is a constant. E is given by equations similar to those of the spiking model except that $R(t)$ replaces $J(1,t)$

$$\frac{dE}{dt} = -\alpha_1 E + H \tag{3.25}$$

with

$$\frac{dH}{dt} = -\alpha_2 H + \alpha_1\alpha_2 R \tag{3.26}$$

by construction $R_0(E_0) = E_0$ so the firing-rate model has an asynchronous solution identical to that of the full model.

What about the transients about this solution? The full description has an infinite number of transient models while the firing-rate model has only three with eigenvalues given by

$$(\lambda + \alpha_0)(\lambda + \alpha_1)(\lambda + \alpha_2) = \alpha_0\alpha_1\alpha_2 R_0^{'}(E_0) \tag{3.27}$$

where $R_0^{'}$ is the derivative of R_0. The best strategy might be to match these eigenvalues to the longest lasting models of the full model ignoring the more rapidly dying transients. Clearly this will only work for parameter values allowing a stable asynchronous state but even so there is a complication. In fact, as n increases the ral part of the eigenvalue tends to become less negative. Indeed, as was mentioned before, the real part of the eigenvalues for large n, in pulse model, was proportional to $-1/n^2$. Therefore, one cannot simply match the eigenvalues of the firing rate model to the modes of the full model with the least negative real parts.

- ***in summary***

We can argue that the firing-rate model provides a coarse-grained description so that transients of $\rho(y,t)$ with rapid variations in y are irrelevant. In this case one would match the eigenvalues of the firing-rate model to the longest wavelength models of the full model. However, one might still worry about the effect of the higher models since they are not damped quickly. To apply the firing-rate description one must be assured either that these hight n modes will not be excited or that they will have no important effects. It is possible to think of two remedies for such conflict:

1 The situation is much less murky when noise is present. Noise makes the real parts of the high-n modes more negative. If the noise level is high enough, a situation can arise, where the $n = 1$ mode is the longest lasting mode. In this case, the justification of a firing-rate description is more straightforward.

2 Taking insight from Treves 1993 and Treves, Rolls and Simmen 1997, one can think of this higher modes as the *gross* structure of the eigenvalue spectrum. Thus, the model itself is not likely to be very informative about the fast dynamics of the real system -those relating to single action potentials or axonal delay. However, it can be safely used to study the phenomena that are dependent on synaptic conductance time constant.

3.5 Spike frequency adaptation

Spike-frequency adaptation, a gradual reduction of the firing frequency in response to a constant input(Kandel et al. 2000; Wilson 1999), is a prominent feature of several types of neurons that generate action potentials, and it is observed in pyramidal cells in cortical slice preparations (Barkai and Hasselmo, 1994; Connors et al. 1982; Lorenzon and Foehring 1992; Mason and Larkman 1990), or in vivo intracellular recordings (Ahmed B et al., 1998). The biophysical mechanisms of spike-frequency adaptation have been extensively studied in vertebrate and invertebrate systems and often involve calcium-dependent potassium conductances (Meech RW 1978; Sah 1996). However, little is known about its computational role in processing behaviorally relevant natural stimuli, beyond filtering out slow changes in stimulus intensity. Recent studies have sought to attach computational significance to this ubiquitous phenomenon in cortical circuits. Spike-frequency adaptation is thought to be important for functions as diverse as short-term memory (Wilson 1999), contrast adaptation (Sanchez-Vives et al. 2000), neuronal plasticity (Koch 1999), attention (Wilson et al. 2000), and the extraction of high-frequency signals from complicated input stimuli (Benda et al. 2005). It has been suggested that adaptation might be a determining factor in setting the oscillation frequency of cortical circuits (Crook et al. 1998; Fuhrmann et al. 2002; van Vreeswijk and Hansel 2001) for the temporal decorrelation of the inputs (Goldman et al. 2002; Wang et al, 2003), for rate-of-change and anticipation computations in cortical circuits (Puccini, Sanchez-Vivesa, and Compte 2007), and in balancing coding and prediction (Treves 2004). Adaptation may also provide the mechanism which drives alternations between dominant and suppressed states of perception such as those found perceptual bistabiliy (Lehky 1988; Wilson et al. 2000; Shpiro et al 2009).

3.5.1 Prediction through adaptation

In (Treves 2004), the author speculates that a small quantitative change in connectivity, followed by special intrinsic feature of granual cells, namely firing rate adaptatin, would be enough to produce a phase transition to an entirely novel computational faculty. Simulating simple formal models, Treves quantifies the evolution of cortical networks in terms of their computations (Treves 2004a). They all dwell on the interrelationship between qualitative and quantitative change. However, they all include, as a necessary ingredient of the relevant computational mechanism, a simple feature of pyramidal cell biophysics: firing rate adaptation. Adaptation thus effectively separates out in time, two information processing operations that occur in different spaces: the retrieval of memories in the abstract space of attractors, and the accurate relay of stimulus position in the physical space of the cortical surface. This simple mechanism of firing frequency adaptation in pyramidal cells was found to be sufficient for prediction, with the degree of adaptation as the crucial parameter balancing retrieval with prediction (Treves 2004b). The key role of adaptation in switching one percept to another in classical bi-stable perceptual paradigms has been explored by many researchers (Shpiro et al. 2009;

3.5.2 Possible mechanisms underlying spike-rate adaptation

In presence of spike-rate adaptation, a stimulated neuron fires at a high rate for some characteristic period of time-generally tens of milliseconds to several seconds-before the neuron adapts to the stimulus, and the firing rate decreases to a steady value. In many cases the mechanisms underlying adaptation are driven by the spiking activity of the neuron itself, leading to a negative feedback on the firing rate (Benda and Herz 2003; Gollisch and Herz 2004). On a cellular or microscopic level, the spiking activity of a neuron depends on various ionic currents which influence spike generation. Negative feedback is generally associated with ionic currents that hyperpolarize a cell membrane, inhibiting further spike generation and decreasing the firing rate (Gollisch and Herz 2004). Three main types of such adaptation currents are known: M-type currents, which are caused by voltage-dependent, high threshold potassium channels (Brown and Adams 1980); after-hyperpolarizing (AHP)-type currents, mediated by calcium-dependent potassium channels (Madison and Nicoll 1984); and slow recovery from inactivation of the fast sodium channel (Fleidervish et al. 1996). One type of AHP-type current, is sodium activated potassium current found in mammalian neocortical neurons (Lorenzon and Foehring 1992), which can be observed also in vivo (Sanchez-Vives et al. 2000). While different modeling studies have focused on the effects of specific adaptation currents, Benda and Herz derived a universal model for the firing-frequency dynamics of an adapting neuron that is independent of the biophysical processes underlying adaptation (Benda and Herz, 2003).

3.6 Gain modulation

An f-I curve, defined as the mean firing rate in response to a stationary mean current input, is one of the simplest ways to characterize how a neuron transforms a stimulus into a spike train output as a function of the magnitude of a single stimulus parameter. The firing rate gain of neurons, defined as the slope of the relation between input to a neuron and its firing rate, has received considerable attention in the past few years. This has been largely motivated by the may experimental demonstrations of behaviour related gain changes in a variety of neural circuits of the CNS.

To what extend modulation of firing rate gain is important in the moment-to-moment operation of the nervous system in its various functions such as attention, sensory processing and motor control remains to be elucidated. To our knowledge, there is no direct evidence that firing rate gain is casual to a class neural processing operation, rather than simply being a consequence of these operations; i.e. gain modulation does not imply purposeful gain control. Clearly, however, many examples in which gain control may be important have been described. For example, in weak electric fish, there is no mechanism for controlling the sensitivity of the electro-receptor for afferents or their synaptic efect. These afferents synapse directly on E-cells of the first sensory relay nucleus. It has been shown that the gain through this nucleus is modified without a significant change of the baseline spontaneous activity of the output E-cells of this nucleus (Bastian 1986). In the motor system, gain control of reflex pathways appear to be important for proper motor performance (Lisberger 1996; Prochazka 1989). More complex gain modulation have been reported in neocortical neurons. For example, the firing rate of posterior parietal neurons is a function of retinal location.

There have been several proposals put forward to model gain regulation in neural circuits. A long-standing and widely cited candidate mechanism for gain control is shunting inhibition (Eccles 1964; Blomfield 1974) such as that mediated by GABA-gated Cl^- conductance, which decreases input resistance with little direct effect on membrane potential (Vm). Shunting inhibition is thought to decrease the gain between neural input and output, and thus is proposed to act divisively on the gain of neural output (Blomfield 1974) and has been invoked in models of receptive field transformations during sensory processing (Torre and Poggio 1978; Carandini and Heeger 1994).

It is well established that mixtures of postsynaptic excitatory and inhibitory inputs into the soma have little, if any, effect on firing rate gain. Based on a single numerical study of a complicated model neuron, Holt and Koch concluded that dendritic shunting inhibition also does not change the firing rate gain; i.e. the divisive impact of shunting inhibition applies only to subthreshold voltages (Holt and Koch, 1997). In response to these results, several groups have investifated more elaborate single cell mechanism for gain control. For instance, Ulrich (2003) showed that shunting conductances act divisively on subthreshold voltages but subtractively on spike rates, in agreement with theoretical predictions (Blomfield, 1974; Holt and Koch 1997). Most of these mechanisms have in common that they rely on stochastic membrane potential fluctuations, referred to as noise, to effect a change in gain. There is disagreement on the nature of gain control by membrane noise. Chance et el. (2002) proposed that excitatory and inhibitory inputs acting in the background control gain, whilst added excitatory inputs produce firing. This gain control mechanism assumes excitatory and inhibitory inputs be balanced so as not to affect the resting membrane potential. It is not clear how this balance, or other noise parameters, may be controlled by neural circuits. Mitchell and Silver (2003) proposed that gain control may be achieved in a simpler way with tonic background inhibition and noisy excitatory inputs. Yet another mechanism was proposed by Prescott and De Koninck (2003), who suggested that noise itself has

a relatively small influence in firing gain, but coupled to a process termed dendritic saturation, gain modulation is considerably enhanced for inputs on dendrite.
One remarkable feature of the model studied in Prescott and De Koninck (2003) us uts similarity to the original work by Holt and Koch (1997), which indicated that in a very similar model inhibitory inputs, arriving at the dendrite doeas not lead to a c change in gain. Therefore, we have a situation in which two numerical studies suggest opposite results for dendritic shunting inhibition. What is needed to disambiguate the situation is an understanding of the possible mechanism that may lead to a change in gain through dendritic shunting inhibition. Capaday and Vreeswijk (2006) has suggested a simple two compartment integrate-and-fire model in which this issue can be studied analytically, and the mechanism can be understood. The main new finding in this report is that dendritic inhibition leads to a direct change of a neuron's firing rate gain for excitatory inputs that arrive at the dendrite, but not for inputs arriving at the soma. The gain change is due to attenuation of current flowing from dendrite to soma, by shunting dendritic inhibition.

Functional differentiation of $GABA_a$ and $GABA_b$ The main inhibitory neurotransmitter in the mammalian forebrain is α-amino butyric acid (GABA), which acts through A and B type receptors. α-Aminobutyric acid (GABA) type A receptors are mainly permeable to chloride ions (Bormann et al. 1987) and the chloride reversal potential in cortical pyramidal cells is close to the membrane resting potential (Owen et al. 1996). In the presence of a negligible driving force, GABAergic inhibition is mediated by increasing membrane conductance, i.e. by introducing a membrane shunt. Douglas et al. (1989) reported that intracellular recording combined with ionophoresis of y-aminobutyric acid (GABA) agonists and antagonists showed that intracortical inhibition is mediated by $GABA_a$ and $GABA_b$ receptors. The $GABA_a$ component occurs in the early phase of the impulse response. It is reflected in the strong hyperpolarization that follows the excitatory response and lasts about 50 ms. The $GABA_a$ component occurs in the late phase of the response, and is reflected in a sustained hyperpolarization that lasts some 200-300 ms. Both components are seen in all cortical pyramidal neurons. However, the $GABA_a$ component appears more powerful in deep layer pyramids than superficial layer pyramids.

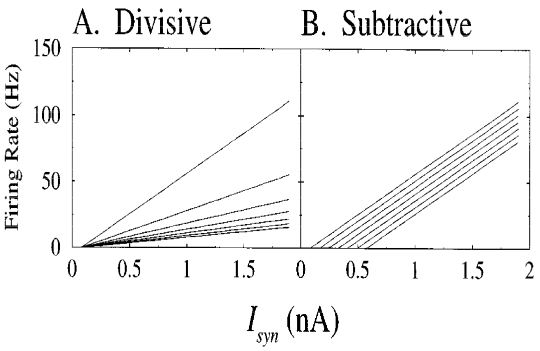


FIGURE 3.4: **A comparison of divisive and subtractive inhibition**. (A) Divisive inhibition changes the slope of the input-output relationship. (B) Subtractive inhibition shifts the curves by subtracting a current.

3.7 Concluding remarks

We have discussed the existing theoretical models that could be applied to study the effects of recent experience on memory retrieval. However, they have not been really applied yet, and before they are, a number of basic issues have to be addressed (in chapter 4).

Chapter 4

Theoretical basis of the model

This chapter has four sections:

- In the first section we define the structure of the model network; its type of unit activity, coupling between units, different dynamical terms i.e. gain, threshold and firing rate adaptation, and the order parameters which will be used later.
- Then, in section 4.2 we provide a simulation setting, which is used in sections 4.3 and 4.4, and also later, in chapter 5 and 6, will be used to model IT data and ”priming” experiments. These two first sections include the theoretical and its corresponding simulation basis for firing rate adaptation, as well. However, we will not explore its function in this chapter and we leave it for chapters 5 and 6, to study the influence of firing rate adaptation in associative networks.
- In section 3, the retrieval properties of various network structures which differ in architecture and input pattern statistics will be compared and discussed. In this section, the storage capacity is studied in models with no firing-rate adaptation.
- Finally, in the last section we study the dynamics of an attractor network which receives ambiguous inputs that are partially correlated with two stored patterns, i.e. morphed inputs. We develop an analytical framework to study how an associated network retrieves a morphed pattern. The analytical solution is given for a zero-adaptation network, which is followed by numerical simulations including non-zero firing rate adaptation values, as well.

4.1 Structure of the model

The major motivation for our model comes from neural structures such as the CA3 region in the mammalian hippocampus or high visual cortical areas. Two aspects of such systems are: (1) the sparseness of connectivity which leads to asymmetry in excitatory connections (Braitenberg and Schuz 1991; Treves and Rolls 1992); (2) the relative rarity of inhibitory interneurons, which implies that large numbers of excitatory cells are inhibited by the same interneuron. However, in Section 3 of these chapter, we will discuss possible relevance of connectivity dilution on retrieval quality of different associative networks. The second aspect allows us to adapt a global inhibition mechanism that applies equally to all units.

4.1.1 General parameters and analytical approach

The input to each unit. We have assumed a simple type of neuron model, the threshold-linear description, in which the geometry of the neuron is reduced to a point that maps the presynaptic input to an amplified level in the output stage, once it crosses the threshold to fire. We considered a network of N units, in which the level of activity of unit i is represented by a variable $r_i \geq 0$. This variable can be taken to represent the firing rate of the neuron averaged over a short time window. This assumption enables the units to assume real continuously variable firing rates, similar to what is found in the brain (Rolls and Treves 1998). In our model, we assume that the input current to each neuron i, is composed of the synaptic currents that enter the cell through the excitatory recurrent synapses, the external input and the adaptation current (due to potassium conductances). The local field (or input current) takes the following form:

$$H_i = \sum_{j \neq i} w_{ij} r_j + I_{in}^i - I a_i \tag{4.1}$$

where the first term is the sum of recurrent input from the other N neurons in the network, which is assumed to be proportional to the firing rate of the presynaptic neuron, weighted by a synaptic efficacy J_{ij}. The second term is the afferent input. In general, the vector I_{in} can be designed to be correlated with either one or several stored patterns, and therefore can signal a "morphed" input (between a set of patterns). In section 3, in which we study the issue of storage capacity, I_{in}, represents a full or partial cue that is correlated only with one of the stored patterns, but in section 4, this term represents a morphed pattern. The third term, finally, is the adaptation inducing current. Later, in section 4.1.2, we explain how this term can be realized through the time-varying potassium conductances.

Stored memory patterns We assume that the network has "learned", i.e. stored, p different patterns of activity. Each pattern of activity is represented by a vector $\xi^\mu = (\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu), \mu = 1, \ldots, p$, in which ξ_i^μ represents the level of activity of neuron i in pattern μ. This implies that synaptic weights w_{ij} have undergone plastic modifications such that the dynamical attractors of the network include these predefined patterns. In a limit case, such patterns are the *only* attractors of the network, but this is not always true, and there may well be other attractors which do not correspond to any single memory pattern.

Synaptic weights. We consider a diluted asymmetric version in which each unit is connected, on average, with C other units in the network. One functional form for the synaptic weights which has been widely used in the literature is a sparsely coded version of the "covariance hebbian learning rule" which assumes a linear summation of contributions from the storage of each pattern

$$w_{ij} = \frac{1}{Ca(1-a)} \sum_{\nu=1}^{p} c_{ij} \xi_i^\nu (\xi_j^\nu - a) \tag{4.2}$$

where ξ_i^ν represents the activity of unit i in memory pattern ν and $c_{ij} \in 0, 1$ is an independent identically distributed random variables (IIDRV) with probability $Pr(c_{ij} = 1) = \frac{C}{N} < 1$. Thus, c_{ij} is equal to 1 if there is a connection running from neuron j to neuron i, and 0 otherwise. Each ξ_i^ν is taken to be a "quenched variable", i.e. a given parameter, drawn independently from a distribution $p(\xi)$, with the constraints $\xi \geq 0$, $\langle \xi \rangle = \langle \xi^2 \rangle = a$, where $\langle \, \rangle$ stands for the average over the distribution $p(\xi)$, and **a** is the level of sparseness.

Temporal evolution of the firing rates. As we discussed in the previous chapter, we can reduce a spiking model to a rate model by keeping only the most relevant time constant which is synaptic conductance one. In our model, the activity of each neuron (firing rate) at time t is assumed to be related to the input current it receives at time t $(h(t))$through the threshold input-output function

$$r(h(t)) = g(h(t) - Tr)\Theta(h(t) - Tr) \tag{4.3}$$

where Tr is the threshold current, g is the gain and Θ is the heaviside function, vanishing for negative values of its argument.

Equation 4.3 is the first step in the construction of a firing-rate model. The second step is to determine the temporal evolution of local field h_i (or dendritic current). We assume that transients that correspond to fluctuations away from the static steady state value of the current, decay exponentially, with time constant τ_h, which relates to synaptic conductance; so that

$$\tau_h \frac{dh(t)}{dt} = -h(t) + H(r, I_{in}, Ia) \tag{4.4}$$

where H is the instantaneous current entering the unit. Equation 4.4 and 4.3 give a complete description of the firing-rate once the static local (dendritic current) $H(r, I, Ia)$ is known. r is the excitatory synaptic input, I_{in} is the external input and Ia is the adaptive current. We can simply write down H as the superposition of different synaptic contributions, as in Eq. 4.1.

$$\begin{aligned} H_i &= \sum w_{ij} r_j + I_{in}^i - Ia_i \\ &= \sum g w_{ij} [h_j - Tr]^+ + I_{in}^i - Ia_i \end{aligned} \tag{4.5}$$

Thus, the final equation for the current will be

$$\tau_h \frac{dh_i(t)}{dt} = -h_i(t) + \sum g w_{ij} [h_j(t) - Tr]^+ + I_{in}^i(t) - Ia_i(t) \tag{4.6}$$

Overlap The relevant order parameters measuring the quality of retrieval are the overlap of the microscopic state of the network and the ν_{th} pattern, "local overlap" with pattern ν, i.e. m^ν, defined as

$$m_i^\nu(t) = \frac{1}{Ca(1-a)} \sum_{j=1}^{N} c_{ij} (\xi_j^\nu - a) r_j(t) \tag{4.7}$$

it is straightforward to show that by substituting (4.2) in (4.5) and using the definition of overlap (4.7), we arrive to:

$$H_i(t) = \sum_{\nu=1}^{p} \xi_i^\nu m_i^\nu(t) + I_{in}^i(t) - Ia_i(t) \tag{4.8}$$

In general, a pattern ν is said to be retrieved if the retrieval overlap for that pattern pattern is macroscopic, i.e. of order $O(1)$ in the thermodynamic limit $C, N \rightarrow \inf$. The rest of the patterns causes a residual noise at each time step of the dynamics. Thus, during the retrieval of one pattern, the local field to each unit can be decomposed into two terms. One is the signal, which is in the direction of keeping the network in a state with large overlap with the retrieved pattern. The second term, which can be called noise, contributes random interference due to the other stored patterns in the network. Without loss of generality, we suppose that I_{in} is signaling the morphed pattern between **a** and **b** (two patterns out of the total P patterns) and the retrieved pattern is either **a** or **b**; meaning that the value of the overlap between the current state of the network and the memory pattern for for either **a** and **b** is the highest, with respect to the other patterns. Thus, we decompose the local field to each unit (due to the recurrent connections) to three components:

$$H_i(t) = \xi_i^a m_i^a(t) + \xi_i^b m_i^b(t) + \sum_{\nu \neq a,b} \xi_j^\nu m_i^\nu(t) \tag{4.9}$$

In equation (4.9) the signal is nothing but the first two terms in the sum on the rhs, whereas the noise is the rest. At the end, replacing H_i in the previous dynamical equation for the current (4.4)), with one in 4.9 gives us

$$\tau_h \frac{dh_i(t)}{dt} = -h_i(t) + \xi_i^a m_i^a(t) + \xi_i^b m_i^b(t) + \sum_{\nu \neq a,b} \xi_i^\nu m_i^\nu(t) + I_{in}^i(t) - I a_i(t) \tag{4.10}$$

The noise term (crosstalk effect) We define $R_i(t)$ as the noise term at each time. Should be clarified here that the term ”noise” refers to the cross-talk between patterns and not stochastic dynamical effects (like spike or synaptic noise). This term can be called ”quenched noise”, as well, and is defined as:

$$R_i(t) = \sum_{\nu \neq a,b} \xi_i^\nu m_i^\nu(t) \tag{4.11}$$

Assuming ξ_i^ν and ξ_j^ν as two independent variables, it could be concluded, based on central limit theorem, that for big p, the value of $R_i(t)$ tends to have a Gaussian distribution. Then we need to calculate the mean and the variance of this Gaussian distribution. It is easy to show that

$$\begin{aligned}
\langle R_i \rangle_\xi &= \frac{1}{Ca(1-a)} \left\langle \sum_{j=1}^{N} c_{ij} r_j \sum_{\nu \neq a,b} (\xi_i^\nu(\xi_j^\nu - a)) \right\rangle \\
&= \frac{1}{Ca(1-a)} \sum_{j=1}^{N} c_{ij} r_j (p-2)(a^2 - a^2) \\
&= 0
\end{aligned}$$

And for variance, we have:

$$\langle R_i^2\rangle_\xi \;=\; \frac{1}{C^2a^2(1-a)^2}\left\langle \sum_{j,k}^{N} c_{ij}c_{ik}r_jr_k \sum_{\nu,\mu\neq a,b}^{p} (\xi_i^\nu\xi_i^\mu(\xi_j^\nu - a)(\xi_j^\mu - a))\right\rangle$$

At the end, the variance will be $\langle R^2\rangle = \frac{A_\eta}{(1-a)}q$; where

$$\begin{aligned} q(t) &= \sum_{j}^{N} r_j^2 \\ A_\eta &= \frac{1}{C^2}(p-2) \end{aligned}$$

The details of the calculation is fully reported in the Appendix.

Going back to the definition of local field, the noise term is presented as a random variable η, with normal distribution of mean zero and standard deviation $\sqrt{\frac{A_\eta}{(1-a)}q}$. Having such formalism for the noise term, the general equation for the current $h_i(t)$ will be

$$\tau_h\frac{dh_i(t)}{dt} = -h_i(t) + \xi_i^a m^a(t) + \xi_i^b m^b(t) + \sqrt{\frac{A_\eta}{(1-a)}q}\eta + I_{in}^i(t) - Ia_i(t) \qquad (4.12)$$

Firing rate probability, as a function of the signal. As it was described in the previous section, the local field of each unit, h_i, at each time, can break down in two parts i.e. $\bar{h}_i$, the signal (averaged over noise across all units) and η_i, the noise. Thus, each r_i, the output of threshold function ($r_i = [h_i - Tr]^+$), is dependent on a certain level of noise, $\sqrt{\frac{\alpha}{(1-a)}q_i}\eta$, which is itself a Gaussian variable, and thus gives us the ability to model r_i as a Gaussian distribution with mean $\bar{r}_i$ and variance of σ_i, when $h_i > Tr$. If $h_i < Tr \Rightarrow r_i = 0$. At this point, we should calculate the probability of finding $h_i < Tr$. For the moment, we consider a case without adaptation i.e. $Ia_i = 0$.

$$\begin{aligned} P\left(hi < Tr\right) &= P\left(\bar{h}_i + \eta_i < Tr\right) \\ &= P\left(\eta_i < Tr - H_i\right) \\ &= \int_{-\infty}^{Tr-H_i} \frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(\eta_i-\bar{\eta}_i)}{2{\sigma_i}^2}}d\eta \\ &= \phi(\frac{Tr-H_i-\bar{\eta}_i}{\sigma_i}) \end{aligned}$$

where

$$\Phi(x) \equiv \int\limits_{-\infty}^{x} \frac{dz}{\sqrt{2\pi}} e^{\frac{z^2}{2}}$$

and finally, we have such description for the probability of firing rate r_i:

$$P(r_i) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma_i} r^{-\frac{(r_i - \bar{r}_i)}{2\sigma_i{}^2}} \Theta(r_i) + \delta(r_i)\Phi(-\frac{Tr - H_i - \bar{\eta}_i}{\sigma_i})$$

here, σ_i is the amplitude of the noise which is $\sqrt{\frac{\alpha}{(1-a)}q}$, and $\bar{r}_i$ is the mean firing rate of unit i that crosses the threshold:

$$\bar{r}_i \quad = \quad \langle [H_i + \eta_i - Tr] \rangle_\eta = (H_i - Tr)$$

However, $\bar{r}_i$ is not equal to $\langle r_i \rangle$. This is due to the zero values of r_i for negative fields which is reflected in the second term in rhs of the $P(r_i)$. Now that we have the probability distribution of r_i, the mean firing rate $\langle r_i \rangle$ and the second moment $\langle {r_i}^2 \rangle$ can be found:

$$\langle r_i \rangle \quad = \quad \sigma_i \rho \Phi(\rho) + \sigma_i \Psi(\rho) \tag{4.13}$$

$$\langle {r_i}^2 \rangle \quad = \quad {\sigma_i}^2 (1 + \rho^2) \Phi(\rho) + {\sigma_i}^2 \rho \Psi(\rho) \tag{4.14}$$

where

$$\rho \quad = \quad \frac{\bar{r}}{\sigma_i} \tag{4.15}$$

$$\Phi(\rho) \quad = \quad \int\limits_{-\infty}^{\rho} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz \tag{4.16}$$

$$\Psi(\rho) \quad = \quad \frac{1}{\sqrt{2\pi}} e^{\frac{-\rho^2}{2}} \tag{4.17}$$

It should be noted that as $\rho \to \infty$, $\Phi(\rho) \to 1$ and $\Psi(\rho) \to 0$, and thus, $\langle r_i \rangle \to \sigma\rho = \bar{r}$ and $\langle {r_i}^2 \rangle \to \sigma^2 \rho^2 = \bar{r}^2$

4.1.2 Modeling the adaptation current

In the previous chapter, we have discussed the nature and implication of firing rate adaptation in different brain structures. Here, we develop a dynamic equation for adaptive current contributing to the local field of each unit, which will be also recruited later in the simulation.

Spiking neuron A spiking neuron in a standard Integrate and Fire model can be endowed with an extra (adaptive) input current, I_a, which can be assumed sodium dependent potassium current that follows an α-function dynamics. This adaptive current is proportional to the instantaneous calcium concentration through a (potassium) conductance g. This type of conductance dynamics is meant as a representation of the summed dynamics of many individual channels, each following a simple dynamics as below:

$$\frac{dg_j(t)}{dt} = -\frac{g_j(t)}{\tau_j} + \Delta g_j \sum_k \delta(t - t_{k,j}) \tag{4.18}$$

g_j is potassium conductance that undergoes an upward jump Δg_j upon spike emission, after which it decays with a characteristic time τ_j, a conductance time constant. $t_{k,j}$ is the time of emission of the kth spike by the same unit j. At $t_{k,j}$ the value of conductance increases with the value of Δg_j and during the inter-spike-interval, it decreases exponentially with time constance τ_j. To produce an adaptive conductance with α-shape, it is enough considering two input synapses with such dynamics.

$$\begin{aligned}
\frac{dg_1(t)}{dt} &= -\frac{g_1(t)}{\tau_1} + \Delta g_1 \sum_k \delta(t - t_k) \\
\frac{dg_2(t)}{dt} &= -\frac{g_2(t)}{\tau_2} + \Delta g_2 \sum_k \delta(t - t_k)
\end{aligned}$$

where $\tau_1 < \tau_2$. Without loss of generality we assume that Δg_1 and Δg_2 are equal to one.
The final adaptation current contributing to the input current of each neuron, is the difference of g_2 and g_1. We define $I_a = \alpha_a \frac{\tau_2 - \tau_2}{\tau_2 \tau_2} (g_2 - g_1)$. Notice that I_a is not anymore conductance, but has the dimension of current. α_a determines the strength of adaptation and time constants τ_1 and τ_2 are giving the speed of adaptation. One can derive the dynamic time course of I_a:

$$\frac{d(g_2 - g_1)}{dt} = -\frac{1}{\tau_2} g_2 + \frac{1}{\tau_1} g_1$$

and using the definition of I_a, one can get

$$\frac{dI_a}{dt} = -\frac{1}{\tau_2} I_a + \alpha_a \frac{(\tau_2 - \tau_1)^2}{\tau_1^2 \tau_2^2} g_1$$

if we calculate the second derivation of equation above, we find

$$\begin{aligned}
\frac{d^2I_a}{dt^2} &= -\frac{1}{\tau_2}\frac{dI_a}{dt} + \alpha_a\frac{(\tau_2-\tau_1)^2}{\tau_1^2\tau_2^2}\frac{dg_1}{dt} \\
&= -\frac{1}{\tau_2}\frac{dI_a}{dt} + \alpha_a\frac{(\tau_2-\tau_1)^2}{\tau_1^2\tau_2^2}(-\frac{1}{\tau_1}g_1 + \sum_K \delta(t-t_k)) \\
&= -\frac{1}{\tau_2}\frac{dI_a}{dt} - \frac{1}{\tau_1}\frac{dI_a}{dt} - \frac{1}{\tau_1\tau_2}I_a + \alpha_a\frac{(\tau_2-\tau_1)^2}{\tau_1^2\tau_2^2}\sum_K \delta(t-t_k)
\end{aligned}$$

if $\tau_1 \rightarrow \tau_2 \Rightarrow$

$$\frac{d^2I_a}{dt^2} = -\frac{2}{\tau}\frac{dI_a}{dt} - \frac{1}{\tau^2}I_a + \alpha_a\frac{(\tau_2-\tau_1)^2}{\tau_1^2\tau_2^2}\sum_K \delta(t-t_k) \tag{4.19}$$

Rate model Now we should go one step further and derive a phenomenological model for the firing frequency of an adapting neuron, whose parameters are independent of the specific adaptation process. To achieve this goal, one can replace the adaptation gating variable g as well as the adaptation current Ia by suitable time averages. All the dependencies on the membrane potential can then be replaced by functions depending on the firing frequency r. In this way, it is possible to absorb the adaptive currents in the $r-I$ curve of firing rate neuron type (Benda and Herz 2003).

$$r = r_0(I - A(1+\beta(r)) \tag{4.20}$$

$$\tau_a(1+\epsilon(r))\frac{A}{dt} = A(\infty) - A \tag{4.21}$$

$$\tag{4.22}$$

The adaptation state **a** generalizes the averaged adaptation gating variable g and relaxes exponentially, with the adaptation time constant τ_a, towards a possibly voltage dependent steady-state adaptation strength A_∞. The input adaptive current, $A[1+\beta(r)]$ depends linearly on A and may be influenced by f through $\beta(r)$. The term $\epsilon(r)$ covers a potential dependence of adaptation time constant τ_a on firing rate. We assume $\epsilon(r) = 0$ for the rest of this chapter, unless otherwise stated.

4.1.3 Mean activity, sparseness, and setting *threshold* and *gain*

In every neural network, the necessity of an activity control system, which tries to keep the activity of the network in the retrieval process the same as the one for the memorized patterns, has been emphasized vastly (Amit et al. 1987; Amari 1989). Sparseness and mean activity are usually two indices by which the level of network activity is measured. In the following, we will briefly explain each of them and at the end suggest a possible network implementation to regulate these two factors.

Mean activity In classical models of attractor neural network, there usually is a uniform input to each unit which is a function of the mean activity of the network and is unrelated to the memory patterns. This term, with an appropriate functional form, can regulate the activity of the network, so that, for example, at any moment in time the mean activity $\frac{1}{N}\sum_i r_i$ may be maintained constant, avoiding the run-away of excitatory reverberation. This term can be thought of as modeling the effect of inhibitory interneurons. While it can take an explicit functional form in the simulations, its exact form is not important for our purpose, in that it is irrelevant to determine fixed-point attractors. This is because its effect is constant across units, when the fixed-point equations are satisfied and $\frac{1}{N}\sum_i r_i = const.$, and it can therefore be absorbed in the threshold. In other words, one may assume that by changing the threshold the mean activity of the network is regulated, and that the threshold is effectively modulated by inhibitory neurons.

Sparseness The sparseness **a** of the representation can be measured, by extending the binary notion of the proportion of neurons that are firing, as

$$a = \frac{\left(\sum_i^N \frac{r_i}{N}\right)^2}{\sum_i^N \frac{{r_i}^2}{N}}$$

The memory pattern with small coding level **a** is called a sparse pattern. Sparsely coded models have attracted a lot of attention in the development of neural networks, not only because the sparse coding is believed to be biologically plausible (Miyashita 1988; Baddeley et al. 1997), but also it is well-known that they have a large storage capacity, which behaves as $1/(alna)$ for **a** small. However, it is clear that the basins of attraction, e.g., should not become too small because then sparse coding is, in fact, useless. An activity control mechanism is requisite for stable retrieval in the sparse coding scheme (Okada, 1996).

In our model, we keep the mean activity and sparseness constant, during learning and retrieval phases. More explicitly, we want to have $\langle r_i \rangle = \langle r_i^2 \rangle = a$. One way to find a proper temporal dynamics for g and Tr is to define a minimizing function L to find the optimum values for g and Tr to satisfy the conditions above.

$$L = (\langle r_i \rangle - a)^2 + K(\langle {r_i}^2 \rangle - a)^2$$

Another approach to set g and Tr is to model these two parameters according to divisive and subtractive gain modulation, respectively, which in a neural system are usually implemented through the mechanisms of fast and slow GABA modulatory system.

Here, we introduce another equation for adjusting "gain", modeling the action of neuromodulator $GABAa$ which at this time, fixes the mean activity of whole network. As we have reviewed in Chapter 2, we know that the $GABA_A$ and $GABA_B$ neurotransmitters, behave differentially in regulating excitatory activity. $GABA_A$ component occurs in the early phase of the impulse response. It is reflected in the strong hyperpolarization that follows the excitatory response and lasts about 50 ms. In this sense, $GABA_A$ plays the role of divisive gain modulation. In contrary, the $GABA_B$ component occurs in the late phase of the response, and is reflected in a sustained hyperpolarization that lasts some 200-300 ms, and affects the activity in a subtractive way(Douglas et al. 1989). We adapt these two mechanism in to the dynamics of I_g and Tr

$$\begin{aligned} \tau_g \frac{dI_g(t)}{dt} &= -I_g(t) + \frac{<r>}{a} \\ \tau_t \frac{dTr(t)}{dt} &= \frac{(\langle r \rangle)^2}{{I_g}^2} - a\frac{\langle r^2 \rangle}{{I_g}^2} \end{aligned}$$

where Ig is $\frac{1}{g}$, inverse of the gain.

Under certain condition, we can assume that the mechanism for setting threshold is much slower with respect to the time scale of gain modulation and change in currents. So, we assume a fixed value for the threshold throughout the whole process.

4.2 Simulation setting

”There is an underlying assumption that the observations express activity of large groups of similarly acting neurons that is the result of a bottom-up scenario in which individual cells, via their synaptic interactions, lead to the large scale phenomena. The connection between the levels must be provided by theory, which must also provide the relevant variables for observation. It is suggested that between the experiment and the full theory there is a creative, mixed role for simulation: both experimental and theoretical. A simulation presents complex dynamics and hence is an empirical board for testing theoretical tools, yet its controlled behaviour can make predictions about the biological system” (Amit 1998).
In this section, we describe the general characteristics of the simulated network.

4.2.1 General characteristics of the simulated network

As we aim to model specific data from monkey IT cortex, and also some visual psychophysics (”priming” and ”adaptation aftereffect”) which are believed to rely mainly on extrastriat structures, we set the basis of the simulated network, in line with previous work of modeling IT (Inferior Temporal) networks (see e.g. Parga and Rolls 1998; Roudi and Treves 2008). We have considered a simple autoassociative network model comprised of two layers, shown schematically in Fig. 4.1, which simulates a cortical patch as a local recurrent network. The first layer functions as an input stage that projects afferent inputs to the second layer; this layer is analogous to the input from earlier visual areas to the second, recurrent layer. Units in the second layer receive inputs both from the first layer, as well as from units in the same layer, and provide outputs to one another (recurrent connections). The second layer is analogous to the cortical patch containing the neurons recorded in this study (neurons recorded from inferotemporal cortex, IT). In our simulation, we consider the dynamics of local interconnected networks in IT, and thus the simulation is focused on the second, output, layer of units with recurrent connections. The units in the network are labeled with an index i, $i = 1...N$, but the connectivity between the units, or the probability that two units are connected, does not depend on their indexes. In our model, units receive feedforward projections from an input layer of another $N = 1000$ units. Each unit in the (output) patch receives C_{ff} feed-forward connections from the input array, and C_{rc} recurrent collateral connections from other units in the patch. Both sets of connections are assigned to each receiving unit at random. Weights are originally set at a uniform constant value, to which is added a random component of similar mean square amplitude, to generate an approximately

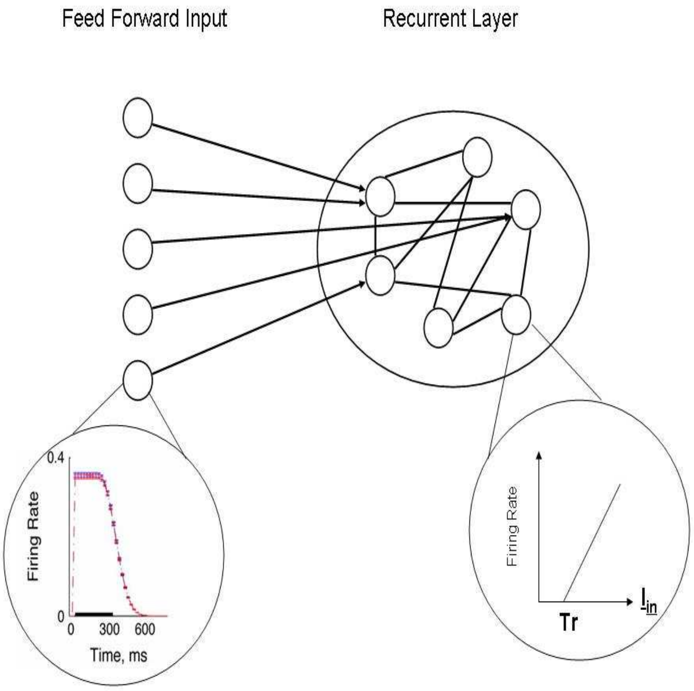


FIGURE 4.1: **Schematic view of the simulated network**. The network includes an input layer, which projects its activity to an output layer (recurrent connections) through sparse FF connections. Different units in first layer receive input, generated using a common truncated logarithmic distribution, with durations drawn at random from a logarithmic distribution; one example is shown in circle at bottom. The units in this layer are active at a certain level for a specific duration, with a gradual transition to zero (example shown in circle at top).

exponential distribution of initial weights onto each unit. Once a pattern is imposed on the input layer, the activity circulates in the network for 80 simulation time steps. Each updating of unit i amounts to summing all excitatory inputs

$$h_i = \sum_k w_{ik}^{ff} r_k^{input} + M \sum_j w_{ij}^{rc} r_j^{rc} + b\frac{1}{N} \sum_k r_k^{rc}$$

The first two terms enable the memories encoded in the weights to determine the dynamics; the third term is unrelated to the memory patterns, but is designed to regulate the activity of the network, so that at any moment in time, $\frac{1}{N}\sum_k r_k$ and $\frac{1}{N}\sum_k {r_k}^2$ both approach the prescribed value **a** (the pattern sparseness mentioned above).

Like the theoretical basis, the simulation assumes a threshold-linear activation function for each unit.

$$r_i = g(h_i - Th)\Theta(h_i - Th)$$

where Th is a threshold below which the input elicits no output and g is a gain parameter.

Dynamics The simulation implements the model defined in section 1, but with slightly different dynamics: cells are updated in parallel according to

$$h(t+\Delta t) = (1-\Delta t)h(t) + \Theta[h(t)-Th)]g(h(t)-Th)\Delta t$$

This is a forward Euler integration scheme for the differential equation 4.12. We can rewrite this equation to

$$\frac{h(t+\Delta t)-h(t)}{\Delta t} = -h(t) + \Theta[h(t)-Th)]g(h(t)-Th)\Delta t$$

In this way, it is easy to make a correspondence with the dynamical equations in our theoretical model, Eq.4.4. Therefore, Δt is indeed in terms of τ_h, synaptic time constant. Thus, in our simulations, with parallel updating, Δt has value of 1 and it means that each time step can be taken to correspond to ca $12.5ms$ (Treves 2004).

Input patterns Each ξ^ν is the projection to the second layer of the input signal from the first layer,ξ^ν_{in} , which is drawn independently from a fixed distribution, with the constraints $\xi > 0$, $\langle\xi\rangle = \angle\xi^2\rangle = a$, where $\langle\rangle$ stands for the average over the distribution. p uncorrelated patterns were generated using a common truncated logarithmic distribution (Fig. 4.2, middle panel) obtained by setting for each input unit

$$\xi_{in} = -\frac{1}{N}\log(1-\frac{x}{a})$$

if $x < a$, and $\xi = 0$ if $x > a$, where x is a random value with a uniform distribution between zero and one.

Weights Feedforward connections, playing the role of afferent signals to IT, are set once, as mentioned above, and kept fixed during the simulation. Recurrent connections, which are the storage site for the memory patterns, have their baseline weight modified according to a model "Hebbian" rule. The specific covariance "Hebbian" learning rule we consider prescribes that the

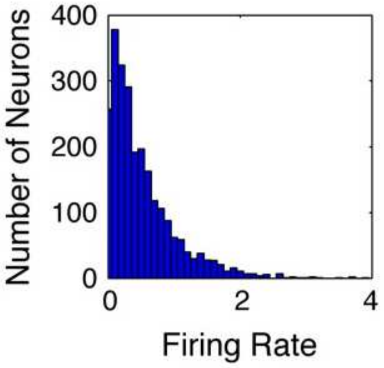


FIGURE 4.2: **distribution of input firing rates** p uncorrelated patterns were generated using a common truncated logarithmic distribution as is shown.

synaptic weight between units i and j, wij, be given as:

$$w_{ij} = \frac{1}{Ca(1-a)} \sum_{\nu=1}^{p} c_{ij} \xi_i^{\nu} (\xi_j^{\nu} - \bar{\xi})$$

where ξ_i^{ν} represents the activity of unit i in memory pattern ν, and c_{ij} is a binary variable equal to 1 if there is a connection running from neuron j to neuron i, and 0 otherwise. $\bar{\xi}$ is the mean activity of unit j over all memory patterns.

Recurrent collaterals do not add signal during the storage In Eq. 4.2.1, M can be any value between 0 and 1, and corresponds to the proportional contribution of collaterals in driving the activity of each unit. But, as previously shown (Menghini et al. 2007; Treves 2004) the best performance is obtained when collaterals are suppressed during pattern storage, in line with the Hasselmo argument about the role of cholinergic modulation of recurrent connections (Barkai and Hasselmo, 1994). The suppression of collaterals during training provides a mechanism for ensuring that during storage, the firing rate of output units, r_i, follows external inputs relayed by afferents to the network. Without this suppression, afferent inputs are represented less accurately in the pattern to be stored in the network, which ends up largely reflecting the previously stored patterns. Therefore, in this simulation, $M = 0$ during storage and $M = 1$ during testing, corresponding to suppression of collaterals during ”training”, and to allowing their full influence during testing.

Active process to regulate *threshold* and *gain* (fixed level of sparseness In the simulations, induced activity in each unit is followed by a competitive algorithm that normalizes the mean activity of the (output) units, and also sets their sparseness to a constant $a = 0.2$ (Treves and Rolls, 1991). It is quite arbitrary to choose the level of sparseness. It is ideal to infer the value of **a** from electrophysiological data from visual cortex, and more specifically, IT cortex. However, pyramidal neurons in IT, due to their hight firing rate (with respect to hippocampal neurons) are in general very difficult to be distinguished from interneurons. On the other hand, due to the limitations of experimental paradigms, it has been almost impossible to monitor a population of neurons (in IT) in response to daily life stimuli. In laboratory adapted experiments, usually, a set of stimuli are shown to the animal while recording from (single) neurons. These set of stimuli can not be a representative of external world. However, there are a few number of studies in which the investigators tried to give an estimate of the neural sparseness in IT cortex (Rolls and Tovee 1995; Franco et al. 2007). Rolls and Tovee (1995) have estimated the sparsity in macaque IT cortex, by measuring the neural responses to a set of 68 visual stimuli during a visual fixation task . The have found that the sparseness of the representation of those 68 stimuli by each neuron (averaged over all neurons) is around 0.68. However, if the spontaneous (baseline) firing rate is subtracted from neuron’s response to each stimulus, therefore only the changes of firing rate, i.e. the responses of the neurons, are used in the calculation of sparseness, then the value of the sparseness has a lower value, with a mean of 0.33 for the population of neurons. Since in our simulations we do not model spontaneous activity, the lower value of sparseness is more biologically relevant. So, referring to the existing body of literature in brain electrophysiology, it is quite difficult and vague how to adapt any biological plausible value for network sparseness. However, we know from theoretically driven hypotheses about the role of sparseness we know the relevance of sparseness for the storage capacity of associative networks (Treves and Rolls 1993), and it has been discussed that the lower levels of sparseness lead to higher storage capacity. Therefor, we have used $a = 0.2$ in our model.

The algorithm that we used to set the mean activity and sparseness, represents a combination of subtractive and divisive feedback inhibition, and operates by iteratively adjusting the gain g and threshold Th of the threshold-linear transfer function.

Implementation of adaptation as firing rate decay In our model, we introduce firing rate adaptation only to the second layer units, i.e. only the output units are prone to get adapted. The derived equation for I_a in section 4.1.2, to be applied in a computer simulation, should be modified from a continuous function in time, to a discrete function.

$$\begin{aligned} \frac{dg(t)}{dt} &= -\frac{g(t)}{\tau} + \sum_K \delta(t - t_k) \\ \frac{g(t+\Delta t) - g(t)}{\Delta t} &= -\frac{g(t)}{\tau} + \sum_K \delta(t - t_k) \\ \Rightarrow g(t+\Delta t) &= -\frac{g(t)}{\tau}\Delta t + g(t) + \Delta t \sum_K \delta(t - t_k) \\ &= g(t)(1 - \frac{\Delta t}{\tau}) + \Delta t \sum_K \delta(t - t_k) \\ &\approx g(t) \exp -\frac{\Delta t}{\tau} + u(t) \end{aligned}$$

where $u(t)$ is the number of spikes in Δt.
for $\Delta t = 1 \Rightarrow$

$$\begin{aligned} g_1(t+1) = g_1(t)e^{-\frac{1}{\tau_1}} + e(t) \\ g_2(t+1) = g_2(t)e^{-\frac{1}{\tau_2}} + e(t) \end{aligned}$$

Where $r_i(t)$ is the activity of unit i at time t. We implemented adaptation by subtracting from the input activation of each unit the difference of g_1 and g_2, which is a term proportional to the recent activation of the unit.

$$r_i(t+1) = g(h_i(t+1) - \alpha_a(g_2(t) - g_1(t))\Theta(h_i(t+1) - \alpha(g_2(t) - g_1(t));$$

$h_i(t)$ is the summed input to the unit at time t. α sets the strength of adaptation. The input to each unit is then affected by its firing rate at all previous time steps. The exponential decay makes its activity at the last time step more influential than the others. The difference of the two exponentials means that the effect of adaptation appears only after the second iteration. Note that this formulation reduces the effectiveness of adaptation when t is small.
The default parameters used in the simulations are listed in table 1, unless otherwise specified.

Size, Sparseness and Time Constants	
Input array	Nin = 1000
Output array	Nout = 1000
(time step)	12.5 ms
FF connections	$C_{ff} = 350$
Recurrent connections	$C_{rc} = 300$
Initial neuronal gain	g = 1
Output sparseness	$a_{out} = 0.2$
Initial neuronal threshold	Th = 0.05
Input sparseness	aout = 0.5
Adaptation time constant	$\tau_1 = 4$
Adaptation time constant	$\tau_2 = 8$
Adaptation strength	$\alpha_a = 5 \times 10^{-3}$

Table 4.1: Default values of network parameters

4.3 Result: Retrieval in different network structures

Now that we have defined the network structure and its corresponding simulation setup, we analyze retrieval properties of such network.

Background Memory is retrieved from the network when neural activities, stimulated with a partial cue, evolve into a pattern strongly correlated with one of those which have been stored. How smoothly can such an operation proceed, and how wide are the basins of attraction of the p memory states, depend critically on whether other attractors exist, that could hinder or obstruct retrieval. In general, the characteristic features of the network during the retrieval process results from a mixed contribution of different factors in storage and retrieval phases. What determines storage specificity, has been studied extensively (see e.g.). All these studies focus on trained networks with desired memories stored as fixed points and recoverable from partial cues in a few time steps. The main issues in studying the retrieval properties of such associative networks are the speed and quality of recall, the required size for cues, and the storage capacity. Considering each of these factors, the relevance of several different concepts such as pattern statistics of the original quenched patterns and the contribution of network architecture on changing the original distributions, and introducing new properties in storage phases, can be studied.

Dilution in connectivity and storage capacity One of the well known terms in determining the storage properties of associative networks is the level of connectivity. In a classic Hopfield model the connectivity is complete, which means every unit in the network receives input from all other units (Hopfield 1982). The connectivity can be sparse, but still independent of the index, as in (Sompolinsky 1986) or in the highly diluted limit considered by (Derrida et al. 1987). Sparse, asymmetric connectivity has been well documented in brain regions such as the CA3 (Amaral et al. 1990; Ishizuka et al. 1990). On the other hand, it was the imposition of symmetry on the synaptic connections (coupling constants) which led to a great clarification of the properties of neural networks (Hopfield 1982 and 1984, Amit et a1 1985 and 1987; Hertz et al. 1991;). But once an initial clarity was obtained, attention turned to the effects of asymmetry and different pressures have acted in this direction to go beyond this simplifying but non-biological assumption; from biological plausibility to questions about the robustness of the results, given that no basic principle enforces symmetry, to a possible cognitive role for asymmetry. Several researchers have studied the dilution of fully connected networks (Derrida and Pomeau 1986; Sompolinsky and Kanter 1986; Derrida et al. 1987; Kree and Zippelius 1991). As some researchers have also pointed out (Crisanti and Sompolinsky 1987; Gutfreund et al. 1988), asymmetrically connected networks have the potential for rich and robust dynamical behavior that might be very useful in modeling complex cognitive processes, especially in their temporal aspect.

This type of model has been thoroughly analyzed in terms of its storage capacity, yielding a relation between the maximum number p_c of patterns that can be turned into dynamical attractors, i.e. that can be associatively retrieved, and the number C of connections per receiving unit. Typically the relationship includes, as the only other crucial parameter, the sparseness of firing **a**, and for sparsely coded patterns (values of **a** close to 0) it takes the form (Treves and Rolls 1991)

$$p_c = k\frac{C}{a}\log{(\frac{1}{a})}$$

where k is a numerical factor of order $0.1 - 0.2$.

Noise Reverberations It is commonly believed that the essential difference introduced by the sparse (i.e. diluted) connectivity is that noise has less of an opportunity to reverberate along closed loops. In fact the signal, which during retrieval is simply contributed by the "condensed" patterns, propagates coherently and proportionally to C, independently of the density of feedback loops in the network. The fluctuations in the overlaps with the uncondensed patterns, which in low fast noise ($T \to 0$) represent the sole source of noise, propagate coherently along feedback loops, giving rise to a decrease in signal to noise ratio for the fully connected case. On the other hand, extreme dilution is a special limit in which only specific kinds of feedback can still exist. For a given load (fixed α), diluted connectivity reduces therefore the influence of this "static" noise, and performance is better than in the fully connected case with $C = N - 1$. However, in the contrary, exactly the same function of recurrent loop may end up to reduce the quenched disorder and sharpen the boundary between retrieved patterns which means a decrease in the number of collapsing patterns.

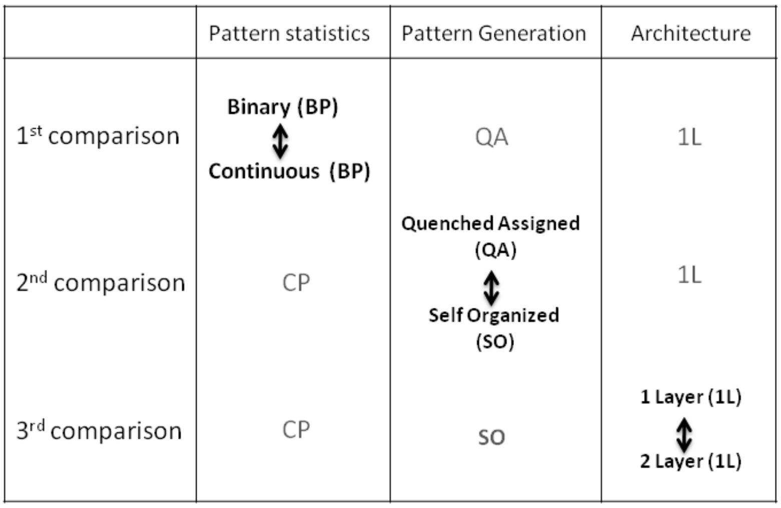
Although, there have been several studies of asymmetrically diluted networks for associative memory, they were focused on storage capacity in the limit of strong dilution, $c = O(1/N)$, (Kree and Zippelius, 1987) or on the stability to a single spin flip in a binary network (Treves and Amit 1988), or on the stability of mixture states in a fully connected continuous network (Roudi and Treves 2003). However, it has remained an open question that how the number of retrieved patterns scales with the level of dilution in a graded-response model, and how a change in connectivity produces a collapse between several patterns with finite levels of correlation.

Here we try to interpolate between the fully recurrent and symmetric attractor network studied by Amit et al (1987), the highly diluted model of Derrida et al (1987) and the strictly feed-forward attractor network studied by Domany et al (1986). To have a quantitative assessment, we measure the effect of dilution, adding the fast-noise and increasing the load in different network structures.

Three comparisons Given our perspective, we have studied a simple but broad class of sparse, asymmetric random networks in which all specific connections are excitatory and inhibition is provided by a single interneuron that receives firing information from all primary neurons and inhibits them all equally at the next time step.

In the rest of this section, first we will discuss how sparseness sets the correlation level between different quenched patterns. Then, we will use the simulated autoassociative network, described in the previous section, to make three comparisons (as is sketched in table 4.3), based on

- **Pattern Statistics**: Binary Distribution (**BP**) vs. a Continuous Distribution (**CP**).
- **Pattern Generation**:a Self-Organized network (**SO**) vs. a network with Quenched Assigned (**QA**) pattern.



	Pattern statistics	Pattern Generation	Architecture
1st comparison	**Binary (BP)** ↕ **Continuous (BP)**	QA	1L
2nd comparison	CP	**Quenched Assigned (QA)** ↕ **Self Organized (SO)**	1L
3rd comparison	CP	SO	**1 Layer (1L)** ↕ **2 Layer (1L)**

Figure 4.3: **Comparison between different structures**

- **Architecture**: a 1-layered network (**1L**) vs. a 2-layered network (**2L**), which receives the external input via feedforward afferents

to see whether the capacity for storage and retrieval with non-negligible basins of attraction vary or not for each comparison. To quantify these differences, we examine the effect of connectivity dilution and fast-noise in limiting storage capacity in each comparison. At the end, we explore the effect of connectivity on the statistical properties of stored and consequently retrieved patterns, in a two layered autoassociative network.

4.3.1 General points: sparseness and cross correlation

Input patterns In each pattern each cell i is taken to code for independent information, i.e. its firing rate during learning η_i^μ is assigned as a random number, where η_i^μ represents the activity of unit i in pattern η^μ. Each η_i^μ is taken to be a quenched variable, drawn independently, for all

i and μ, from a probability distribution $P_\eta(\eta)$.
In our model, once the input pattern is drawn from an original distribution, it may propagate through different stages, before being stored in the synaptic weights of the recurrent connections. Indeed, it is assumed that p final patterns (labeled $\xi = 1...p$), that may be different from the original ones are embedded with equal strength in the synaptic efficacies. Thus, we need to distinguish the original input patterns from the final transformed patterns which then will be stored in the network through Hebbian learning. We adapt the notation η^μ and ξ^μ to represent the original pattern and the to-be-stored patterns, respectively.

Parameters of the quenched patterns To study how the different behaviour of the network depends on the macroscopic features of the statistical distribution of patterns, one has to vary the parameters characterizing P_η. As η is a firing rate, $P_\eta(\eta) \geq 0$ only for $\eta \geq 0$, and $P_\eta(\eta) = 0$, otherwise. Moreover, $\int P_\eta d\eta = 1$. Within these constrains, the first free parameter is the average firing rate

$$a = \int P_\eta(\eta)\eta d\eta \equiv \langle\eta\rangle_\eta \tag{4.1}$$

where $\langle . \rangle_\eta$ denotes averages over P_η. This average pattern activity does not, however, affect memory encoding in the model, as the contribution of each pattern to the efficacies J is normalized, in units of **a** itself. Nor it does affect retrieval, because with a threshold-linear transfer function, the information retrieved does not depend on the absolute scale of the neuronal outputs. Therefore, the average firing activity of the encoded patterns is irrelevant in determining the performance of the network, and the first relevant parameter is the average *square* activity. Imposing that also

$$a = \int P_\eta(\eta)\eta^2 d\eta \tag{4.2}$$

turns **a** into a parameter giving the ratio $\frac{\langle\eta\rangle_\eta^2}{\langle\eta^2\rangle_\eta}$, which in fact is a measure of the sparseness of the coding scheme. The sparseness of the coding scheme turns out to be the most crucial factor on which the performance of the models considered depends (Rolls and Treves 1991). The parameter **a** measures the sparseness of the stored representation, the one used to code information in the learning phase. A different representation, whose degree of sparseness need not be the same, is generated by the retrieval process. As the response of neuron i during retrieval is V_i, the sparseness of the retrieved representation can be quantified by $a_r = \frac{\langle V\rangle^2}{\langle V^2\rangle}$ where now the average $\langle . \rangle$ is both over the random assignment of patterns and over the dynamical process of retrieval.

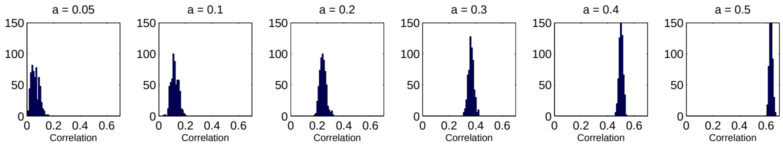


FIGURE 4.4: Sparseness and pairwise correlation. Each plot shows the histogram of pairwise correlation between stored pattern, for different levels of sparseness.

Different distributions with the same sparseness. The remaining features of P_η (the structure of the code) also affect the performance as was fully discussed in, for three different forms of pattern distribution, binary, trinary, and logarithmic. The maximum of the number of patterns that can be encoded on the synaptic strength and individually retrieved (p_c) can be derived for different pattern distributions (Treves and Rolls 1991).

One can consider various forms of P_η, all parameterized in terms of **a**. In the following we will consider one that is consistent with some experimental data ([3]), which is a distribution with exponentially scare high rates $P_\eta(\eta) = (1-2a)\delta(\eta) + 4ae^{-2\eta}$. This example on the one hand will demonstrate that autoassociative memories can function perfectly well with pattern distributions that are not bimodal or n-modal (Abeles 1991; Treves and Rolls 1991), and on the other will offer quantitative insight into some of the effects of a continuous distribution.

Result: sparseness sets correlation In our model, by assuming a fixed level of sparseness during storage and retrieval, for the output layer, (which is satisfied by regulating gain and threshold), we change the level of sparseness in the input patterns η^μ and measure the correlation between different patterns. In figure 4.4, it is shown that how the cross correlation distribution depends on the level of sparseness. It is trivial to see how for independent patterns, correlation is equal to the sparseness. However, we emphasize on this concept mainly because later, we want to compare ability of different network structures (i.e. self-organized networks versus imposed networks) to change the correlation between stored patterns.

4.3.2 Three comparisons: binary vs. continuous, quenched assigned vs. self organized, and 1-layer vs. 2-layer.

All the network simulations in this section, are done with fixing the level of sparseness at $a = 0.2$.

Introducing dilution in our model Introducing the factor $\gamma = C/N$, as dilution, first we measure the storage capacity of each network, when the size of the network is fixed ($N = 1000$). In this case, increasing γ, is equivalent to increase in the number of connections per unit.

Introducing fast noise – The ability to retrieve distorted patterns One should note that although the network may successfully retrieve a pattern, once given a full cue as the input, it may fail to evolve to the right configuration if the initial input is only partially correlated with the stored pattern. That is, depending on the size of basin of attraction, the network may or may not be able to complete a distorted pattern. Thus, the ability of the network in completing and not only retrieving can be tested, with different levels of connectivity. To do this, we have presented the network with noisy version of initial patterns. We call this noise "fast-noise", opposite to the "quenched noise" that is contributed through synaptic weights that code the other patterns. We introduce fast-noise, basically by substituting a portion of the original input vector by random values. Therefore, in the following figures, the fast-noise has a value between 0 and 1, showing the distorted proportion. Putting these two parameters (fast-noise, and connectivity) together, one can study the thermodynamic properties that can be presented by means of (storage capacity versus fast-noise, connectivity, and number of stored patterns) phase diagrams.

1^{st} **comparison: binary vs. continuous pattern** We have stored several Quenched Assigned (**QA**) patterns in the network. Figure 4.5, the top panel shows the phase-diagram for storage capacity versus fast-noise and connectivity for a network with $p = 40$ stored patterns (left: binary, right: continuous). The lower panel shows storage capacity versus storage load and connectivity (tested in zero noise limit). It can be concluded that

- In general, increase in connectivity enhances the storage capacity.
- Continuous pattern is highly fragile to noise
- In zero noise limit, with the same level of connectivity, γ, a network with continuous pattern has a higher storage capacity compared to a network with binary patterns.

- In the binary case, the storage capacity enhances by increase in connectivity, whereas in the case with continuous patterns, this enhancement is limited only to the diluted networks.

2^{nd} **comparison: self-organized vs. quenched assigned** In this part, the input pattern is drawn from a logarithmic distribution like in 4.7. We test i) a 1-layer Hopfield like network, in which the inputs are directly imposed to the collateral layer (**QA**) and ii) a similar network, except that before setting the hebbian weights, the network is allowed to reverberate activity through its recurrent connections, and hens regulate the gain and threshold (**SO**).

We measure the same quantities as in Comparison 1 and the results are reported in figure 4.6.

- The same continuous patterns that had high fragility to the fast noise, in a network with imposed patterns (**QA**, left) become way more robust against the addition of noise if the patterns are instead self organized (**SO**, right).
- However, in zero noise limit, the storage capacity of the network with imposed patterns, is as good as the self organized network.

Two-layered network with continuous patterns In this model, we restrict our study to a self-organized network with continuous patterns (**2L-SO-CP**). Due to two main factors we avoid the two other possibilities in the 2-layer network: 1) We will not study the Quenched Patterns in this architecture, because: we only change the recurrent weights in the second layer and the feedforward weights are set randomly (and kept fixed during the simulation). The addition of this first layer introduces a Gaussian noise and the logarithmic distribution of the input patterns, after they project to the second layer, will not be preserved (it's instead a Gaussian distribution). Using such Gaussian patterns as the Quenched patterns, although is an interesting topic, is out of our goals here. 2) We do not study Binary patterns, again due to such random feedforward weights that changes the values of output units from being binary. Still, one can ”binarize” the input patterns after their projection to the second layer. But, it is beyond our concerns for the moment.

In the 2-layer network, on one hand the feedforward layer plays the role of a decorrelator and on the other hand the recurrent synapses, modified based on Hebb-rule, forms properly distinct attractor structures. In Figure 4.7 we have plotted the distribution of pairwise correlation between 40 pairs of original patterns drown from the exponentially distribution in Eq. 4.7 (a), immediately after being projected to the recurrent layer (b) and after 10 iterations of propagating through the recurrent connections (c). Comparing these three stages of pattern generation, one can detect the influence of each step, and can understand how the recurrent collaterals, together with feedforward afferents can change the ”correlation”. In this figure, different rows show correlation values when the initial level of sparseness in producing the original patterns, before projecting to the second layer, is different; either bigger or equal or smaller than the sparseness in the recurrent layer. The third column of this figure depicts the distribution of correlation between configurations that will be stored in the recurrent weights through the Hebbian learning. For the rest of the simulation studies in this section we set the level of input sparseness to $a = 0.2$, unless otherwise stated.

3^{rd} **comparison: 1-layer vs. 2-layer** . We apply the same quantitative comparison to contrast the features introduced by an additional layer (Fig 4.8).

- In low connectivity region, the 1-layer network is more robust to noise compared to the 2-layer network
- Whereas in high connectivity region, the slope of the storage capacity (versus noise and connectivity), is almost the same for both networks
- In zero noise region, the 1-layer network has higher storage capacity compared to the 2-layer network

Number of connectivity, versus number of units Based on the results up-to now, one can conclude that the recurrent connections and therefore the number of closed loops are more of an advantage in reducing the quenched noise. This is, however, an issue of definition because in a different region in which the ratio of recurrent inputs to each unit does change by varying the number of total units, instead of varying the number of connectivity, different mechanism could lead to an opposite effect. Namely, as it was discussed in Treves and Rolls 1991, for different levels of sparseness and different size of the network, either C or N could play the main role in defining the storage capacity of the network.

Thus, we rerun the same simulations as before (**2L-SO-CP**), with one difference: this time we kept number of connection per unit as a fixed parameter $C = 700$ and instead we have varied the size of the network N. The result is shown in figure 4.9. The left panels show the percent of successfully retrieved patterns in a network with fixed number of connections per unit ($C = 700$) and of different size, whereas the right panels show the result of a network with fixed number of units ($N = 1000$) and varying number of connections per unit. Different rows of this figures relate to different level of sparseness.

4.3.3 Time course of retrieval

Before finding an explanation for what the degree of connectivity does in terms of changing the statistical properties of stored patterns, we would like to point at another relevant issue, related to the time course of retrieval process.

Autocorrelation during retrieval – The ratio of the external input currents to the recurrent collateral currents To assess the quality of the retrieval process, we measure autocorrelation values, which are simply defined as the cosine between the two population vectors ξ, the to-be-stored pattern, and ζ, the retrieved pattern.

In every type of network architecture the configurational state of the network at each time, is a function of the input current that enters each unit, which itself is resulting from the contribution of both feedforward (external) input, and recurrent connections with different weights. The competition between these two type of input (feedforward vs. recurrent inputs) is a crucial factor in determining the network trajectory in its phase-space: to fall into a basin of attraction relevant to the signal contributed by recurrent collaterals, or to stay in a configuration close to the one dictated by the external input. This issue has been also been explored by Parga and Rolls (1998) in developing a network model for view invariant recognition. They have shown that different synaptic values (between different views of an object, or within each view) can lead to different state configurations, either coding the identity (ignoring the view) or the identity plus the viewpoint. In our model, an interesting finding is that although at long time, the basin of attraction of some of stored patterns may shrink and several patterns collapse into a very few number of stable attractors, there is a time after input removal at which almost for all patterns the network approaches a configuration near the stored pattern. However, the exact time of retrieval and the quality of retrieval, in this phase, depends on how strong are the attractor states to overcome the external input. The ratio of successfully retrieved patterns at each time steps, for different strength of recurrent connections, is plotted in figure (4.10, for **2L-SO-CP**). Interestingly, during the early times, a network with larger synaptic factors evolves faster to its final asymptotic state. This asymptotic state, when the external input is totally washed out, is the same for any finite value of recurrent strength, above 0.01, not only in terms of number of uncollapsed patterns, but also in the exact distribution of retrieved patterns. In the next analysis, we set the reference time for capacity measures at time 40 and studied the effect of connectivity dilution and fast noise on storage capacity, at each comparison.

4.3.4 Why increasing the number of connectivity improves the storage capacity?

At this stage it begs the question that under which mechanisms an increase in the number of recurrent connections results in an increase in the number of stable attractors. We limit our investigation only to **2L-SO-CP** model. To tackle this issue, we first looked at the pair-wise correlation between stored patterns and retrieved ones. One could expect that the higher the correlation between the stored patterns, the higher the probability of collapsing together during retrieval.

- Correlation between stored patterns can not predict the level of correlation among the retrieved patterns in low-connectivity region.

Stable versus Unstable patterns Figure 4.11, top panel shows the final autocorrelation between the same 40 patterns after 120 time steps of iteration, when the network is asymptotically stable. This graph shows that while some of the patterns have been retrieved successfully, the network has failed to correctly retrieve part of the stored patterns. Setting a threshold for the value of autocorrelation at $\rho = 0.8$, we then divided the set of 40 patterns into two groups the stable and unstable patterns. The middle panel in figure 4.11 shows the distribution of pairwise correlation between stable patterns (in red) and between unstable patterns (in blue). Clearly, the mean of the distribution for unstable patterns has a positive shift with respect to the other distribution. This reflects that on average a bigger proportion of unstable patterns, with respect to stable ones, have been attracted to a similar configuration. From the bottom panel which shows the pairwise correlation between the stable and unstable, it can be understood that several unstable patterns indeed have been collapsed onto a stable one.

Retrieval correlation versus Stored correlation Does this separation between the distribution of stable and unstable patterns reflect a correlation inherited from the original patterns? In the other words, given the two final retrieved configurations, ζ_i and ζ_j, which are on average more correlated if the patterns are unstable, one can speculate that perhaps the two original patterns ξ_i and ξ_j are also more correlated, on average, than stable pattern. However, figure 4.12 clearly shows that this is not the case.

- **I**ncrease in connectivity, γ, reduces the influence of less-correlated patterns

The raster plot for retrieved correlation versus stored correlation is shown in figure 4.13, for different levels of γ. Interestingly, while for low γ, the retrieved correlation for originally low-correlated pairs of patterns is as high as those that are highly correlated, for high values of γ, the retrieved correlation is mostly close to one, only for those pair of patterns that are a priori more correlated than the others. This could suggest that increasing the number of recurrent connections plays a role in averaging out the quenched noise, in favour of the ”condensed” pattern.

Quenched noise The quenched noise due to the storage of several patterns in the network, using the Hebbian learning, gives rise to overlapping synaptic weights that at the end causes the collapse of some patterns that could not form a well separated basin of attraction. These patterns although they have the same level of correlation, a priori, as the final stable ones, probably receive more of such quenched noise through the recurrent connections.

The next step is to actually measure the quenched noise for different values of γ. To better understand the function of noise on the stability of different patterns, we separately study patterns

that *a*) are stable for any value of connectivity; *b*) are never stable for any values of connectivity; *c*) are stable, and do not attract any of the other patterns; *d*) are unstable for low values of γ, but stable for larger γ.
Then, for each pattern in any of the four groups we have calculated

$$R_i(t) = \sum_{\nu \neq a} \xi_i^\nu m_i^\nu(t) \tag{4.3}$$

We found that the noise term for patterns depends highly on the level of connectivity. e.g. for those patterns the stability of which changes with the number of connections.

In the next section, we develop a theoretical framework to monitor the dynamics of an attractor network in response to morphed patterns.

4.3.5 Summary and conclusion

The main issue we aimed to explore was to see the effect of different network structure and pattern statistics on retrieval of morphed patterns. These three comparisons in fact, puts forward three steps in making the network modelings, in general, more realistic; i.e. i) replacing a binary representation with a graded distribution of activity, ii) allowing the input patterns to self organize in the output layer, before the storage in the recurrent weights, and iii) inserting additional layers. We found that

- Almost in every structure, the storage capacity increases by the number of connectivity (as it was well known already)
- The ability of the network to retrieve distorted patterns is higher if patterns are binary.
- In zero noise limit, network with graded patterns has higher storage capacity.
- Binary patterns gain more from the increase in connectivity, compared to graded patterns.
- The same continuous patterns that had high fragility to the fast noise, in a network with imposed patterns become more robust against the addition of noise if the patterns are instead self organized.

- The addition of and input layer only changes the function of the network from the quantitative point of view (unless for the highly diluted case, that in fact can be a finite size effect rather).

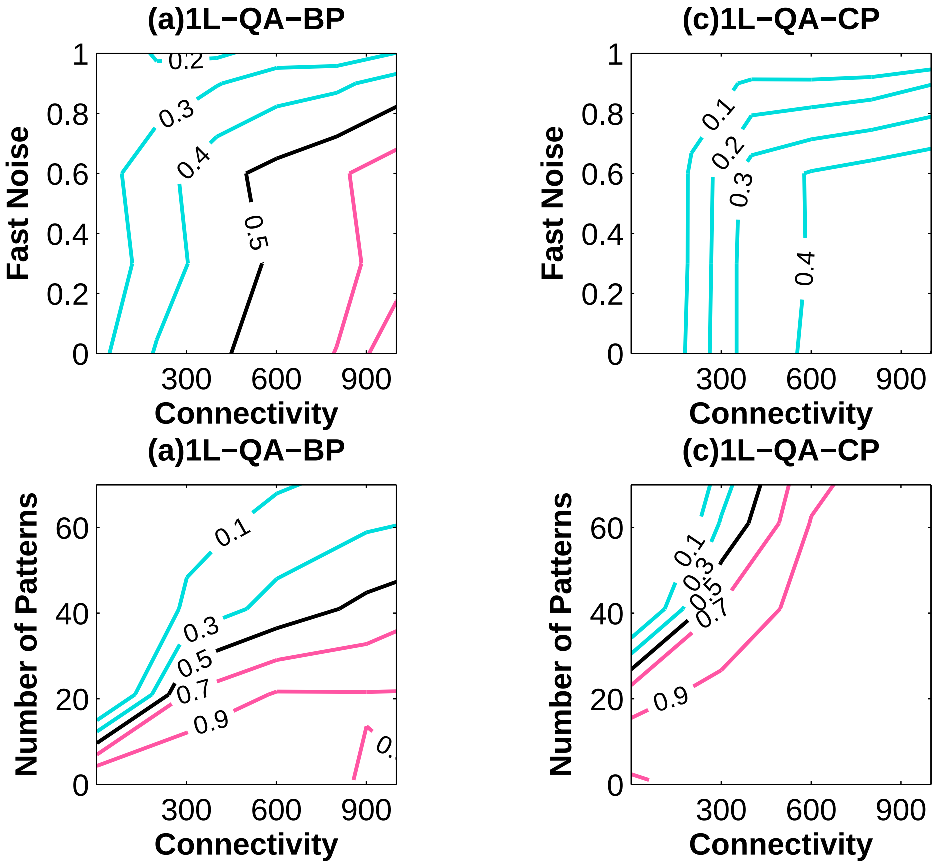


FIGURE 4.5: **Storage capacity, Binary vs. Continuous.** The stored patterns that are imposed to the recurrent collaterals (quenched assigned, **QA**) are drawn from either a binary (a, and c) or a continuous distribution (b, and d). The top panel shows the phase-diagram for storage capacity versus fast-noise and connectivity, with $p - 40$ stored patterns. The level of fast-noise, shown by a number between 0 and 1, means the proportion of the original input pattern that is set to random values. The lower panel shows storage capacity versus storage load and connectivity in zero noise limit. Red (blue) is adapted for storage capacity above (below) 0.5. Black curves show storage capacity equal to 0.5

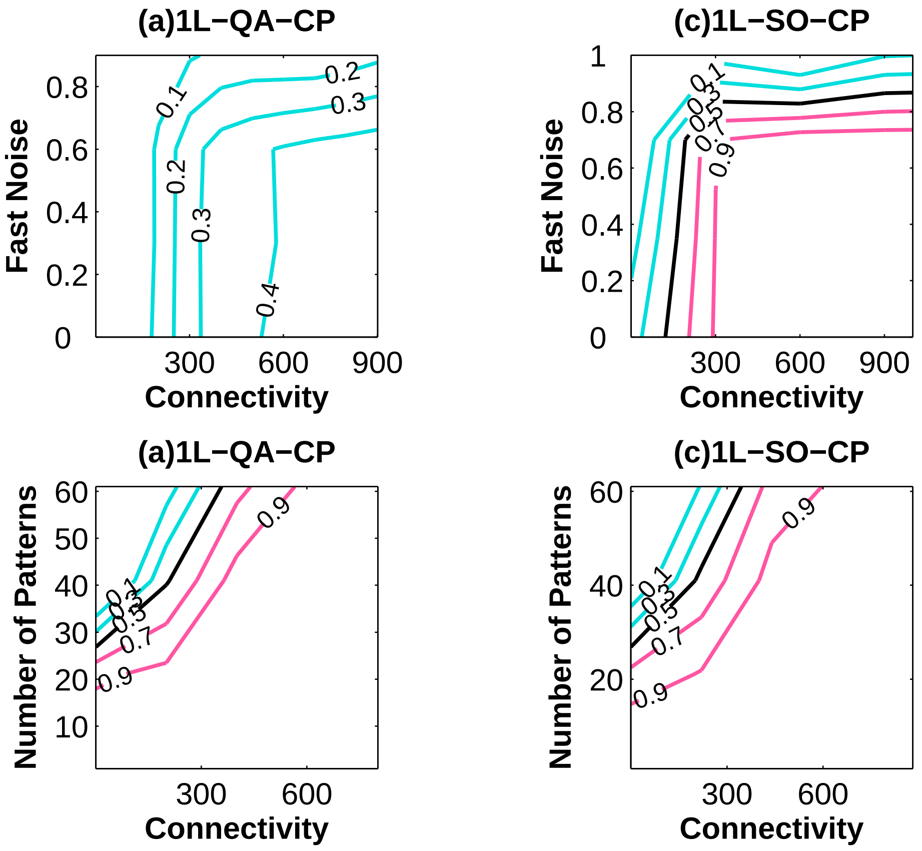


FIGURE 4.6: **Storage capacity, Self organized vs. Quenched Assigned** A set of continuous patterns (drawn from a logarithmic distribution) are either imposed to the recurrent collaterals (quenched assigned, **QA**, left) or are reverberated through the recurrents while the level of mean activity and sparseness are regulated (self organized **SO**, right). The top panel shows the phase-diagram for storage capacity versus fast-noise and connectivity, with $p - 40$ stored patterns. The lower panel shows storage capacity versus storage load and connectivity in zero noise limit

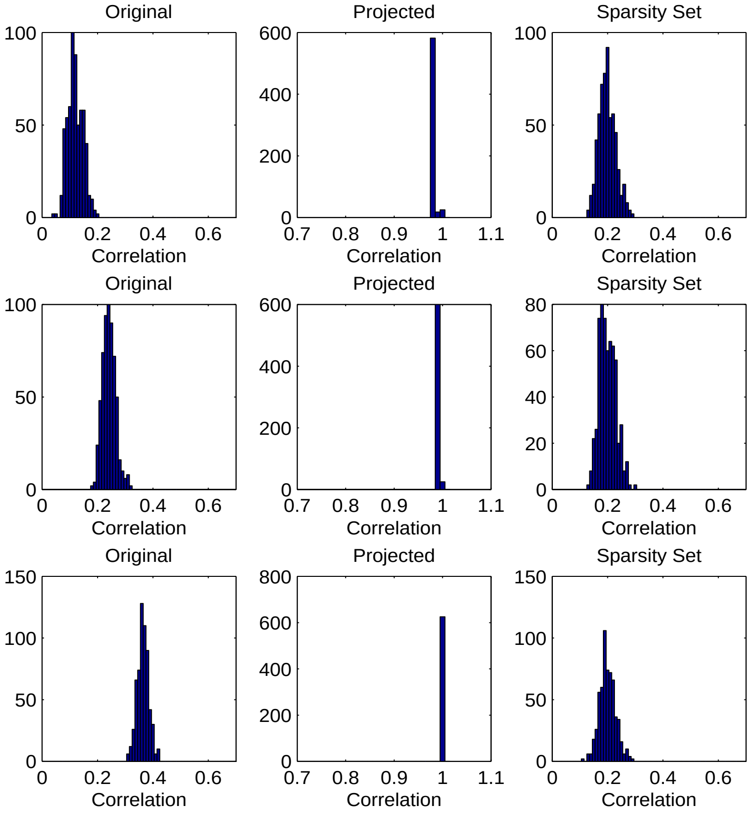


FIGURE 4.7: **Self organized network regulates mean activity and sparseness** Distribution of pairwise correlation for 25 random patterns that are drawn from the logarithmic distribution as in) (patterns η, left), then are projected through the random feedforward connections to the second layer (middle), and then are propagated through the recurrent connections while the mean activity and the sparseness of the network is regulated (patterns ξ, right). The difference between rows comes from the initial level of sparseness in the input layer(top,$a = 0.1$; middle $1 = 0.2$,bottom $a = 0.3$

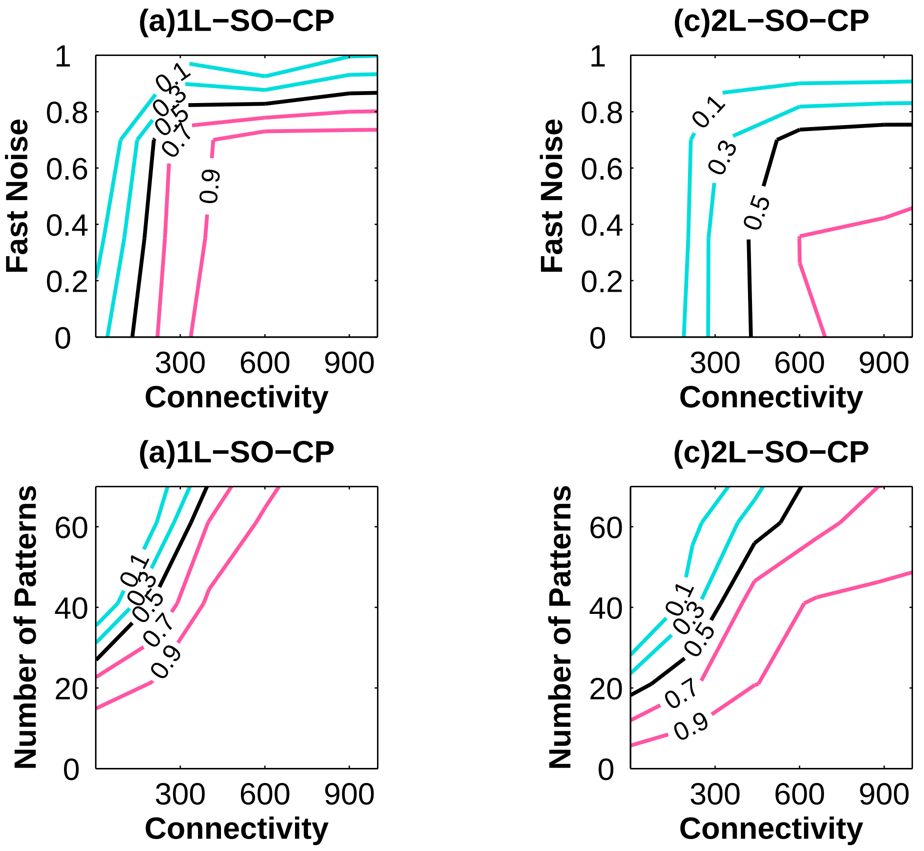


Figure 4.8: **Storage capacity, 1-layer vs. 2-layer network** The same quantitative measurements as in Fig. 4.5 and 4.6 have been applied to compare the role of an additional layer that send feedforward connections to the second layer. (1-layer on left, 2-layer on right). The top panel shows the phase-diagram for storage capacity versus fast-noise and connectivity, with $p-40$ stored patterns. The lower panel shows storage capacity versus storage load and connectivity in zero noise limit

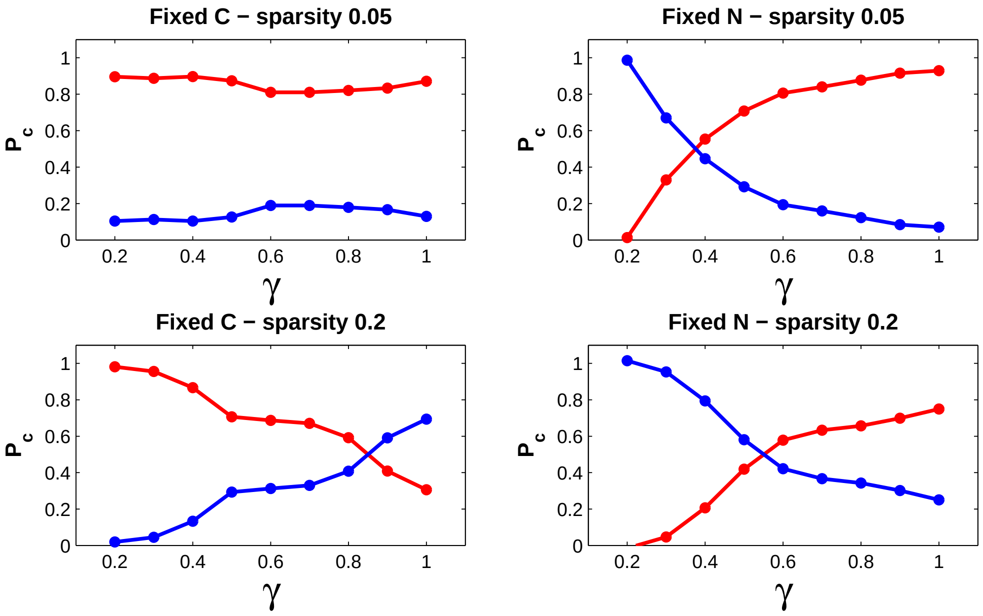


FIGURE 4.9: **Number of units vs. number of connectivity**. Number of retrieved patterns (P_C) as a function of γ when the number of connectivity is fixed ($C = 700$) and the total number of units changes (left), or when the size is fixed ($N = 1000$) and the number of connectivity changes (right). Top: storage capacity in very spars limit $a = 0.05$; Bottom: $a = 0.2$

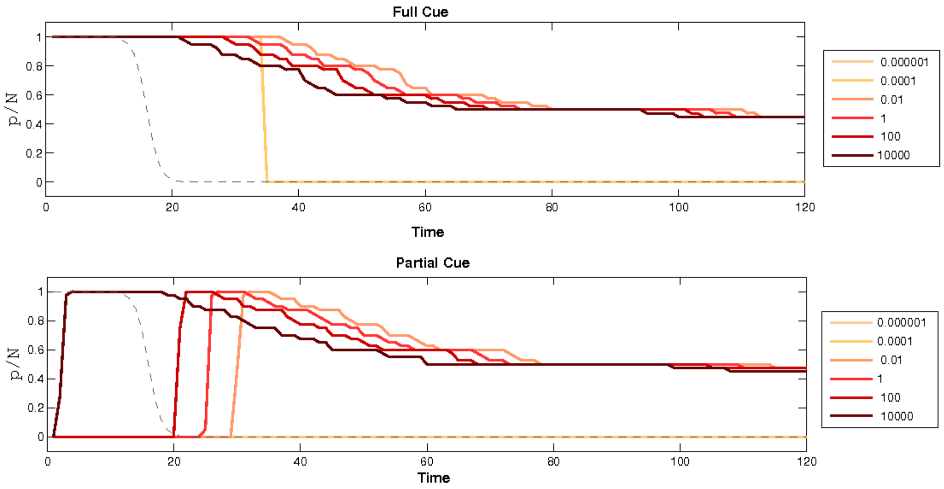


FIGURE 4.10: **Time course of retrieval** Time course of percent retrieved patterns when the initial input is a full (top) or partial cue (bottom). $P = 40$, $C = 300$, $a = 0.2$

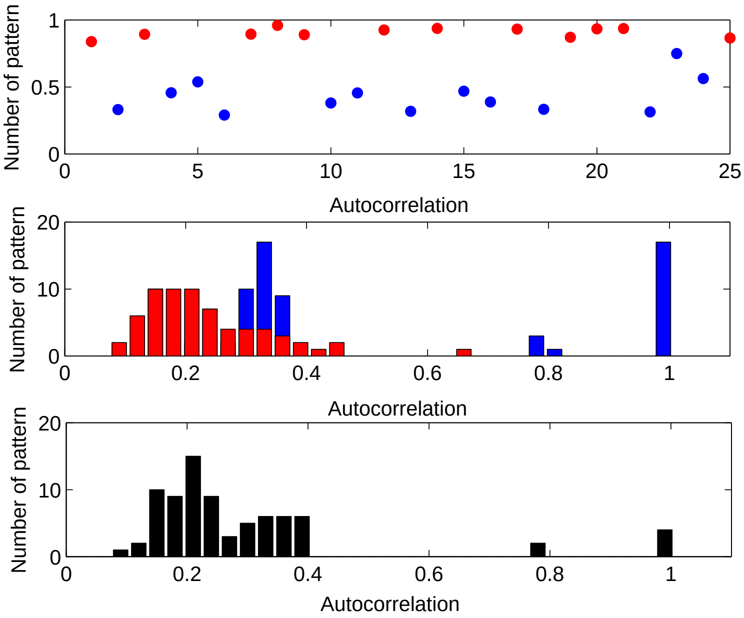


FIGURE 4.11: **Autocorrelation distribution for stable vs. unstable patterns**Top: Autocorrelation values between stored and retrieved patterns, Middle: pairwise correlation distribution among stable patterns, in red and unstable patterns, in blue. Bottom: Correlation between stable and unstable patterns

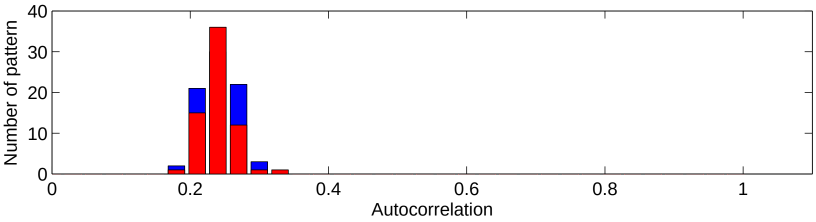


FIGURE 4.12: **Statistics of stable vs. unstable patterns** The distribution of pairwise correlation for the group of stable stored patterns, in red, and unstable stored patterns, in blue.

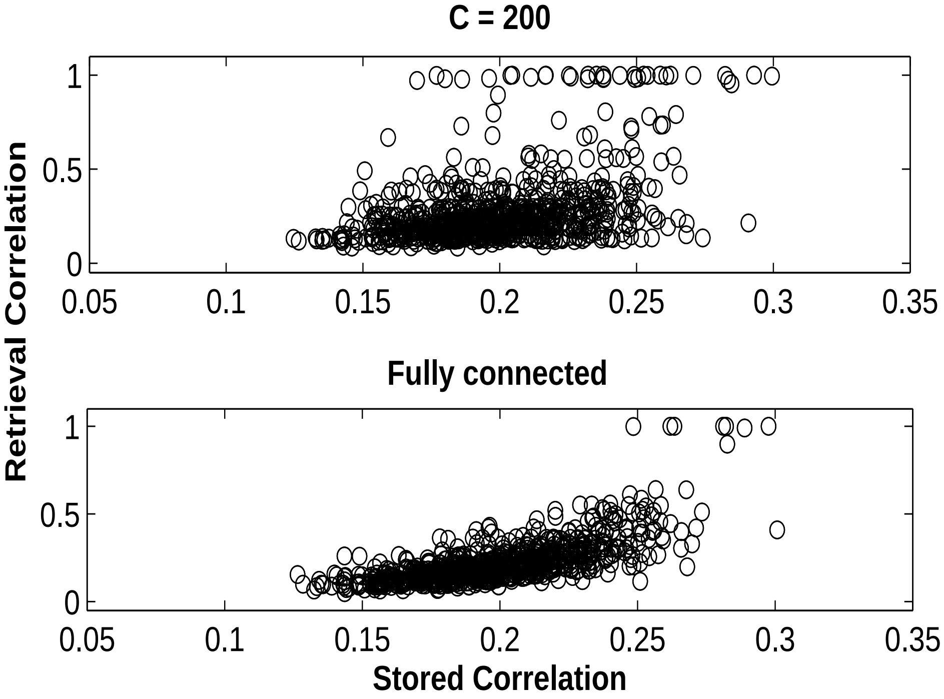


FIGURE 4.13: γ **reduces the influence of less-correlated patterns** Raster plot of pairwise correlation between retrieved pattern-versus stored pattern. Top: very diluted network $N = 1000$, $C = 700$. Bottom: the same network, but fully connected

4.4 Network dynamics in response to ambiguous patterns

In the previous sections, we discussed different properties of pattern statistics and their effect on memory retrieval in an associative network. The input to the network was a full or partially distorted version of a stored pattern. Here, we study network dynamics, when the input is an ambiguous pattern, correlated with more than one of the stored patterns, namely a "morphed" pattern. We would like to see the retrieval of morphed patterns (between a pair of stored pattern), that are drown from either of the two extreme of pattern distribution; binary and continuous distribution. We limit our analytical treatment to study the situation where the external input to the network is equally correlated with two of the stored patterns. The question is what determines the final state of the network given such external input? There is a large body of literature on noise/adaptation induced alternations in networks without quenched memory patterns (Moreno-Bote et al. 2007). We have already reviewed the general characteristics of such memory-less models in the first chapter of the Thesis. Studying the dynamics of a network endowed with memory that is encoded in asymmetric synaptic couplings is not trivial due to lack of a time-dependent Hamiltonian description. In general, the switch from one attractor to another is an non-perturbative phenomena and thus to do the perturbation, one should study a system at the edge between the two attractors, instead of a system that is already settled into the bottom of the attractor. On the other hand, under such condition, the usual signal to noise analysis, is not effective because at the edge between two attractors the signal is too small. Thus, one can adapt the perturbation theory at the edge, which is an unstable steady-state and use the signal to noise analysis only at the final steady state, once the system is close enough to an attractor state.

4.4.1 Rationale for an approximate description

Conceptually, we can draw an abstract framework for dynamics governing the evolution of the network trajectory after the onset of a specific stimulus. Although in our network, several variables affects the dynamics, for the sake of the argument we can assume that the vector field of the system can be depicted in a two dimensional space with only two nullcline whose intersections determine the equilibrium states of the network, either stable or unstable (figure 4.14). The main question becomes what determines the fate of trajectories in this space. The middle cross-section is an unstable point, however, it is instructive to consider small perturbations of this, which then determine the fate of their trajectories as shown in figure 4.14. The fate depends only on whether

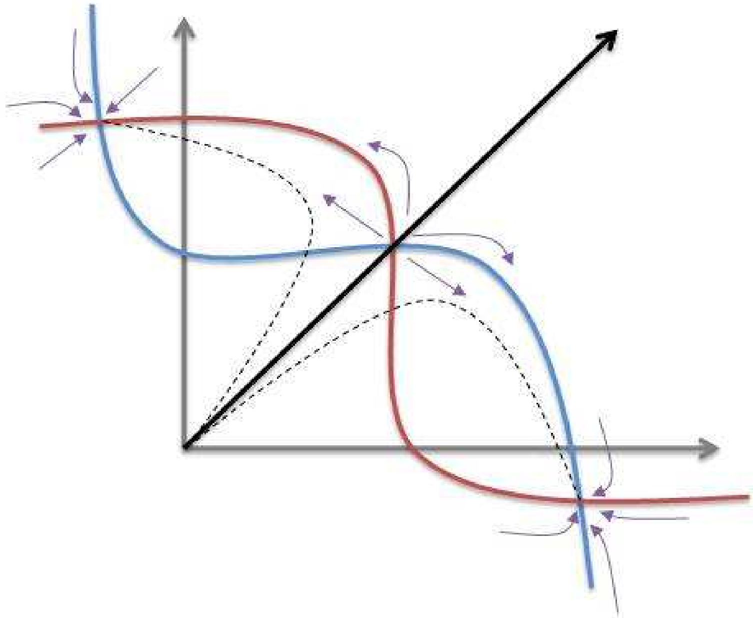


Figure 4.14: **Network trajectory** An abstract schematic view of the dynamics underlying convergence of the network state into an attractor state

the initial state (at onset of the ambiguous input) is below or above the diagonal. In this chapter, we rather use a simple model to study the time evolution of an attractor network, for which we will find the minimal system of differential equation that describes the dynamics of input currents and output currents. We consider a network at steady-state (a configuration partially correlated with two attractors of the network) and will find the condition under which this steady state un-stabilized an moves toward the bottom of an attractor. We define the "default" network, as a network with P number of patterns stored in its recurrent weights, which receives external input from time 0 upto t_{off}. At time t_{off}, external input starts to decrease exponentially. This external input is resembling one of the "morphed" patterns between **a** and **b**. At the first approximate, the "default" network does not include firing rate adaptation. We assume that this network arrives at a steady-state before the external input starts to fade away. We aim to look at the stability of such a system, around its steady-state, under the perturbation that happens at the time of removing the external input. The question is whether such a perturbation is able to produce a bifurcation, and how.

6 groups of units			
G1	$\xi^a = 1$	$\xi^b = 1$	$I_{in}^{\mu_{ab}} = 1$
G2	$\xi^a = 0$	$\xi^b = 0$	$I_{in}^{\mu_{ab}} = 0$
G3	$\xi^a = 1$	$\xi^b = 0$	$I_{in}^{\mu_{ab}} = 1$
G4	$\xi^a = 1$	$\xi^b = 0$	$I_{in}^{\mu_{ab}} = 0$
G5	$\xi^a = 0$	$\xi^b = 1$	$I_{in}^{\mu_{ab}} = 1$
G6	$\xi^a = 0$	$\xi^b = 1$	$I_{in}^{\mu_{ab}} = 0$

Table 4.2:

4.4.2 Storing binary patterns

In dealing with large networks, once the temporal dynamics of each unit is derived, a mean-field description can be obtained by summing up the equations describing the dynamics of individual units to get (fewer) equations that describe collective behaviour (Frolov and Medvedev 1986). One type of network in which cells are grouped into a large number of classes (possibly as large as N the number of cells) on the basis of the strengths of their mutual connections is an autoassociative memory with graded response units coding graded value patterns. Thus grouping Eq. 4.12 results in N functional equations describing the evolution in time of the fraction of cells of a particular class that at a given instant have a given firing rate. However, at the same time, if one allow the patterns presented to the network to have only binary values, (while the units are kept of threshold-linear type), the number of actual groups reduces from N, total number of units, to only 6.

6 groups of neurons encoding pattern a, b and *ab* Assuming binary patterns, units can be divided into 6 distinct groups, based on their response to binary patterns **a**, **b** and ab, a morphed pattern between these two.

then the local field to each group will be (for $i = 1 \dots 6$):

$$\begin{aligned}
\tau_h \frac{dh_i(t)}{dt} &= -h_i(t) + \frac{\alpha}{Ca(1-a)} \xi_i^a \sum_{k=1}^{N_g} n_k (\xi_k^a - a) r_k + \\
&\quad \frac{\alpha}{Ca(1-a)} \xi_i^b \sum_{k=1}^{N_g} n_k (\xi_k^b - a) r_k + \\
&\quad \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta + I_{in}^i - Ia_i
\end{aligned}$$

n_k and $\bar{r}_k$ are the normalized number of units belonging to group k, and their mean activity, respectively $(n_1 + n_3 + n_4 = n_1 + n_5 + n_6 = Ca)$.

Since the different contributions to the local field of each of the two neurons i and j, that belong to the same group k, is the same, then $\bar{r}_k$ is similar to $r_i = r_j$. For simplicity we can replace $\bar{r}_k$ with r_k.

The final set of differential equations for our 6 groups will be

$$\begin{aligned}
\tau_h \frac{dh_1(t)}{dt} &= -h_1(t) + m^a + m^b + I_{in}^1 + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_1 \\
\tau_h \frac{dh_2(t)}{dt} &= -h_2(t) + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_2 \\
\tau_h \frac{dh_3(t)}{dt} &= -h_3(t) + m^a + I_{in}^3 + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_3 \\
\tau_h \frac{dh_4(t)}{dt} &= -h_4(t) + m^a + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_4 \\
\tau_h \frac{dh_5(t)}{dt} &= -h_5(t) + m^b + I_{in}^5 + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_5 \\
\tau_h \frac{dh_6(t)}{dt} &= -h_6(t) + m^b + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_6
\end{aligned} \tag{4.1}$$

Moreover, in addition to these 6 equations, the equations for gain and threshold should be taken into account, as well, as it was discussed in the previous section (4.23).

$$\tau_g \frac{dI_g(t)}{dt} = -I_g(t) + \frac{\sum_k n_k < [h_k - Tr]^+ >}{Na} \tag{4.2}$$

$$\tau_t \frac{dTr(t)}{dt} = \frac{(\sum_k n_k \langle r_k \rangle)^2}{N^2} - a \frac{\sum_k n_k \langle r_k \rangle^2}{N} \tag{4.3}$$

$$\tag{4.4}$$

Up to now, We have defined a set of differential equations, describing the firing rate output of 6 cell assemblies, together with their gain and threshold, that can be implemented through inhibitory GABAergic synapses. At this point, these 8 differential equations should be numerically integrated to give us the full picture of temporal change in each group's firing rate. This system of equations, once solved, can be used to define the time evolution of overlap between network state and memory state for patterns **a** and **b** which is one of the most relevant order parameters for our purpose. In the following sections, we are going to first find the mean firing rate of each cell assembly at its steady-state, and then analyze the stability of such steady-state perturbed by different relevant factors.

4.4.3 Steady-state response

From the system of differential equation (Eq.4.1) , developed in section 4.4.2, the output firing rate of each group, when the adaptive current is zero, at steady state will be:

$$\begin{aligned}
h_1 &= \tau_e g[m^a + m^b + I_{in}^1 + \sqrt{\frac{A_\eta}{(1-a)}}q\eta - Tr]^+ \\
h_2 &= \tau_e g[\sqrt{\frac{A_\eta}{(1-a)}}q\eta - Tr]^+ \\
h_3 &= \tau_e g[m^a + I_{in}^3 + \sqrt{\frac{A_\eta}{(1-a)}}q\eta - Tr]^+ \\
h_4 &= \tau_e g[m^a + \sqrt{\frac{A_\eta}{(1-a)}}q\eta - Tr]^+ \\
h_5 &= \tau_e g[m^b + I_{in}^5 + \sqrt{\frac{A_\eta}{(1-a)}}q\eta - Tr]^+ \\
h_6 &= \tau_e g[m^b + \sqrt{\frac{A_\eta}{(1-a)}}q\eta - Tr]^+
\end{aligned}$$

in which

$$\begin{aligned}
m^a = \frac{\omega}{Ca(1-a)}(\frac{n_1}{<}r_1 > + \frac{n_3}{<}r_3 > + \frac{n_4}{<}r_4 > -a\sum_k \frac{n_k}{<}r_k >) \\
m^b = \frac{\omega}{Ca(1-a)}(\frac{n_1}{<}r_1 > + \frac{n_5}{<}r_5 > + \frac{n_6}{<}r_6 > -a\sum_k \frac{n_k}{<}r_k >)
\end{aligned}$$

and $n_1 + n_3 + n_4 = n_1 + n_5 + n_6 = Ca$. The ω factor is a parameter to scale the recurrent weights to a desired proportion of feedforward weights. (By absorbing this factor into the definition of overlaps m_a and m_b, they are no longer span from -1 to 1, but instead $m_\nu \exists [-\alpha\alpha]$).

Statistics for mean firing rate of each group In the first section, once we did consider the effect of quenched noise as a Gaussian variable on changing the firing probability of different units. Here, we use the same concept once more, to resume the variability of firing rate within each group. We assume a Gaussian distribution for the values of h_i in each group. The probability of the group to fire above the threshold is equal to

$$Pr(r_i > 0) = Pr(h_i > Tr) = \int\limits_{Tr}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(h_i - \bar{h_i})}{2{\sigma_i}^2}} dh$$

Based on such property, we can estimate $\langle r_i \rangle$ and $\langle e^2 \rangle$ as

$$\langle r_i \rangle = \sigma_i \rho \Phi(\rho) + \sigma_i \Psi(\rho) \tag{4.5}$$

$$\langle {r_i}^2 \rangle = {\sigma_i}^2 (1 + \rho^2) \Phi(\rho) + {\sigma_i}^2 \rho \Psi(\rho) \tag{4.6}$$

where $\rho_i = \frac{h_i - Tr}{\sigma_i}$, and σ_i is the standard deviation of the current h_i of each group. The definition of functions $\Phi(\rho)$ and $\Psi(\rho)$ are given in the first Section of this chapter. In figure 4.15 we have plotted the distribution of h_k (on the left) and r_k (on the right) for 6 different groups. The vertical red line depicts the threshold.

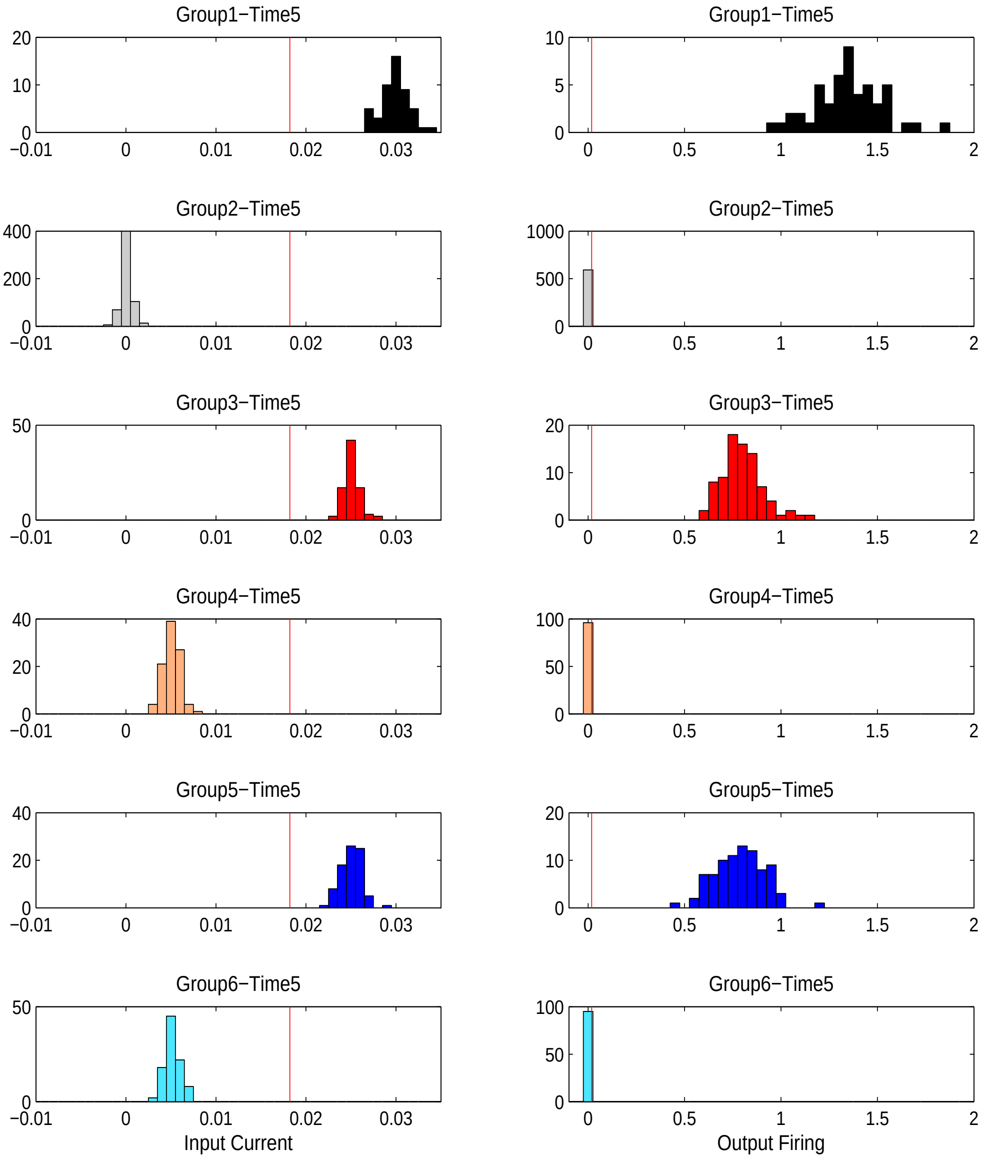


FIGURE 4.15: **Distribution of activity in 6 Binary groups** Input current (left) and output firing rate (right) distribution for different groups, in the presence of external input. The vertical red line shows the threshold.

4.4.4 Minimal system of differential equation

For simplicity, we start from a situation in which the external input signals a configuration almost equally correlated with **a** and **b**, with a slight deviation toward **a**. Once considering the network in such equilibrium state, in the presence of an external input, one can assume that only groups 1, 3 and 5 are active and the other groups are still silent. In other words, the currents that enter groups 2, 4 and 6 are below the threshold and thus insufficient to make them fire. Therefore, we can limit our probe only to the activity of these groups from the moment that the external input start to fade away, until the moment that group 4 (and perhaps 6) start to build up an activity and the stability of such steady-state is destroyed and following that the system runs into a trajectory toward the most attractive state close by. In such reduced model, it is an acceptable assumption to replace $\langle r_i \rangle$ with $h_i - Tr$, for $i = 1, 3, 5$. In figure 4.15 the distribution of the current and its corresponding firing rate is plotted for 6 different groups, using the simulated network, when the external input is fixed. Clearly the simulation result confirms the above assumption. However, reducing the external input gives rise to a positive shift in the distribution of h_4 and h_6, due to an increase in the gain of the system $g = \frac{1}{I_g}$. Thus, for these two groups, we estimate the mean firing rate of the group using only the tail of the Gaussian distribution (which is above threshold), whereas for groups 1, 3 and 5 we use the linear dependency on input current $(h - Tr)$.

The difference h_{35} and h_{46} Considering this reduced model, we define the variable h_{35} as the difference between the current entering group 3 and group 5, i.e. $h_3 - h_5$. Similarly, can be defined h_{46}. Now, the steady-state can be found not anymore using all the 8 variables, but only based on the two new variables h_{35} and $h_{46}6$. The time evolution of them is like

$$\begin{aligned}
\frac{dh_{35}}{dt} &= \frac{dh_3}{dt} - \frac{dh_5}{dt} \\
&= -h_{35} + m_{ab} + I_{35} + QN \\
\frac{dh_{46}}{dt} &= \frac{dh_4}{dt} - \frac{dh_6}{dt} \\
&= -h_{46} + m_{ab} + QN
\end{aligned}$$

where $m_{ab} = m_a - m_b = \frac{\beta}{1}(n_1 r_1(1-a) + n_3 r_3(1-2a) + n_5 r_5(1-2a)$ and I_{35} is the differential input to group 3 and 5. At the steady state, when $\frac{dh_{35}}{dt} = 0$, the overlap of the system with pattern **a** and **b** does not change anymore (figure 4.16).

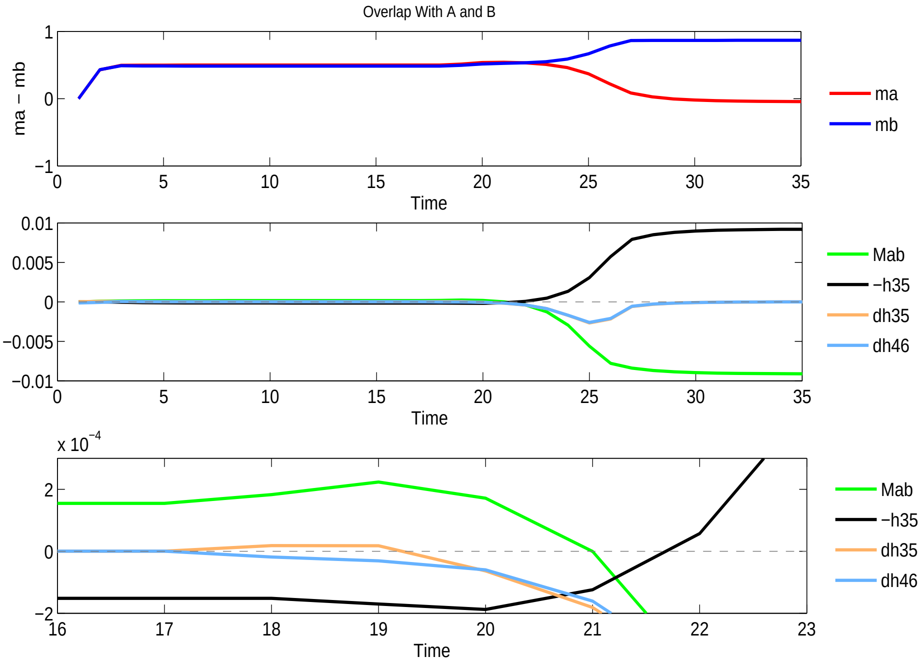


FIGURE 4.16: **The time course of m_{ab}, h_{35}, and $\frac{dh_{35}}{dt}$** A. the overlap with **a** and **b** is plotted. At time $t = 10$ the external input starts to fade away exponentially. The yellow line depicts $\frac{dh_{35}}{dt}$. Before the input starts to change, $\frac{dh_{35}}{dt}$ has touched the zero line

Critical value of gain At this point, by linearizing the equation for h_{35} one can find the critical condition under which $\frac{dh_{35}}{dt}$ grows starts to exponentially. This is the moment that the network leaves the unstable steady-state midway between the two attractor states and the difference $m_a - m_b$ grows exponentially, as well. Linearizing equation 4.7 gives a Jacobian of dimension 1 as

$$J_{h_{35}} = -1 + \beta g n_3 \tag{4.7}$$

This gives us the critical value of the gain g above which the network state becomes unstable (we call the term $\beta g n_3$, critical gain, g_c for the rest of the discussion). Now, we should find the value of g_c at the steady state which is a function of external input.

Steady-state equation as a function of the external input. The minimal steady-state equations, which are two, can be derived from a linear relation between h_1, h_3 and h_4, which is

$$\begin{aligned} h_3 &= \frac{h_1 + I_i n}{2} \\ h_4 &= \frac{h_1 - I_i n}{2} \end{aligned}$$

and the equations for gain and sparseness

$$\begin{aligned} g &= \frac{aN}{n_1(h_1 - Tr) + 2n_3(h_3 - Tr) + 2 < r_4 >} \\ a &= \frac{(n_1(h_1 - Tr) + 2n_3(h_3 - Tr) + 2 < r_4 >)^2}{N(n_1((h_1 - Tr)^2 + {\sigma_3}^2) + 2n_3((h_3 - Tr)^2 + {\sigma_3}^2) + 2 < {r_4}^2 >)} \end{aligned}$$

we remain with two equations for h_1 and sparseness in which there is also a non-algebraic expression for $< r_4 >$ and $< {r_4}^2 >$ (the gain can be absorbed into the equations, with two free parameters h_1 and Tr).

Eq_1: $-g\alpha(1 - 2a)(2n_1 + n_3)(h_1 + Tr) + h_1 + (g\alpha n_3 - 1)I_{in} + n_4 < r_4 > (1 - 2a) = 0$

Eq_2: $((n_1 + n_3)(h_1 - Tr) - n_3 I_{in} + 2n_4 < r_4 >)^2 =$
$Na(n_1((h_1 - Tr)^2 + \sigma_1^2) + 2n_3((h_3 - Tr)^2 + \sigma_3^2) + 2n_4 < r_4^2 >)$

Figure 4.18 shows the numerical solution of Eq_2, in violet, and Eq_1 in brown, for different values of input (at different times). Out of several solutions, only one is relevant to our network dynamics (the other gives a negative gain).
In figure 4.17, considering an exponentially decreasing form for the external input I_{in}, the values of the threshold and gain are plotted, for the simulation results (in black dashed line) and analytical calculation (in black solid line).

Different input time course - simulation results We have tested the network with different time scales of the external input. In figure 4.19, the top panel shows the time evolution of external input with 4 different exponentially decaying time course. The dots, superimposed on each curve, depict the moment at which g_c crosses zero (from negative values to positive). In the lower panel

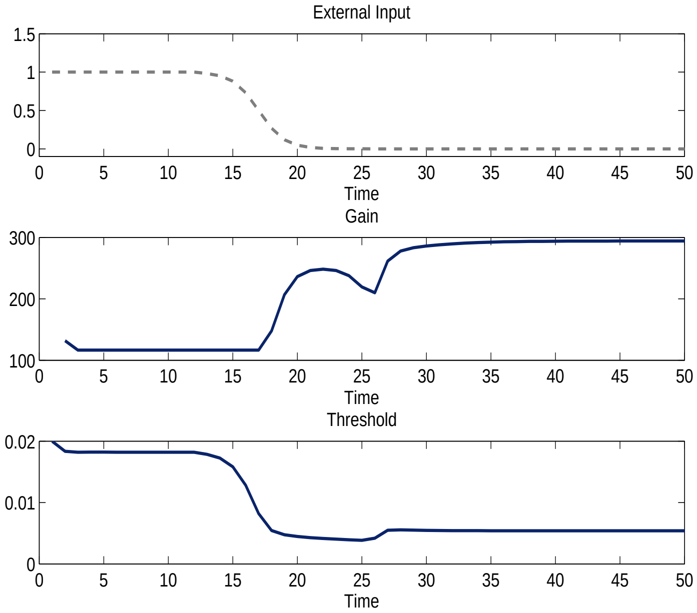


FIGURE 4.17:

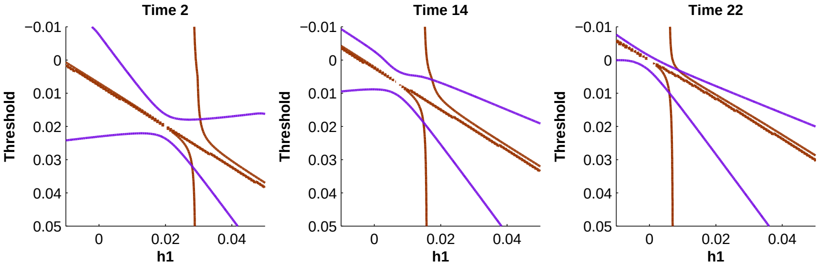


FIGURE 4.18: **Steady-state solutions**. The numerical solution of Eq_2, in violet, and Eq_1 in brown, for three different values of input (at three different times). Out of the two possible solutions, at each time, only one is relevant for us

the overlap with patterns **a** and **b** is shown (again the time for crossing to zero of g_c is marked). The predictive criterion based on equation 4.7 is well matched with the instant of bifurcation in the overlap values.

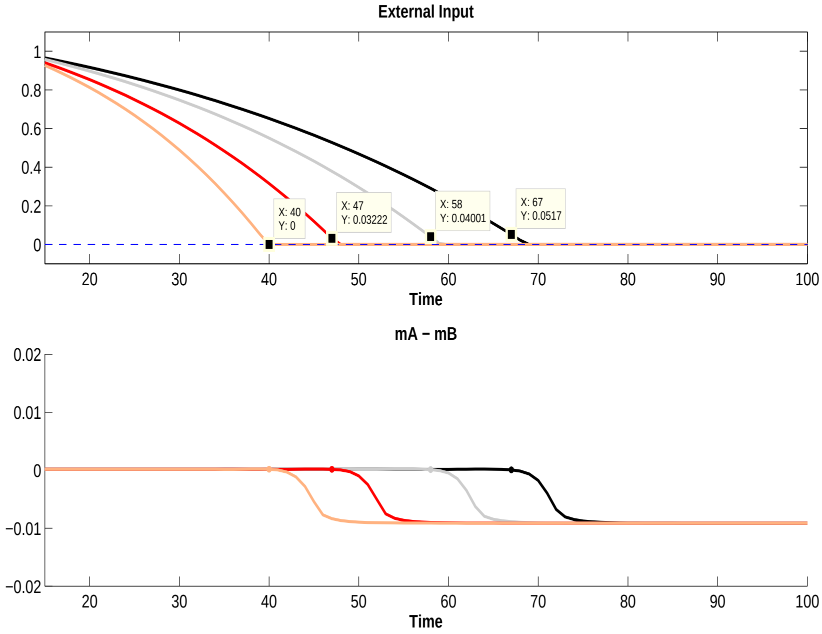


FIGURE 4.19: **Different input time scale**. The top panel shows the time course of external input with four different time constant. Each curve is marked with moment at which the g_c crosses the threshold for stability. The lower panel shows the difference overlap m_{ab} again marked with the critical time of g_c.

4.4.4.1 Including firing rate adaptation

To find the solution for the input dependent steady-state equations, we have ignored the firing rate adaptation. However, as long as the time course of firing rate adaptation is much slower than the time scale for transition to zero of the external input, this solution of the steady-state is valid. In this view, what adaptation does is to determine the state of the network after the external input subsides.

It should be noted that the network is subject to possible switches between the two attractors of **a** and **b**, before the external input subsides. This does depend on duration of input onset, its strength, and the time course of adaptation. We rather leave this subject open for the feature studies and thus will not discuss it more here.

4.4.4.2 Jacobian matrix of 8 x 8 system of differential equation

It is possible also to study the full system of 8 differential equations. In the appendix, we first linearized the system around the steady-state and then studied the properties of the Jacobian matrix to find the critical eigenvalues, whose real part can change sign depending on the level of external input. Out of all 8 eigenvalues, there are 4 $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = -1$ and one of the other three eigenvalues, $\lambda_6 = -1 + C_3 + C_4$, depends on the values of system parameters at the steady-state, and thus is function of external input.

Vanishing External Input: Change in the sign of λ_1 Here, we find the region in which all eigenvalues remain negative. Out of 8 eigenvalues, only the last three may change sign.

$$\lambda_6 \quad = \quad -1 + C_3 + C_4$$

For stability, we need to have $-1 + C_3 + C_4 > 0$. where
$C_3 = \frac{N}{Ca}\frac{\Phi({\rho_3}^*)}{\delta I_g}\bar{c}_3$
$C_4 = \frac{N}{Ca}\frac{\Phi({\rho_4}^*)}{\delta I_g}\bar{c}_4$
and
$\bar{c}_1 + \bar{c}_3 + \bar{c}_4 = \bar{c}_1 + \bar{c}_5 + \bar{c}_6 = a$

These steady-states are a function of *external input*. Thus, one can find directly the dependency of each λ on the level of *external input*, either analytically or numerically. Comparing the temporal evolution of firing rate for each of the 6 cell assemblies, with the temporal evolution of λ_6 gives the same critical value for the external input.

4.4.5 Different initial steady-states and different input values - simulation results

We have shown how a balanced steady-state -in which the network state is equally correlated with two attractors- can become unstable, due to a change in the level of external input. However, this analytical result is limited to one specific steady-state, and also to binary patterns. Thus, we run a network simulation to first verify the analytical result, and then see how stability of different steady-states are subject to changes in external input which can be either drown from a continuous or binary distribution. In the following simulations, whose properties are set as described in Section 3, first the two patterns **a** and **b** are stored in the recurrent weights of the second layer, according to the Hebbian learning rule. Afterward, during the test phase, the network receives intermediate morphs. The original patterns were combined into pairs, and 9 ”morphed” intermediate versions of each pair of patterns were set by gradually changing their correlation level with the two original patterns. This was achieved simply by taking one of the original patterns and setting the firing rate of a randomly chosen 10%, 20%, 90% of the input units to their firing rate in the second original pattern.

We hypothesized that in a system of interest which has been stable in a state with overlap m^a with pattern **a**, and m^b with pattern **b** (in an ideal case $m^b = 1 - m^a$), the external input signals a specific morph level (between **a** and **b**), with a particular strength with respect to recurrent inputs. Based on the type and the strength of external input, the network will either experience a bifurcation or stay in the same state, depending on the starting steady-states.

To simulate different steady-states, we set the initial condition of the network at different intermediate morphs, and then probed it by a fixed input, signaling the desired morphed pattern, for 100 time steps. In Fig. 4.20, different insets show the simulation results of such network probed by different morphed patterns, while each of them depicts the basin of attraction in the space of external input and initial condition; i.e. the abscissa is external input and ordinate shows different initial condition, measured in the level of overlap with pattern A (0 means $m^a = 1$ and 1 means $m^b = 1$), and the contour plot shows the difference in the overlap with pattern **a** at time-step 100 and the initial overlap with pattern **a**, $m^a(t = 100) - m^a(t = 0)$. We have repeated the same simulation, this time for binary patterns imposed to a one-layer recurrent network. The general outcome is similar to the simulation with graded patterns in a self-organized network, however, the basins of attractions are sharper.

4.5 Concluding remarks

The initial motivation in studying such model was to ascertain its dynamic behaviour, because the conditions under which a starting configuration evolves to a specific pattern define how such network may contribute in producing priming/aftereffect effects. Although the assumption of the initial steady-state at the edge between the two attractors is a simplified version of what is happening in a real experimental setup, since the model can be analyzed (at least in the zero-noise limit), it could serve as a convenient future playground for calculating stationary state properties of non-symmetric initial states, including fast adaptation as well.

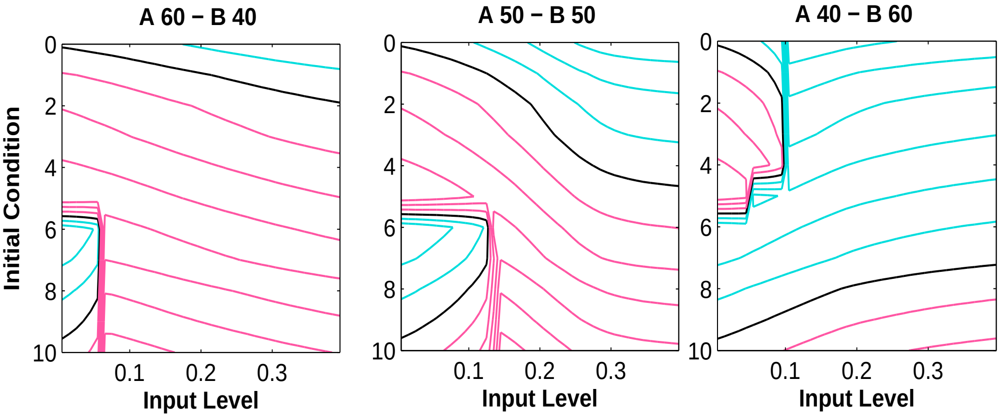


FIGURE 4.20: **Portrait of basin of attraction**. Different insets show the simulation results of the network probed by different morphed patterns: (a) more correlated with **a** than with **b**, (b) equally correlated with **a** and **b**, (c) more correlated with **a** than with **a**. In each plot, the abscissa is external input and the ordinate shows different initial condition, measured in the level of overlap with pattern **a** (0 means $m^b = 1$ and 1 means $m^a = 1$), and the contour plot shows the difference in the overlap with pattern **a** at time-step 100 and the initial overlap with pattern **a**, $m^a(t = 100) - m^a(t = 0)$. The red codes for any positive shift toward pattern **a**, whereas blue stands for the negative shifts.

Chapter 5

Attractor dynamics and Long Term Memory

In this chapter, we present an experimental study which was carried out in collaboration with Prof. Bharathi Jagadeesh and Yan Liu in the University of Washington, Seattle. In this experiment, aiming to explore the neural basis of attractor dynamics in visual cortex, we have recorded from IT cortex in two monkeys while the animal was presented with morphed stimuli. The neural data is suggestive of a possible contribution of attractor dynamics in producing categorical boundaries. This hypothesis, together with another alternative one (firing rate adaptation) was then tested, by simulating a simple autoassociative network model, and found to be consistent with observations.

The recording experiment was carried out by Yan Liu, and I did the data analysis and network simulation.

5.1 Experiment: convergence in neuroal activity of IT cortex, in response to morphed images

The inferior temporal (IT) cortex is thought to play an important role in visual categorization (Wilson and DeBauche 1981; Afraz et al. 2006). IT neurons can be selective for complex visual stimuli including people, places, and objects (Desimone et al. 1984; Allred et al. 2005). In some cases this selectivity corresponds more strongly to exemplar-specific than to category-specific information (Vogels 1999; Rolls et al. 1977; Freedman et al. 2003). However, IT neurons can also be sensitive to the features that distinguish categories and are influenced by experience (Sigala and Logothetis 2002). At the population level, neural responses may reflect performance in behavioral classification tasks (Vogels 1999; Allred and Jagadeesh 2007. Single IT neurons encode different kinds of information about visual stimuli in their temporal firing pattern, suggesting that the dynamics of visual responses may reflect different kinds of processing of a visual image (Sugase et al. 1999; Matsumoto et al. 2005). These results suggest that IT can represent both stimulus-specific information and categories. To further understand this dual representation, we recorded IT activity in monkeys performing a visual categorization task and examined the dynamic conversion of incoming information from complex visual stimuli into categories.

One way to extract relatively stable features from the flow of sensory information to form associations and categorize information is through the operation of attractor-based neural networks. In Chapter 3 and 4, we extensivly discussed the dynamics of an attractor network. In principle, attractor dynamics might also be expressed in IT cortex, where associative long-term visual memories are stored, to extract visual category information. However, it is unclear which aspects of IT activity might represent category boundaries and whether these boundary representations are interpretable in terms of the basins of an IT attractor network (Sakai and Miyashita 1991; Amit et al. 1997). Here, we present evidence of converging neural activity in macaque IT cortex representing the conversion of graded visual information into a category. We then simulate a local neural network to assess the possible relevance of attractor states and spike-rate adaptation to the observed neural dynamics. These simulations support a contribution by distributed local attractor networks, modulated by firing rate adaptation, to drive neural response convergence to reflect perceptually relevant categories.

5.1.1 Experimental procedure

The data in this study was recorded from two male adult monkeys (Macacca Mulatta). Neural recordings were targeted near the center of their recording chamber (Monkey L: 17 L, 17.5 A; Monkey G: 16 L, 17.5 A), in between the perirhinal sulcus and the anterior middle temporal sulcus. The sites might include cells from both TE and perirhinal cortex. However, the selection criterion for recording locations was the presence of cells that responded selectively to one of the 12 image pairs used in this study, and recording locations were altered until such selectivity was found. The experimenter usually found an apparently selective neuron (one included in the population presented in this report) after sampling one to three sites in a session. Recording procedures can be found in detail, in Appendix.

5.1.1.1 Stimuli

Images consisted of photographs of people, animals, natural and man-made scenes, and objects (Supplementary Fig. 1). All images were $90x90$ pixels, and were drawn from a variety of sources, including the World Wide Web, image databases, and personal photo libraries. Image pairs were organized prior to recording sessions into 12 pairs of stimuli. From these predefined lists of image pairs, selective neurons were found (see Recording Procedures) for a total of 12 unique image pairs used in the analysis. At the viewing distance used, stimuli were presented on a computer monitor with $800x600$ resolution (refresh rate $100Hz$), and images subtended 4°. Cells selective for each of the 12 image pairs were found.

5.1.1.2 Effective and Ineffective images

Based on the average response over trials during the sample presentation epoch, offset by a latency (i.e., over the 75- to 375-ms period) we assigned the image in the pair that provided a stronger response to be the "Eff" image, whereas the other was deemed the "Ineff" image. Because we recorded from multiple neurons with the same stimulus sets, either of the 2 images in a pair could serve as the Eff image during a particular recording session. Across the sample included in the study, each image in the pair was the Eff image in approximately half the sessions using that pair.

5.1.1.3 Image morphing and ranking

Each of the 12 pairs of images was morphed using MorphX, a freeware, open source program for morphing between 2 photographic images. We constructed 9 intermediate images in between the 2 original images, as described in Liu and Jagadeesh (2008); the images and their morph variants are presented in Figure 2 of Liu and Jagadeesh (2008). These 9 intermediate images, along with the 2 images in the pair were used as samples in the 2AFC-DMS task described above. The particular pair used in a recording session depended on observing selectivity for one of the images in the pair.

The morphing algorithm used by MorphX cannot be presumed to be linear; nevertheless we assigned a level to each morph variant corresponding to the ordering of each morph variant between the 2 original images from which they were morphed; There are 11 possible sample images (the 2 original images and 9 morph variants). The original image that produced a weaker response in the cell in a particular session (Ineff, as defined above) was assigned morph level 0; the original image that produced a stronger response in the experiment (Eff, as defined above) was assigned morph level 10. The 9 intermediate morph variants were assigned levels 1-9. Of these, morph variants 1-4 were closer to the Ineff image, and therefore, images 0-4 are collectively referred to as the Ineff morphs. Morph variants 6-9 were closer to the Eff image, and therefore, images 6-10 are considered Eff morphs. Morph variant level 5 was a priori defined as the midpoint of the morph continuum between the 2 images. These designations matched the behavioral reward contingencies, described below.

5.1.2 Behavioral tasks

5.1.2.1 Two-Alternative-Forced-Choice Delayed-Match-to-Sample

On each day, the monkey performed the 2AFC-DMS task (Liu and Jagadeesh 2008) with 2 sample images and 9 morph variants of those images. In each trial, one sample image (or one of its morph variants) was presented, followed by a delay and then followed by a pair of choice images ("choice array"). The monkey's task was to saccade to the image in the choice that most resembled the sample image. An example image pair and associated trials are illustrated in Figure 1. In each trial a red fixation spot ($0.3°x0.3°$) appeared at the center of the monitor, and was the cue for the trial to begin. After the monkey acquired fixation, there was a variable delay (250-500 ms) before the onset of the sample image. The sample was presented for 320 ms. After a delay period (700-1100 ms), the choice array (which consisted of both sample images from which the morph variants were created, the Eff and Ineff image) was presented. The choice images were presented 5° up (or down) and to the left of the fixation spot. Location of individual choice images was randomized between the 2 positions (up and down), so the monkey could not determine the location of correct saccade before choice array onset. The different morph variants were presented as samples in random order, until 5-17 trials were recorded for each image.

When the original image pairs were presented as the sample, the monkey's task was to pick the identical sample image from the choice array (Fig. 5.1.a). When the morph variants were presented, the monkey's task was to classify the morphed sample as one image in the choice pair by judging the similarity between the morphed image and the original images (which were presented as choices). The monkey was rewarded for picking image 0 (the Ineff) image when morph variants 0-4 were presented as the sample image and the monkey was rewarded for picking image 10 (the Eff) when morph variants 6-10 were presented. For morph variant 5, the monkey was rewarded randomly, resulting in 50 % reward for either choice.

The monkeys were trained over apperiod of 6 months before the recording sessions began, with the 12 pairs of images and their morphed exemplars, as described in Liu and Jagadeesh (2008), so that both the morphed images and image pairs were not novel to the animal before the beginning of the recording session.

5.1.3 Analysis of neural data

Neurons were included in the population for analysis based on post hoc analysis of selectivity for the image pair selected during the recording session (averaged responses over the sample presentation period to the effective image, Eff, are to be at least 110% of those to Ineff), yielding a neural population of 132 experiments. Four experimental sessions were also discarded because of poor performance by the monkey, resulting in a neural population of 128 experiments. Choosing different populations of cells (all 154) or only cells which pass a selectivity criterion (P value between the 2 original images < 0.01) does not change the results shown in the neural data figures.

Average spike rates were calculated by aligning action potentials to the onset of the sample stimulus presentation, and analyzing the data from 100 ms before the onset of the image to the period 1000 ms after the onset of the image. The peristimulus time histogram (PSTH) for each cell was calculated by averaging the rate functions across the repeated trials of presentation of the same stimulus. The population PSTH was calculated by averaging the PSTHs across the set of 128 selective cells. All completed trials were included in the analyses; trials were excluded if the monkey did not make a choice from the 2 possible choice stimuli. Both correct and incorrect trials were included.

All the tests of significance were performed on firing rate functions FR(t). FR(t) was calculated for each neuron, for each sample image, by averaging firing rate across multiple presentations of each sample in overlapping time bins (also called epochs) of 100 ms, shifted in time steps of 10 ms (Zoccolan et al. 2007). This procedure smoothes the data. The average FR(t) was plotted at the middle of the 100-ms bin. Therefore, average responses at time 0 consist of the average of responses from -50 to 50 ms after stimulus onset. To calculate the dependence of the neural responses on morph level, we performed a regression analysis for each cell for each epoch. We regressed the spike rate in an epoch against the morph level, separately for Eff and Ineff images (Fig. 5.4). To compare the response to Eff with its 4 variants and also Ineff with the other 4 ineffective variants (Fig. 5.5), we applied an unbalanced 2-way ANOVA. In this ANOVA, we treated the cell as one factor (128 level), and stimulus as the second factor (2 level: Eff vs. 9, Eff vs. 8, Eff vs. 7, Eff vs. 6, Ineff vs. 1, Ineff vs. 2, Ineff vs. 3, Ineff vs. 4). Morphs 2 levels apart were also compared using an unbalanced one-way ANOVA, considering again ”stimulus” and ”cell” as 2 factors with 2 levels (2 vs. image Ineff, 4 vs. 2, 6 vs. 4, 8 vs. 6 and Eff vs. 8) and 128 levels, respectively (Fig. 5.5).

5.1.4 Experimental results: convergence in neural activity in response to Eff stimuli

We recorded the responses of 154 IT cells in 2 macaque monkeys while the monkeys performed a 2AFC-DMS task. In a previous report, we compared the discrimination capacity of single neurons calculated in a fixed response epoch for morphed photographic images to behavior with those same images during the sessions in which the neurons were recorded (Liu and Jagadeesh 2008). In this report, we examined the changing dynamics of neural responses during a fixed presentation of a static images morphed between 2 exemplars. The subset of cells in which one of the 2 choice images produced a response at least 10% greater than the other (n = 128) are presented in this analysis. Data are combined for the 2 monkeys, and no detectable differences between the 2 monkeys were found.

The behavior in the task was linear for intermediate morphs but essentially categorical for the morphs most similar to the choice images. Figure 5.1 shows the proportion of trials in which the monkey chose the effective image (Eff) by making a saccade to it, across all sessions and all stimuli. The trend is linear in the central region, morph levels 2 to 8, but levels off at the extremes and their nearest neighbors, morph levels Ineff and 1 and morph levels 9 and Eff. The central region can in fact be fitted by a 1-parameter sigmoid (df = 6, chi2 = 0.43) but is even better fit by a straight line (df = 6, chi2 = 0.33). Is a similar pattern evident in the responses of individual IT units?

Most cells were also modulated by the degree of morph. Immediately after stimulus onset (100 - 200 ms after stimulus onset) the response of 93/128 cells decreased as the response was morphed away form the Eff image, and the response of 100/128 cells increased as the response was morphed away from the Ineff image, as tested by the slope of the linear regression of the responses for each cell. The pattern of modulation differed among individual neurons, however. Six example cells are shown in Figure 5.2.a,b. By definition, the average response to the Eff stimulus (solid line) was greater than the average response to the Ineff stimulus (dashed line) (Fig. 5.2.a). Most cells responses increased systematically between the Eff and Ineff image with the mean response to the mid-morph stimulus lying somewhere between the 2 extremes (Fig. 5.2.b). Eff images were defined on the basis of the response during the stimulus epoch (75 - 375 ms after stimulus onset). Immediately after stimulus onset (100 - 200 ms), 124/128 cells had bigger responses to the Eff image (level 10) than the Ineff image (level 0), replicating the response difference based on the longer epoch in which the Eff and Ineff image were defined. In the epoch immediately after the stimulus onset, 99/128 cells had smaller responses to the middle morph image (level 5) than to the Eff image (level 10); 106/124 cells had bigger responses to the middle morph image (level 5) than to the Ineff image (level 0) The responses of a smaller number (n = 35) of individual cells to the intermediate morphs, on the other hand, did not vary linearly along the morphing dimensions and did not increase monotonically with morph level (Fig. 5.2a,b, bottom right). The range of firing rates across the morphs and the time course of the responses was also variable, from 1020 Hz for some cells, and up to 120 Hz for others.

Classification ability and neural response Classification ability depended linearly on morph level for intermediate morph levels but not morph levels close to either familiar images, which were classified nearly perfectly (Fig. 5.1.b). A hallmark of this behavior might be reflected in the neural responses, if neural responses were also linearly dependent on morph level, except for those morph exemplars that were similar to the Eff or Ineff image. Some neurons did appear to follow this pattern, whereas others did not. One cell shown in Figure 5.2.b produced responses that roughly follow the pattern seen in the behavior (Fig. 5.2.b, top left), whereas others produced linear responses for all morph levels (Fig. 5.2.b, top right, bottom left and middle).

Symmetry in firing rate differences but not in firing rate values In order to further examine the relationship between these neural responses and the morph level of the stimulus, we calculated population averages across the neurons (n = 128). Because the behavioral response is symmetric around morph level 5 (Fig. 5.1.b), we initially took such symmetry for granted and compared average responses to morphed images ”equidistant” to morph level 5 by calculating the difference between them (Kreiman et al. 2000; Allred and Jagadeesh 2007; Liu and Jagadeesh 2008). However, averaged across the population, firing rate differences did not replicate the plateaus shown in the behavior for morph levels close to the Eff or Ineff images (Fig. 5.3.a). Instead, average firing rate differences decreased smoothly, almost linearly, with decreasing distance along the morph continuum, throughout the response to the sample stimulus (morph level main factor, 50 - 550 ms ($P < 0.02$), nonparametric 2-way ANOVA (Friedman test)).
Symmetry in responses to Eff and Ineff morphs is not preordained, however: subtracting the response to Ineff images from the response to Eff images might obscure the time course of the separate responses to Eff and to Ineff images. If either the Eff or Ineff responses were strongly dependent on morph level, the difference between the firing rates might mask the lack of dependence on morph level of the other images. Image Eff, moreover, had been selected from the pool of image pairs for being visibly effective for that particular cell, across thee group of images used, all of which were generically effective for some IT cells; although the ”ineffective” image, Ineff, produced a smaller response, but was not necessarily ineffective in driving the cell. Frequently, cells responded to ”ineffective” images producing responses substantially higher than the baseline response. Therefore, the apparent linear trend in Figure 5.3a could result from a strong quasi-linear dependence of either the Eff or Ineff images on morph level, masking the other half of the dependence.

Asymmetric convergence in neural activity The asymmetry between responses to Eff morphs and Ineff morphs is visible when responses to each individual morph level are plotted separately (Fig. 5.3c). The Ineff morph responses were linearly dependent on the morph level throughout and after the sample presentation nearly until the responses return to baseline. The Eff morph responses, in contrast, are linearly scaled as a function of morph level only for a brief time at the first peak of the response, centered around 120-ms poststimulus onset. By 200 ms post-stimulus (average response in the 150- to 250-ms epoch) the dependence of Eff morph responses was decreasing, and in the linear regression as a function of morph level, the slope decreases more rapidly than the one describing the dependence of Ineff morph responses on morph level. There is a second, lower response peak around 270 ms, where levels Eff, 9 and 8 are together but significantly above levels 7 and 6, and at 320 ms, at the end of the sample presentation time, all morph levels 6 to 10 are within the 95% confidence interval of each other. The response to morph level 5, which was not consistently classified as either stimulus choice and was randomly rewarded is different from the response to both the Ineff and Eff variants until 700-ms poststimulus, late in the delay period. There are 3 behaviorally defined groups in the morph continuum, the Ineff group, which must be classified as the Ineff choice, the Eff group, which must be classified as the Eff choice, and the image corresponding to morph level 5, which belongs to neither Eff nor Ineff group, and can be classified as either Eff or Ineff, with random reward for each possible choice. These 3 groups remain distinct until at least 700 ms after stimulus onset; the Eff group of stimuli, the morph level 5 (middle morph stimulus), and the Ineff group of stimuli, even as the responses to the individual images in the Eff group become indistinguishable.

The flattening of the linear relationship with respect to stimulus morph level can be seen in Figure 5.3.d, where the average firing rates to the 11 morph variants are shown, over 4, 100-ms time periods. The data for Eff and Ineff morphs are fit with separate lines. The slope of the linear fit for the Eff morphs gradually drops off compared with the slope for Ineff images. For the first 2 time windows (100 - 200 ms and 200-300 ms) the slopes are not significantly different from each other ($P = 0.85$ and $P = 0.20$, for the 2 windows respectively), but are both significantly different from zero (Eff: t-test, $P = 0.02$ and $P = 0.01$, Ineff 0.02 and 0.01, for the 2 time windows, respectively). During the later epoch (400 - 500 ms) the slope of the linear fit for the Eff morphs was not significantly different from zero (ttest, $P = 0.34$ 400 - 500 ms) and is significantly different from the slopes for Ineff morphs ($P = 0.01$). Thus across the entire population of neural data, unlike the behavioral data, the response ”plateau” appears to extend over the whole Eff range, and does not extend over the Ineff range. Note, however, that the notion of a response plateau is an oversimplification, which does not really describe the response of individual cells (see below). Morph level is a main factor affecting the Eff responses from 70 - 220 ms after stimulus onset (unbalanced one-way ANOVA shows that for Eff stimuli, the $P < 0.05$), whereas the Ineff morphs responses remain significantly different from each other for the entire sample presentation and into the delay period (70- to 590-ms poststimulus onset, $P < 0.05$). Responses to the Eff images, as a group, remain significantly above those to Ineff images (2-way ANOVA, $P < 0.05$) until 900 ms, when responses to both Eff and Ineff images are back at spontaneous level. The similarity of the firing rates for the Eff image and its 4 nearing morphs could be a hallmark of the morphs having been attracted to the basin of attraction of image Eff. These data show that subtracting the response to Eff images, in Figure 5.3.a, had obscured the time course of the convergence among responses to Eff images. This can be interpreted, presumably, as an indication that there is no convergence of neural responses to Ineff images, at least on average across the cells in our dataset, whereas there is, on average, a convergence of neural responses to Eff images.

Is the population average a faithful reperesentation of most cells? In going from single neurons to the population average, it is important to take into account the potential differences among responses of individual cells (Fig. 5.2). Is the population average a faithful representation of most cells or does it reflect the behavior of a few highly active cells? To address this concern, we applied the same analysis used for the population average in Figure 5.3.d to each individual cell. We fit with a line the firing rate of individual neural response as a function of morph level for 4 different response epochs of 100-200 ms, 200 - 300 ms, 300 - 400 ms, and 400-500 ms. In each epoch the linear regression was applied separately for Eff and Ineff images, giving a fit slope for the 2 subgroups of stimuli. Slope is expressed in units of spikes/second/morph level and reports how well the response to the Eff and Ineff images was modulated as a function of morph level. Figure 5.4 shows scatter plots of the slopes in the late time window (400 - 500 ms) with respect to the early one (100 - 200 ms), for Eff (Fig. 5.4.a) and Ineff (Fig. 5.4.b) morphs, respectively, for each individual cell in the population. Across the population of neurons, the slope is significantly higher in the early epoch than in the late epoch (sign test, $P < 0.0001$) for the Eff images (Fig. 5.4.a). Most of the points lie below the diagonal line indicating equal slopes. This effect is not found for Ineff morphs, for which individual neurons are uniformly distributed around the diagonal (Fig. 5.4.b, sign test, $P = 0.1329$). Furthermore, slopes for both the Eff and Ineff morphs are significantly different from zero in the early time window (sign test, $P < 0.0001$, population significance), whereas in the late time window, only slopes for the Ineff morphs are significantly greater than 0 (sign test, $P < 0.0001$). Figure 5.4.c shows time course of averaged slope (over all cells) for Eff (red curve) and Ineff (blue curve) morph stimuli. In the time bin 380 - 480 ms after stimulus onset, Eff responses no longer depend on the morph level of the individual stimulus (t-test, difference from 0, $P > 0.05$). At those same time periods, Ineff responses still depend significantly on the morph level (t-test, difference from 0, $P < 0.0001$) and Eff and Ineff response slopes are different from each other (paired t-test, $P < 0.01$). Ineff slopes remain significantly different from zero until the 700- to 800-ms time bin, when they are no longer significantly different from 0 (t-test, difference from 0, $P > 0.05$). Thus, the pattern of response dynamics seen in the population average in Figure 5.3.c,d is present in the individual cells (Fig. 5.4.a-c). Both the individual cells responses and the average population show that response to the Eff morphs and the Ineff morphs depends on morph level in the period immediately following the onset of the stimulus. Over time, however, the responses evolve so that neural responses to different Eff images all converge to similar values. The responses to Ineff variants remain separated, however, and the response remains dependent on the morph level.

Regression analysis only for middle morphs The response to the Eff and Ineff images (morph levels 0 and 10) were used to classify the 2 images, raising the possibility that these images might skew the regression analysis. Therefore, we performed the linear regression shown in Figure 5.4 for each cell, after first eliminating the Eff and Ineff images from the regression (i.e., morph levels 0 and 10). The analysis on this limited data set confirms the analysis shown in Figure 5.4. In the time bin 430530 ms after stimulus onset, the responses to Eff morphs no longer depend on the morph level of the individual stimulus (t-test, difference from 0, $P > 0.05$). At those same time periods, the responses to Ineff morphs still depend significantly on the morph level (t-test, difference from 0, $P < 0.0001$). The Ineff slopes remain significantly different from zero until the 660- to 770-ms time bin, when they are no longer significantly different from 0 (t-test, difference from 0, $P > 0.05$).

The simple regressions used in this analysis do not completely represent the patterns seen in individual cells. Among several alternative analyses, one may fit regression lines only to parts of the entire morph range. The behavioral data raises the expectation that convergence might be expected only for morph levels 8-10 (the original, and the 2 variants close to it), suggesting the possibility that only some of the stimuli for which a particular behavioral classification was required consistently converge to the same node. Stimuli closer to the response boundary may not always converge to the same node (different stimuli may behave differently, and the same stimuli may behavior differently in different trials or different sessions). To address this question, linear regression can be performed with data for morphs 0 - 2 and 8 - 10, corresponding to the ”plateaus” seen in the behavioral data, corresponding to stimuli for which the behavior was roughly similarly (each was classified consistently, respectively, as the Ineff or Eff image). This regression analysis changes the time course of dependence on morph level for the Eff and Ineff images. Eff responses converge faster, resulting in zero slopes sooner after the onset of the image. In addition, the Eff slopes actually turn negative, with greater morph levels resulting in a slightly smaller response at response onsets. Ineff slopes remain significantly different from zero until late in the delay period, with the difference between Eff and Ineff slopes increasing.

Part of the variability among cells may be due to the diversity of image pairs used in the experiment, but the significant trends shown in Figure 5.4 are replicated for individual images, with the similarity to the population increasing with the number of recorded units for each image.

"Gradual" convergence Neural responses evolve, or change dynamically over the course of the presentation of the constant, unchanging stimulus to show convergence to different response levels. Does the dynamics of this response evolution depend on the morph level? To address this question, we performed an analysis of variance to examine how separated responses to different morphed images remained as a function of time (Fig. 5.5). The times of convergence to the Eff (or Ineff) stimulus for each morph variant are shown in Figure 5.5a. This graph shows the time at which the responses to the morph variants were no longer significantly different form the Eff (or Ineff) stimulus (P value of ANOVA < 0.01 Supplementary Fig. 5.5.a,b). The response to the Ineff variants remains different from the Ineff until long after the stimulus presentation, until approximately 600 ms after stimulus onset. Eff variants, on the other hand, take progressively shorter times to converge to the Eff response as the morph level increases (indicating higher similarity to the Eff stimulus).

Figure 5.5.a illustrates the timing of convergence of the response to each morph toward that to the Eff or Ineff image. We can also assess convergence between pairs of other morph levels. The analysis comparing how quickly different pairs of morph levels converge is shown in Figure 5.5.b, using the same analysis of variance used in Figure 5.6.a, but comparing other pairs of morphed stimuli (Fig. 5.5.c). We compared the response to morphs 20% apart, by running an unbalanced one-way ANOVA analysis for responses to morphs 2 versus image Ineff, 4 versus 2, 6 versus 4, 8 versus 6 and Eff versus 8. The data in Figure 5.6.b shows that the Ineff morphs remained separated for durations longer than the presentation of the sample stimulus (greater than 500 ms) (Fig. 5.6.b, left 2 points). The pair of morphed stimuli that lie across the behaviorally defined classification border (morphs 4 and 6, remain separated for over 700 ms). The Eff morphs, on the other hand, converged to one another at durations close to sample duration, or even shorter (Fig. 5.5.b, right 2 points).

Improvement in behavioral performance, in correlation with change in degree of convergence Although the monkeys were trained before the beginning of the recording session, improvements can be seen in behavioral performance over the course of the multiple recording sessions in the study. Behavioral performance was significantly better during the second half of the recording sessions compared with the first indicating that the monkeys performance continued to improve over the course of the sessions (Fig. 5.7.a, paired ttest, $P < 0.01$ for morph levels 1 - 4 and 6 - 9). Performance for the 2 original images was stable over the course of the recording sessions ($P = 0.52$). An improvement in behavioral performance might suggest that neural representations were also changing over the course of the study. To examine whether the dynamics of the response as shown in Figure 5.4 changed over the course of the study, we separately examined the neural response dependence on morph level (shown for the entire data set in Fig. 5.4.c) for the first and second half of the sessions (Fig. 5.7.b,c). The results suggest that the dynamics of the response convergence (the pattern of results shown in Fig. 5.4) changed over the course of the recording sessions. The difference in slope for Eff stimuli at stimulus onset compared with stimulus offset was significant only during the 2nd half of the sessions (n = 62 first half, n = 66 second half, Eff slope at 100-200 ms compared with slope at 400-500 ms, P = 0.0160, early sessions, P = 0.0959). Furthermore, the difference between slopes for Eff and Ineff images was significantly different only in the second half of the sessions (slope at 400-500 ms, compare Eff vs. Ineff slopes, P = 0.0105, early sessions P = 0.2374). The trends were compatible with storage of the patterns (as described in the network model) improving over sessions.

5.2 Theory: simulation results of an autoassociative memory model

What is the neural mechanism underlying this convergence? Can the observed convergence express the outcome of visual signal processing within IT cortex, or must it be driven by afferent inputs that have already converged before reaching IT, or top-down, by signals from more advanced processing stages (Bar et al. 2006)? If the convergence can result from local processing within IT, what is the contribution of network interactions, that is, of dynamical attractor states determined by the structure of recurrent connections in IT? Or, could the convergence reflect, in part, simple firing rate decay of individual IT neurons, expressed as a gradual decay, rather than network dynamics? Firing rate adaptation effects are conceptually quite different from those arising out of genuine network interactions, but may in practice be difficult to distinguish. If firing rate adaptation progressively suppresses the responses to Eff and to its closest morphs $9, 8, ...$ effectively squashing them onto each other, the functional consequences may resemble the convergence posited to result, in network models that do not include firing rate adaptation, from synaptically mediated attractor dynamics. We addressed these possibilities by simulating a simple local network model of cortical activity, with and without firing rate decay, to assess the relative contribution of attractor dynamics and of adaptation. Note that this simulation does not rule out all forms of adaptation, and a sufficiently complicated form might replicate response dynamics in this individual data set, even if the simple form does not.

Our model simulates a single hypothetical local network within the IT cortex. The network includes an input station, simulating afferent inputs from earlier visual areas, which projects its activity to an output layer, simulating an IT patch, through sparse FF connections. The units in the output layer receive both FF and recurrent connections at random, with unstructured baseline weights (see Fig. 5.8.a and the initial section of this chapter).

When $a_{out} = 0.2$, theoretical calculations indicate that the storage capacity of the model is around 0.2-0.4 times the number C_{rc} of recurrent connections per neuron (in our simulations C_{rc} varies between 200 to 999), Thus, although finite size effects make the notion of storage capacity less well defined for a network that is small, it is expected to be able to retrieve on the order of 100 - 200 patterns.To assess the storage capacity of our model, for each value of p we gave the trained network a full cue, corresponding to one of the stored patterns, and after 80 synchronous updates we measured the final overlap of the network state with the presented pattern. If the final overlap is larger than 0.8, retrieval was deemed successful. Repeating this process for 4 different seeds of the random number generator and p different patterns, the maximum value of p at which success still reaches 50% is around 250 patterns, higher than but consistent with the theoretical expectation.

We ran simulations in which we stored 20 or 160 patterns, which correspond to conditions where the network is far below its storage capacity and near its storage capacity.

Different units in the first layer receive inputs of variable duration, drawn at random, for each unit from a logarithmic distribution. Inputs were not removed sharply, but gradually, with a linear decay to zero. The distribution of input offset latency is shown in Fig. 5.8.b (top right panel). The output of units in this layer is a step-like function, active at a certain level for a specific duration (Fig. 5.8.b-top left), with a gradual transition to zero. The average activity across all input units, for one pattern, is shown in Fig. 5.8.a (top left panel).

Once either $p = 20$ or 160 original patterns had been stored, the network was tested with intermediate morphs. Original patterns were then combined into pairs, and 9 ”morphed” intermediate versions of each pair of patterns were set by gradually changing their correlation level with the two original patterns. This was achieved simply by taking one of the original patterns and setting the firing rate of a randomly chosen 10%, 20%, 90% of the input units to their firing rate in the second original pattern. To simulate experimental procedures, for each output unit we assign a pair of patterns to which the unit has a different response during stimulus onset (a pair of ”effective” and ”ineffective” stimuli). In some simulations, intermediate morphs were produced not between two stored patterns, but between one pattern that was stored and one other pattern which was not stored in the network.

To test the network, we measured the time evolution of all output units, over 80 time steps, after presenting a morphed pattern (or the original images from which the patterns were morphed) in the input layer.

5.2.1 Adaptation alone can not replicate experimental results - simulation 1

In simulation 1 we assessed the effect of firing frequency adaptation on responses, modeled as a decay term, a linear decay as a function of the recent activity of the cell. Can convergence result from such firing rate adaptation over time?

The simulated network is a simple approximation of inputs and recurrent connections to a patch of cortex. It consists of 2500 units that receive FF projections from an input layer containing another 2500 units (Fig. 5.8.a). Each unit in the (output) patch receives approximately 750 FF connections from the input layer, and 500 recurrent connections from other units in the output patch (Fig. 5.8.a). The connections are assigned at random and, because there is no storage of activity patterns in this first version of simulation, weights are not modified to reflect memory storage. The weights are instead set to a random value and then normalized to generate an approximately exponential distribution of initial weights onto each unit (see Chapter 4). Once a pattern is imposed on the input layer, the activity circulates in the network for 80 simulation time steps, each corresponding to ca 12.5 ms (in total ca 1 s. Fig. 5.8.b top left panel) (Treves 2004). The details of the network, including signals receiving by each unit and their activity functions, together with the default values for parameters used in the model are reported in the first section of Chapter 4.

The input patterns simulate the hypothetical input produced by 20 unrelated visual images, and morph variants of them. Inputs consisted of 20 uncorrelated input patterns combined into pairs. Nine "morphed" intermediate versions of each pair of patterns were set by gradually changing the correlation of one original pattern with the other pattern. To simulate experimental procedures, for each output unit we assign a pair of patterns to which the unit has a different response during stimulus onset (a pair of "Eff" and "Ineff" stimuli). Then, the response of individual output units to these selected patterns and to their morphs was monitored. The duration of the inputs was variable, to simulate the potentially variable duration of different input streams. Input offset time for each unit is driven from a sharp logarithmic distribution, peaked at time ca. 300 ms (Fig. 5.8.b top right panel).

In the first simulation, we examined the effect of firing rate decay. Firing rate decay was modeled (see Chapter 4, section 1) by subtracting from the sum of FF inputs and recurrent connections to each output unit a fraction of its own recent output activity. The trace of its "recent" activity is calculated with a convolution kernel, expressed as a difference of 2 exponentials, with inverse time constants $\beta_1 = 2\beta_2 = 0.2$ (time steps)-1. This form of firing rate decay applies to all output units from the second time step, for all the succeeding 100 time steps.

The average network dynamics shows that linear decay of responses over time produced by adaptation does not produce a response convergence similar to that observed (Fig. 5.8.c). In the early phase of the simulated neural response, during which the network is mostly driven by afferent inputs, average network responses to all morphing levels are well separated. Then, as the afferents are gradually removed, the population response to all the morph levels decrease, but there is no tendency for the Eff responses to group or squeeze together (Fig. 5.3.b).

Convergence might require ”memories” to be stored within the network. Unlike the simple local network above, with no stored patterns, in autoassociative networks memories can be stored as stable network activity states, called attractors (Hopfield 1982; Treves and Rolls 1992; Amit 1994; Brunel 1996). A stored pattern, may be then be retrieved when a noisy or occluded version of it (a partial cue) is provided as input. This ability is due to the formation of dynamical attractors that capture network activity, if an input is sufficiently close to one of the patterns stored. The formation of the attractor landscape is achieved by creating overlapping patterns of synaptic modifications adhering to the Hebbian paradigm (Hebb 1949) such that each synapse is involved in the storage of multiple related memories. Inevitably, this common synaptic representation implies interactions between memories stored in the same network. The putative presence of long-term memories in IT and the observation of increasing stimulus selectivity, through learning, in individual neurons (Sakai and Miyashita 1991) suggest that attractor dynamics may be plausibly expressed in IT cortex, where visual object memories are likely stored, and may drive the extraction of visual category information. Therefore, in our next simulation, we considered an autoassociative network, with memory patterns stored in RC connections through a realistic synaptic modification mechanism.

5.2.2 Attractor dynamics, a potential mechanism underlying behavioral categorization - simulation 2 and 3

Does the addition of stored patterns produce convergence in the network? The properties of the network were identical to those used in simulation 2; the only addition is that in simulation 2 the recurrent weights are modified to store memory patterns before testing the response of the network to patterns and their morphs. First we produced 200 uncorrelated patterns, using a common truncated logarithmic distribution, from which the firing rate of each unit is driven independently (see Chapter 4, section 1). Then we stored in the network P = 160 of these patterns, by modifying the RC weights of the output layer with a ”Hebbian” learning rule (see Chapter 4, section 1). Either 2 stored patterns or a stored pattern and a novel one, from other 40 unused patterns, are then combined into pairs, and 9 ”morphed” intermediate versions of each pair of patterns are presented as the input. As indicated in the Chapter 4 (first section), this results in a network that is below its storage capacity.

Simulation 2 - morph between two stored patterns In the first simulation with stored representations, we analyzed network activity in response to stimuli obtained by morphing between 20 sets of 2 stored patterns (40 out of 160 stored patterns). Single units show a variety of behaviors in response to morphs (Fig. 5.9.a), resembling the diversity observed with real cells. Averaged population activity qualitatively mirrors the convergence seen in the population average of IT neural responses, for Eff morphs (Fig. 5.3.c). Figure 9b shows the mean responses in the simulation, after averaging over all units. In the first phase, whereas sufficiently many units in the network still receive afferent inputs, all the morphs are well separated. Rather abruptly, as the average ratio of RC to FF activation increases beyond a critical level, network activity was determined by the attractors embedded in the RCs. The responses to the Eff morphs are all attracted to the full memory pattern Eff. However, in this simulation, unlike in the data, the responses to the Ineff morphs are also attracted to a basin of attraction for the Ineff morphs, a feature not seen in the data (Fig. 5.3.c). Adding adaptation to this simulation (as described for the first simulation), does not produce the lack of convergence for Ineff morphs seen in the neural data (Fig. 5.9.c).

Simulation 3 - morph between one stored and one non-stored pattern Testing with morphs between 2 stored patterns corresponds to the assumption that both images used in the real experiment had memory representations in the local network that includes the particular unit being recorded from. The experimental procedure for picking image pairs for each cell, on the other hand, might introduce a bias, where the Eff image is more likely to be represented by a ”neural assembly” to which the unit belongs, than the Ineff image, which might have its own representation elsewhere over IT (Haxby et al. 2001; Kiani et al. 2007). To model this situation, we tested network activity in response to morphs obtained between an (Eff) stored pattern and one that had not been stored. Figure 9d shows that the convergence of the mean responses, again averaged over all units, is now limited to the Eff patterns. We found that the morph level, above which responses converge, is strongly dependent on the storage load. With low load (Fig. 5.10, 20 stored patterns), all morphs converge to the stored pattern Eff, whereas when many patterns are stored (Fig. 5.9.d, 160 stored patterns), the basin of attraction effectively shrinks, and only the morphs closer to the Eff pattern showed convergence.

We then combined stored attractors with firing rate decay over time. In this version of simulation, we used the same network as in Simulation 2, with 160 stored patterns, and applied firing rate adaptation, modeled as in simulation 1. We again monitored the firing activity of output units in response to intermediate morphs produced between one stored pattern and one novel pattern. In this way we could assess the effect of response decay on the simulation in Figure 5.9.d (or Fig. 5.9.b). Introducing this form of firing rate adaptation did not change the qualitative behavior of the network (Fig. 5.9.e or Fig. 5.9.c): in the first phase of the response, when the network is mostly driven by afferent inputs, different morphs are linearly separated; in the second, ”memory” phase, when afferent inputs have been largely removed, the network activity converges to either one or 2 attractor states, depending on whether both patterns are stored (Fig. 5.9.c) or only one is stored (Fig. 5.9.e). Adaptation however introduces a third phase, in that after some time it brings the network out of the current attractor state and makes single units fire in a somewhat erratic manner to the different morphs and brings all responses close together. This disorderly behavior imitates the population average of neural responses (Fig. 5.3.c). The simulation also shows a crossover between the responses to Eff and to morphs 9, 8 and those to morphs 7, 6, which is an effect of adaptation (Fig. 5.9.e,c). This crossover is also visible in the experimental data, in that around 450-ms poststimulus onset the average firing rate of the response to stimulus Eff drops below the responses to the rest of the Eff images (Fig. 5.3.c).

The simulations can thus replicate the linear dependence of the response on morph level at stimulus onset (Fig. 5.3.c,d) for both Eff and Ineff images, and the selective convergence of the Eff responses, whereas the Ineff morphs remain separated long after the stimulus has been turned off. The simulation best matches the data if we assume that 1) many patterns are stored in the network (close to storage capacity), that 2) the sampled cells belong to the representation of only one of the 2 images that are morphed into one another, and if we add some degree of response adaptation. Even with these characteristics, the simulation cannot replicate another feature of the data: the gradual convergence over time among Eff responses (Fig. 5.3.c). In addition, compared with the real data the onset transient is less peaked in the simulation, and the delay activity returned to a common value for both sets of morph sooner (Fig. 5.9.e) than in the real data (Fig. 5.3.c).

5.3 Concluding remarks

In our model, the intermediate morph stimuli do not contribute to synaptic modifications during learning. In other words, there is no attractor individually assigned to each morph level. In this sense our model is different from those discussing "attractor collapse", in which each intermediate morph patterns contribute equally to synaptic modification (Blumenfeld et al. 2006). Stimuli and task demands, in other experimental studies, may also differ in many ways from our experimental setting. In one particular study (Blumenfeld et al.), a fundamental difference is that they used faces as visual stimuli, which allows them to generate a whole morphing stream equally meaningful for the subjects of each individual morph image has a specific identity and is perceptually recognizable as a face, and the subject should report whether each morph face is a Friend or non-Friend. In our study, instead, the intermediate morph images are rather ambiguous and nonmeaningful, and the monkey is not asked to recognize each morph independently. It seems more justified to assume a synaptic plasticity effect for perceptually meaningful faces than for nonmeaningful images, which in our experiment which must be classified in 2 groups by the monkey, based on their similarity to either Eff or Ineff. Even if we take synaptic plasticity into account for intermediate morphs and one could easily implement it in the current model, having the morphs stored in the network by changing the weights with a β factor an order of magnitude weaker than the original patternswe still believe that the storage of the 2 end point images (Eff and Ineff) would dominate the ensuing attractor dynamics.

The simulations show that a very simple model of an IT patch, with memory attractors stored on recurrent connections by associative plasticity, responds with a convergent dynamics similar to that seen in the data. Furthermore, the simulations suggest that such local networks may be loaded with memories close to their storage capacity, which would be consistent with the expectation of an efficient utilization of the available memory resources in the synaptic weights (Braitenburg and Schuz 1998).

Although the network simulation suggests that attractor dynamics could explain the dynamics of the responses seen in IT, the simulation cannot address the question of whether this attractor network plays out in IT itself, or if it is inherited from another visual area with all of the dynamics preserved. For example, some simulations of learned categories suggest that an interactive feedback between IT and prefrontal cortex might provide initial information separating the categories. Over time, this feedback information changes synaptic weights in IT, enhancing the representation of features that differentiate between the categories (Sigala and Logothetis 2002; Sigala 2004; Szabo et al. 2006). Alternatively stimulus frequency has been proposed as a method for adjusting synaptic weights to produce categorical boundaries (Rosenthal et al. 2001). These computations could play a role in the dynamics reproduced here with attractor networks, and could precede the attractor dynamics demonstrated here.

5.3.1 Summary

IT We recorded from individual neurons in IT cortex while monkeys performed a classification task on morphed visual images. We report here a population of IT neurons whose responses evolve gradually over the course of a trial,

- first representing parametrically the morphed image,
- and later converging to represent one of the 2 categories.
- The convergence of IT activity from a stimulus-based representation to a category-based representation was **asymmetric**, in that only responses to the morphed images that resemble the effective stimulus for an individual cell converge, whereas responses to morphed images that resemble the ineffective stimulus remain segregated by morph level.

Model network An asymmetric convergence may result from multiple mechanisms, of course. We have tried to assess 2 possible underlying mechanisms:

- A gradual decay of the response over time, With our first simulation, we could rule out the possibility that the convergence was the result of simple linear decay of neural responses. A linear decay of responses had an equivalent effect at each individual morph level, and did not produce a change from the linear dependence visible at stimulus onset (simulation 1, Fig. 5.8.c)
- Attractor dynamics in the local recurrent networks. the operation of a simple attractor network produced qualitatively similar convergence to that observed in the neural data, allowing a more detailed interpretation of the observations.

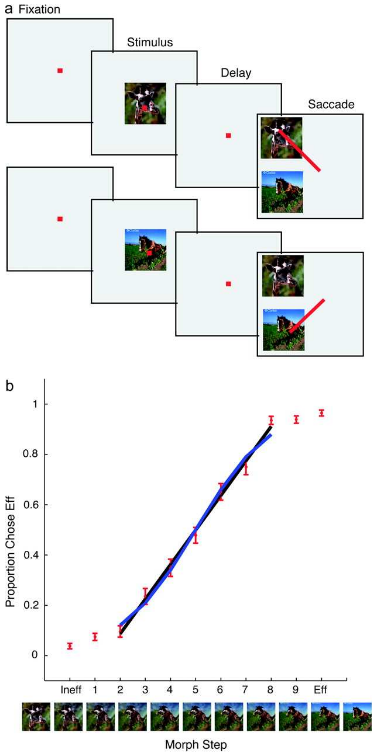


FIGURE 5.1: **Behavioral paradigm** (a) Classification task. After the monkey achieved fixation on a fixation point, a sample, chosen at random among the 9 morphed images or the pair of photographs from which the morphs were made, was presented for 320 ms. Then, after a delay, the photographs appeared together as possible choices (targets). The monkey's task was to pick the target choice that more closely resembled the sample, and make a saccade to it. (b) Behavioral performance. The data are plotted as the proportion of times the monkey chose one of the images (the "effective" image for the cell (see section 5.1.1), or Eff) of the 2 original photographs, as a function of the different samples. The trend is linear in the central region between morphs 2 and 8, but performance levels off at the extremes and their nearest neighbors, images Ineff (0) to 2 and 8 to Eff (10). The data are fit with a sigmoid (blue line) and a line (black line). Error bars are standard errors of the mean across different sessions. Images are examples used in one session, where the giraffe was the Ineff image, and the horse the Eff image.

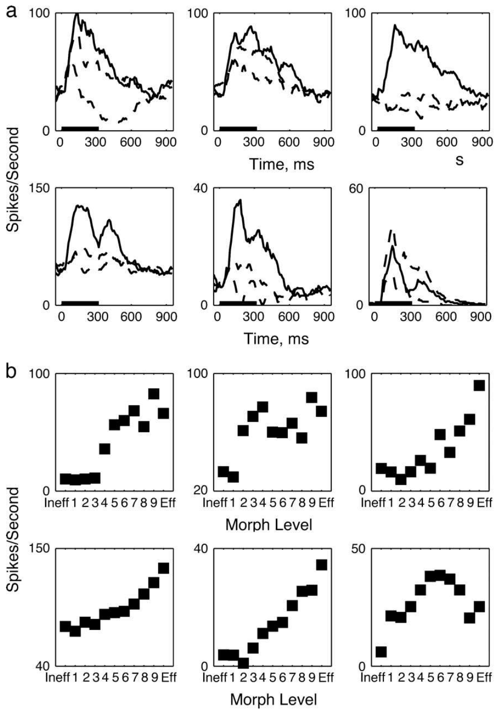


FIGURE 5.2: **Single cell activity** Single cells show a variety of neural responses to different morphed images. (a) Response time course of 6 different cells to the 2 end point images Eff (black) and Ineff (black dashed) and to the midlevel morph (blue dashed). (b) Firing rates to the Eff and Ineff and 9 morph variants computed over time period 100-200 ms. The black horizontal line shows the period of sample presentation (320 ms).

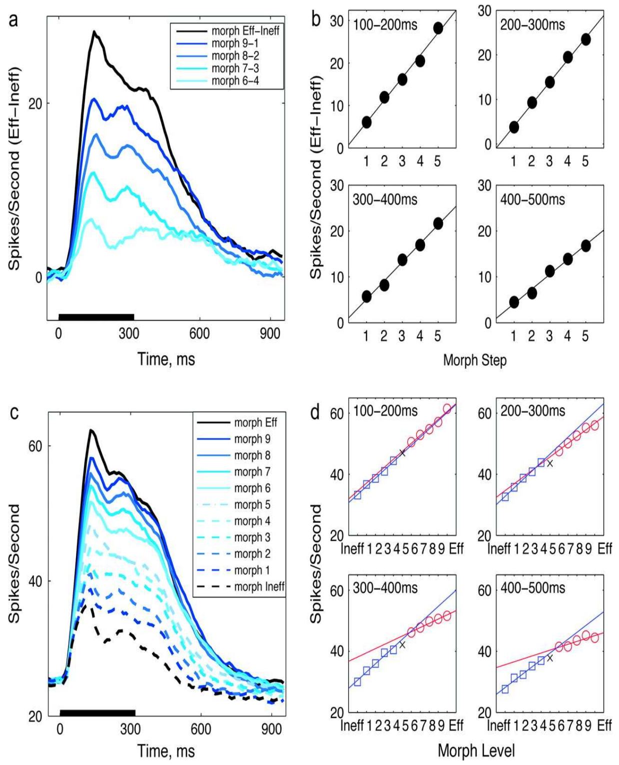


FIGURE 5.3: **Time course of population responses to morphed images**. (a) Time course of average differences between the responses to images Eff and Ineff (black) and to morphs successively different from the images (Eff and Ineff) between which they were morphed, averaged across the population of cells n=128. In both panels, as in Figure 5.2 spike counts are binned into 100 ms bins, which slide every 10 ms from stimulus onset, and are averaged across 1520 trials per unit and morph step. (b) Mean response difference between Eff and Ineff morphs in successive 100 ms epochs after sample onset. (c) Time course of firing rate to Eff and Ineff, and each morph variant, as in (a). (d) Mean response to Eff and Ineff image and morph variants in successive 100 ms epochs as a function of morph level.

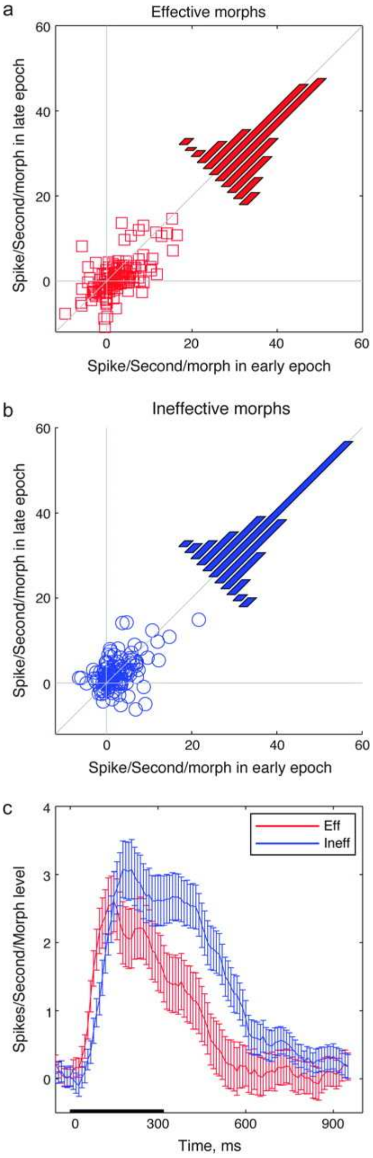


FIGURE 5.4: **Linear regression for individual neurons**. (a, b) Scatter plots of slope of linear regression (in Spikes/Second/Morph Level) in late versus early epoch (100-200 ms vs. 400 - 500 ms after sample onset) for each individual cell in the population. Histograms are the distributions of slopes for individual cells in early and late epochs. n = 128 experiments. (a) Slope of Eff image and Eff morphs (Eff, 6 - 9), (b) slope of Ineff image, and Ineff morphs (Ineff, 1 - 4). (c) Time course of slope (across population) as a function of time. One hundred millisecond bins, stepped 10 ms.

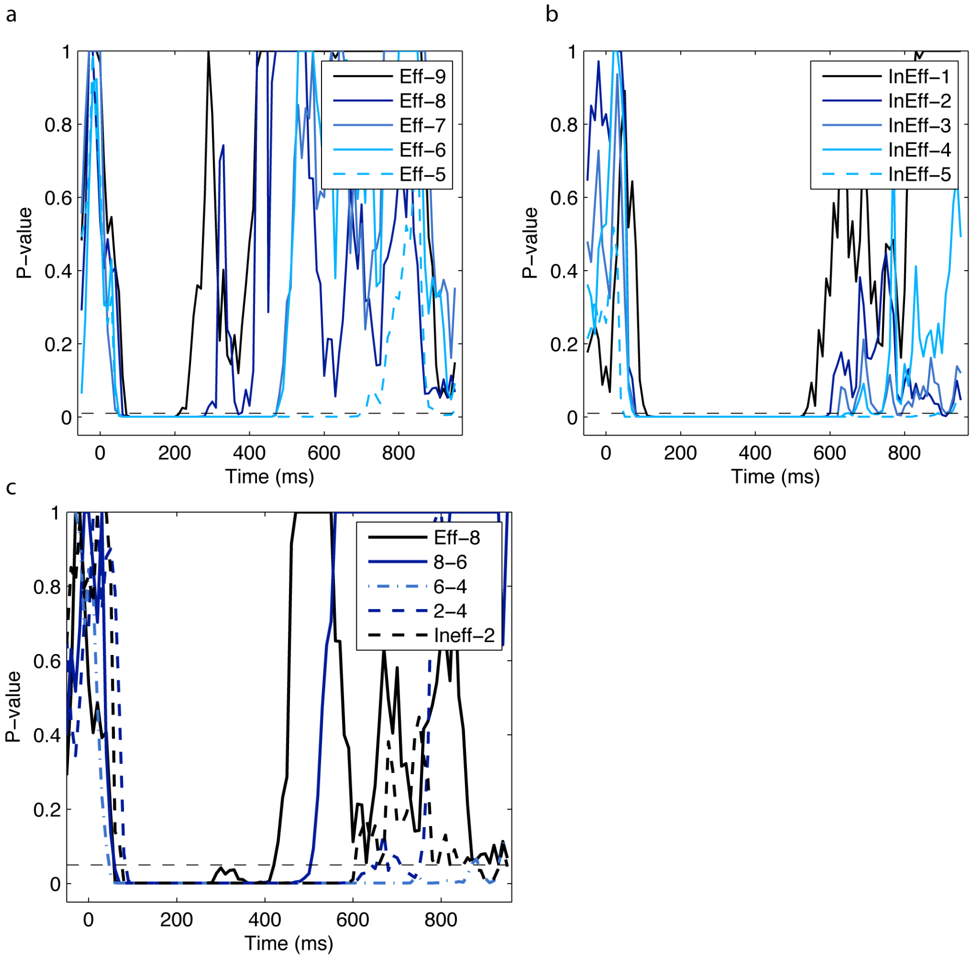


FIGURE 5.5: **Within-category vs. cross-category** p-values of an unbalanced one-way ANOVA, in logarithmic scale, comparing the mean of responses to a) image Ineff with its four neighbor morphs, b) image Eff with 4 ineffective morphs, and c) variants that are 2 morph levels apart, Eff vs. 8, 8 vs. 6, 6 vs. 4, 4 vs. 2 and 2 vs. Ineff. Gray dashed line presents the significance level 0.01.

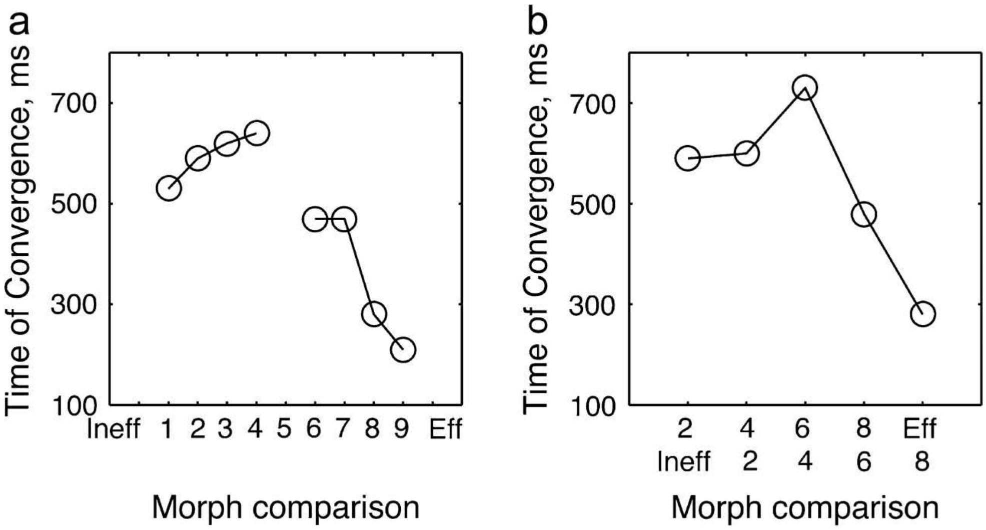


Figure 5.6: **Gradual convergence** (a) The times of convergence to the Eff (or Ineff) response for each morph variant. This graph shows the time at which the responses to the morph variants were first no longer significantly different form those to Eff (or Ineff) stimulus (taken from the ANOVA, when P > 0.01). (b) The same ANOVA-based analysis as in a, but comparing the response to morphs 2 level apart, that is, to morph 2 versus Ineff, 4 versus 2, 6 versus 4, 8 versus 6, and Eff versus 8.

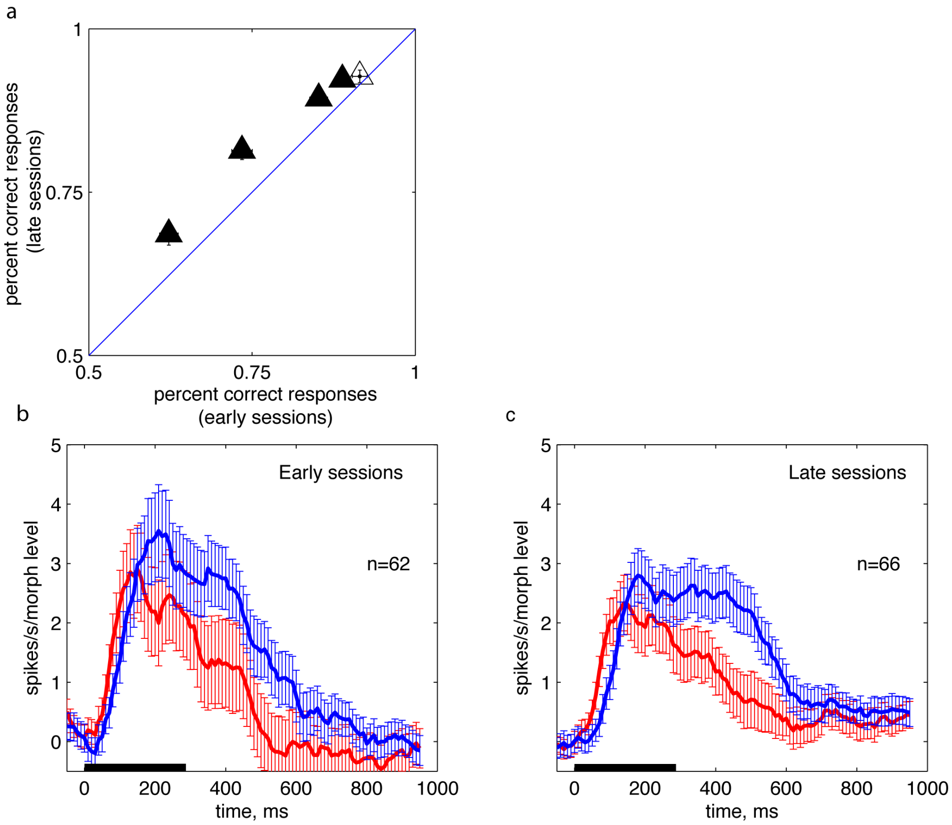


FIGURE 5.7: **Behavioral improvement, correlated with degree of convergence** and a) Proportion correct responses for Eff and Ineff images in early sessions (n=62) and in late sessions (n=66). Performance is significantly different for the filled points. Error bars, over sessions, are smaller than the points. Behavior at four different morph levels showed significant increases in performance (filled triangles, t-test, p¡0.05). b-c) Slope of linear regression to Eff and Ineff images as a function of time. Eff images (images 6-10), red line; Ineff images (images 1-4), blue lines. Error bars are the standard error of the mean acorss sessions. Black horizontal line is the stimulus duration. a) Early sessions, n=62 b)late sessions, n=66.

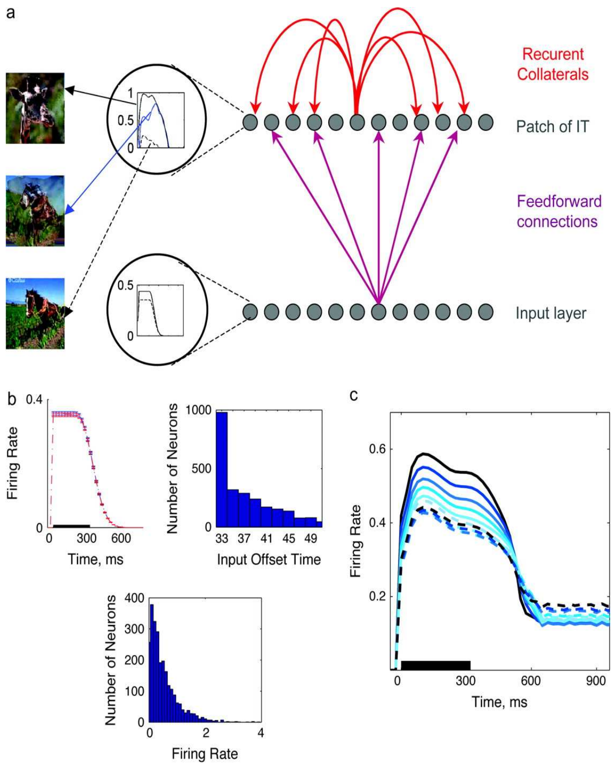


Figure 5.8: **Simulation setup** (a) Schematic view of the simulated network, including an input layer, which projects its activity to an output layer (recurrent connections) through sparse FF connections. Different units in first layer receive input, generated using a common truncated logarithmic distribution (b, bottom middle), with durations drawn at random from a logarithmic distribution (b, top right); one example is shown in circle at bottom. The units in this layer are active at a certain level for a specific duration, with a gradual transition to zero (example shown in circle at top). (b) Simulated input activity pattern: the average activity across all of the input units, for one pattern, is shown in top left; distribution of input offset times in top right, and firing rates in bottom middle. (c) Average network activity, in response to morphs obtained between 2 nonstored patterns, including a linear decay of firing frequency. Because there are no stored patterns, no attractors appear in this simulation.

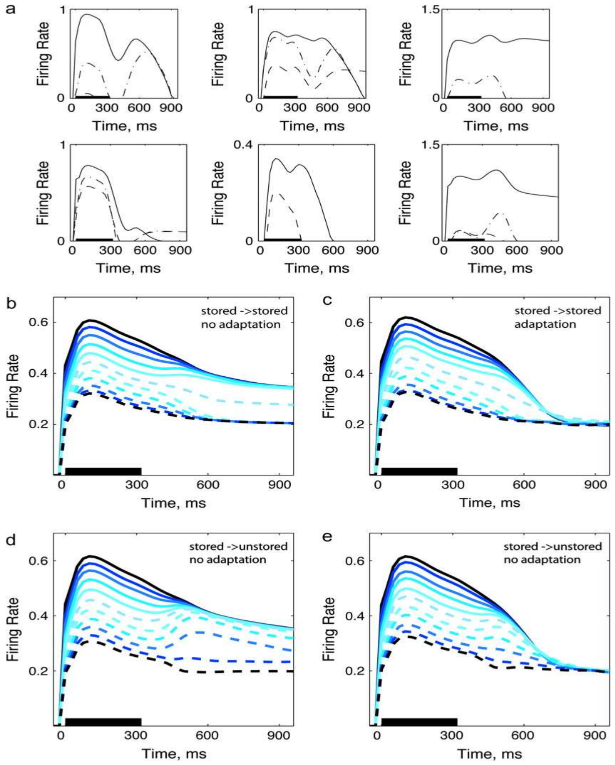


FIGURE 5.9: **Sample of single unit activities in the simulated network** (a) Sample of single units from the model; Eff: solid, Ineff: dashed, mid-morph: dashed-dots (b) Simulation: 160 stored patterns, tested with morphs between stored patterns, no adaptation. (c) Simulation: 160 stored patterns, tested with morphs between stored patterns, adaptation. (d) Simulation: 160 stored patterns, tested with morphs between one stored and one unstored pattern, no adaptation. (e) Simulation: 160 stored patterns, tested with morphs between one stored and one unstored pattern, adaptation.

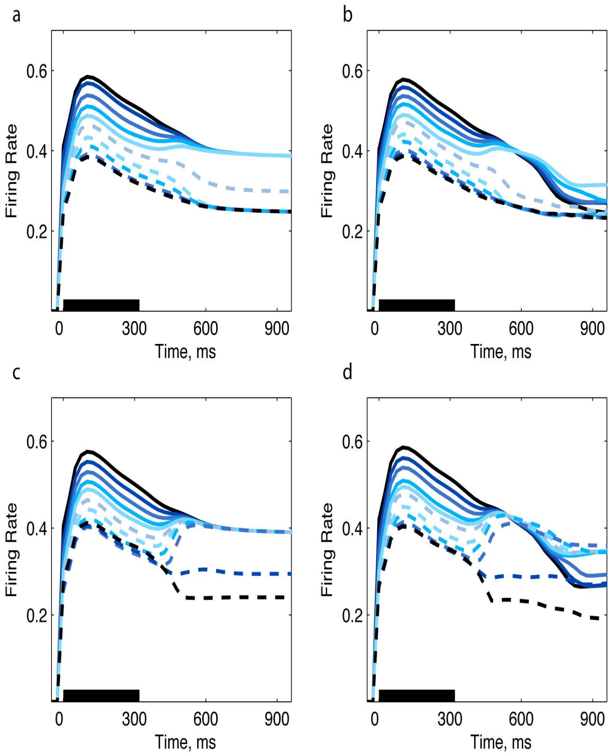


FIGURE 5.10: **Simulation results** Output of simulation with 20 stored patterns rather than 160, as in Figure 3.9 a) tested with morphs between stored patterns, no adaptation. b) tested with morphs between stored patterns, adaptation. c) tested with morphs between one stored and one unstored pattern, no adaptation. d) tested with morphs between one stored and one unstored pattern, adaptation

Chapter 6

How Attractor Dynamics could shed light on short term effects of visual experience

In chapters 4 and 5 we dealt with a situation in which the "memory faculty", either in the simple simulated network, or in the animal, has to retrieve the relevant memory once presented with the ambiguous stimulus (target), without being influenced by the preceding experience of an adapter (prime). However, in more ecological cases, no sensory stimulus is isolated; rather, it can only be properly interpreted in light of the stimuli that surround it in space and time. The perception of, and neurophysiological responses to, a target input depend strongly on both its spatial context (what surrounds a given object or feature) and its temporal context (what has been observed in the recent past) (Schwartz et al. 2007). Perceptual history (events that have been recently experienced), builds up expectations and can help resolve the interpretation of ambiguous or degraded visual stimuli (Dolan et al. 1997).

Here, we would like to extend the model and see whether a generic model that once could replicate the electrophysiologycal observation, can be used to model more complicated behavioural phenomena like "priming" and "adaptation aftereffectss". We try to show that elaborations of such models that have been successful in reproducing the electrophysiological data in monkeys can also reproduce many types of behavioral findings on priming experiments in humans.

Main questions The main issue we would like to explore is to understand the required properties of an associative network that can produce either priming or adaptation aftereffectss, and in particular the observed shift from one to another.

6.1 Modified autoassociative memory network to model priming

Regarding the dynamics of a perceptual process, a theoretical framework is used that provides us with proper tools to fully pin down the contribution of possible dynamical factors. A novel analytical approach allows studying the dynamics of networks of threshold-linear model neurons, reciprocally connected through synaptically activated conductances. This model includes, as a necessary ingredient of the relevant computational mechanism, a simple feature of pyramidal cell biophysics: firing rate adaptation. Analysis of such model yields the attractor states of the network and the full spectrum of time constants of the transients associated with different steady stated. Studying the transient dynamics, affected by previous activity of the network, could shed light on the possible contribution of attractor dynamics to perceptual boundary shifts. The results that will be discussed in this chapter show that firing rate adaptation plays the main role to produce adaptation aftereffects in our network, without which one could only get priming effect, if any. Moreover, the observed shift from adaptation aftereffects to priming in human psychophysics (Daelli and Treves 2009, accepted), is only possible in a network endowed with attractor states. The relative duration of target length and adaptation time scale is one of the crucial terms determining the dynamics. The strength of recurrent connection with respect to feedforward inputs is another relevant parameter. In the following sections, we study the effect of short term memory in different associative networks that vary based on a) the number of stored pattern (zero, two, or several); b) the degree of firing rate adaptation; and c) the type of the "adapter", whether it is one of the two end points of the morph continuum, (**Full-adapter**), or one of the intermediate morphed patterns (**Morph-adapter**). We study the effect of adapter length (T_a), delay duration (T_d), and target length (T_t).

6.1.1 Is "Attractor Dynamics" necessary?

Do aftereffectss need attractor formation? It has always been debated whether adaptation-aftereffects/priming tap into low-level (say, memory-independent) processes or they rely on some sort of long-term memory. Several recent demonstrations using visual adaptation have revealed high-level aftereffectss for complex patterns including faces. While traditional aftereffectss involve perceptual distortion of simple attributes such as orientation or color that are processed early in the visual cortical hierarchy, face adaptation affects perceived identity and expression, which are thought to be products of higher-order processing (Webster and MacLin 1999). And, unlike most simple aftereffectss, those involving faces are robust to changes in scale, position and orientation between the adapting and test stimuli. These results can not simply be explained based on a combination of adaptation to low-level features.

Here we would like to test our model of autoassociative network with zero-stored patterns to see whether a memory-less network, with distributed patterns over the units that are endowed with firing rate decay can show either a positive or negative bias in responding to a target morphed pattern, after the network has experienced an adapter stimulus.

We test the "with" adapter conditions when the network is presented with morphs generated by two unlearned patterns. With mid values of adaptation there is a minimal tendency toward adaptation aftereffects, mostly during the delay between the adapter and the target. If the delay is short enough, the effect extends over the target, as well. Increasing the amplitude of the adaptive current, produces a larger effect (Fig. 6.1). From figure 6.1, one can conclude that

- Adaptation aftereffects, in a memory less network, need a very strong firing rate adaptation magnitude.
- A network primed with A, shows suppressed overlap with A, without any increase in overlap with B.
- Under any choice of "adapter", "delay", and "target" length, the network did not show any positive bias toward the adapter (priming).

6.1.2 Attractor dynamics and *priming* vs. *adaptation-aftereffects*

In chapter 4, we have analyzed the network dynamics underlying the time evolution of trajectory toward an attractor state, due to a change in external input. Here, we would are interested in understanding the effect of short term memory, induced by our adapter shown before presenting the local memory network with an ambiguous target. From what we found in Chapter 4, it can be hypothesized that what is important, in the response to an ambiguous target, is the network state at the moment that the external input is about to subside below a critical value. Given the initial state of the network and the external input, the network trajectory in the basin of attraction can be predicted.

Initial state and "adapter". Whether this initial state, in the moment at which the "target" fades away, is closer to A or B depends on type of the "adapter" input and possibly on events occurring between the "adapter" and the "target", and during the "target" itself. In our model, there are two possible mechanisms by which the presence of an "adapter" similar to A could lead to the final attraction of the morphed target to the attractor A (priming):

- **Type A: Perceptual bias**. After presenting the network with adapter A, the network settles down in the attractor A and keeps the same configuration till the external input signals the intermediate "target" partially correlated with A. At this moment, if the external input is not too strong to move the network out of the basin of attraction of A, the network will be attracted again to A after the target subside –if the target duration is short with respect to the time scale of firing rate changes (Fig. 6.2.a). We discuss the relevance of adaptation and target duration later.

- **Type B: "Double (or multiple) to an aftereffects**. In another scenario, the final shift of the network state in favor of the attractor related to the adapter, may come from multiple switches from one attractor to another, due to the presence of firing rate adaptation.

In the next part we elaborate on these two mechanisms and discuss the portion of parameter space in which any of these two may unfold.

6.1.2.1 Type A priming: perceptual bias

In this scenario, the initial state of the network at the time of target presentation directly reflects the bias acquired with the experience of the "adapter".

Low adaptation - low noise region. In a network, in which the mean firing rate is kept fixed, once the network is settled into stable fixed point attractor, if adaptation is weak, then it will remain there, unless the external input does signal a configuration different from that attractor state (Fig. 6.2.a). Thus, with weak adaptation, the priming effect emerges roughly independent of the delay length between the "adapter" and the "target", and with zero firing rate adaptation, is incapable of producing adaptation aftereffects. One should notice that if the updating is asynchronous (and the network has several attractor states), or at each time step of updating there is strong enough fast-noise, then, the probability that the network is pulled out of the initial attractor is no longer negligible.

The ratio of the external input to the recurrent collateral inputs Here again the balance between the feedforward inputs and the recurrent currents plays an important role in determining the network trajectory in its phase space, in response to an external input (Section 4.3.3; Parga and Rolls 1998). One of the conditions under which this type of priming has the possibility to occur, is when the ratio of the recurrent collateral inputs to the external input is too high, to let the network to be derived by external feedforward inputs. In other words, if we decrease this ratio (and thus weaken the attractor state), then during the target presentation the network state is moved toward a configuration very close to the one that is signaled by the external input (the ambiguous state). In conclusion, if external input is much more stronger than the recurrent inputs, then this type of priming will not happen. This issue asks for future quantitative study to find the parameter region in which type A priming is possible.

The effect of a "mask" As we have discussed in the first chapter, many electrophysiological studies have suggested that backward masking influences neural processing either by stopping neurons in IT cortex from receiving feedforward signals, or by disrupting reentrant processing within recurrent connections. The aim of the simulation is to investigate the effect of backward masking, expressed as a combination of these two mechanisms, in the formation of basins of attraction. In chapter 4 we have discussed the condition under which a network, at its steady-state and in the presence of external input, leaves the current state and follows a trajectory toward the nearest attractor state. We derived threshold value for the external input, below which such steady-state becomes unstable. This critical value of the input is also relevant for studying the effect of backward masking: i.e. if a mask appears before that critical value is reached, the convergence to the basin of attraction is prevented, by the change in initial condition. This simulation result suggests that backward masking, if the mask is presented before the recurrent connections dominate the dynamics, can stop the network from settling down into the attractor state close to the adapter. The model then indicates that in a network with sufficient firing rate adaptation, in the presence of a mask, indeed, the only memory of the past, is through firing rate adaptation and thus a) no place is left for type A priming, and b) adaptation aftereffects become possible. In the text below, we further describe these different possibilities.

- The adaptation-less network only shows **priming** effect
- But adaptation opens the door to richer behaviour

6.1.2.2 Type B Priming: adapting to an aftereffects

We saw that in very low adaptation regime it is only possible to observe a positive shift toward the adapter and adaptation aftereffects cannot emerge from the network dynamics, and on the other hand, in the presence of a mask, even this positive shift can be eliminated. With a stronger firing rate adaptation, however, such a picture changes: not only adaptation aftereffects are possible (6.2.b), but also priming can occur with an additional mechanism: multiple transitions between A and B due to ongoing firing rate adaptation (6.2.c).

To discuss these effects, it is useful to consider the influence of firing rate adaption on the network state during a) the "adapter, T_a b) the delay between the adapter and target, T_d and c) the target, T_t. We start with an investigation of the delay period.

Spontaneous transitions during the delay period. One of the features, emerging from an associative network with multiple stored patterns, is the sequential transitions between correlated attractors, when firing rate adaptation is strong enough. We further discuss this particular phenomenon later (in section 6.1.4). What interests us here is the time course of the network overlap with the two patterns A and B, when the external input is actually removed i.e. during the delay period. Figure 6.3 shows the time course of network overlap with patterns A and B, in black and red, respectively, and with the other possible stored patterns in other colors. The external input (in gray) signals pattern A and gradually fades away. As one can see, in the case with only two stored patterns, the network, after removal of the external input, is sequentially alternating between A and B (with a frequency that depends on the adaptation time scale). This happens, because in our model, the mean activity of the network is kept fixed (by regulating the gain and the threshold; see section 4.1.3). Thus, starting from a situation in which the network is in attractor A, gradually the units active in such state adapt and the network instead moves toward its only other stable state, attractor B. This sequential retrieval of patterns A and B in the absence of external input is an artifact, in fact, and we avoid it by storing several patterns in the network (Fig. 6.3 b,c). In this way, the network has potentially several fixed point attractors and during the delay period, the possibility that the network moves toward any of them is the same.

The influence of varying "adapter", "delay", and "target" duration. We first define an index to measure the distance between two network states at any time: when the external input as adapter is pattern A or pattern B:

$$\mathcal{R} = m^a_a(t) - m^a_b(t);$$

which is the network overlap with pattern A when A is the adapter ($m^a_a(t)$), minus the network overlap with A when B is the adapter ($m^a_b(t)$). Then, we can summarize the effects in a relevant parameter space –adaptation, adapter duration, target duration, delay duration.

Increasing delay period In figure 6.4 we have simulated a network that receives the **Full-adapter** for 10 time steps, and after a variable delay period, the external input signals the middle morphed pattern between A and B, for $T_t = 6$ (left) or $T_t = 20$ (right). In this figure the "prime index", defined in 6.1.2.2, is plotted at 5 time steps after the removal of the target ($m^a_a(T_a + T_d + T_t + 5) - m^a_b(T_a + T_d + T_t + 5)$). Red shows a positive shift toward A (priming), and blue shows the opposite (aftereffects). This simulation shows that in a portion of the phase space:

- If the target does not last for a long time, increasing the delay priod produces Type B priming; i.e. while short delays lead to a negative bias, increasing the delay allows the network to make a second transition and therefore produces priming.

- If the target last for long time, the network shows and opposite behaviour: short gaps between the adapter and the target result in priming, whereas long ones produces adaptation aftereffects.

Increasing target length In figure 6.5 we have simulated a network that receives the **Full-adapter** for 10 times steps, and after either a short delay ($T_d = 2$ left panel), or long delay ($T_d = 20$, right panel), the external input signals the middle morphed pattern between A and B. Again, if firing rate adaptation is bigger than a threshold:

- If the delay is short, prolonging the target duration produces Type B priming; i.e. while short target lengths result in adaptation aftereffect, longer targets produce priming.
- With a long gap between the adapter and the target, instead, the network shows the opposite effect (only for strong enough adaptation values). For mid range of firing rate adaptation, the network produces negative bias for short targets and an increase in target duration abolishes any bias altogether.

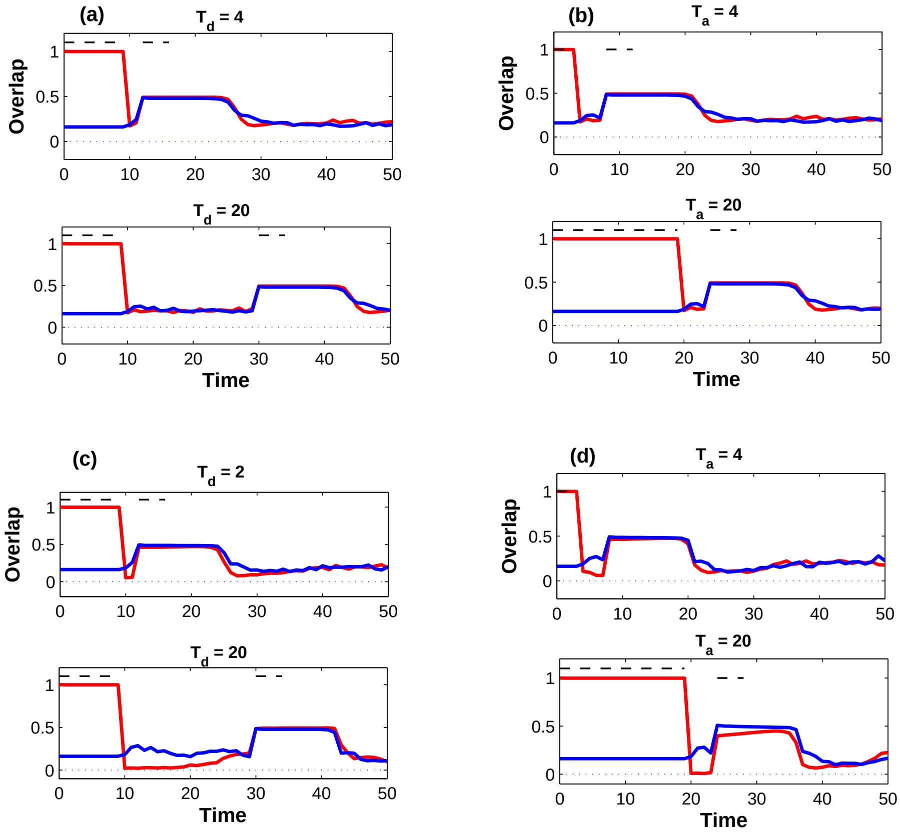


FIGURE 6.1: **Memory-less network**. Simulation results of a network with zero stored pattern, tested with an intermediate morph pattern, equally correlated with pattern A and B. Red line shows the time course of network overlap with pattern A, when the adapter is A (m_a^b), and blue shows the overlap with B (m_a^b). (a) $T_a = 10$, $T_t = 10$ for two different delay durations, $T_d = 2$ in the top panel, $T_d = 22$ in the lower panel (adaptation magnitude $= 0.39$). (b)$T_a = 10$, $T_d = 4$ for two different adapter length, $T_a = 4$ in the top panel, $T_a = 20$ in the lower panel (adaptation magnitude $= 0.39$). (c) and (d) replicate the same simulations of (a) and (b), but with stronger adaptation(adaptation magnitude $= 0.39$)

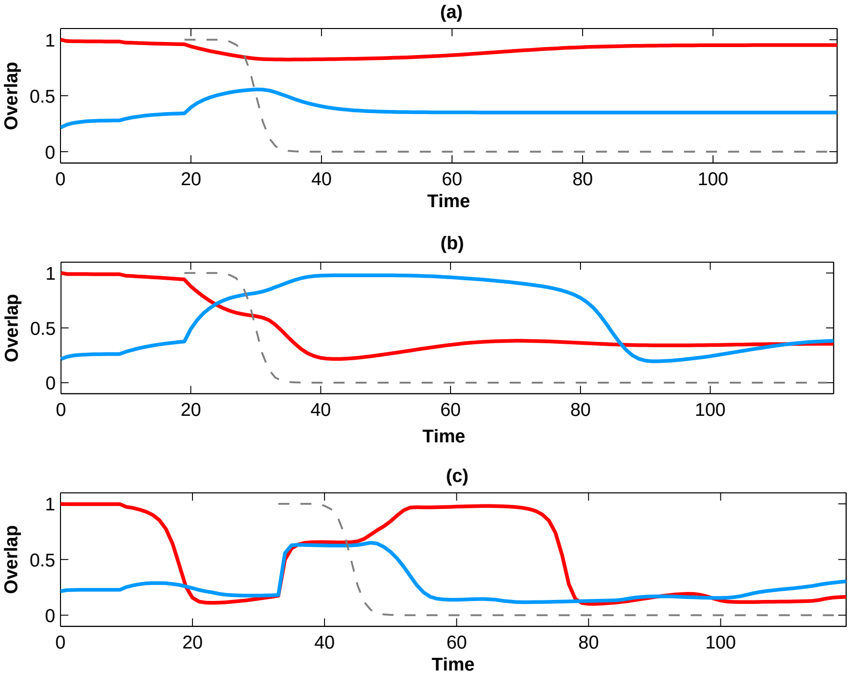


FIGURE 6.2: **"Type A" and "Type B" effects in an attractor neural network**. In this figure, red line shows difference between the network overlap with pattern A when A is the adapter, minus the network overlap with A when B is the adapter $(m_a^a(t) - m_b^a(t))$ and blue shows the opposite$(m_a^b(t) - m_b^b(t))$. (a) Type A priming: perceptual bias, (b) adaptation aftereffects (c) Type B priming: double aftereffects due to the prolongation of the delay period (with respect to (b))

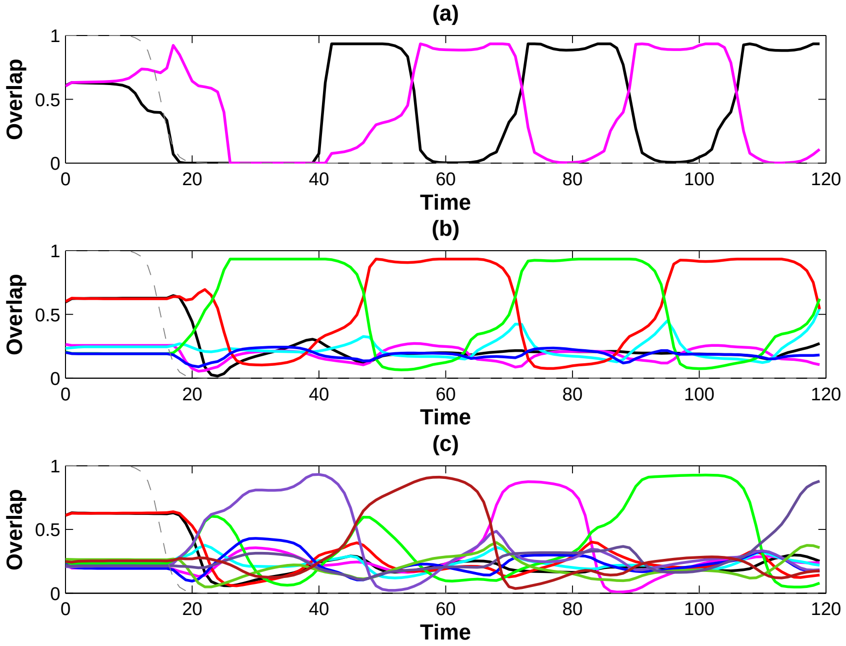


FIGURE 6.3: **Spontaneous transitions during the delay period.** Time course of network overlap with (a) 2 stored patterns, (b) 6 stored patterns, and (c) 15 stored patterns in (c). In each graph the gray dashed line shows the time course of external input, the red and black depict the network overlap with pattern A, and B, respectively; and the other colors show network overlap with other stored patterns

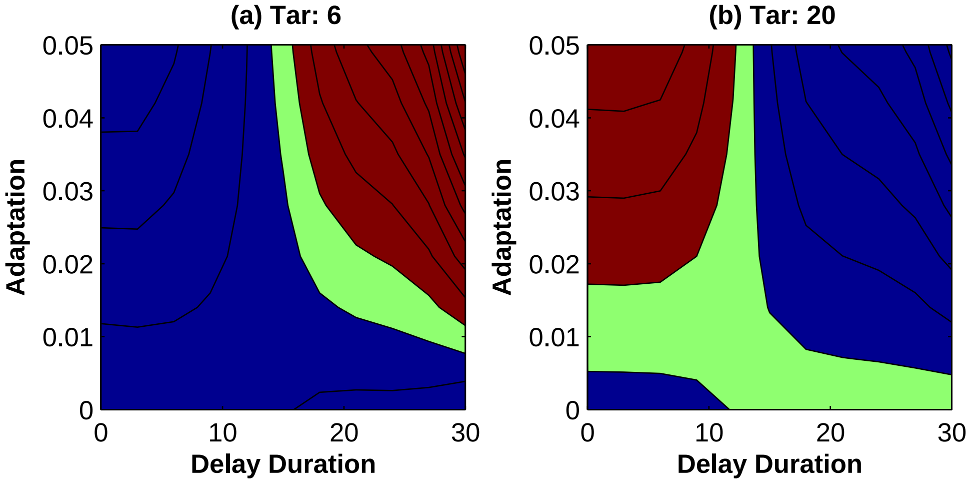


FIGURE 6.4: **Increasing delay period.** Phase diagram of prime index ($\mathcal{R} = m_a^a(t_1) - m_b^a(t_1)$) for a network presented with **Full-adapter** for $T_a = 10$ time steps and varying delay periods, where $t_1 = T_a + T_d + T_t + 5$. Red codes for positive values of $\mathcal{R}$ (priming) and blue instead shows adaptation aftereffects. (a) $T_t = 10$; (b) $T_t = 20$.

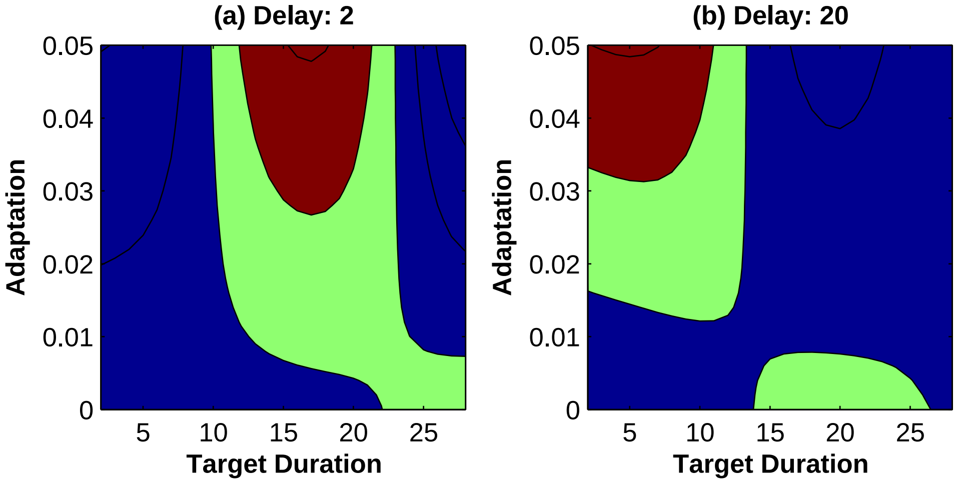


FIGURE 6.5: **Increasing target length** $\mathcal{P}$ for a network presented with **Full-adapter** for $T_a = 10$ time steps, and varying delay target periods. (a) $T_d = 2$; (b) $T_d = 20$

6.1.3 Adapter: an intermediate morph between A and B

We have studied the network response to a morphed patterns, once presented with an ”adapter” which is either of the two end points. However, as it was mentioned in the second chapter, Valentina Daelli has observed the replacement of negative (aftereffects) with positive (priming) bias when the adapter is chosen from one of the intermediate morphed patterns that has high similarity with both of the two end points. Thus, we decided to explore the network response to a morph adapter.

Morph-adapter: higher transition during the prolonged ”adapter” The first difference that replacement of the **Full-adapter** with an intermediate pattern (**Morph-adapter**) brings is an increased probability of having at least one transition during the presentation of adapter.

Response to the prolonged A vs. morph Figure 6.6 shows the $\mathcal{P}$ index, during the presentation of a constant input that signals either full pattern A (a), or a morphed pattern (b, and c).

> A network that receives an external input which has high correlation with two attractors, is more prone to show several transitions.

Increasing the adapter length, Morph-adapter vs. Full-adapter Now we test **Morph-adapter** in our priming paradigm (Fig. 6.7.b). In particular, we examine the effect of the adapter duration (Fig. 6.7). Here keep the target and the delay period fixed ($T_d = 2$, $T_t = 10$) and vary the length of the adapter. The network is tested with **Full-adapter** (left panel), or **Morph-adapter**. We found that We found that

> - If the adapter duration is long enough, then the **Morph-adapter** behaves differently from the adapter that is highly correlated with either of the two patterns.
> - In particular, for short delay durations, increasing

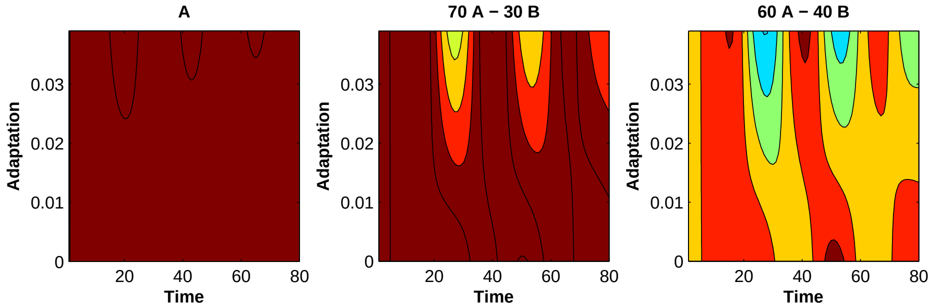


FIGURE 6.6: **Response to a constant input:** A **vs. morph** shows the time course of difference overlap, $m^a - m^b$, when the network is presented with a constant, prolonged (120 time steps) input that signals pattern A in (a), an intermediate morph that shares 70% similarity with A and 30% with B (b), or even more similar to both end points (60% A - 40% B), in (c)

6.1.4 Latching

Interestingly, with strong adaptation, one specific feature emerges in our network in which multiple patterns are stored: after removing the external input, when the feedforward afferents subsides below a particular value, and the dynamics is mainly driven by recurrent connections, sequential transitions between different attractor occur. Latching dynamics emerges as a consequence of incorporating two crucial elements: neuronal adaptation and correlation among attractors. Intuitively, latching may follow from the fact that all units active in the successful retrieval of a pattern tend to adapt, leading to a drop in their activity and a consequent tendency of the corresponding memory units to drift away from their local attractor state. At the same time, though, the residual activity of several units can act as a cue for the retrieval of patterns correlated to the current attractor. As usual with autoassociative memory networks, however, the retrieval of a given pattern competes, through an effective inhibition mechanism, with the retrieval of other patterns. One can then imagine a scenario in which two conditions are fulfilled simultaneously: the activity associated with a decaying pattern is weak enough to release in part the inhibition preventing convergence toward other attractors; but, as an effective cue, it is strong enough to trigger the retrieval of a new, sufficiently correlated pattern. In such a regime of operation, after the first, externally cued retrieval, the network state experiences the concatenation in time of successive memory patterns, i.e. it latches from attractor to attractor.

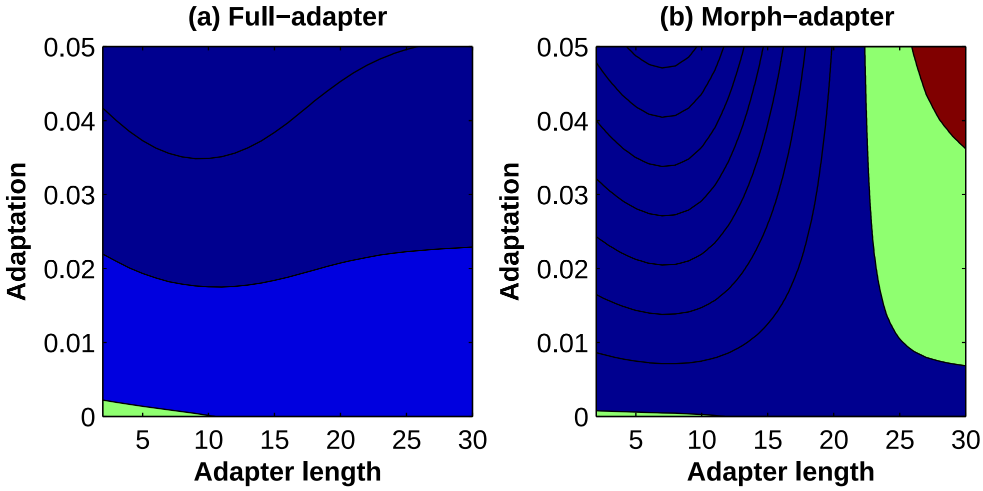


Figure 6.7: **Increasing adapter duration, Full-adapter vs. Morph-adapter.** The prime index ($\mathcal{P}$) for a network presented with **Full-adapter** for varying time steps. (a) $T_d = 2$, $T_t = 10$; (b) $T_d = 2$, $T_t = 10$

In figure 6.8, we have tested the network with nine morphed intermediate versions of the pair of patterns A and B, plus the two original patterns. Each subplot shows the overlap with 6 stored pattern, for different morph level. It is interesting to see different sequences of latching for different morph levels.

The frequency of oscillation and minimum adaptation value to drift the network in a latching phase, as well as the sequential retrieval of different patterns, depend on the adaptation time scale and on the correlation between the patterns.

The emergence of latching dynamics in "Potts" models of cortical networks, has been treated, quite extensively, as a simplified model of a recursive process (Russo et al. 2008; Kropff and Treves 2007). It is interesting to see such a phenomenon repeated in a simple single associated network as ours, though we did not aim to fully address it in our work.

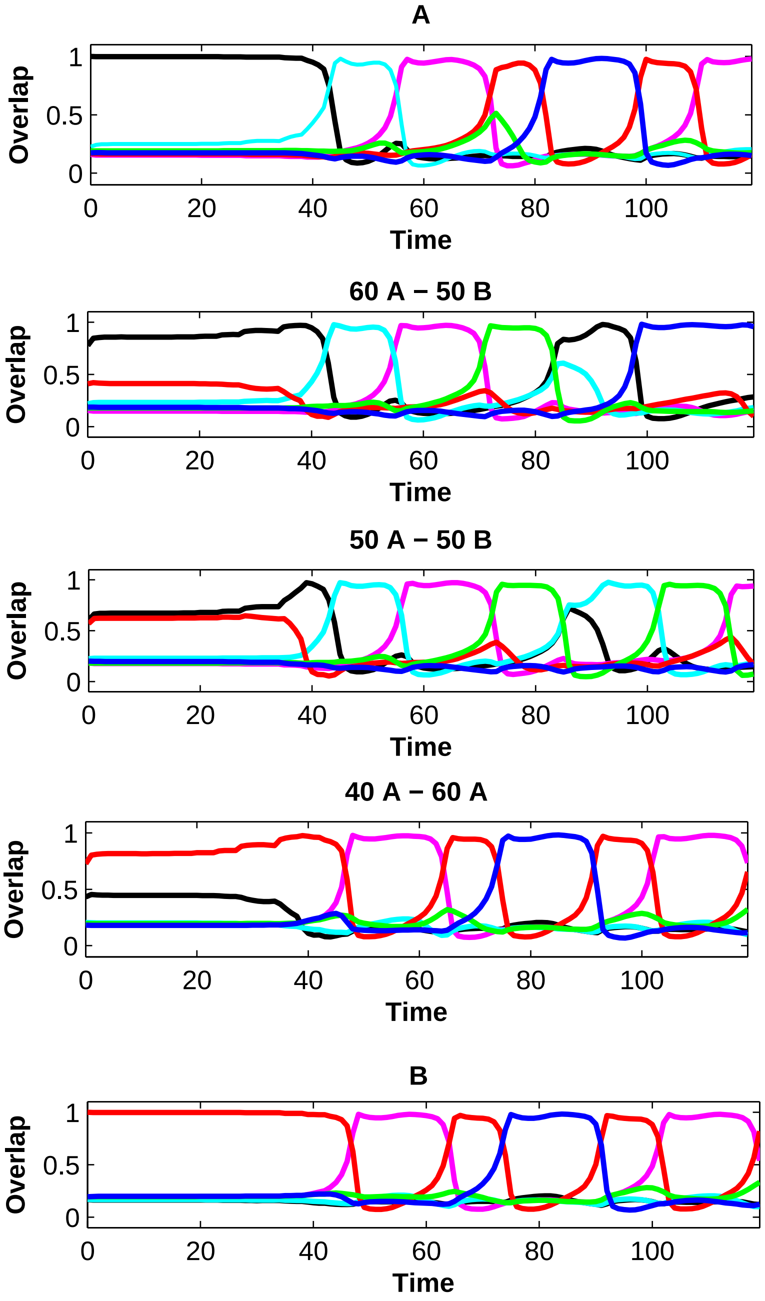


FIGURE 6.8: **Latching**. The network was tested with a pattern whose correlation with patterns A and B gradually changes from top to the bottom. In each panel, the network overlap with all the stored patterns ($p = 6$) is plotted. Red curves show the overlap with A, and black curves show the overlap with B.

6.2 Concluding remarks, and discussion

The theoretical aim of the present study was to provide evidence in support of the hypothesis that memory mechanisms are not dissociated from perceptual mechanisms and involve shared neuronal systems. Our network modeling puts forward two different possible mechanism underlying priming effect and at the same time is able to produce negative and positive bias in different region of phase-space. We have discussed the possible contribution of firing rate adaptation in producing different perceptual shifts along a morphing continuum, due to the short term experience of an adapter. The relevance of attractor states was also clarified; it was shown that a network with no memory states is not able to produce priming. Our results suggest that the minimal model required to show this kind of effect (aftereffects) includes an explicit model of neural fatigue, although it does not dictate the exact form of neural fatigue. Aftereffects do not strictly require, then, attractor formation, however, their exact shape, and values, does depend on the formation of attractors. In a network with stored memories, instead, both delay and target durations are important to produce priming.

6.2.1 Predictions

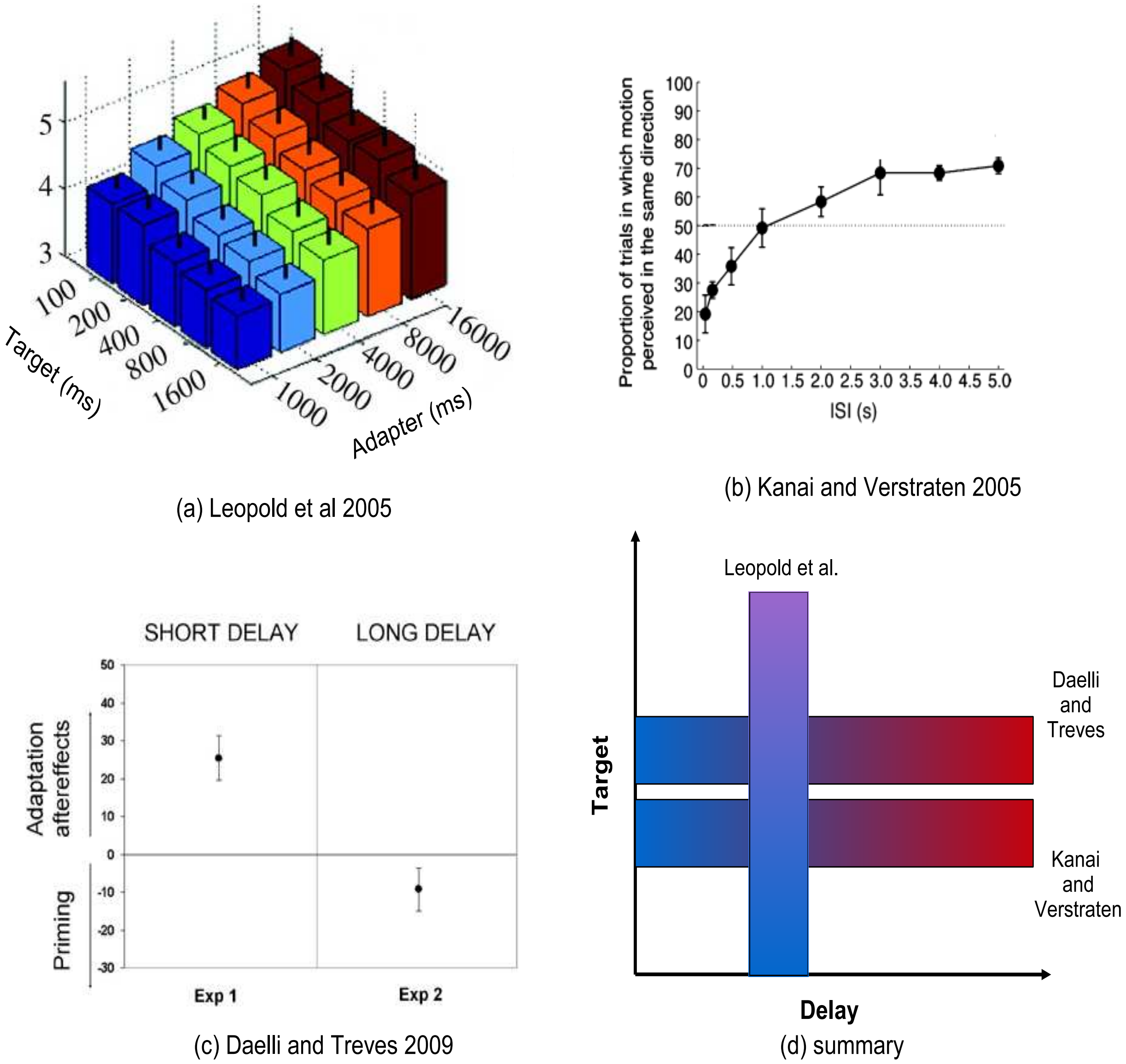
Our model puts forward some immediate predictions, since in terms of temporal parameters of the network, we have found that different combinations of the adapter, delay, and target duration can produce different biases, through the **Type B** priming, which is double (or multiple) adaptation aftereffects. Some of the findings of our model is in agreement with the previous behavioral studies that show that the positive and negative priming occur in the same stimulus configuration, producing opposite effects due to a slight change in temporal setting of the task:

- **target and adapter duration** There have been some classical aftereffect stdudies (for motion aftereffect see Hershenson 1989 and Patterson et al. 2009; for tilt aftereffect see Magnussen and Johnsen 1986), in which the relation between the aftereffect strength and logarithm of the adaptation time is nearly linear. This logarithmic build-up has been replicated for adaptation to faces, as well (Leopold et al. 2005). The face identity aftereffect grew stronger as a function of adaptation time. Moreover, in Leopold et al. 2005 the effect of target duration has been studied, as well. It was found that adaptation aftereffect gets weaker as a function of target duration. This result is However, in none of these studies reducing adaptation aftereffect ends in a positive shift in the direction of the adapter (priming effect). One possibility is that all of these experiments span the adapter and target duration only in a limited time range: the adapter is always bigger 1 second and the target lasts less than 2 seconds. We postulate that decreasing the duration of adapter (or increasing the target length) could produce a positive shift. Indeed, there is one study related to motion aftereffect (Kanai and Verstraten 2005), in which decreasing the adapters to less than 150 ms (and a particular range of adapter-target gap), the perceptual effect switches from negative to positive priming.

- **delay duration** In the same study, Kanai and Verstraten 2005, showed that while brief gap durations between the adapter and the target results in adaptation aftereffect, increasing the delay gradually shifts the effect from negative to positive. The same phenomena was produced, in our group, with more complex visual stimuli in a morphing experiment in which the category boundary shifts due to the recent experience of the adapter (Daelli and Treves 2009, accepted).

- ambiguous prime Our model also predicts different network dynamics in response to a **Full-adapter** or a **Morph-adapter** (Fig. 6.7). This result is in agreement with a bunch of behavioral studies on bistable perception and binocular rivalry (Pearson and Clifford 2005). Also, Kanai and Verstraten (2005) in the same experiment showed that an ambiguous adapter produces different bias in motion perception from a directional adapter (Fig. 6.10.a). Daelli and Treves (2009) have found the similar effect in object aftereffect paradigms. However, our model predicts the type of prime plays a role if long enough. In motion and object aftereffect studies the tested adapter was presented for relatively long time. The prediction of our model opens a new door to expand those behavioral studies with manipulating adapter duration.

In addition, our model, based on the multiple switch phenomena, puts forward series of experimentally testable predictions that have not been examined yet; e.g. **Switch from priming to adaption aftereffect, with increasing the delay duration, while the target is long**, or **non-monotonic change in the bias by lengthening the target duration**. In our model, another portion of the phase-space depends on biophysical properties, such as firing rate adaptation and its time scale that cannot be tested directly.



(a) Leopold et al 2005

(b) Kanai and Verstraten 2005

(c) Daelli and Treves 2009

(d) summary

Figure 6.9: **Switch from Adaptation Aftereffect to Priming – different behavioral studies**. (a) Adaptation aftereffect index as a function of target and adapter duration, in a face aftereffect study, by Leopold et al. 2005. Delay duration is 250 ms. (b) Proportion of trials in which motion has been perceived in the same direction of the adapter, in a motion aftereffect study, by Kanai and Verstraten 2005. Adapter duration and target duration were fixed at 320 ms. (c) Category boundary shift in an object aftereffect study, by Daelli and Treves 2009 (see Chapter 2). Exp1: The adapter is one of the end points (Full adapter), long delay. Exp2: Full adapter, short delay. (d) Summary of different experiments (blue and red stand for adaptation aftereffect and priming, respectively): Increasing delay (with a fixed target duration), produces a switch from adaptation aftereffect to priming; and increasing adapter duration decreases AA.

6.2.2 Future developments

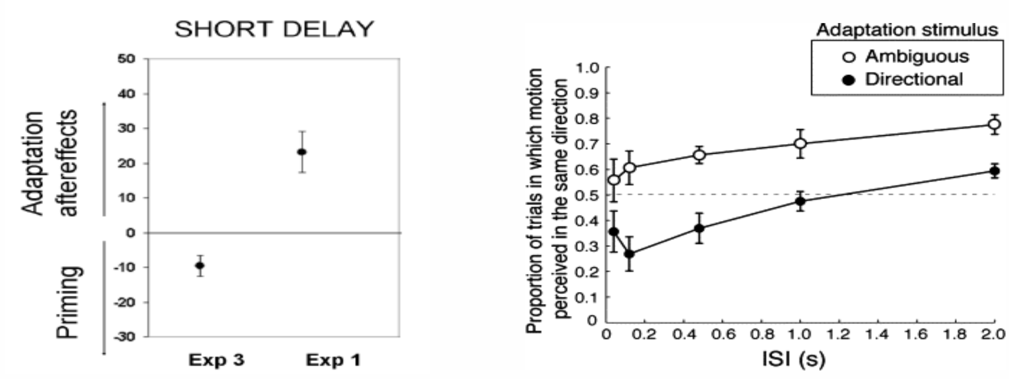
Silent delay activity We have discussed in chapter 2, the experimental and theoretical suggestions that memory traces can implicitly be stored in synapses and only explicitly read out in spiking form at the time of a memory recall signal or new visual input. It is also possible that cortical mechanism underlying the contextual effects like priming recruit the same strategy to store short term memory traces. In particular, in an ongoing priming experiment with monkeys, which is running in Prof. Bharathi Jagadeesh lab in the University of Washington, the same Delayed-Match-To-Sample paradigm that we have discussed in Chapter 5 is adapted to test the effect of an additional adapter before the morphed image. They have observed priming results (at least with short adapter durations). From the point of view of neuronal processing, out of many interesting findings, one particular feature, that challenges the current model, is the absence of delay activity during the gap between the adapter and the target. i.e. on average the neural population relaxes to the baseline firing rate level, although the response to the target is enhanced.

Learning There have been several studies in which the priming effect was suggested to imply some form of ”learning”. Therefore, a mechanism to allow the network to readjust its weights during (and after) the presentation of the adapter can be integrated into the current model that lacks any synaptic changes due to the presence of the adapter.

Multiple local networks We hypothesized that the complexity reported in different priming paradigms may stem from the unfolding in time of the same basic mechanisms at multiple cortical loci, and we asked what determines the direction and magnitude of the effects in priming experiments using complex naturalistic images, which are putatively analyzed in advanced visual cortices and under the influence of multi-modal semantic memories. However, in our current modeling we have restricted ourselves to only one neural population. It remains a challenge for the future expansion of the model to include several interconnecting local networks.

Summary

- Weak adaptation-aftereffects in a memory-less network, with much stronger values of firing rate adaptation (compared to a network with stored patterns),
- No priming was observed in the network with no stored memory (under any duration of the adapter, the target and the gap between them),
- Adaptation aftereffct needs firing rate fatigue, whereas priming attractor dynamics,
- Switch from *priming* to *adaptation aftereffects* happens in a network endowed with firing rate adaptation and attractor state.
- Depending on the exact choices of the temporal parameters (duration of the adapter, target, and delay), the network produces different biases.
- **Morph-adapter** is different from **Full-adapter**, if long enough.



(a) Daelli and Treves 2009

(b) Kanai and Verstraten 2005

FIGURE 6.10: **Ambiguous adapter vs. directional adapter - motion perception**. (a) adapted from Daelli and Treves (2009) Exp1: The adapter is one of the end points (Full adapter), long delay. Exp3: The adapter is one of the ambiguous morphs. (b) adapted from Kanai and Verstraten (2005).

Chapter 7

Appendix

7.1 The noise term

$$\begin{aligned}
\langle R_i \rangle_\xi &= \frac{1}{Ca(1-a)} \left\langle \sum_{j=1}^{N} c_{ij} E_j \sum_{\nu \neq a,b} (\xi_i^\nu (\xi_j^\nu - a)) \right\rangle \\
&= \frac{1}{Ca(1-a)} \sum_{j=1}^{N} c_{ij} E_j \sum_{\nu \neq a,b} \left\langle \xi_i^\nu (\xi_j^\nu - a) \right\rangle \\
&= \frac{1}{Ca(1-a)} \sum_{j=1}^{N} c_{ij} E_j \sum_{\nu \neq a,b} \left\langle \xi_i^\nu \xi_j^\nu \right\rangle - a^2 \\
&= \frac{1}{Ca(1-a)} \sum_{j=1}^{N} c_{ij} E_j (p-2) \left\langle \xi_i^\nu \xi_j^\nu \right\rangle - a^2
\end{aligned}$$

on the other hand

$$\begin{aligned}
\langle \xi_i^\nu \xi_j^\nu \rangle &= P(\xi_i^\nu = 1) \times 1 \times (P(\xi_j^\nu = 1) \times 1 + P(\xi_j^\nu = 0) \times 0) + P(\xi_i^\nu = 0) \times 0 \times (P(\xi_j^\nu = 1) \times 1 + P(\xi_j^\nu = 0) \times \\
&= a \times 1(a \times 1 + 0) \\
&= a^2
\end{aligned}$$

then

$$\begin{aligned} \langle R_i \rangle_\xi &= \frac{1}{Ca(1-a)} \sum_{j=1}^{N} c_{ij} E_j (p-2)(a^2 - a^2) \\ &= 0 \end{aligned}$$

It must be noted that if ξ_i^ν and ξ_j^ν are independent, we could immediately put $\langle \xi_i^\nu \xi_j^\nu \rangle = \langle \xi_i^\nu \rangle \langle \xi_j^\nu \rangle$ which is equal to a^2.
Thus $\langle R_i \rangle$ is equal to zero, and for variance, we have:

$$\begin{aligned} \langle R_i^2 \rangle_\xi &= \frac{1}{C^2 a^2 (1-a)^2} \left\langle \sum_{j,k}^{N} c_{ij} c_{ik} E_j E_k \sum_{\nu,\mu \neq a,b}^{p} (\xi_i^\nu \xi_i^\mu (\xi_j^\nu - a)(\xi_j^\mu - a)) \right\rangle \\ &= \frac{1}{C^2 a^2 (1-a)^2} \sum_{j,k}^{N} c_{ij} c_{ik} E_j E_k \sum_{\nu,\mu \neq a,b}^{p} \big\langle (\xi_i^\nu \xi_i^\mu \xi_j^\nu \xi_k^\mu - a \xi_i^\nu \xi_i^\mu \xi_j^\nu - a \xi_i^\nu \xi_i^\mu \xi_k^\mu - a^2 \xi_i^\nu \xi_i^\mu \big\rangle \end{aligned}$$

for $\nu \neq \mu$ and $j \neq k$ one gets

$$\begin{aligned} \langle \xi_i^\nu \xi_i^\mu \xi_j^\nu \xi_k^\mu \rangle &= \\ a \times 1\,(a \times 1\,(a \times 1\,(a \times 1 + 0) + 0) + 0)) + 0 &= a^4 \end{aligned}$$

with similar reasoning we have

$$\begin{aligned} \langle \xi_i^\nu \xi_i^\mu \xi_j^\nu \xi_k^\mu \rangle = \langle a \xi_i^\nu \xi_i^\mu \xi_j^\nu \rangle = \langle \xi_i^\nu \xi_i^\mu \xi_j^\nu \xi_k^\mu \rangle = \langle a \xi_i^\nu \xi_i^\mu \xi_k^\mu \rangle = \langle a^2 \xi_i^\nu \xi_i^\mu \rangle & \\ &= a^4 \end{aligned}$$

So, for $\nu \neq \mu$ and $j \neq k$, $\langle R_i^2 \rangle_\xi = 0$; and for $\nu = \mu$ and $j = k$ we have

$$\begin{aligned}\langle R_i^2\rangle_\xi &= \frac{1}{C^2a^2(1-a)^2}\sum_j^N e_j^2 \sum_{\nu\neq a,b}^p \langle \xi_i^{\nu 2}(\xi_j^\nu - a)^2\rangle \\ &= \frac{1}{C^2a^2(1-a)^2}\sum_j^N e_j^2 (p-2)(a^2-a^3) \\ &= \frac{1}{C^2(1-a)}(p-2)\sum_j^N e_j^2\end{aligned}$$

We now define a new variable q as the mean square of the firing rate

$$\begin{aligned}q(t) &= \sum_j^N e_j^2 \\ \alpha &= \frac{1}{C^2}(p-2)\end{aligned}$$

In this way, we can set the variance as $\langle R^2\rangle = \frac{\alpha}{(1-a)}q$.

7.2 dynamic equation for firing rate, instead of current

From equation 3 we find that

$$\frac{dF(t)}{dt} = \frac{dh}{dt}[g\Theta(h-Tr) + g(h-Tr)\delta(h-Tr)] \tag{7.1}$$

using 4 we find

$$\begin{aligned}\tau\frac{dF(t)}{dt} &= [-h + h0(E, I, Ia)]\Theta(h - Tr) \\ &= g(-h + Tr - Tr + h0)\Theta(h - Tr) \\ &= -F + g[h0(E, I, Ia) - Tr]\Theta(h - Tr)\end{aligned}$$

If we replace the h inside the Θ-function with h0 and define

$$F0(E, I, Ia) \quad = \quad [h0(E, I, Ia) - Tr]\Theta(h0 - Tr)$$

then, we can write and equation for the firing rate:

$$\tau_e\frac{dE_i(t)}{dt} \quad = \quad -E_i(t) + F0_i$$

$$\tau_e\frac{dE_i(t)}{dt} = -E_i(t) + g(h0_i - Tr)\Theta(h0_i - Tr) \tag{7.2}$$

7.3 Stability analysis near the steady-state for the system of 8 differential equations

We study the satability property for differential equations of the currents:

$$
\begin{aligned}
\tau_h \frac{dh_1(t)}{dt} = f_1 &= -h_1(t) + m^a + m^b + I_{in}^1 + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_1 \\
\tau_h \frac{dh_2(t)}{dt} = f_2 &= -h_2(t) + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_2 \\
\tau_h \frac{dh_3(t)}{dt} = f_3 &= -h_3(t) + m^a + I_{in}^3 + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_3 \\
\tau_h \frac{dh_4(t)}{dt} = f_4 &= -h_4(t) + m^a + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_4 \\
\tau_h \frac{dh_5(t)}{dt} = f_5 &= -h_5(t) + m^b + I_{in}^5 + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_5 \\
\tau_h \frac{dh_6(t)}{dt} = f_6 &= -h_6(t) + m^b + \sqrt{\frac{\alpha_\eta}{(1-a)}} q\eta - Ia_6 \\
\tau_g \frac{dI_g(t)}{dt} = f_7 &= -I_g(t) + \frac{\sum_k n_k < e_k >}{Na} \\
\tau_t \frac{dTr(t)}{dt} = f_8 &= \frac{(\sum_k n_k \langle e_k \rangle)^2}{{I_g}^2 n^2} - a \frac{\sum_k {n_k}^2 \langle e_k \rangle^2}{{I_g}^2 n}
\end{aligned}
$$

in which

$$
\begin{aligned}
m^a &= \frac{1}{Ca(1-a)} \left(\frac{n_1}{I_g} < e_1 > + \frac{n_3}{I_g} < e_3 > + \frac{n_4}{I_g} < e_4 > - a \sum_k \frac{n_k}{I_g} < e_k > \right) \\
m^b &= \frac{1}{Ca(1-a)} \left(\frac{n_1}{I_g} < e_1 > + \frac{n_5}{I_g} < e_5 > + \frac{n_6}{I_g} < e_6 > - a \sum_k \frac{n_k}{I_g} < e_k > \right)
\end{aligned}
$$

and $n_1 + n_3 + n_4 = n_1 + n_5 + n_6 = Ca$.

7.3.0.1 Jacobian Matrix

The steady-state solution can be described in terms of the variable h_i, I_g, and Tr. To investigate whether the steady-state is stable, we expand around this solution, writing

$$\dot{X} = F(X) \tag{7.1}$$

$$X = \bar{X} + Z(t) \tag{7.2}$$

where $\bar{X}(h_1, h_2, h_3, h_4, h_5, h_6, I_g, Tr)$ is a steady state solution for the system $F(\bar{X}) = 0$. To examine the stability to small fluctuations about the stable firing state, we expand $F(X)$ to first order in the quantities $h_{i=1\ldots6}$, I_g and Tr. Once linearized, the system reduces to $\dot{Z} = JZ$, in which $J({h_1}^*, {h_2}^*, {h_3}^*, {h_4}^*, {h_5}^*, {h_6}^*, {I_g}^*, Tr^*)$
is the jacobian matrix around the steady-state. This reduced system tells us whether the fluctuations around the steady-state, vanish or grow exponentially with time. One way to study the effect of such fluctuations is to derive the eigenvalues and (their corresponding eigenvectors) of the linearized jacobian matrix. The condition for stability is satisfied by having negative values for the real parts of all the eigenvalues. The stability limit is when some of the eigenvalues then can change sign in their real part, due a change in one of the parameters of interest, e.g. value of the external input, or the adaptation strength.

In general, the jacobian matrix is of the form:

$$\begin{bmatrix} \frac{\partial f_1}{\partial h_1} & \cdots & \frac{\partial f_1}{\partial I_g} & \frac{\partial f_1}{\partial Tr} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\partial f_8}{\partial h_1} & \cdots & \frac{\partial f_8}{\partial I_g} & \frac{\partial f_8}{\partial Tr} \end{bmatrix}$$

In our model, the form of the Jacobian Matrix can be written:

$$\begin{pmatrix}
-1+2C_1-2aC_1 & -2aC_2 & C_3-2aC_3 & C_4-2aC_4 & C_5-2aC_5 & C_6-2aC_6 & d_1 & t_1 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
C_1-aC_1 & -aC_2 & -1+C_3-aC_3 & C_4-aC_4 & -aC_5 & -aC_6 & d_3 & t_3 \\
C_1-aC_1 & -aC_2 & C_3-aC_3 & -1+C_4-aC_4 & -aC_5 & -aC_6 & d_3 & t_3 \\
C_1-aC_1 & -aC_2 & -aC_3 & -aC_4 & -1+C_5-aC_5 & C_6-aC_6 & d_5 & t_5 \\
C_1-aC_1 & -aC_2 & -aC_3 & -aC_4 & C_5-aC_5 & -1+C_6-aC_6 & d_5 & t_5 \\
u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 \\
l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8
\end{pmatrix}$$

where

$$\begin{aligned}
C_i &= \frac{1}{Ca(1-a)}\frac{\Phi({\rho_i}^*)}{I_g}n_i \\
d_1 &= -\frac{1}{Ca(1-a){I_g}^2}(\sum_{i=1,..,6}\langle e_i\rangle n_i + \langle e_1\rangle n_1 - 2a\sum_{i=1,..,6}\langle e_i\rangle n_i) \\
d_3 &= -\frac{1}{Ca(1-a){I_g}^2}(\sum_{i=1,3,4}\langle e_i\rangle n_i - a\sum_{i=1,3,4}\langle e_i\rangle n_i) \\
d_5 &= -\frac{1}{Ca(1-a){I_g}^2}(\sum_{i=1,5,6}\langle e_i\rangle n_i - a\sum_{i=1,5,6}\langle e_i\rangle n_i) \\
u_{i=1,..,6} &= \frac{n_i\Phi({\rho_i}^*)}{N\tau_g a} \\
u_7 &= -\frac{1}{\tau_g} \\
u_8 &= -\frac{\sum_i n_i\phi(\rho_i)}{\tau_g Na} \\
t_1 &= \frac{1}{Ca(1-a)I_g}(-\sum_i n_i\phi(\rho_i) - n_1\phi(\rho_1) + 2a\sum_i n_i\phi(\rho_i)) \\
t_3 &= \frac{1}{Ca(1-a)I_g}(-\sum_{i=1,3,4} n_i\phi(\rho_i) + a\sum_{i=1,3,4} n_i\phi(\rho_i)) \\
t_5 &= \frac{1}{Ca(1-a)I_g}(-\sum_{i=1,5,6} n_i\phi(\rho_i) + a\sum_{i=1,5,6} n_i\phi(\rho_i)) \\
l_{i=1,..,6} &= \frac{2}{\tau_t n^2{I_g}^2}n_i\phi(\rho_i)\sum_k n_k\langle e_k\rangle - \frac{2a}{\tau_t n{I_g}^2}{n_i}^2\langle e_i\rangle\phi(\rho_i) \\
l_7 &= -\frac{2}{\tau_t n^2{I_g}^3}(\sum_k n_k\langle e_k\rangle)^2 + 2a\frac{2}{\tau_t n^2{I_g}^3}\sum_k {n_k}^2\langle e_k\rangle^2 \\
l_8 &= -\frac{2}{\tau_t n^2{I_g}^2}\sum_i n_i\phi(\rho_i)\sum_i n_i\langle e_i\rangle + \frac{2a}{\tau_t n{I_g}^2}\sum_i {n_i}^2\langle e_i\rangle\phi(\rho_i)
\end{aligned}$$

${\rho_i}^*$ is the steady-state value of ρ_i.

7.3.1 Eigenvalues and Eigenvectors

To make the analysis simple, here we consider a situation in which the external input to the system, the morphed pattern, is equally correlated with pattern A and pattern B. Under this situation, holds the conditions $C_3 + C_4 = C_5 + C_6$ and $\phi(\rho_3) + \phi(\rho_4) = \phi(\rho_3) + \phi(\rho_4)$ hold. These two assumptions make the analytical treatment simple. Then, the stability analysis in such situation finds the necessary condition for the stability of a steady-state; i.e. the condition under which there is a balanced competition between attractor A and attractor B, that produces an equal overlap with each of them. As was reported in the main text, we find that the negativity of the real parts of the eigenvalues, and consequently the stability of the steady-state, depends on the value of the firing rates at the steady-state, which itself is a function of the level of external input. Thus, starting from a system stable at a specific firing rate value, if the external input subsides a particular value, the system loses its stability and slides in to the basin of attraction of one of the relevant attractor states. We tested these analytical result with network simulations.

In the following, we first find the eigenvalues of the system of differential equation in 4.1. The jacobian matrix, found in the previous section, has 8 real eigenvalues.

Then, it is easy to see that for such a matrix, if an eigenvalue $\lambda \neq -1$, then a) the second element of the eigenvector is zero, and b) the eigenvector has the form:

$$v = \begin{pmatrix} \alpha + \beta, & 0, & \alpha, & \alpha, & \beta, & \beta, & \gamma_1, & \gamma_2 \end{pmatrix}$$

Moreover, if $\alpha \neq \beta$, then λ is equal to $-1 + C_3 + C_4$. Moreover, the matrix has 4 eigenvalues equal to -1. Their corresponding eigenvectors are of no interest, as their corresponding fluctuations decay slowly. The other 4 eigenvalues, due to their dependence on steady-state firing rates, can change sign in their real part, and so are of potential interest to study the stability of the system.

7.4 Experimental Method

Surgery on each animal was performed to implant a head restraint, a cylinder to allow neural recording, and a scleral search coil to monitor eye position(Fuchs and Robinson 1966; Judge et al. 1980). Materials for these procedures were obtained from Crist Instruments (Hagerstown, MD) or produced in-house at the University of Washington. Responses of single IT neurons were collected while monkeys performed a delayed-match to sample task (Liu and Jagadeesh 2008Go). Spikes were recorded using the Alpha-Omega spike sorter (Nazareth: Israel). Coded spikes were stored on a PC at a rate of $1000Hz$ using CORTEX, a program for neural data collection and analysis developed at the National Institutes of Health (Bethesda, MD). Eye movements were monitored and recorded (at 500 Hz) using an eye coil based system from DNI (Newark, DE). All animal handling, care, and surgical procedures were performed in accordance with guidelines established by the National Institutes of Health and approved by the Institutional Animal Care and Use Committee at the University of Washington.

Chamber Placement The chambers were placed over the right hemisphere, using stereotaxic coordinates. Neural recordings were targeted near the center of the chamber (Monkey L: 17L, 17.5 A; G: 16 L, 17.5A); this location is in between the perirhinal sulcus and the anterior middle temporal sulcus, in reference to reconstructions from the structural MRI. Recording depths ranged from 27 to 32 mm for Monkey L and 30 to 33 mm for Monkey G. Depth measurements are from the dural surface, measured during an early recording session. The recording locations are identical to those in Liu and Jagadeesh (2008).

Recording Procedures To isolate neurons, we moved the electrode while monkeys performed the passive fixation task with a set of 24 images arranged in 12 pairs (Supplementary Fig. 1). When the experimenter judged that a neuron responded better to one of the 2 images in the 12 pairs of images, she recorded from that neuron while the monkey performed the 2-alternative-forced-choice delayed-match-to-sample ($2AFC - DMS$) task with that stimulus pair.

We repeatedly sampled a single location until we could no longer isolate cells with selectivity for one of the 12 pairs used in the experiment. We moved the electrode location only when selectivity was not detectable over 2-3 days of recording, and moved only slightly across the surface (less than 1 mm). The range of sampled cites spanned a 4 mm diameter circle centered on the stereotaxis locations above. Using this procedure, we found potential selectivity for the 12 image pairs in 75% of the attempted sessions; thus, the cells included in this sample were found frequently. The recorded neurons might include samples from both TE and perirhinal cortex. No anatomical confirmation of recording sites is available from these monkeys because the monkeys continue to be used in other experiments.

Bibliography

[1] L. F. Abbott and Carl van Vreeswijk. Asynchronous states in networks of pulse-coupled oscillators. *Phys. Rev. E*, 48(2):1483–1490, Aug 1993.

[2] L.F. Abbott, Edmund T. Rolls, and Martin J. Tovee. Representational capacity of face coding in monkeys. *Cerebral Cortex*, 6, 1996.

[3] M. Abeles. *Corticonics: Neuronal Circuits of the Cerebral Cortex.* Cambridge University Press, Cambridge, England, 1st edition, 1991.

[4] E. Adelson. Lightness perception and lightness illusions, 1999.

[5] S.R. Afraz, R. Kiani, and H. Esteky. Microstimulation of inferotemporal cortex influences face categorization. *Nature*, 442(7103):692–695, 2006.

[6] B. Ahmed, JC Anderson, RJ Douglas, KA Martin, and D. Whitteridge. Estimates of the net excitatory currents evoked by visual stimulation of identified neurons in cat visual cortex. *Cerebral cortex*, 8(5):462, 1998.

[7] A. Akrami and A. Treves. Neural basis of perceptual expectations: insights from transient dynamics of attractor neural networks. *BMC Neuroscience*, 10(Suppl 1):P174, 2009.

[8] S. Allred, Y. Liu, and B. Jagadeesh. Selectivity of inferior temporal neurons for realistic pictures predicted by algorithms for image database navigation. *Journal of neurophysiology*, 94(6):4068, 2005.

[9] S.R. Allred and B. Jagadeesh. Quantitative comparison between neural response in macaque inferotemporal cortex and behavioral discrimination of photographic images. *Journal of Neurophysiology*, 98(3):1263, 2007.

[10] DG Amaral, N. Ishizuka, and B. Claiborne. Neurons, numbers and the hippocampal network. *Progress in brain research*, 83:1, 1990.

[11] S. Amari. Characteristics of sparsely encoded associative memory. *Neural Networks*, 2(6):451–457, 1989.

[12] S.I. Amari. Characteristics of randomly connected threshold-element networks and network systems. *Proceedings of the IEEE*, 59(1):35–47, 1971.

[13] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55(14):1530–1533, Sep 1985.

[14] D.J. Amit. The Hebbian paradigm reintegrated: local reverberations as internal representations. *Behavioral and brain sciences*, 18(4):617–625, 1995.

[15] D.J. Amit. Simulation in neurobiology: theory or experiment? *Trends in Neurosciences*, 21(6):231–237, 1998.

[16] DJ Amit, A. Bernacchia, and V. Yakovlev. Multiple-object working memory–A model for behavioral performance. *Cerebral Cortex*, 13(5):435, 2003.

[17] Dj Amit and N. Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, 7(3):237–252, April 1997.

[18] DJ Amit, N. Brunel, and MV Tsodyks. Correlations of cortical Hebbian reverberations: theory versus experiment. *Journal of Neuroscience*, 14(11):6435, 1994.

[19] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Information storage in neural networks with low levels of activity. *Physical Review A*, 35(5):2293–2303, 1987.

[20] Y. Amit, D. Geman, and B. Jedynak. Efficient focusing and face detection. *Face Recognition: From Theory to Applications*, 163:124–156, 1997.

[21] B. Anderson, R. E. Mruczek, K. Kawasaki, and D. Sheinberg. Effects of familiarity on neural activity in monkey inferior temporal lobe. *Cerebral cortex (New York, N.Y. : 1991)*, 18(11):2540–2552, November 2008.

[22] J.A. Anderson, J.W. Silverstein, S.A. Ritz, and R.S. Jones. Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, 84(5):413–451, 1977.

[23] J.R. Anderson. A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22(3):261–95, 1983.

[24] T.J. Andrews and D. Purves. Similarities in normal and binocularly rivalrous viewing. *Proceedings of the National Academy of Sciences of the United States of America*, 94(18):9905, 1997.

[25] SM Anstis and RL Gregory. The after-effect of seen motion: The role of retinal stimulation and of eye movements. *The Quarterly Journal of Experimental Psychology*, 17(2):173–174, 1965.

[26] A. Baddeley. Working memory. *Comptes Rendus de l'Academie des Sciences Series III Sciences de la Vie*, 321(2-3):167–173, 1998.

[27] R. Baddeley, LF Abbott, MC Booth, F. Sengpiel, T. Freeman, EA Wakeman, and ET Rolls. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society B: Biological Sciences*, 264(1389):1775, 1997.

[28] E. Barkai and ME Hasselmo. Modulation of the input/output function of rat piriform cortex pyramidal cells. *Journal of Neurophysiology*, 72(2):644, 1994.

[29] H. B. Barlow. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, 1(4):371–394, 1972.

[30] HB Barlow. A theory about the functional role and synaptic mechanism of visual after-effects. *Vision: Coding and efficiency*, 363375, 1990.

[31] M.S. Bartlett and T.J. Sejnowski. Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network: Computation in neural systems*, 9(3):399–417, 1998.

[32] J. Bastian. Gain control in the electrosensory system mediated by descending inputs to the electrosensory lateral line lobe. *Journal of Neuroscience*, 6(2):553, 1986.

[33] F.P. Battaglia and A. Treves. Stable and rapid recurrent processing in realistic autoassociative memories. *Neural Computation*, 10(2):431–450, 1998.

[34] W.F. Battig and W.E. Montague. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, 80(3 Pt 2):1–46, 1969.

[35] D. Bavelier and M.I. Jordan. A dynamical model of priming and repetition blindness. In *Advances in Neural Information Processing Systems 5,[NIPS Conference]*, page 886. Morgan Kaufmann Publishers Inc., 1992.

[36] P.J. Bayley, J.C. Frascino, and L.R. Squire. Robust habit learning in the absence of awareness and independent of the medial temporal lobe. *Nature*, 436(7050):550, 2005.

[37] G. C. Baylis, E. T. Rolls, and C. M. Leonard. Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Res*, 342(1):91–102, September 1985.

[38] GC Baylis and ET Rolls. Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Experimental Brain Research*, 65(3):614–622, 1987.

[39] Behrmann M Joordens S. Becker S, Moscovitch M. Long-term semantic priming: a computational account and empirical evidence. *J Exp Psychol Learn Mem Cogn.*, 23(5):1059–82., 1997.

[40] J.A. Bednar and R. Miikkulainen. Tilt aftereffects in a self-organizing model of the primary visual cortex. *Neural Computation*, 12(7):1721–1740, 2000.

[41] J. Benda and A.V.M. Herz. A universal model for spike-frequency adaptation. *Neural computation*, 15(11):2523–2564, 2003.

[42] J. Benda, A. Longtin, and L. Maler. Spike-frequency adaptation separates transient communication signals from background oscillations. *Journal of Neuroscience*, 25(9):2312, 2005.

[43] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychol Rev*, 94(2):115–147, Apr 1987.

[44] R. Blake and N.K. Logothetis. Visual competition. *Nature Reviews Neuroscience*, 3(1):13–21, 2002.

[45] C. Blakemore and F. W. Campbell. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal image. *The Journal of Physiology*, 203:237–260, 1969.

[46] T. V. P. Bliss and G. L. Collingridge. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407):31–39, January 1993.

[47] S. Blomfield. Arithmetical operations performed by nerve cells. *Brain Research*, 69(1):115–124, 1974.

[48] J. Bormann, OP Hamill, and B. Sakmann. Mechanism of anion permeation through channels gated by glycine and gamma-aminobutyric acid in mouse cultured spinal neurones. *The Journal of Physiology*, 385(1):243, 1987.

[49] A. Borsellino, A. Marco, A. Allazetta, S. Rinesi, and B. Bartolini. Reversal time distribution in the perception of visual ambiguous stimuli. *Biological Cybernetics*, 10(3):139–144, 1972.

[50] V. Braitenberg and A. Schuz. *Anatomy of the cortex.* Springer, 1991.

[51] V. Braitenberg and A. Schuz. *Cortex: Statistics and Geometry of Neuronal Connectivity*, volume 249. Springer, Berlin, Germany, 1998.

[52] J.W. Brascamp, T.H.J. Knapen, R. Kanai, A.J. Noest, R. van Ee, and A.V. van den Berg. Multi-timescale perceptual history resolves visual ambiguity. *PLoS One*, 3(1), 2008.

[53] AS Brown, TC Jones, and DB Mitchell. Single and multiple test repetition priming in implicit memory. *Memory*, 4(2):159–174, 1996.

[54] DA Brown and PR Adams. Muscarinic suppression of a novel voltage-sensitive K&plus; current in a vertebrate neurone. 1980.

[55] N. Brunel. Hebbian learning of context in recurrent neural networks. *Neural Computation*, 8(8):1677–1710, 1996.

[56] N. Brunel and F. Lavigne. Semantic priming in a cortical network model. *Journal of Cognitive Neuroscience*, 21(12):2300–2319, 2009.

[57] Nicolas Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J Comput Neurosci*, 8(3):183–208, 2000.

[58] Mark J. Buckley and David Gaffan. Perirhinal cortex ablation impairs visual object identification. *Journal of Neuroscience*, 18(6):2268–2275, March 1998.

[59] RL Buckner, SE Petersen, JG Ojemann, FM Miezin, LR Squire, and ME Raichle. Functional anatomical studies of explicit and implicit memory retrieval tasks. *Journal of Neuroscience*, 15(1):12, 1995.

[60] J.A. Bullinaria. Modelling lexical decision: Who needs a lexicon. *Neural computing research and applications*, 3:62–69, 1995.

[61] C. Capaday and C. Van Vreeswijk. Direct control of firing rate gain by dendritic shunting inhibition. *Journal of Integrative Neuroscience*, 5(2):199–222, 2006.

[62] M. Carandini and D.J. Heeger. Summation and division by neurons in primate visual cortex. *Science*, 264(5163):1333–1336, 1994.

[63] Cave CB. Very long-lasting priming in picture naming. *Psychol. Sci*, 1997.

[64] C. W. Clifford and G. Rhodes, editors. *Fitting the Mind to the World Adaptation and After-Effects in High-Level Vision*. Oxford University Press, 2005.

[65] C.W.G. Clifford. Perceptual adaptation: motion parallels orientation. *Trends in Cognitive Sciences*, 6(3):136–143, 2002.

[66] A.M. Collins and E.F. Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407–428, 1975.

[67] M Coltheart. Visual feature-analyzers and after-effects of tilt and curvature. *Psychological review*, 78:2:114–21, 1971.

[68] Silva LR. Connors BW, Malenka RC. Two inhibitory postsynaptic potentials, and gabaa and gabab receptor-mediated responses in neocortex of rat and cat. *J Physiol*, 406:443–468, 1988.

[69] G.S. Cree and K. McRae. Analyzing the Factors Underlying the Structure and Computation of the Meaning ofChipmunk, Cherry, Chisel, Cheese, andCello (and Many Other Such Concrete Nouns). *Journal of Experimental Psychology: General*, 132(2):163–201, 2003.

[70] G.S. Cree and K. McRae. Analyzing the Factors Underlying the Structure and Computation of the Meaning ofChipmunk, Cherry, Chisel, Cheese, andCello (and Many Other Such Concrete Nouns). *Journal of Experimental Psychology: General*, 132(2):163–201, 2003.

[71] O.D. Creutzfeldt. Generality of the functional structure of the neocortex. *Naturwissenschaften*, 64(10):507–517, 1977.

[72] A. Crisanti and H. Sompolinsky. Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model. *Physical Review A*, 36(10):4922–4939, 1987.

[73] S. M. Crook, G. B. Ermentrout, and J. M. Bower. Spike frequency adaptation affects the synchronization properties of networks of cortical oscillations. *Neural Comput*, 10(4):837–854, May 1998.

[74] W.H. Griffith D.A. Brown, B.H. Gahwiler and J.V. Halliwell. Membrane currents in hippocampal neurons. *Progr. Brain Res.*, 83:141–160, 1990.

[75] S. Dehaene and J. P. Changeux. A hierarchical neuronal network for planning behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13293–13298, November 1997.

[76] JB Demb, JE Desmond, AD Wagner, CJ Vaidya, GH Glover, and JD Gabrieli. Semantic encoding and retrieval in the left inferior prefrontal cortex: a functional MRI study of task difficulty and process specificity. *Journal of Neuroscience*, 15(9):5870, 1995.

[77] B. Derrida, E. Gardner, and A. Zippelius. An exactly solvable asymmetric neural network model. *EPL (Europhysics Letters)*, 4:167–173, 1987.

[78] B. Derrida and Y. Pomeau. Random networks of automata: a simple annealed approximation. *EPL (Europhysics Letters)*, 1:45–49, 1986.

[79] R. Desimone. Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, 3(1):1–8, 1991.

[80] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995.

[81] S.L. Ding, G.W. Van Hoesen, M.D. Cassell, and A. Poremba. Parcellation of human temporal polar cortex: A combined analysis of multiple cytoarchitectonic, chemoarchitectonic, and pathological markers. *The Journal of Comparative Neurology*, 514(6), 2009.

[82] Foley H. Schacter D. L. Dobbins, I. G. and A. D. Wagner. Executive control during episodic retrieval: Multiple prefrontal processes subserve source memory. *Neuron*, 35:989–996, 2002.

[83] I.G. Dobbins, D.M. Schnyer, M. Verfaellie, and D.L. Schacter. Cortical activity reductions during repetition priming can result from rapid response learning. *Nature*, 428(6980):316–319, 2004.

[84] RJ Dolan, GR Fink, E. Rolls, M. Booth, A. Holmes, RSJ Frackowiak, and KJ Friston. How the brain learns to see objects and faces in an impoverished context. *Nature*, 389(6651):596–599, 1997.

[85] E. Domany, R. Meir, and W. Kinzel. Storing and retrieving information in a layered spin system. *EPL (Europhysics Letters)*, 2:175–185, 1986.

[86] RJ Douglas and KA Martin. A functional microcircuit for cat visual cortex. *The Journal of Physiology*, 440(1):735, 1991.

[87] Rodney J. Douglas, Kevan A. C. Martin, and David Whitteridge. A canonical microcircuit for neocortex. *Neural Comput.*, 1(4):480–488, 1989.

[88] David M. Eagleman, John E. Jacobson, and Terrence J. Sejnowski. Perceived luminance depends on temporal context. *Nature*, 428(6985):854–856, April 2004.

[89] J.C. Eccles. The physiology of synapses. 1964.

[90] H. Eichenbaum, A. P. Yonelinas, and C. Ranganath. The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, 30(1):123–152, 2007.

[91] S. Eifuku, W. C. De Souza, R. Tamura, H. Nishijo, and T. Ono. Neuronal correlates of face identification in the monkey anterior temporal cortical areas. *Journal of neurophysiology*, 91(1):358–371, January 2004.

[92] R.A. Epstein, W.E. Parker, and A.M. Feiler. Two kinds of fMRI repetition suppression? Evidence for dissociable neural mechanisms. *Journal of Neurophysiology*, 99(6):2877, 2008.

[93] C.A. Erickson and R. Desimone. Responses of macaque perirhinal neurons during and after visual stimulus association learning. *Journal of Neuroscience*, 19(23):10404, 1999.

[94] E. N. Eskandar, B. J. Richmond, and L. M. Optican. Role of inferior temporal neurons in visual memory. i. temporal encoding of information about visual images, recalled images, and behavioral context. *J Neurophysiol*, 68(4):1277–1295, October 1992.

[95] M.J. Fenske, E. Aminoff, N. Gronau, and M. Bar. Top-down facilitation of visual object recognition: object-based and context-based contributions. *Visual Perception: Fundamentals of awareness: multi-sensory integration and high-order perception*, page 3, 2006.

[96] I.A. Fleidervish, A. Friedman, and MJ Gutnick. Slow inactivation of Na current and slow cumulative spike adaptation in mouse and guinea-pig neocortical neurones in slices. *J Physiol*, 493(1):83–97, 1996.

[97] K.I. Forster. Form-priming with masked primes: The best match hypothesis. *Attention and performance XII*, pages 127–146, 1987.

[98] L. Franco, E.T. Rolls, N.C. Aggelopoulos, and J.M. Jerez. Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biological Cybernetics*, 96(6):547–560, 2007.

[99] D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, 23(12):5235, 2003.

[100] D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cerebral Cortex*, 16(11):1631, 2006.

[101] A.W. Freeman. Multistage model for binocular rivalry. *Journal of Neurophysiology*, 94(6):4412, 2005.

[102] G. Fuhrmann, H. Markram, and M. Tsodyks. Spike frequency adaptation and neocortical rhythms. *Journal of neurophysiology*, 88(2):761, 2002.

[103] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J Neurophysiol*, 61(2):331–349, February 1989.

[104] N. Furl, N.J. van Rijsbergen, A. Treves, K.J. Friston, and R.J. Dolan. Experience-dependent coding of facial expression in superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 104(33):13485, 2007.

[105] J. M. Fuster and G. E. Alexander. Neuron activity related to short-term memory. *Science*, 173(997):652–654, August 1971.

[106] J.M. Fuster. *Memory in the cerebral cortex.* MIT press Cambridge, MA, 1995.

[107] JM Fuster and JP Jervey. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *Journal of Neuroscience*, 2(3):361, 1982.

[108] JDE Gabrieli. Cognitive neuroscience of human memory. *Annual Review of Psychology*, 49(1):87–115, 1998.

[109] J.D.E. Gabrieli, D.A. Fleischman, M.M. Keane, S.L. Reminger, and F. Morrell. Double dissociation between memory systems underlying explicit and implicit memory in the human brain. *Psychological Science*, pages 76–82, 1995.

[110] R.F. Galán, S. Sachse, C.G. Galizia, and A.V.M. Herz. Odor-driven attractor dynamics in the antennal lobe allow for simple and rapid olfactory pattern classification. *Neural computation*, 16(5):999–1012, 2004.

[111] EJ Gibson. Principles of perceptual learning. *New York: Appleton-Century-Crofts*, 1969.

[112] J.J. Gibson. Adaptation with negative after-effect. *Psychological Review*, 44(3):222–244, 1937.

[113] J.J. Gibson and M. Radner. Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *J. exp. Psychol*, 20:453–467, 1937.

[114] C.D. Gilbert and T.N. Wiesel. Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. *Nature*, 280(5718):120–125, 1979.

[115] M.S. Goldman, P. Maldonado, and LF Abbott. Redundancy reduction and sustained firing with stochastic depressing synapses. *Journal of Neuroscience*, 22(2):584, 2002.

[116] T. Gollisch and A.V.M. Herz. Input-driven components of spike-frequency adaptation can be unmasked in vivo. *Journal of Neuroscience*, 24(34):7435, 2004.

[117] P. Graf, L.R. Squire, and G. Mandler. The information that amnesic patients do not forget. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):164–178, 1984.

[118] MW Greenlee and S. Magnussen. Saturation of the tilt aftereffect. *Vision Research*, 27(6):1041, 1987.

[119] K. Grill-Spector. Visual priming. *In H. Eichenbaum (Ed.), Memory Systems, Learning and Memory: A Comprehensive Reference*, 3:219–236, 2008.

[120] H. Gutfreund, JD Reger, and AP Young. The nature of attractors in an asymmetric spin glass with deterministic dynamics. *Journal of Physics A: Mathematical and General*, 21:2775–2797, 1988.

[121] H. Haarmann and M. Usher. Maintenance of semantic information in capacity-limited item short-term memory. *Psychonomic Bulletin and Review*, 8(3):568–578, 2001.

[122] Olkkonen M. Walter S. Gegenfurtner K.R. Hansen, T. Memory modulates color appearance. *Nature Neuroscience*, 9:1367–1368, 2006.

[123] B. Hauptmann and A. Karni. From primed to learn: the saturation of repetition priming and the induction of long-term memory. *Cognitive Brain Research*, 13(3):313–322, 2002.

[124] J. Hawkins and S. Blakeslee. *On intelligence*. Owl Books, 2005.

[125] C.A.G. Hayman and E. Tulving. Contingent dissociation between recognition and fragment completion: The method of triangulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2):228–240, 1989.

[126] R. Henson, T. Shallice, and R. Dolan. Neuroimaging evidence for dissociable forms of repetition priming. *Science*, 287(5456):1269, 2000.

[127] RN Henson, A. Rylands, E. Ross, P. Vuilleumeir, and MD Rugg. The effect of repetition lag on electrophysiological and haemodynamic correlates of visual object priming. *Neuroimage*, 21(4):1674–1689, 2004.

[128] RNA Henson and MD Rugg. Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia*, 41(3):263–270, 2003.

[129] M. Hershenson. Duration, time constant, and decay of the linear motion aftereffect as a function of inspection duration. *Perception & psychophysics*, 45(3):251–257, 1989.

[130] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the theory of neural computation*. Perseus Books Group, 1991.

[131] C. Holscher, E.T. Rolls, and J. Xiang. Perirhinal cortex neuronal activity related to long-term familiarity memory in the macaque. *European Journal of Neuroscience*, 18(7):2037–2046, 2003.

[132] Gary R. Holt and Christof Koch. Shunting inhibition does not have a divisive effect on firing rates. *Neural Comput.*, 9(5):1001–1013, 1997.

[133] J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc Natl Acad Sci U S A*, 81(10):3088–3092, May 1984.

[134] JJ Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554, 1982.

[135] K.L. Horlitz and A. O'Leary. Satiation or availability? Effects of attention, memory, and imagery on the perception of ambiguous figures. *Perception & Psychophysics*, 53(6):668–681, 1993.

[136] David H. Hubel and Torsten N. Wiesel. Uniformity of monkey striate cortex: A parallel relationship between field size, scatter, and magnification factor. *The Journal of Comparative Neurology*, 158(3):295–305, 1974.

[137] D.E. Huber. Immediate priming and cognitive aftereffects. *JOURNAL OF EXPERIMENTAL PSYCHOLOGY GENERAL*, 137(2):324, 2008.

[138] D.E. Huber and R.C. O'Reilly. Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cognitive Science*, 27(3):403–430, 2003.

[139] N. Ishizuka, J. Weber, and D.G. Amaral. Organization of intrahippocampal projections originating from CA3 pyramidal cells in the rat. *J Comp Neurol*, 295(4):580–623, 1990.

[140] L.L. Jacoby and M. Dallas. On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, 110(3):306–340, 1981.

[141] L.L. Jacoby, J.P. Toth, and A.P. Yonelinas. Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology General*, 122:139–139, 1993.

[142] D. Z. Jin, V. Dragoi, M. Sur, and H. S. Seung. Tilt aftereffect and adaptation-induced changes in orientation tuning in visual cortex. *J Neurophysiol*, 94(6):4038–4050, December 2005.

[143] W. Joung, R. van der Zwan, and CR Latimer. Tilt aftereffects generated by bilaterally symmetrical patterns. *Spatial vision*, 13(1):107–128, 2000.

[144] B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981.

[145] GJ Kalarickal and JA Marshall. Neural model of temporal and stochastic properties of binocular rivalry. *Neurocomputing*, 32(1):843–854, 2000.

[146] R. Kanai and F.A.J. Verstraten. Perceptual manifestations of fast neural plasticity: Motion priming, rapid motion aftereffect and perceptual sensitization. *Vision Research*, 45(25-26):3109–3116, 2005.

[147] M.M. Keane, J.D.E. Gabrieli, H.C. Mapstone, K.A. Johnson, and S. Corkin. Double dissociation of memory capacities after bilateral occipital-lobe or medial temporal-lobe lesions. *Brain*, 118(5):1129, 1995.

[148] Y.J. Kim, M. Grabowecky, and S. Suzuki. Stochastic resonance in binocular rivalry. *Vision Research*, 46(3):392–406, 2006.

[149] Scott Kirkpatrick and David Sherrington. Infinite-ranged models of spin-glasses. *Phys. Rev. B*, 1978.

[150] E. Kobatake, G. Wang, and K. Tanaka. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *Journal of Neurophysiology*, 80(1):324, 1998.

[151] J. Kornmeier and M. Bach. The Necker cubean ambiguous figure disambiguated in early visual processing. *Vision Research*, 45(8):955–960, 2005.

[152] J. Kornmeier and M. Bach. Bistable perceptionalong the processing chain from ambiguous visual input to a stable percept. *International Journal of Psychophysiology*, 62(2):345–349, 2006.

[153] R. Kree and A. Zippelius. Continuous-time dynamics of asymmetrically diluted neural networks. *Physical Review A*, 36(9):4421–4427, 1987.

[154] R. Kree and A. Zippelius. Asymmetrically diluted neural networks. *Springer Physics Of Neural Networks*, pages 193–212, 1991.

[155] E. Kropff and A. Treves. The complexity of latching transitions in large scale cortical networks. *Natural Computing*, 6(2):169–185, 2007.

[156] K. Kubota, M. Tonoike, and A. Mikami. Neuronal activity in the monkey dorsolateral prefrontal cortex during a discrimination task with delay. *Brain Research*, 183(1):29–42, 1980.

[157] L.F. Lago-Fernandez and G. Deco. A model of binocular rivalry based on competition in IT. *Neurocomputing*, 44:503–508, 2002.

[158] C.R. Laing and C.C. Chow. A spiking neuron model for binocular rivalry. *Journal of computational neuroscience*, 12(1):39–53, 2002.

[159] F. Lavigne and S. Denis. Attentional and Semantic Anticipations. *International Journal of Computing Anticipatory Systems*, 8:74–95, 2001.

[160] M.A. Lebedev, A. Messinger, J.D. Kralik, and S.P. Wise. Representation of attended versus remembered locations in prefrontal cortex. *PLoS Biol*, 2(11):e365, 2004.

[161] S.R. Lehky. An astable multivibrator model of binocular rivalry. *Perception*, 17(2):215–228, 1988.

[162] D.A. Leopold and I. Bondar. Adaptation to complex visual patterns in humans and monkeys. *Fitting the mind to the world: Adaptation and after-effects in high-level vision*, pages 189–211, 2005.

[163] D.A. Leopold and N.K. Logothetis. Multistable phenomena: changing views in perception. *Trends in Cognitive Sciences*, 3(7):254–264, 1999.

[164] D.A. Leopold, G. Rhodes, K.M. M
"uller, and L. Jeffery. The dynamics of visual adaptation to faces. *Proceedings of the Royal Society B: Biological Sciences*, 272(1566):897, 2005.

[165] L. Li, EK Miller, and R. Desimone. The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of neurophysiology*, 69(6):1918, 1993.

[166] S.G. Lisberger. Motor learning and memory in the vestibulo-ocular reflex: the dark side. *Annals of the New York Academy of Sciences-Paper Edition*, 781:525–531, 1996.

[167] WA Little and GL Shaw. Analytic study of the memory storage capacity of a neural network. *Math. Biosci*, 39(3-4):281–290, 1978.

[168] T. Liu, J. Larsson, and M. Carrasco. Feature-based attention modulates orientation-selective responses in human visual cortex. *Neuron*, 55(2):313–323, July 2007.

[169] Y. Liu and B. Jagadeesh. Neural selectivity in anterior inferotemporal cortex for morphed photographic images during behavioral classification or fixation. *Journal of Neurophysiology*, 100(2):966, 2008.

[170] Y. Liu, S.O. Murray, and B. Jagadeesh. Time course and stimulus dependence of repetition-induced response suppression in inferotemporal cortex. *Journal of Neurophysiology*, 101(1):418, 2009.

[171] GM Long and TC Toppino. Enduring interest in perceptual ambiguity: Alternating views of reversible figures. *Psychological Bulletin*, 130:748–768, 2004.

[172] G.M. Long, T.C. Toppino, and G.W. Mondin. Prime time: Fatigue and set effects in the perception of reversible figures. *Perception and Psychophysics*, 52:609–609, 1992.

[173] NM Lorenzon and RC Foehring. Relationship between repetitive firing and afterhyperpolarizations in human neocortical neurons. *Journal of neurophysiology*, 67(2):350, 1992.

[174] A. Lueschow, E.K. Miller, and R. Desimone. Inferior temporal mechanisms for invariant object recognition. *Cerebral Cortex*, 4(5):523, 1994.

[175] AV Lukashin, BR Amirikian, VL Mozhaev, GL Wilcox, and AP Georgopoulos. Modeling motor cortical operations by an attractor network of stochastic neurons. *Biological cybernetics*, 74(3):255–261, 1996.

[176] JS Lund. Anatomical organization of macaque monkey striate visual cortex. *Annual Review of Neuroscience*, 11(1):253–288, 1988.

[177] L. Maccotta and R.L. Buckner. Evidence for neural effects of repetition that directly correlate with behavioral priming. *Journal of Cognitive Neuroscience*, 16(9):1625–1632, 2004.

[178] DV Madison and RA Nicoll. Control of the repetitive discharge of rat CA 1 pyramidal neurones in vitro. *The Journal of Physiology*, 354(1):319, 1984.

[179] S. Magnussen and T. Johnsen. Temporal aspects of spatial adaptation. a study of the tilt aftereffect. *Vision Res*, 26(4):661–672, 1986.

[180] A. Maier, M. Wilke, N.K. Logothetis, and D.A. Leopold. Perception of temporally interleaved ambiguous patterns. *Current Biology*, 13(13):1076–1085, 2003.

[181] D. Marr. A theory for cerebral neocortex. *Proceedings of the Royal Society (London) B*, 176:161–234, 1970.

[182] A. Mason and A. Larkman. Correlations between morphology and electrophysiology of pyramidal neurons in slices of rat visual cortex. II. Electrophysiology. *Journal of Neuroscience*, 10(5):1415, 1990.

[183] M.E.J. Masson. A distributed memory model of semantic priming. *Journal of Experimental Psychology-Learning Memory and Cognition*, 21(1):3–22, 1995.

[184] N. Matsumoto, M. Okada, Y. Sugase-Miyamoto, and S. Yamane. Neuronal mechanisms encoding global-to-fine information in inferior-temporal cortex. *Journal of Computational Neuroscience*, 18(1):85–103, 2005.

[185] J. L. McClelland and D. E. Rumelhart. *A distributed model of human learning and memory.* MIT Press, Cambridge, MA, USA, 1986.

[186] JL McClelland and DE Rumelhart. Amnesia and distributed memory. In *Parallel distributed processing*, page 527. MIT Press, 1986.

[187] D. A. Mccormick, B. W. Connors, J. W. Lighthall, and D. A. Prince. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *J Neurophysiol*, 54(4):782–806, October 1985.

[188] R. McEliece, E. Posner, E. Rodemich, and S. Venkatesh. The capacity of the Hopfield associative memory. *IEEE Transactions on Information Theory*, 33(4):461–482, 1987.

[189] D.B.T. McMahon and C.R. Olson. Repetition suppression in monkey inferotemporal cortex: relation to behavioral priming. *Journal of neurophysiology*, 97(5):3532, 2007.

[190] K. McRae. Semantic memory: Some insights from feature-based connectionist attractor networks. *The psychology of learning and motivation: Advances in research and theory*, 45:41–86, 2004.

[191] K. McRae, G.S. Cree, M.S. Seidenberg, and C. McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, 37:547–559, 2005.

[192] K. McRae, VR de Sa, and MS Seidenberg. On the natureandscopeoffeaturalrepresentationsofwordmeaning. *Journal of Experimental Psychology: General*, 126:99–130, 1997.

[193] RW Meech. Calcium-dependent potassium activation in nervous tissues. *Annual review of biophysics and bioengineering*, 7(1):1–18, 1978.

[194] E.M. Meyers, D.J. Freedman, G. Kreiman, E.K. Miller, and T. Poggio. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology*, 100(3):1407, 2008.

[195] M. Mezard, G. Parisi, and M.A. Virasoro. Spin glass theory and beyond. 1987.

[196] E.K. Miller and J.D. Cohen. ANI NTEGRATIVE T HEORY OF P REFRONTAL C ORTEX F UNCTION. *Annual review of Neuroscience*, 24(1):167–202, 2001.

[197] E.K. Miller and R. Desimone. Parallel neuronal mechanisms for short-term memory. *Science-AAAS-Weekly Paper Edition-including Guide to Scientific Information*, 263(5146):520–522, 1994.

[198] E.K. Miller, C.A. Erickson, and R. Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, 16(16):5154, 1996.

[199] E.K. Miller, P.M. Gochin, and C.G. Gross. Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. *Brain Research*, 616(1-2):25–29, 1993.

[200] EK Miller, L. Li, and R. Desimone. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *Journal of Neuroscience*, 13(4):1460, 1993.

[201] B. Milner. Visual recognition and recall after right temporal-lobe excision in man. *Epilepsy Behav*, 4(6):799–812, December 2003.

[202] B. Milner, S. Corkin, and HL Teuber. Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of HM. *Neuropsychologia*, 6(3):215–234, 1968.

[203] M. Mishkin. A memory system in the monkey. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 298(1089):85–95, 1982.

[204] Y. Miyashita. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(6193):817–820, 1988.

[205] G. Mongillo, O. Barak, and M. Tsodyks. Synaptic theory of working memory. *Science*, 319(5869):1543, 2008.

[206] R. Moreno-Bote, J. Rinzel, and N. Rubin. Noise-induced alternations in an attractor network model of perceptual bistability. *Journal of Neurophysiology*, 98(3):1125, 2007.

[207] J Morton. Interaction of information in word recognition. *Psychological Review*.

[208] HE Moss, ML Hare, P. Day, and LK Tyler. A distributed memory model of the associative boost in semantic priming. *Connection Science*, 6(4):413–428, 1994.

[209] JP Nadal, G. Toulouse, JP Changeux, and S. Dehaene. Networks of formal neurons and memory palimpsests. *EPL (Europhysics Letters)*, 1:535–542, 1986.

[210] K. Nakano. Associatron-a model of associative memory. *IEEE Transactions on Systems, Man, and Cybernetics*, 2(3):380–388, 1972.

[211] Y. Naya, K. Sakai, and Y. Miyashita. Activity of primate inferotemporal neurons related to a sought target in pair-association task. *Proceedings of the National Academy of Sciences*, 93(7):2664, 1996.

[212] Y. Naya, M. Yoshida, and Y. Miyashita. Backward spreading of memory-retrieval signal in the primate temporal cortex. *Science*, 291(5504):661, 2001.

[213] AJ Noest, R. Van Ee, MM Nijs, and RJ Van Wezel. Percept-choice sequences driven by interrupted ambiguous stimuli: A low-level neural model. *Journal of Vision*, 7(8):10, 2007.

[214] K A Martin O Bernander, R J Douglas and C Koch. Synaptic background activity influences spatiotemporal integration in single pyramidal cells. *Proc Natl Acad Sci U S A.*, 88(24):1156911573, 1991.

[215] D. O'Kane and A. Treves. Short-and long-range connections in autoassociative memory. *Journal of Physics A: Mathematical and General*, 25:5055–5069, 1992.

[216] D.F. Owens, L.H. Boyce, M.B.E. Davis, and A.R. Kriegstein. Excitatory GABA responses in embryonic and neonatal cortical slices demonstrated by gramicidin perforated-patch recordings and calcium imaging. *Journal of Neuroscience*, 16(20):6414, 1996.

[217] A. J. Oxenham. Forward masking: adaptation or integration? *J Acoust Soc Am*, 109(2):732–741, February 2001.

[218] DN Pandya and EH Yeterian. Prefrontal cortex in relation to other cortical areas in rhesus monkey: architecture and connections. *Progress in brain research*, 85:63, 1990.

[219] Biella G. Rolls E. T. Skaggs W. E. & Treves A. Panzeri, S. Speed, noise, information and the graded nature of neuronal responses. *Network*, 7:365–370, 1996.

[220] M.A. Paradiso, S. Shimojo, and K. Nakayama. Subjective contours, tilt aftereffects, and visual cortical organization. *Vision Research*, 29(9):1205–1213, 1989.

[221] N. Parga and E. Rolls. Transform-invariant recognition by association in a recurrent network. *Neural computation*, 10(6):1507–1525, 1998.

[222] A. Pastukhov and J. Braun. A short-term memory of multi-stable perception. *Journal of Vision*, 8(13):7, 2008.

[223] R. Patterson, L. Tripp, J.A. Rogers, A.S. Boydstun, and A. Stefik. Modeling the simulated real-world optic flow motion aftereffect. *Journal of the Optical Society of America A*, 26(5):1202–1211, 2009.

[224] J. Pearson and C.W.G. Clifford. Mechanisms selectively engaged in rivalry: Normal vision habituates, rivalrous vision primes. *Vision Research*, 45(6):707–714, 2005.

[225] CG Phillips, S. Zeki, and HB Barlow. Localization of function in the cerebral cortex: past, present and future. *Brain*, 107(1):328, 1984.

[226] A. PINKUS and A. PANTLE. Probing visual motion signals with a priming paradigm. *Vision Research*, 37(5):541–552, 1997.

[227] D.C. Plaut. Semantic and associative priming in a distributed attractor network. In *Proceedings of the seventeenth annual conference of the Cognitive Science Society: July 22-25, 1995, University of Pittsburgh*, page 37. Lawrence Erlbaum, 1995.

[228] R.A. Poldrack and J.D.E. Gabrieli. Characterizing the neural mechanisms of skill learning and repetition priming: evidence from mirror reading. *Brain*, 124(1):67, 2001.

[229] S.A. Prescott and Y. De Koninck. Gain control of firing rate by shunting inhibition: roles of synaptic noise and dendritic saturation. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):2076, 2003.

[230] A. Prochazka. Sensorimotor gain control: a basic strategy of motor systems? *Progress in Neurobiology*, 33(4):281, 1989.

[231] G.D. Puccini, M.V. Sanchez-Vives, and A. Compte. Integrated mechanisms of anticipation and rate-of-change computations in cortical circuits. *PLoS Comput Biol*, 3(5), 2007.

[232] G. Rainer and E. K. Miller. Timecourse of object-related neural activity in the primate prefrontal cortex during a short-term memory task. *Eur J Neurosci*, 15(7):1244–1254, April 2002.

[233] G. Rainer, S.C. Rao, and E.K. Miller. Prospective coding for objects in primate prefrontal cortex. *Journal of Neuroscience*, 19(13):5493, 1999.

[234] B. Randall, H.E. Moss, J.M. Rodd, M. Greer, and L.K. Tyler. Distinctiveness and correlation in conceptual structure: Behavioral and computational studies. *Journal of Experimental Psychology Learning Memory and Cognition*, 30(2):393–406, 2004.

[235] C. Ranganath and M. DEsposito. Directing the mind's eye: prefrontal, inferior and medial temporal mechanisms for visual working memory. *Current opinion in neurobiology*, 15(2):175–182, 2005.

[236] A. Raposo, HE Moss, EA Stamatakis, and LK Tyler. Repetition suppression and semantic enhancement: an investigation of the neural correlates of priming. *Neuropsychologia*, 44(12):2284–2295, 2006.

[237] D. Reimann and H. Haken. Stereo vision by self-organization. *Biological cybernetics*, 71(1):17–26, 1994.

[238] A. Renart, R. Moreno, J. de la Rocha, N. Parga, and E.T. Rolls. A model of the IT-PF network in object working memory which includes balanced persistent activity and tuned inhibition* 1. *Neurocomputing*, 38:1525–1531, 2001.

[239] M. Riani and E. Simonotto. Stochastic resonance in the perceptual interpretation of ambiguous figures: A neural network model. *Physical review letters*, 72(19):3120–3123, 1994.

[240] A. Richardson-Klavehn and J.M. Gardiner. Retrieval volition and memorial awareness in stem completion: An empirical analysis. *Psychological research*, 57(3):166–178, 1995.

[241] J.L. Ringo. Stimulus specific adaptation in inferior temporal and medial temporal cortex of the monkey. *Behavioural Brain Research*, 76(1-2):191–197, 1996.

[242] I. Rock, A. Gopnik, and S. Hall. Do young children reverse ambiguous figures? *PERCEPTION-LONDON-*, 23:635–635, 1994.

[243] E.T. Rolls. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27(2):205–218, 2000.

[244] ET Rolls and GC Baylis. Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Experimental Brain Research*, 65(1):38–48, 1986.

[245] E.T. Rolls, G. Deco, and G. Deco. *Computational neuroscience of vision.* Oxford university press Oxford, 2002.

[246] E.T. Rolls and M.J. Tovee. Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings: Biological Sciences*, 257(1348):9–15, 1994.

[247] E.T. Rolls and M.J. Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.*, 73:713726, 1995.

[248] E.T. Rolls, A. Treves, and E.T. Rolls. *Neural networks and brain function.* Oxford University Press Oxford, 1998.

[249] Treves A. & Tovee M. J. Rolls, E. T. The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Exp. Brain Res.*, 114(149-162), 1997.

[250] R. Romo, C. D. Brody, A. Hernández, and L. Lemus. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399(6735):470–473, June 1999.

[251] O. Rosenthal, S. Fusi, and S. Hochstein. Forming classes by stimulus frequency: Behavior and theory. *Proceedings of the National Academy of Sciences*, 98(7):4265, 2001.

[252] Y. Roudi and A. Treves. Disappearance of spurious states in analog associative memories. *Physical Review E*, 67(4):41906, 2003.

[253] Y. Roudi and A. Treves. Representing where along with what information in a model of a cortical patch. *PLoS Comput Biol*, 4:e1000012, 2008.

[254] J.B. Rowe, I. Toni, O. Josephs, R.S.J. Frackowiak, and R.E. Passingham. The prefrontal cortex: response selection or maintenance within working memory? *Science*, 288(5471):1656, 2000.

[255] M.F.S. Rushworth, P.D. Nixon, M.J. Eacott, and R.E. Passingham. Ventral prefrontal cortex is not essential for working memory. *Journal of Neuroscience*, 17(12):4829, 1997.

[256] P. Sah. Ca2+-activated K+ currents in neurones: types, physiological roles and modulation. *Trends in Neurosciences*, 19(4):150–154, 1996.

[257] K. Sakai and Y. Miyashita. Neural organization for the long-term memory of paired associates. 1991.

[258] E. Salinas. Background synaptic activity as a switch between dynamical states in a network. *Neural computation*, 15(7):1439–1475, 2003.

[259] M.V. Sanchez-Vives, L.G. Nowak, and D.A. McCormick. Cellular mechanisms of long-lasting adaptation in visual cortical neurons in vitro. *Journal of Neuroscience*, 20(11):4286, 2000.

[260] H. Sawamura, G.A. Orban, and R. Vogels. Selectivity of neuronal adaptation does not match response selectivity: A single-cell study of the fMRI adaptation paradigm. *Neuron*, 49(2):307–318, 2006.

[261] R. Sayres and K. Grill-Spector. Object-selective cortex exhibits performance-independent repetition suppression. *Journal of Neurophysiology*, 95(2):995, 2006.

[262] D. L. Schacter. Understanding implicit memory. a cognitive neuroscience approach. *Am Psychol*, 47(4):559–569, April 1992.

[263] Daniel L. Schacter and Randy L. Buckner.

[264] D.L. Schacter, J. Bowers, and J. Booker. Intention, awareness, and implicit memory: The retrieval intentionality criterion. *Implicit memory: Theoretical issues*, pages 47–65, 1989.

[265] D.L. Schacter and E. Tulving. *Memory systems 1994*. MIT Press Cambridge, MA, 1994.

[266] Odelia Schwartz, Anne Hsu, and Peter Dayan. Space and time in visual context. *Nature Reviews Neuroscience*, 8(6):522–535, July 2007.

[267] W. B. Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. 1957. *J Neuropsychiatry Clin Neurosci*, 12(1):103–113, 1957.

[268] TJ Sejnowski. Storing covariance with nonlinearly interacting neurons. *Journal of mathematical biology*, 4(4):303–321, 1977.

[269] SI Shapiro and D.S. Palermo. An atlas of normative free association data. *Psychonomic Monograph Supplements*, 2(12):219–250, 1968.

[270] A. Shpiro, R. Curtu, J. Rinzel, and N. Rubin. Dynamical characteristics common to neuronal competition models. *Journal of neurophysiology*, 97(1):462, 2007.

[271] Asya Shpiro, Ruben Moreno-Bote, Nava Rubin, and John Rinzel. Balance between noise and adaptation in competition models of perceptual bistability. *Journal of Computational Neuroscience*.

[272] N. Sigala. Visual categorization and the inferior temporal cortex. *Behavioural brain research*, 149(1):1–7, 2004.

[273] N. Sigala and N. Logothetis. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 155:318320.

[274] S Sobotka and J L Ringo. Mnemonic responses of single units recorded from monkey inferotemporal cortex, accessed via transcommissural versus direct pathways: A dissociation between unit activity and behavior. *Journal of Neuroscience*, 16:4222–4230, 1996.

[275] S. Sobotka and JL Ringo. Stimulus specific adaptation in excited but not in inhibited cells in inferotemporal cortex of macaque. *Brain research*, 646(1):95, 1994.

[276] H. Sompolinsky and I. Kanter. Temporal association in asymmetric neural networks. *Physical Review Letters*, 57(22):2861–2864, 1986.

[277] IM Spigel. The effects of differential post-exposure illumination on the decay of a movement after-effect. *Journal of Psychology*, 50:209–210, 1960.

[278] L.R. Squire, N.J. Cohen, and L. Nadel. The medial temporal region and memory consolidation: A new hypothesis. *Memory consolidation: Psychobiology of cognition*, pages 185–210, 1984.

[279] L.R. Squire, J.T. Wixted, and R.E. Clark. Recognition memory and the medial temporal lobe: A new perspective. *Nature Reviews. Neuroscience*, 8(11):872, 2007.

[280] L. Stollenwerk and M. Bode. Lateral neural model of binocular rivalry. *Neural computation*, 15(12):2863–2882, 2003.

[281] Y. Sugase, S. Yamane, S. Ueno, and K. Kawano. Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400(6747):869–873, August 1999.

[282] Y. Sugase-Miyamoto, Z. Liu, M.C. Wiener, L.M. Optican, and B.J. Richmond. Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. *PLoS Computational Biology*, 4(5), 2008.

[283] W.A. Suzuki, E.K. Miller, and R. Desimone. Object and place memory in the macaque entorhinal cortex. *Journal of Neurophysiology*, 78(2):1062, 1997.

[284] M. Szabo, M. Stetter, G. Deco, S. Fusi, P.D. Giudice, and M. Mattia. Learning to attend: modeling the shaping of selectivity in infero-temporal cortex in a categorization task. *Biological cybernetics*, 94(5):351–365, 2006.

[285] M. Takeda, Y. Naya, R. Fujimichi, D. Takeuchi, and Y. Miyashita. Active maintenance of associative mnemonic signal in monkey inferior temporal cortex. *Neuron*, 48(5):839–848, 2005.

[286] JG Taylor, N. Schmitz, K. Ziemons, M.L. Grosse-Ruyken, O. Gruber, H.W. Mueller-Gaertner, and NJ Shah. The network of brain areas involved in the motion aftereffect. *Neuroimage*, 11(4):257–270, 2000.

[287] P.G. Thompson and J.A. Movshon. Storage of spatially specific threshold elevation. *Perception*, 7(1):65–73, 1978.

[288] DJ Tolhurst and PG Thompson. Orientation illusions and aftereffects: Inhibition between channels. *Vision Research*, 15(8-9):967–972, 1975.

[289] A.S. Tolias, S.M. Smirnakis, M.A. Augath, T. Trinath, and N.K. Logothetis. Motion processing in the macaque: revisited with functional magnetic resonance imaging. *Journal of Neuroscience*, 21(21):8594, 2001.

[290] T.C. Toppino and G.M. Long. Selective adaptation with reversible figures: Don't change that channel. *Perception & Psychophysics*, 42(1):37–48, 1987.

[291] V. Torre and T. Poggio. A synaptic mechanism possibly underlying directional selectivity to motion. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 202(1148):409–416, 1978.

[292] M. J. Tovée, E. T. Rolls, A. Treves, and R. P. Bellis. Information encoding and the responses of single neurons in the primate temporal visual cortex. *J Neurophysiol*, 70(2):640–654, August 1993.

[293] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[294] A. Treves. Mean-field analysis of neuronal spike dynamics. *Network: Computation in Neural Systems*, 4(3):259–284, 1993.

[295] A. Treves. Computational constraints between retrieving the past and predicting the future, and the CA3-CA1 differentiation. *HIPPOCAMPUS-NEW YORK-CHURCHILL LIVINGSTONE-*, 14:539–556, 2004.

[296] A. Treves. Learning to predict through adaptation. *Neuroinformatics*, 2(3):361–365, 2004.

[297] A. Treves. Frontal latching networks: A possible neural basis for infinite recursion. *Cognitive Neuropsychology*, 22(3-4):276–291, 2005.

[298] A. Treves and DJ Amit. Metastable states in asymmetrically diluted Hopfield networks. *Journal of Physics A: Mathematical and General*, 21:3155–3169, 1988.

[299] A. Treves and E.T. Rolls. Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, 2(2):189–200, 1992.

[300] A. Treves, E.T. Rolls, and M. Simmen. Time for retrieval in recurrent associative memories. *Physica D: Nonlinear Phenomena*, 107(2-4):392–400, 1997.

[301] A Treves and Rolls E T. What determines the capacity of autoassociative memories in the brain. *Network*, 2:371–397, 1991.

[302] E. Tulving and DL Schacter. Priming and human memory systems. *Science*, 247(4940):301, 1990.

[303] E. Tulving, D.L. Schacter, H. Stark, C. Fund, and H. Stark. Priming effects in word-fragment completion are independent of recognition memory. *Learning, Memory*, 8(4):336–342, 1982.

[304] S. Ullman. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32(3):193–254, 1989.

[305] D. Ulrich. Differential arithmetic of shunting inhibition for voltage and spike rate in neocortical pyramidal cells. *European Journal of Neuroscience*, 18(8):2159–2165, 2003.

[306] R. Van Der Zwan and P. Wenderoth. Mechanisms of purely subjective contour tilt aftereffects. *Vision Research*, 35(18):2547–2557, 1995.

[307] N. van Rijsbergen, A. Jannati, and A. Treves. Aftereffects in the Perception of Emotion Following Brief, Masked Adaptor Faces. *Open Behavioral Science Journal*, 2:36–52, 2008.

[308] M. Virasoro. The effect of synapses destruction on categorization by neural networks. *EPL (Europhysics Letters)*, 7:293–298, 1988.

[309] R. Vogels. Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *European Journal of Neuroscience*, 11(4):1239–1255, 1999.

[310] C. Vreeswijk and D. Hansel. Patterns of synchrony in neural networks with spike adaptation. *Neural Computation*, 13(5):959–992, 2001.

[311] N.J. Wade, M.T. Swanston, and C.M.M. De Weert. On interocular transfer of motion aftereffects. *PERCEPTION-LONDON-*, 22:1365–1365, 1993.

[312] M. T. Wallace, G. E. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan, and J. A. Schirillo. Unifying multisensory signals across time and space. *Exp Brain Res*, 158(2):252–258, September 2004.

[313] X.J. Wang. Synaptic reverberation underlying mnemonic persistent activity. *TRENDS in Neurosciences*, 24(8):455–463, 2001.

[314] X.J. Wang, Y. Liu, M.V. Sanchez-Vives, and D.A. McCormick. Adaptation and temporal decorrelation by single neurons in the primary visual cortex. *Journal of neurophysiology*, 89(6):3279, 2003.

[315] E.K. Warrington and L. Weiskrantz. The effect of prior learning on subsequent retention in amnesic patients. *Neuropsychologia*, 12(4):419–428, 1974.

[316] M. A. Webster, M. A. Georgeson, and S. M. Webster. Neural adjustments to image blur. *Nat Neurosci*, 5(9):839–840, September 2002.

[317] M. A. Webster, D. Kaping, Y. Mizokami, and P. Duhamel. Adaptation to natural facial categories. *Nature*, 428(6982):557–561, April 2004.

[318] M.A. Webster and O.H. MacLin. Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review*, 6(4):647–653, 1999.

[319] C. Wheatstone. Contributions to the Physiology of Vision.–Part the First. On Some Remarkable, and Hitherto Unobserved, Phenomena of Binocular Vision. *Philosophical transactions of the Royal Society of London*, 128:371–394, 1838.

[320] G.S. Wig, S.T. Grafton, K.E. Demos, and W.M. Kelley. Reductions in neural activity underlie behavioral components of repetition priming. *Nature Neuroscience*, 8(9):1228–1233, 2005.

[321] C.L. Wiggs and A. Martin. Properties and mechanisms of perceptual priming. *Current opinion in neurobiology*, 8(2):227–233, 1998.

[322] T.J. Wills, C. Lever, F. Cacucci, N. Burgess, and J. O'Keefe. Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873, 2005.

[323] H. R. Wilson and J. D. Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Biological Cybernetics*, 13(2):55–80, 1973.

[324] H.R. Wilson, B. Krupa, and F. Wilkinson. Dynamics of perceptual oscillations in form vision. *Nature neuroscience*, 3(2):170–176, 2000.

[325] M. Wilson and BA DeBauche. Inferotemporal cortex and categorical perception of visual stimuli by monkeys. *Neuropsychologia*, 19(1):29, 1981.

[326] A Wohlgemuth. On the after-effect of seen movements. *British Journal of Psychology*.

[327] J.M. Wolfe. Reversing ocular dominance and suppression in a single flash. *Vision Res*, 24(5):471–478, 1984.

[328] Luke Woloszyn and David L. Sheinberg. Neural dynamics in inferior temporal cortex during a visual working memory task. *Journal of Neuroscience*, 29(17):5494–5507, April 2009.

[329] K.F. Wong and X.J. Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314, 2006.

[330] R.A. Wyttenbach, M.L. May, and R.R. Hoy. Categorical perception of sound frequency by crickets. *Science*, 273(5281):1542, 1996.

[331] J.Z. Xiang and MW Brown. Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology*, 37(4-5):657–676, 1998.

[332] Y. Xu, N.B. Turk-Browne, and M.M. Chun. Dissociating task performance from fMRI repetition attenuation in ventral visual cortex. *Journal of Neuroscience*, 27(22):5981, 2007.

[333] Laure Zago, Mark J. Fenske, Elissa Aminoff, and Moshe Bar. The rise and fall of priming: How visual exposure shapes cortical representations of objects. *Cerebral Cortex*, 15(11):1655–1665, November 2005.

[334] D. Zaksas and T. Pasternak. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *Journal of Neuroscience*, 26(45):11726, 2006.