Modeling the neural mechanisms underlying rule-based behavior

Mattia Rigotti

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2010

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Stefano Fusi, Principal Advisor
(Department of Neuroscience)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Larry F. Abbott
(Department of Neuroscience)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Vincent P. Ferrera
(Department of Neuroscience)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Daphna Shohamy
(Department of Psychology)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Earl K. Miller
(Department of Brain and Cognitive Sciences, MIT)

Approved for the Department of Neuroscience:

_____

Steven A. Siegelbaum
Chair

ABSTRACT

Modeling the neural mechanisms underlying rule-based behavior

Mattia Rigotti

Sophisticated goal-directed behavior requires the memory of recent events, the knowledge of the context in which they occur, the goals we intend to reach, and the actions we can perform to obtain them. These are all elements characterizing our mental state. The execution of complex behavior can then be viewed as a flexible rule-based navigation of mental states. Here I develop a theoretical framework which assumes that mental states are neurally represented as stable patterns of sustained activity. I show that one of the conditions necessary to unify the stability of the actively maintained activity patterns with the possibility of a flexible event-driven switch, is the existence of an extensive number of neurons displaying mixed selectivity to conjunctions of mental states and external events. Remarkably, this particular selectivity can be naturally obtained through Randomly Connected Neurons (RCNs). I characterize the capacity, scaling, and response properties of several classes of networks displaying mixed selectivity through RCNs. This analysis suggests that a distributed heterogeneous mode of activity is the neural hallmark of a robust and computationally efficient execution of complex rule-based behavior. This motivates a detailed analysis aimed at characterizing the degree of mixed selectivity in neurophysiological recordings in the orbitofrontal cortex (OFC) and

amygdala of monkeys executing a context-dependent trace conditioning task. These areas display extensive mixed selectivity to the task-relevant stimuli and the affective value they bear. Starting from this result, I proceed to illustrate a theory of how this kind of mixed selectivity neurons can accommodate the creation of new mental states encoding the behavioral context. This theory postulates the presence of two hierarchically organized learning systems. The first one rapidly learns the value of the presented stimuli on a feedback-based manner. It then relays this information to a slower learning system which displays mixed selectivity to the stimuli and their value. This system modifies the synaptic connections between neurons on the basis of the temporal contiguity of their activation. The result of these synaptic modifications is the creation of new patterns of reverberating activity encoding the temporal information of the task contingencies. In turn, these new self-sustained activity patterns participate in sequences of activation on longer time-scales and mediate the iterative creation of new patterns of self-sustained activity. This process of fusion of self-sustained activity patterns is termed *attractor concretion*. The mental states obtained as a result of attractor concretion represent the temporal context on relevant behavioral time-scales and contain information allowing an animal to unambiguously assign a value to the events that initially appeared in different situations with different meanings, thereby contributing to goal-directed behavior.

Contents

List of Figures

Acknowledgments

This manuscript is the result of much more inspirations and influences than I could ever properly acknowledge. I would like to thank my family, for being a great family, and in particular my parents, for being great parents. I want to thank them for a house on a beach I know I will always be welcome to.

I would like to thank Michela, with whom I share so many interests and inclinations, like the aversion for profuse acknowledgements.

I'd like anyway to acknowledge my roommates, who wanted to be mentioned in these acknowledgements (which means they deserve it).

I want to thank Daniel for about 2^{10} coffee breaks and maybe as many discussions, scientific and not, and for the 2^{12} miles we flew to get here. I am also in debt with Emanuele for working and spending an enriching summer with us in Zurich when this project was just at the beginning. I want to acknowledge him for introducing me to sport (or at least trying to), for his effortless friendship, for embarking on this journey with us, a journey he had the imprudence to abandon.

I want to thank the other members of the Fusi Group (in pseudo-random order):

Alex, Srdjan, Anthony, Omri, Lorenzo, Kio, and Fabian for sharing every day their opinions, insights and creativity.

I want to acknowledge the great scientists I had the pleasure to collaborate with during my thesis: Maurizio Mattia, Xiao-Jing Wang, Sara Morrison and Daniel Salzman. I want to thank the Neurotheory people at Columbia University for the unconditioned enthusiasm they put in what they do.

I want to thank the members of my thesis committee: Drs. Daphna Shohamy, Earl Miller, Vincent Ferrera and Larry Abbott, for agreeing on being part of this ritual which, from time immemorial, scientists carry out from generation to generation.

Finally, I want to thank my advisor Stefano for being a great model of how to build models, for his contagious optimism, for his effortless openness, and his sharp insights into science and the scientific community. I also want to thank him for introducing me to what he enthusiastically refers to as 'American pragmatism', which I still haven't totally grasped, but without which this dissertation might have been even longer.

À ceux qui s'donnent à fond,

à ceux qui sont au fond.

Chapter 1

Introduction

Non so se mi crederete.
Passiamo metà della vita a deridere ciò in cui altri credono,
e l'altra metà a credere in ciò che altri deridono.

I don't know if you're going to believe me.
We spend half of our lives mocking what others believe,
and the other half believing what others mock.

(Stefano Benni)

Historically, the organization and functions of the brain and mind have been described by means of explanatory metaphors incorporating the most salient technological experience of the time (Daugman, 2001). So for example, the water technology of antiquity originated Hippocrates's and Galen's theory of the four humours, and Descartes's fascination for clockworks inspired his mechanistic description of animals's instincts. In the same way, images of churning steam machines underlaid the psychodynamic theories of the Belle Époque, and telegraph networks served Helmholtz's nerve metaphor.

These conceits probably tell us more about human thought than about the human brain. But, in the same way one may feel inclined to mock what others believed in the past, one may wonder about how one's beliefs will be perceived in

the future.

The present in the meantime sees us nonchalantly contemplating brain functions through computational tropes, where cells exchange information, areas elaborate inputs, regions receive outputs. But these modern metaphors comparing the brain to a computer set a strange loop in motion. Because the computer was itself conceived as a metaphor of the brain in the first place. At John von Neumann's time, the architecture which bears his name was in fact elaborated as a design model replicating the knowledge about the brain's functionally segregated structure. Even the very term 'computer' refers to Turing's original abstraction of a person computing the results of sequences of operations on an infinitely long piece of paper (and with infinite patience). It is as if the knowledge we gathered is rivaling with the ideas we invented, and reality has caught up with the metaphors we use to describe it. Whatever this might mean, it may be a positive motivational spur for those whose role is to come up with new metaphors, new theories and new models of how it may all work.

Correspondingly, this dissertation employs many descriptive metaphors. So for instance, at some point, the evolution of brain activity will be compared to the dynamics of a sphere rolling down a bumpy landscape, or thermal vibration will be likened to neuronal noise. Such metaphors are a useful interpretive component of any theory, since they allow us to gain intuition about a complex system using what we already know about the world we have experienced.

However, quantitative and mathematically sound models have a critical scientific role that supersedes that of mere descriptive or interpretative metaphors. As it has been pointed out (Abbott, 2008) such models have several inherent benefits. Mathematical models require a precise, self-consistent formulation. Moreover,

they allow a complete characterization of their consequences in the form of predictions and explanations (for some reasons sometimes referred to with the reductive portmanteau of 'postdictions').

This is in essence the main motivation of this dissertation: trying to formulate some ideas about brain functioning, specifically about prefrontal cortical functioning, as mathematical models by translating, where possible, words into equations and concepts into formulas. The aim will be to put ourselves in a position which allows us to work out the consequences of these theories, and eventually formulate predictions, and elaborate explanations.

1.1 Background and overview of dissertation

The computational metaphor describing the brain, nobody will deny, is in many respects a rather shaky one. If in fact the processor of a common desktop machine can perform billions of arithmetical operations per second, while the effort of 6 neuroscientists may not be enough to correctly estimate an 18% tip at the local deli, it is also true that humans still largely outperform any implemented algorithm in a variety of mundane tasks. This gap between *in vivo* and *in silico* computation has resisted the predictions of the most brilliant and shrewd visionaries of the field (Turing, 1950), and is still so vast that even the most influent and fervent advocates of Artificial Intelligence admit defeat, though generally in unpublished form (Sutton, 2001).

Artificial Neural Networks

In the rest of this **Introduction chapter** we will rapidly go through some of the basics of what we could see as the connectionist spin on Artificial Intelligence: Artificial Neural Networks. The brief historical exposition of these topics will lay down some of the technical and conceptual developments which influenced and inspired the rest of this dissertation, both through the successes and achievements in the field, and the failures and unresolved puzzles.

Cognitive control and Prefrontal Cortex

One of the still unresolved puzzles, but which the brain evidently solved in an ingenious way, is how to arrange the activity of billions of neurons to acquire, elaborate and execute a concert of complex behaviors. Various neuropsychological, neuroimaging and neurophysiological studies seem to agree that the Prefrontal Cortex (PFC) is the neocortical region which is responsible for this coordination of sensory information and expression of cognitive control. Situated at the top of the sensory and motor hierarchies, the PFC has been proposed to be responsible for temporally integrating the information for the attainment of prospective behavioral goals, and carry out executive function through a top-down interaction with subcortical areas and other parts of associative cortex (Fuster, 2001).

One of the most influential theories of the implementation of Prefrontal Cortex top-down control of cognitive processes and behavior was laid down in Miller and Cohen (2001), which postulated that the function of PFC is to actively maintain patterns of activity representing goals and the procedures to obtain them. These patterns of activity are hypothesized to exert a bias effectively carrying out the mentioned top-down goal-directed coordinative functions.

Neural substrate for rules representation

In **Chapter 2** of this dissertation we will present a neural implementation inspired by previously elaborated ideas (Miller and Cohen, 2001), and developed within the framework of **Attractor Neural Networks** (ANN), a theoretical formalism mostly due to the pioneering contributions of Amari (1977), Little and Shaw (1978), Hopfield (1982) and Amit (1989). In our efforts to unify the work of Miller and Cohen (2001) with the ANN formalism we had some surprising results. If in fact we assume that the PFC actively sustains activity patterns in a stable manner, and can additionally flexibly switch between them, then the monotonicity of the neural input-output relation will force us to postulate the existence of an extensive number of neurons with *mixed selectivity* to combinations of the sustained activity and the external events triggering the switch. Because of the putative goal-directed coordinative function of the PFC activity patterns, we will operatively postulate that these are the neural correlates of *mental states*, i.e. a subject's propositional attitudes and dispositions to behavior. Within such a representational theory, neurons with mixed selectivity therefore respond to combinations of mental states and events triggering transitions between them. As such, they may at first seem like an abstruse and unlikely theoretical concoction. In fact we will see that this is not the case, since, as we will prove, mixed selectivity neurons arise naturally from random synaptic connectivity.

We will next consider some of the advantages in the use of Randomly Connected Neurons (RCNs) in neural networks. In particular, we will investigate *scaling* relations of this class of networks, one of the thorniest issues in AI, Machine Learning and Neural Networks. According to Richard Sutton, for instance (Sutton, 2001), one of the main reasons for the failure of AI is due to the "anti-theoretic, or 'engi-

neering stance' " which tended to neglect the practical issue of scaling in favor of the solution of specific problems, and resulting in the conception of systems "too reliant on manual tuning" which "will not be able to scale past what can be held in the heads of a few programmers". Analogously the question of scaling turned out to be essential in neural networks memory models, when it was realized that the inclusion of realistic bounds in synaptic efficacy resulted in a catastrophic reduction of the capacity of the established paradigms (Amit and Fusi, 1994), urging therefore to rethink the theory of synaptic plasticity (Fusi et al., 2005).

We will subsequently expose the concrete construction of neural networks representing and implementing rule-based behavior within our framework. This kind of system will be analyzed from different perspectives and we will formulate predictions about the level of coding, their selectivity, and the level of heterogeneity and distribution in the responses, all topics which seem to be of very recent interest (Jun et al., 2010). We will then close Chapter 2 with a Discussion section relating our theory to other approaches and some known relevant experimental observations.

The formation of mental states for context representation

After having characterized the neural representation of mental states and their relationship to dispositions to behavior and rule-based behavior, in **Chapter 3** we will investigate the question of how these representations can be created. Since an essential component of a mental state is the temporal relationship between the sensory, cognitive and decisional elements which characterize it, we will explore the consequences of the assumption that temporal contiguity is the drive of such a creation. Specifically, our hypothesis is that a pool of mixed selective neurons constantly receives information about the sensory events and the ensuing decisions

by the animal and, through a temporal asymmetric Hebbian-like plasticity mechanism, it strengthens the connections between cells which tend to be sequentially activated. This process, we assume, is functional to the creation of patterns of self-sustained activity which will therefore display in their correlation structure the temporal information which mediated their creation. These new self-sustained activity patterns will in turn participate in sequences of activation on longer timescales and mediate the iterative creation of new patterns of self-sustained activity. We name this process of fusion of self-sustained activity patterns *attractor concretion*, and we will show how it can be used to create mental states effectively representing information about a behavioral context, thereby conveniently biasing decisions in the case of ambiguous external evidence.

The signature of temporal contiguity in neural correlations

The development of our framework based on temporal contiguity-mediated concretion of attractors was inspired by theoretical work demonstrating the biologically plausible construction of neural networks encoding simple temporal contexts in the spatial structure of patterns of reverberating neural activity (Brunel, 1996; Griniasty et al., 1993; Yakovlev et al., 1998; Rougier et al., 2005). In turn, these contributions were for the most part motivated by experimental observations.

For example Miyashita, by recording in inferior temporal (IT) cortex of monkeys which were extensively trained on a Delayed matching-to-sample task, discovered representations of sequences of visual stimuli in form of increased correlations of the activity of cells selective to temporally neighboring stimuli (Miyashita and Chang, 1988). This work motivated the development of several models on the neural mechanisms underlying context representation in the brain. For example,

Griniasty et al. (1993) interpreted this data as being the expression of the recurrent dynamics of cortical circuits, whose connectivity is determined by the presentation of sensory stimuli in a fixed temporal order. This work therefore suggests that the neural activity elicited by the presentation of a stimulus represents the temporal context in which the sensory stimulus appears. A more detailed model of the learning process that is responsible for tuning the synaptic weights has been proposed by Brunel (1996) and some of the predictions have been experimentally verified in Yakovlev et al. (1998).

More recently, Kobatake et al. (1998) showed a similar effect on the selectivity of IT cells in anesthetized adult monkeys trained on a shape-discrimination task. Essentially the selectivity of IT neurons for specific visual objects was shown to change as a function of the exposition to a fixed set of visual objects, that is with experience. This study also offered a very extensive population analysis and a comparison with the selectivity patterns observed in untrained control monkeys.

Lately, Li and DiCarlo (2008) provided striking evidence that these modifications of the selectivity of IT neurons can be extremely fast in *in vivo* experiments of monkeys experiencing natural viewing. These modification seem to be due to an activity-mediated unsupervised process, driven by the temporal contiguity of the stimuli presentation and may be what underlies the formation of invariant visual object representation.

Creation of mental states representing context in paradigmatic task

We will illustrate the proposed mechanism of attractor concretion to create mental states representing the behavioral context in the case of a trace conditioning task recently used in neurophysiological recordings (Paton et al., 2006; Salzman et al.,

2007; Belova et al., 2008; Morrison and Salzman, 2009). In such a Pavlovian task monkeys learned whether an abstract fractal image (conditioned stimulus, CS) predicted a liquid reward or an aversive air-puff (unconditioned stimulus, US) after a brief time (trace) interval. After a variable number of trials, the CS-reinforcement contingencies were reversed and monkeys had to learn the new contingencies. In order to develop appropriate responses to the presentations of the CSs (defensive blinking in anticipation of negative CS predicting the air-puff, and appetitive licking in anticipation of positive CS predicting liquid reward) the monkeys where required to acquire the contingencies defining two distinct contexts, i.e. the CS-US associations before and after the reversal.

What the cited papers report is that single unit recordings in the amygdala and orbitofrontal cortex (OFC) reveal the existence of cells encoding the learned value of the CSs (rewarded or punished). Notice that, together with the information of the presented CSs, this affective information about the predicted US is enough to determine the behavioral context the monkeys should learn to minimize punishment (by blinking at the air-puff) and maximize reward (by licking at the delivery of liquid reward). In order to accommodate the representation of the contexts, our theory therefore postulates the presence of an extensive number of mixed selectivity neurons which respond to conjunctions of CSs and predicted USs. This prediction is indeed verified by a statistical analysis of the recordings.

What our model assumes in particular is the presence of a fast learning network which can rapidly learn the value of the CSs through a Reinforcement Learning mechanism, and relay this affective information to a second network of mixed selective neurons where attractor concretion can occur in a slower unsupervised manner.

A similar architecture has also been put forward in McClelland et al. (1995), where, motivated by theoretical considerations, the authors influentially suggest that fast learning in the hippocampus interacts and "trains" slower cortical connections, in order to combine rapid information acquisition with generalization over longer time-scales.

Imaging (Poldrack et al., 2001) and double dissociation studies of the performance of amnesiac and Parkinson subjects in probabilistic learning (Shohamy et al., 2008) suggest a similar relationship between the mediotemporal lobes and Basal Ganglia (BG), where the former are engaged in rapid associative learning, while the latter incrementally integrate experience over longer time-scales in a feedback-based manner.

Motivated by the result of simultaneous electrophysiological recordings (Pasupathy and Miller, 2005) during reversal learning of a visuomotor association task, Miller and Buschman (2008) on the other hand, attribute to the BG the role of a supervised fast learning module "training" the slower and less supervised Prefrontal Cortex in charge of creating more abstract cognitive structures.

Chapter 3 can therefore be seen as a study of a concrete mechanism implementing these proposed interactions between supervised feedback-based learning and unsupervised learning systems realizing simultaneous learning over different time-scales.

Conclusion and future directions

Finally, **Chapter 4** will be a conclusive chapter briefly listing some future directions in which the presented material could be developed.

1.2 Artificial neural networks: a brief historical overview

Here we will briefly review some of the theoretical tools which will serve as conceptual and technical basis for the rest of this dissertation. The exposition will be mostly historical, with some excursions into formal details when needed.

1.2.1 The first connectionists

The inception of the research field which is now known as Artificial Neural Networks (ANN) is commonly attributed to McCulloch and Pitts (1943). In this paper Warren McCulloch and Walter Pitts introduced a valuable mathematical abstraction of the computational function of biological neuron in a connectionist perspective. Neurons were modeled as simplified units, whose output is evaluated as an activation function of the weighted sum of the incoming inputs. Apart from being the template of most of the subsequent formal treatments this contribution was a fertile ground for other innovative lines of research. Taking the McCulloch-Pitts paper as a starting point, Kleene (1956) developed a formulation of McCulloch-Pitts nets which is closer to the modern one and analyzed them as finite-state machines. In the same collection of papers and using the same concepts revolving around finite-state machines or automata, John von Neumann proposed *redundancy* as a principle to synthesize "reliable organisms from unreliable components" (von Neumann, 1956), an approach which later lead to developing the idea of *distributed* redundant representations (Winograd and Cowan, 1963).

1.2.2 Rosenblatt's Perceptron

The ANN community welcomed very enthusiastically the work of Rosenblatt, who proposed at the beginning of the 60's a learning algorithm for single-layer neural networks known as the Perceptron (Rosenblatt, 1962). Rosenblatt was able to prove the convergence of the Perceptron algorithm, which was therefore guaranteed to find a synaptic weight configuration performing a desired computation.

The excitement raised about Rosenblatt's *Perceptrons* faded suddenly following the publication of the homonymous book by Minsky and Papert (1969). In this work Minsky and Papert demonstrated that it is in fact true that the Perceptron algorithm converges when a solution to an heteroassociation problem exists, but the problem is that such a solution does not exist in general. What was pointed out very forcefully in their book is that the Perceptron is not able to represent a solution to even extremely simple problems like the implementation of an exclusive OR (XOR) operation.

This strike against what seemed like a promising paradigm for neural networks threw the whole field into a period of discomfort. It seemed in fact clear that one of the ways to overcome the limitations intrinsic in the one-layer architecture of the Perceptron was to consider multi-layer structures. The problem was that these complex architectures were difficult to train, and their general capability was hard to characterize. As Hertz et al. (1991) put it: "With this most of the computer science community left the neural network paradigm for almost 20 years.".

1.2.3 Recurrent neural networks and associative content-addressable memory

One offshoot that researchers continued to study during the 70's was the theme of **associative content-addressable memory**, a paradigm in which different input patterns become associated with one another if sufficiently similar. Some great conceptual advances in this field pioneered by Amari (1977) and Little and Shaw (1978), just to cite some of the most important authors, were very successfully synthesized by Hopfield (1982), who managed to formulate the idea of memories as dynamically stable attractors which are energy minima of a system of interconnected units.

The physical insights gained by Hopfield's reformulation opened up the way to the applications of the formal and conceptual machinery developed in statistical mechanics. So for instance, Amit et al. (1985) applied mathematical methods developed for spin glasses and disordered magnetic systems to calculate the memory capacity of the Hopfield model; Gardner (1988) from the so-called Edinburgh group overthrew the point of view applying the same methods (the replica theory) to calculate the maximal capacity achievable by any model for a given pattern distribution and a local stability parameter, a concept whose importance was recognized by Krauth and Mézard (1987). The role of the stability parameter in determining the basin of attraction of a memory was then investigated numerically by Forrest (1988) and analytically by Abbott and Kepler (1989). Because of their importance for the rest of this thesis the idea of attractor of the neural dynamics, and the related concepts of basin of attraction and stability will be developed below.

Attractors and basins of attraction

The concept of basin of attraction, despite the suggestive name, can be difficult to grasp. The Hopfield model offers a good playground for this, which is probably one of the reasons for its success. Given a set of patterns of activity of a neural network, the Hopfield model gives a neural dynamics and a (Hebbian) storage prescription to create a synaptic matrix so that the given patterns are stable fixed points of the neural dynamics, in one word, attractors. This means that the collective configuration of the neural activity evolves until it encounters one of the attractors, and then it gets stuck there. Attractors of the Hopfield model act as content-addressable memories, because as soon as the pattern of neural activity resembles one attractor, it gets directed towards it. This offers an operative definition of a basin of attraction. In general, the basin of attraction of a given attractor is, loosely speaking, a neighborhood in phase space from which the system's dynamics is lead to approach that attractor. In the case of an attractor neural network, the basin of attraction of an attractor are the activity patterns which evolve towards the considered attractor. The most common figurative description (which is very accurate in the case of the Hopfield model with symmetric synaptic matrix) is the metaphor of a ball falling into a dent on the ground which rolls to the lowest point in the depression (see Fig. 1.1A). This metaphor clarifies what is generally meant by 'width' or 'radius' of a basin of attraction. It is simply the maximal difference an activity pattern can have from an attractor, in order to be still attracted towards it see (Fig. 1.1B).

The basin of attraction can have an important role in the case of a noisy dynamics. Noise, that is stochastic excursions of the activity pattern around an average configuration, can be damped by the attractor dynamics (see Fig. 1.1C). If

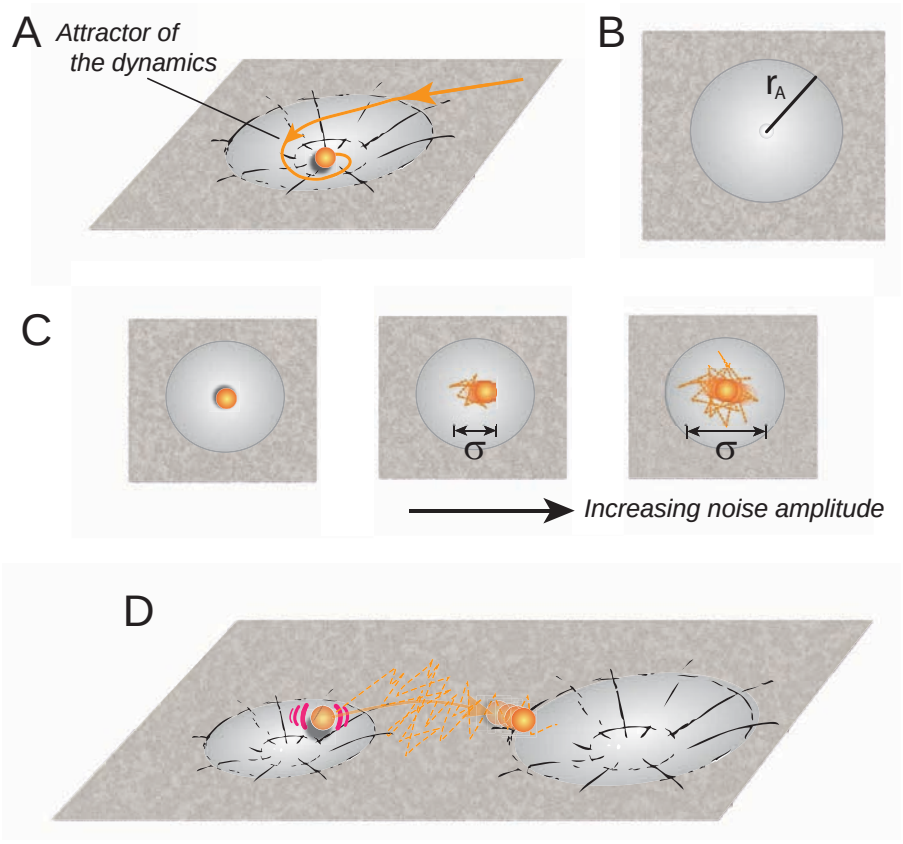

FIG. 1.1: Metaphoric illustration of the concept of 'basin of attraction'. **A**, The dynamics of the network's activity is represented by a ball rolling on a surface subjected to gravity. As soon as the ball falls within the depression of the surface, it is attracted towards the bottom (the attractor of the dynamics). **B**, In this metaphor the radius r_A of the basin of attraction corresponds to the width of the depression. This gives a way to quantify how "attractive" the attractor is, that is, how close the activity has to be in order to be attracted by it. **C**, Noise can be thought of as thermal vibrations kicking around the activity pattern. Increasing noise causes excursions of increasing size around the attractor. As long as the width σ of the average noise excursion is lower than the width r_A of the basin of attraction the activity stays around the attractor. **D**, If the noise amplitude goes above the width of the basin of attraction the activity is kicked out of it. In this case it can fall within a neighboring basin of attraction, and, if its width is now larger than the noise amplitude, the activity will evolve towards the corresponding attractor.

the basin of attraction r_A is wider than the average excursions σ induced by noise, the activity pattern will be consistently pulled back to the attractor, despite the noise. If however the noise level is high enough ($\sigma > r_A$), the stochastic excursions will overwhelm the attractor dynamics. The pattern of activity will therefore deviate from the attractor, until it is caught by another larger basin of attraction (Fig. 1.1D).

The interplay between the noise amplitude σ and the width of the basin of attraction r_A can be exploited to create interesting ANN dynamics. The noise level σ can for instance be increased enough to essentially erase the effect of small unwanted spurious attractors, while still preserving the basins of the wanted attractors (Amit, 1989). Alternatively, the noise can be set high enough to trigger stochastic transitions between attractors (Buhmann and Schulten, 1987b), and promote exploratory behavior (Buhmann and Schulten, 1987a).

In Appendix A we will present a brief but some more technical treatment of the concept of basin of attraction in Attractor Neural Networks. An important result of that section is how to increase the width of a basin of attraction by requiring the presence of a so-called *learning margin*. We will also explore its relationship to *stability parameters*.

1.2.4 Back to feed-forward with Backpropagation

In parallel to the advancements in recurrent associative neural networks, the publication in Rumelhart et al. (1986) of the **Backpropagation algorithm** marked a "renaissance" of multilayer feed-forward neural networks. Even if a similar algorithm appears to have been discovered already in 1969 by Arthur Earl Bryson and Yu-Chi Ho it wasn't until David E. Rumelhart, Geoffrey E. Hinton and Ronald J.

Williams popularized their Backpropagation algorithm that a method to train and use multilayer neural networks was brought to the mainstream. A few years later Cybenko (1989) crucially proved that a single hidden layer feed-forward neural network is capable of approximating any continuous function, as long as it is provided with enough neurons in the hidden layer.

At this point everything was pretty much in place for a full-blown development of the connectionist paradigm: multilayer network were known to possess the potential to represent any possible input-output relation, and the Backpropagation offered a technically transparent way to train them to perform the desired function. In fact this paradigm has been a successful playground for many disciplines interested in connectionist approaches over many years: from cognitive sciences (Poggio and Edelman, 1990; McClelland et al., 1995; Gluck et al., 1996) to statistical data analysis (Geman et al., 1992), from artificial intelligence (Hinton, 1989), to industrial applications (Bernasconi, 1988).

1.3 Modern approaches

After a period of heuristics and experimentation it became somehow clear to many that the Backpropagation program wasn't suited for much more than that.

One problem with Backpropagation is its biological implausibility, due to the need to propagate the error of the output neurons back through the same synapses which conveyed their synaptic input.

In addition to that, the major problems with Backpropagation are probably algorithmic and ontological. Backpropagation's main utility is its applicability to train multi-layer networks. But the problem is that multi-layer networks are a

rather difficult formal entity to control and analyze quantitatively. They lack in fact any discernible mathematical structure. If on one hand they allow a direct and very malleable realization of heuristic insights of specific problems, on the other hand not much is known about general capacity and scaling properties. In this sense the use of multi-layer networks encounters very quickly a difficulty mentioned earlier as one of the main problems of AI: since heuristics and human intervention is constantly needed to keep these systems under control, one cannot expect them to be applicable to larger problems than human control can permit (Sutton, 2001). Moreover, the Backpropagation algorithm, being a form of gradient-descent, is prone to technical problems like slow convergence in deep-networks, convergence to local minima of the error function, overfitting, and poor generalization performance.

For these and other reasons Backpropagation has almost generally been substituted in technical applications by other more algorithmically sound approaches like restricted **Boltzmann machines** (Hinton, 2007) and **Support Vector Machines** (Cortes and Vapnik, 1995). In particular this last paradigm invented by Vladimir Vapnik in the 90's has been very successful, because of its generality, amenability to quantitative analysis and the focus on the minimization of generalization error, all major deficiencies of Backpropagation.

1.3.1 Support Vector Machines

Support Vector Machines (SVM) are able to realize intuitive requirements like the maximization of classification margin through very sophisticated methods of convex optimization. Moreover, SVM reconcile mathematical rigor in their conceptual development within the Probably Approximately Correct (PAC) learning framework (Valiant, 1984), with flexibility in the practical implementations, which make them

a very useful computational tool.

Some of the concepts developed by Vapnik to implement SVM are going to be important in the course of this thesis. In particular the idea of exploiting an **embedding in a higher dimensional space** to increase the separability of data, has a formal parallel in our use of randomly connected neurons as a means to achieve pretty much the same thing (see Chapter 2). In addition, our framework will also focus on **margin maximization** to achieve a form of generalization, that is, robustness to noise and distractors (see Section 2.2.5). This, in our framework, will find a parallel in the enlargement of basins of attraction to increase the domain of stable neural activity patterns. For a detailed treatment of SVM see Scholkopf and Smola (2002).

1.3.2 Recurrent networks

To be fair one should mention that Backpropagation was not completely abandoned after the invention of Support Vector Machines and Boltzmann machines (some Boltzmann machines actually even perform a Backpropagation post-learning refinement step (Hinton and Salakhutdinov, 2006)). In fact, the idea known as **unfolding of time** originally proposed by Minsky and Papert (1969), allowed Backpropagation to survive outside the realm of multilayer feed-forward structures and be applied to the training of recurrent networks with a procedure which is known as **backpropagation through time** (Hertz et al., 1991). The idea is to see the activity of the network at one time as the consequence of the activity at a previous time, as if it was the activity on one layer being determined by the activity on the previous one. This gives rise to a training procedure of the synaptic connections in view of the temporal sequence one wants to achieve, as if training

subsequent layers of a feed-forward architecture.

Some of the problems of this kind of approach are due to the spontaneous activity of the network while it is learning. One of the most difficult situations to overcome is for instance given by the danger of bifurcations in the network's activity in the course of learning (Doya, 1992). On the other hand, the chaotic activity arising by bifurcating across a critical value of the average gain parameter (Sompolinsky et al., 1988) can be a potentially beneficial reservoir of temporal structure (Maass et al., 2002; Jaeger and Haas, 2004). The importance of this balance and other general issues encountered with on-line training of recurrent networks are elegantly summarized in Sussillo and Abbott (2009), a work which sets the state of the art of the field.

1.3.3 Attractor networks of spiking neurons

The concepts developed by Attractor Neural Network theory reach their highest points in the implementation of networks of realistic spiking neurons. The development of sophisticated mean-field theories and of a Fokker-Planck formalism made it possible to control the activity of these complicated systems by providing an indispensable toolbox to tune their parameters (Fusi and Mattia, 1999). This opened the way to the investigation of several interesting regimes arising through the interaction of large networks of spiking neurons, like for instance spontaneous activity (Amit and Brunel, 1997), global oscillations (Brunel and Hakim, 1999a), and elevated persistent activity (Wang, 1999). It also became possible to explore the functional effect on the global neural activity of realistic synaptic properties like neuromodulation, slow synaptic currents (Brunel and Wang, 2001), and short-term synaptic plasticity (Mongillo et al., 2005) and their putative influence on simplified

models of cognitive phenomena like working memory (Compte et al., 2000; Mongillo et al., 2008), and decision-making (Wang, 2002).

1.3.4 Computational advantages of randomness

Another theoretical setting which has to be mentioned for inspiring some of the ideas in this dissertation is Compressed Sensing theory, a recent signal reconstruction theory bearing the names of Emmanuel Candès, David Donoho, Justin Romberg, Terence Tao, Michael Wakin. Compressed Sensing (CS) exploits the fact that many interesting signals possess a structure which renders them very sparse when represented in the proper basis. While classical information theory would encode these signals by sampling at a rate above the Nyquist-Shannon criterion to subsequently compress the signal, CS proceeds by directly sampling the signal at a low rate by projecting it onto a basis which is maximally incoherent with respect to the natural sparse basis of the signal, in order to obtain a representation which is already compressed. This allows high-resolution acquisition with low-resolution sensors. CS works because, if there exists a basis in which the signal is sparse, then the product of an incoherent sensing matrix with the matrix of basis vectors is roughly orthogonal when restricted to the space of sparse signals (the so-called Restricted Isometry Property (Candes and Tao, 2005)). Since orthogonal means invertible, the signal can be exactly reconstructed. The reconstruction technique doesn't involve inversion, though, but rather a minimization extremizing the sparseness of the signal. At this point CS exploits properties discovered by Gelfand relating the l_0 norm (the norm which is minimized by sparseness, but it's difficult to handle) to the l_1 norm (a norm which is easy to minimize with convex minimization techniques).

One remarkable fact about CS is that the Restricted Isometry Property is

almost surely satisfied by several classes of random matrices drawn from a so-called Johnson-Lindenstrauss-favorable distribution (i.e., satisfying the premises of the Johnson-Lindenstrauss lemma, (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2002)). The surprising consequence is that sensing through random matrices via random projections is a data acquisition technique which is incredibly efficient. First of all, it is very *simple* to implement. Second, it is *universal*, in the sense that the same random projections can be used for any compressible signal class, which means in particular that it's a technique that can be implemented in a *non-adaptive* way. Additionally, it's very robust, as each measurement is equivalent to every other measurement and carries the same amount of information.

1.4 Final remarks

The theoretical approaches which were mentioned in this introduction all lent something which inspired the rest of this dissertation. The framework of Attractor Neural Networks is for instance the starting point of our attempt to lay down a formal framework for the active maintenance of neural activity patterns encoding mental states and goal-directed responses. We will show that attractors can only be combined with the necessary flexibility to carry out context-dependent behavior if an extensive fraction of neurons displays *mixed selectivity*. This type of neural selectivity can be understood in the perspective of Support Vector Machines (SVM) as an embedding of the neural response in a higher dimensional space, which essentially increases the modality of response of the network. The training procedure that we use to build such networks is also in part borrowed from the SVM program as well as from the insights gained through the Perceptron learning rule. Moreover,

the selectivity pattern of the *mixed selective* neurons will be shown to be conveniently accommodated by a network with random connectivity. In this sense, we borrow the idea of non-adaptive filters from the setting of Compressed Sensing and Random Projections.

Chapter 2

Neural substrate for rules representation [1]

The brain works in a completely different computational modality than modern computers. The components of a computer are orderly and efficiently organized and each element is engineered to perform a specific task. On the other hand, the computational elements of a brain, the neurons, display a high level of functional heterogeneity. Neurons can respond to the most diverse combinations of sensory stimuli and events, they can encode recent memories, intentions, emotions, motivation, and, more in general, the information which is pertinent to the execution of a behaviorally relevant task.

What is the computational advantage of using such an heterogeneous system? In this chapter we will argue that this kind of response diversity is necessary for the execution of flexible rule-based behavior. In particular, we will show that the brain needs neurons displaying *mixed selectivity* to sensory stimuli and inner mental states in order to execute context-dependent tasks. Mixed selectivity can be

[1] This chapter is an extension of the following reference:
Rigotti, M., Ben Dayan Rubin, D., Wang, X.-J. & Fusi, S.,
Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses (Submitted to Frontiers in Computational Neuroscience)

obtained by designing and training a specific neural circuit, as done in traditional neural networks. Alternatively, we will show that, surprisingly, random synaptic connectivity naturally provides the functional diversity that is needed for arbitrarily complex rule-based tasks. Apart from being computationally inexpensive, this way of endowing a neural network with mixed selectivity has the advantage of providing it with an heterogeneous pool of neurons which are selective to features of the task that, even if not currently relevant, could become important in future behavioral circumstances. We will argue that the availability of such a pool of functional diversity allows the neural system to accommodate stability of acquired neural representations by preserving at the same time the flexibility to potentially acquire new ones. Moreover, we will show that neural diversity can play a crucial role in accelerating the learning of new association and the acquisitions of new tasks.

2.1 Introduction

Neurons in the mammalian brain are highly heterogeneous (Soltesz, 2005; Marder and Goaillard, 2006) and show diverse responses to sensory stimuli and other events. This diversity is especially bewildering with regard to the prefrontal cortex, a brain structure that has been shown to be critically important for higher cognitive behaviors in numerous lesion (Petrides, 1982; Passingham, 1993; Murray et al., 2000), clinical (Petrides, 1985), and imaging (Boettiger and D'Esposito, 2005) studies. Indeed, single-neuron experiments from the prefrontal cortex have yielded a rich phenomenology: neurons have been found to respond to sensory stimuli and show persistent activity during working memory (Fuster and Alexander, 1971; Funahashi et al., 1989; Miller et al., 1996; Romo et al., 1999), reflect animals' decisions or intended actions (Tanji and Hoshi, 2008) and causally bias them (Opris et al., 2005), and encode contexts, task rules (Wallis et al., 2001; Genovesio et al., 2005; Mansouri et al., 2006; 2007) and abstract concepts like numbers (Nieder and Miller, 2003). Typically, a single prefrontal cell is not merely responsive to a single event but shows selectivity to a combination of different aspects of the task being executed, a response property which we will call *mixed selectivity*. These findings naturally pose the question, whether such diversity of responses plays a constructive computational role in complex cognitive tasks.

We found a possible answer in attempting to build a biologically plausible model of a neural circuit that can implement rule guided behavior. Rules are prescribed guides for problem solving and flexible decision making and they vary in the degree of abstraction. Examples include arbitrary conditional sensorimotor associations (if red light, then stop), task rules (respond if two stimuli match),

strategies for decision making (if win, stay; if lose, switch). We assumed that the rule in effect is actively maintained by a recurrent neural circuit (Miller and Cohen, 2001). In particular, we hypothesized that the neural correlate of a rule is a self-sustained persistent pattern of activity, which was additionally required to be stable to small perturbations that were assumed to be dampened by the neural interaction. Attractor Neural Networks (ANN) theory naturally and straightforwardly incorporates these dynamics features. Attractor network models have been previously studied for associative (Hopfield, 1982) and working memory (Amit, 1989; Wang, 2001) of sensory stimuli. In these models a sensory stimulus activates one of the strongly interacting populations of neurons and the memory of stimulus identity is maintained by the persistent activity of the activated population.

Our intention was to extend these models so that every attractor encodes a specific cognitive state and the rules to select other cognitive states on the basis of the external stimuli, generalizing in this way approaches in which an attractor merely codes for a set of response decisions (see e.g. Amit (1988); O'Reilly and Munakata (2000); Wang (2001); Loh and Deco (2005)). Concretely, we assumed that, starting from a pattern of stable activity, every event generates a driving force that steers the neural activity toward a different stable pattern, in a way which depends on both the external event and the previous rule in effect.

As we will show, whenever the rules for committing the course of action contain a dependence on the context, such a scenario cannot in general be implemented with a realistic neural network lacking neurons with mixed selectivity. An example of such a situation where mixed selectivity is indispensable to cope with context-dependence is offered by the Wisconsin Card Sorting Test (WCST), a neuropsychological test commonly used to assess flexibility in the face of changing reinforce-

ment schedules and crucially relying on frontal lobe integrity. Our framework gives therefore some hints as to why the WCST is such a successful clinical test to asses damages of the prefrontal cortex (Milner, 1963), an area which is characterized by a large proportion of cells displaying mixed selectivity (see e.g. Asaad et al. (1998)).

We will then show that neurons with mixed selectivity and diverse response properties not only are necessary in our scenario to perform context dependent tasks, but they are also sufficient to solve arbitrarily complicated tasks. Mixed selectivity is readily obtained by connecting cells with random connections to both the neurons in the recurrent circuit and to the neurons representing the external events. Surprisingly, it turns out that the number of randomly connected neurons needed to implement a particular task is not much larger than the minimal number of neurons required in an ad hoc designed neural circuit. Moreover, randomly connected neurons possess response properties that are more diverse than required in a minimal circuit, as they respond to both necessary and unnecessary combinations of mental states and events. Such response properties are predicted to be pre-existent and universal in the sense that they should be observable before the learning process, and independently from the task to be learned. Our work suggests that the observed diversity of the neural responses plays an important computational role in any task in which our decision or our actions depend on the context.

2.2 Results

2.2.1 Modeling complex cognitive tasks: the general framework

Rule-based tasks as event-driven transitions between mental states

Our framework is based on the assumption that subjects performing rule-based tasks go through a series of inner mental states, each representing an actively maintained disposition to behavior. Each state contains information about task-relevant past events and internal cognitive processes representing reactivated memories, emotions, intentions and decisions, and in general all factors that will determine or affect the current or future behavior. Some of the states may express the execution of motor acts. A particular behavioral rule is therefore simply represented as a relationship between mental states, in term of event-driven transitions between them.

Fig. 2.1A illustrates this scenario in the case of a simplified version of the Wisconsin Card Sorting Test (WCST). In a typical trial, the subject sees a sample stimulus on a screen and, after a delay, two test stimuli are presented. The subject is required to touch the test stimulus matching either the shape or the color of the sample, depending on the rule in effect. The rule can be determined only by trial and error; a reward confirms that the rule was correct, while an error signal prompts the subject to switch to the alternative rule. Every task-relevant event like the appearance of a visual stimulus or the delivery of reward is hypothesized to induce a transition to a different mental state.

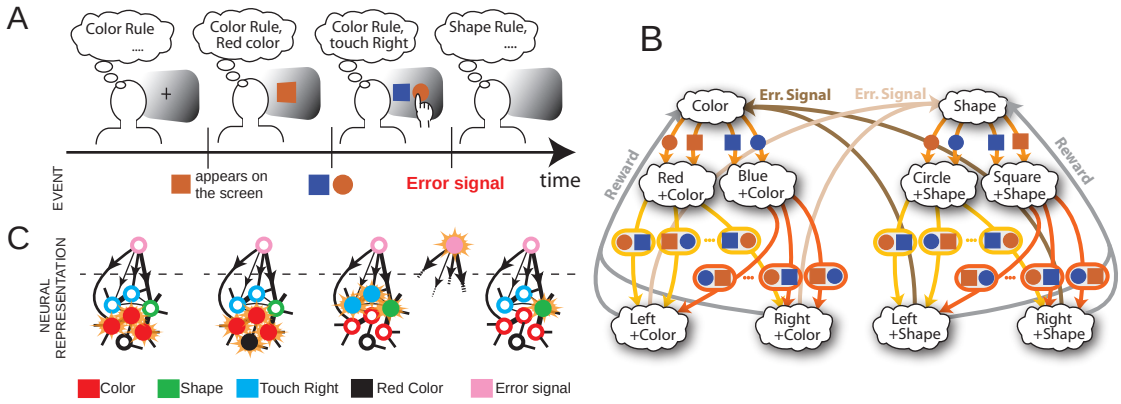

FIG. 2.1: A context-dependent task. **A**, A typical trial of a simplified version of the WCST, similar to the one used in the monkey experiment (Mansouri et al., 2006; 2007). The subject has to classify visual stimuli either according to their shape or according to their color. Before the trial starts, the subject keeps actively in mind the rule in effect (color or shape). Every event, like the appearance of a visual stimulus, modifies the mental state of the subject. An error signal indicates that it is necessary to switch to the alternative rule. **B**, Scheme of mental states (thought balloons) and event-driven transitions (arrows) that enables the subject to perform the simplified WCST. **C**, Neural representation of the mental states shown in A: circles represent neurons, and colors denote their response preferences (e.g. red units respond when Color Rule is in effect). Filled circles are active neurons and black lines are synaptic connections. For simplicity, not all neurons and synaptic connections are drawn.

Neural correlate of mental states

The neural representation of a mental state is assumed to be a stable pattern of activity of a recurrent neural circuit. The same neural circuit can sustain multiple stable patterns corresponding to different mental states. Every event selects one of these predefined patterns, and hence activates one particular mental state. In particular we assume that events like sensory stimuli, reward, or error signal steer the neural activity toward a different stable pattern representing a new mental state. Such a pattern will in general depend on both the external event and the previous mental state. Conversely, every stable pattern of activity implicitly specifies the

relationship with other patterns by encoding which external events will initiate a transition towards them, given the present activity.

Attractor Neural Networks (ANN) offer the ideal theoretical framework within which the ideas that we just sketched can be developed. In particular ANN theory (Amit, 1989) offers a way to formulate the mathematical conditions that the synaptic couplings of a recurrent network have to satisfy in order to accommodate stable patterns of sustained activity. Such activity patterns can in fact be obtained as attractor states of the neural dynamics.

Interestingly, we found that the mathematical conditions which have to be satisfied by the synapses to realize attractors are incompatible with those which would implement arbitrary context-dependent event-driven transitions. As we will show, this turns out to be a very general obstacle that prevents the implementation of even simple context-dependent tasks, and which is actually rooted in the geometrical concept of non-linear separability (Minsky and Papert, 1969) as it was already investigated in the case of pattern completion in semantic networks (Hinton, 1981).

2.2.2 Fundamental difficulties in context-dependent tasks

To illustrate the problem caused by context-dependence, consider a task switching induced by an error signal in the simplified WCST (see Fig. 2.2A). The simplified WCST encompasses two general behavioral contexts corresponding to the *Color* and *Shape Rules*. Every rule is represented by the persistent activation of a specific pool of neurons. The first panel of Fig. 2.2A illustrates this situation and points out that, in order for a neuron encoding *Color Rule* (red) to be active, the total afferent recurrent synaptic input ha to be above its firing threshold (indicated with an arrow labeled with '+'). On the other hand, in order for the network to be

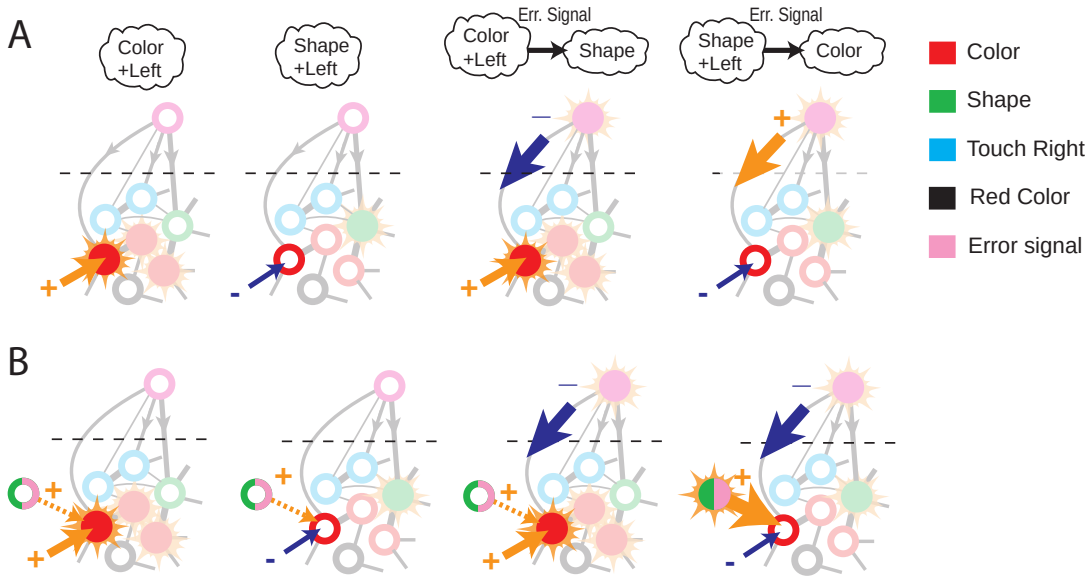

FIG. 2.2: **A**, Impossibility of implementing a context-dependent task in the absence of mixed selectivity neurons. We focus on one neuron encoding *Color Rule* (red). In the attractors (two panels on the left), the total recurrent synaptic current (arrow) should be excitatory when the *Color Rule* neuron is active, inhibitory otherwise. The implementation of a rule switching by the *Error Signal* neuron (pink) now faces the problem that the same external input should be inhibitory (dark blue) when starting from *Color Rule* and excitatory (orange) otherwise (two panels on the right). **B**, The effect of an additional neuron with mixed selectivity that responds to the *Error Signal* only when starting from *Shape Rule*. Its activity does not affect the attractors (two panels on the left), but it excites *Color Rule* neurons when switching from *Shape Rule* upon an *Error Signal*. In the presence of the mixed selectivity neurons, the current generated by the *Error Signal* can be chosen to be consistently inhibitory.

encoding the *Shape Rule* (second panel of Fig. 2.2A), the recurrent synaptic input onto the *Color Rule* neurons has to be below the firing threshold (indicated with an arrow labeled with '-'). Now, in order to induce a rule switch, the additional synaptic input generated by the error signal is required to be inhibitory enough to overcome the recurrent input and inactivate the *Color Rule* neurons when starting from the *Color Rule* mental state, and excitatory enough to activate them when starting from the *Shape Rule* state (Fig. 2.2A, second and third panel). This is

impossible to realize because the neural representation of the *Error Signal* is the same in the two contextual cases.

This kind of problems is encountered whenever the same external event must activate a neural population in one context, and inactivate it in another. It is interesting to point out that these situations have a straightforward geometrical interpretation in terms of non-linear separability of vectors. In particular, the scenario exemplified in Fig. 2.2A is the *exclusive (XOR) problem* in disguise, Minsky and Paperts workhorse in their critics against Rosenblatt's perceptron. All these technical considerations will be developed in Appendix B.1, where we will also show that the probability of not encountering such a problem decreases exponentially with the number of transitions and with the number of neurons in the network, if the patterns of activities representing the mental states are random and uncorrelated. This result indicates that it is very likely to encounter this problem every time our action or, more in general, our next mental state, depends on the context, unless the neural representations of the external events and the inner mental states are carefully chosen. On a positive note, we will show in the next sections that neurons with mixed selectivity solve the problem in the most general case and for any neural representation.

2.2.3 The importance of mixed selectivity

The main problem of the example illustrated in Fig. 2.2A arises from the fact that each neuron is selective either to the inner mental state (*Color* or *Shape Rule*) or to the external input (such as the *Error Signal*). Consider indeed if, on the other hand, we introduce an additional neuron with a more complex selectivity which responds to the *Error Signal* only when the neural circuit is in the state

corresponding to the *Shape Rule*. Such a neuron exhibits mixed selectivity as it is sensitive to both the inner mental state and the external input. Its average activity is higher in the trials in which *Shape Rule* is in effect compared to the average activity in *Color Rule* trials. In particular, the average activity in time intervals during and preceding the *Error Signal* is higher when starting from *Shape Rule* than when starting from *Color Rule*. At the same time it is also selective to the *Error Signal* when we average across the two initial inner mental states corresponding to *Color* and *Shape Rule*. Neurons with such selectivity are commonly observed in prefrontal areas (see Section 2.1) and from Fig. 2.2B one can readily see that their participation in the network dynamics solves the context-dependence problem. The mixed selectivity neuron just introduced is indeed inactive in the absence of external events, and therefore does not affect the mental state dynamics in the first two panels of Fig. 2.2B. However, it responds differently depending on the initial state preceding an *Error Signal* transition. This allows us to design the circuit in such a way that the *Error Signal* is consistently inhibitory. In this way, when starting from *Color Rule*, the external input inactivates the *Color* neurons, as required to induce a transition to the *Shape Rule* state. When starting from the *Shape Rule*, the mixed selectivity neuron is activated by the *Error Signal* and its excitatory output to the *Color* neuron can overcome the inhibitory current of the *Error Signal* and activate the *Color* neuron. Notice that it is possible to find analogous solutions every time the neuron has mixed selectivity to the *Error Signal* and to the rule in effect. A mixed selectivity neuron which would be active only at the presentation of the *Error Signal* in the *Color Rule* could be for instance chosen to project an inhibitory current to the *Color* neuron to perform the *Color Rule* $\rightarrow$*Shape Rule* transition, while the opposite transition can be implemented

by a positive synapse from the *Error Signal* population. A development of these considerations will be presented in Appendix B.2 for general tasks).

2.2.4 Randomly connected neurons exhibit mixed selectivity

Creating mixed selectivity by training multi-layer networks

Most of the known neural networks able to overcome the non-linear separability problem rely on the presence of intermediate layers of hidden units. The downside of increasing the system with hidden layers is that it considerably complicates the training procedure. Although several learning algorithm are known to train multi-layer networks (Hertz et al., 1991; Hinton, 2007), there is in general no guarantee that the they will converge to a solution. In fact algorithms like back-propagation (Rumelhart et al., 1986),which are essentially gradient descent procedures, cannot avoid to get stuck in local minima of the error function they try to minimize. This makes them very sensitive to the initial condition of the synaptic weights, which, please notice, is often taken to be random. Notice also that, in perspective to what was discussed in the previous section, the scope of these convoluted algorithms can be thought of being the creation of mixed selectivity so that it can be utilized by the output units of the network. Rather than seeing it as a byproduct of a complex learning algorithm (Zipser and Andersen, 1988), it may therefore be convenient to think of mixed selectivity as a functional building block to achieve a desired neural construction. But is it possible to obtain such building blocks without having to go through a lengthy training procedure?

Creating mixed selectivity with random connectivity

We found that there is a simple and surprisingly general way to create mixed selectivity to solve the context-dependence problem, which does not require any training. It is based on the observation that mixed selectivity is naturally exhibited by neurons which receive inputs from the recurrent network and the external neurons with random synaptic weights (Randomly Connected Neurons, or RCNs).

Let us see how the participation of RCNs in the network dynamics is sufficient to implement the scheme of attractors and transitions corresponding to the rule switch in Fig. 2.2. Consider a neuron receiving random synaptic inputs from the recurrent neurons encoding the rule and from the external neurons. In order for such an RCN to have the same selectivity as the mixed selectivity neuron of Fig. 2.2B, it is sufficient that synaptic inputs from the *Shape Rule* recurrent neurons and the *Error Signal* external neurons sum to a contribution which is above the firing threshold. Moreover, the synaptic input in the other cases has to be lower than the threshold. As we will show in Appendix B.3, it turns out that the probability that an RCN responds as a mixed selectivity neuron useful to solve the problem of Figure 2.2 can be as large as $1/3$ when the firing threshold θ is centered around the mean of the random synaptic inputs. Surprisingly, this result implies that the number of RCNs needed to solve a context-dependent problem is on average only three times larger than the number of neurons needed in a purposely designed neural circuit.

Firing threshold and coding level

As just pointed out, the fraction of RCNs which solve the non-linear separability problem corresponding to a given context-dependence depends on the firing threshold θ of the RCNs. This is simply because θ determines the selectivity of the RCNs

by determining the fraction of pre-synaptic inputs to which the RCNs will respond to. Say for instance that the RCNs see a distribution of presynaptic inputs centered evenly around 0. In this case, if the firing threshold is $\theta = 0$, every RCN would respond on average to half of all possible input patterns, as the total synaptic current is either positive or negative with equal probability. We refer to this situation as the *dense coding* regime, since in this case an extensive number of randomly connected neurons is active at every time. If the firing threshold θ increases, every RCN will tend to respond to a progressively decreasing fraction f of input patterns. We refer to this case as the *sparse coding* regime. Notice that, because of the symmetry implicit in our formalism (we can code information either with active or inactive neurons), we refer to the coding level as being low in both cases, when the fraction of active neurons is very low ($f \to 0$), or very high ($f \to 1$). This means that we have the highest coding level for $f = 1/2$, that is when a neuron responds to half of the possible inputs.

Coding level and RCN capacity

RCN patterns of activities corresponding to dense coding ($f = 1/2$) are more efficient than sparse representations ($f \to 0$ or $f \to 1$), regardless of the coding level f_0 of the representations of the mental states and the external inputs (f_0 is defined as the fraction of active neurons in the recurrent and the external network). This is proved in Appendix B.3 and illustrated in Fig. 2.3B where the probability that an RCN responds as a mixed selectivity neuron is plotted against f for three values of f_0. The proof is valid for patterns representing mental states and events that are random and uncorrelated. Notice that all curves of Fig. 2.3B have a maximum in the probability in correspondence of $f = 1/2$, around which the plots are relatively

flat for a wide range of f values. The maximum decreases gently as f_0 approaches zero (approximately as $\sqrt{f_0}$) because the overlap between different mental states and external inputs progressively increases, and this makes it difficult for an RCN to discriminate between different initial mental states, or different external inputs. For the same reasons, the maximum decreases in the same way as f_0 tends to 1.

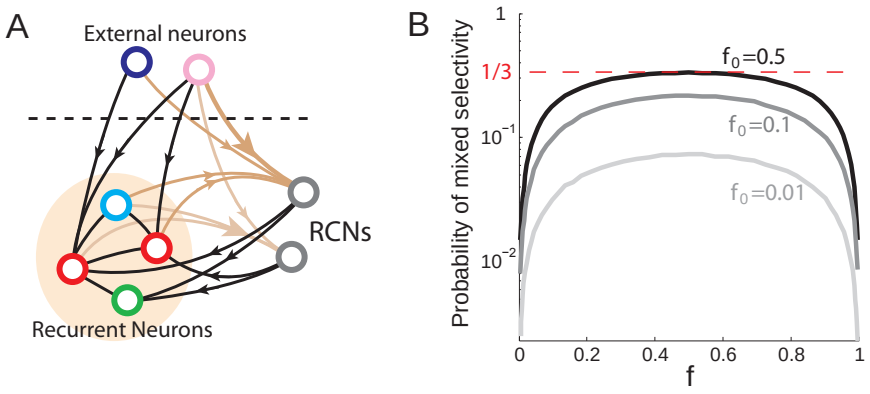

FIG. 2.3: **A**, Neural network architecture: randomly connected neurons (RCN) receive fixed random synapses (brown) from both the recurrent and the external neurons. Each RCN projects back to the recurrent network by means of plastic synapses (black). Not all connections are shown. **B**, Probability of an RCN displaying mixed selectivity (on log scale) and hence solving the problem of Fig. 2.2 as a function of f, the average fraction of input patterns to which each RCN responds. Different curves correspond to different coding levels f_0 of the representations of the mental states and the external inputs. The peak is always at $f = 1/2$ (dense RCN representations), and the curve is relatively flat for a wide range of f.

Effects of correlations on RCN capacity

As the patterns representing mental states and events become progressively more correlated, the number of needed RCNs increases. The case of correlated representations is treated analytically in in Appendix B.3, and the results are illustrated in Fig. 2.4. In particular, in Fig. 2.4A we show the probability of mixed selectivity as a function of f of the RCNs for different correlation levels between the patterns rep-

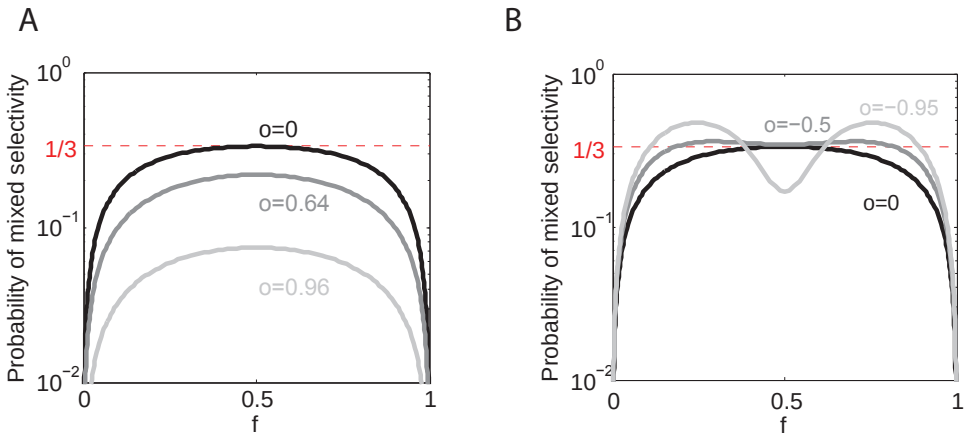

FIG. 2.4: **A**, Probability that an RCN has mixed selectivity as a function of f, for different positive values of the overlap o between the two initial mental states, and the two external inputs corresponding to the spontaneous activity and the event. Again the peak is always at $f = 1/2$ and the curve decays gently as o goes to 1. **C**, As in **A**, but for negative values of the overlap o. There are now two peaks that remain close to $f = 1/2$ for all values of o.

resenting mental states and external events. The degree of correlation is expressed as the average overlap o between the two patterns representing the initial mental states (the same overlap is used for the two external events). o varies between -1 and 1, and it is positive and close to 1 for highly similar patterns (Fig. 2.4A) or negative (Fig. 2.4B), for anti-correlated patterns. The value $o = 0$ corresponds to the case of random and uncorrelated patterns. As o increases, it becomes progressively more difficult to find an RCN that can have a differential response to the two initial mental states. This is reflected by a probability that decreases approximately as $\sqrt{1-o}$. For all curves plotted in Fig. 2.4A, the maximum is always in correspondence of $f = 1/2$. Interestingly, for anti-correlated patterns, the maximum splits in two maxima that are slightly above 1/3 (see Fig. 2.4B). The maxima initially move away from $f = 1/2$ as the patterns become more anti-correlated, but then, for $o < -5/6$, they move back toward the mid point. Notice in summary that the

optimal value is always in realized for a fairly high coding level, for f between 0.3 and 0.7.

In this section we analyzed the probability that an RCN solves a single, generic, context-dependence problem. How does the number of needed RCNs scale with the complexity of an arbitrary task with multiple context-dependencies? In order to answer this question, we first need to explicitly construct a neural circuit that harnesses RCNs to implement an arbitrary scheme of mental states and event-driven transitions.

2.2.5 A general procedure for constructing recurrent networks that implement rule-based tasks

In order to implement the ideas we just exposed and combine them with the classical attractor neural networks paradigm we propose a network architecture with the three following populations of McCulloch-Pitts neurons (see Fig. 2.3A):

1. External neurons representing events
2. Recurrent neurons encoding the mental state
3. Randomly connected neurons (RCNs).

The recurrent neurons receive inputs through plastic synaptic connections from all the neurons in the three populations. The RCNs receive connections from both the external and the recurrent neurons through synapses with fixed, Gauss distributed random weights. Given a scheme of mental states and event-driven transitions like the one of Fig. 2.1B, the weights of the plastic synaptic connections are modified according to a prescription that guarantees that the mental states are stable

patterns of activity (attractors) and that the events steer the activity toward the correct mental state. In particular, for each attractor encoding a mental state, and for each event-driven transition we modify the plastic synapses through an iterative process which is basically a perceptron learning rule with margin (Rosenblatt, 1962; Krauth and Mézard, 1987; Freund and Schapire, 1999). Such a procedure is illustrated in Fig. 2.5 where we reconsider the example of a transition from the mental state *Shape+Left* to *Color* induced by an *Error Signal*. As shown in Fig. 2.5A the idea is to start clamping the recurrent neurons to the pattern of activity corresponding to the initial state (*Shape+Left*). We then compute the ensuing activity of the RCNs. We isolate a in turn every recurrent neuron and we modify its afferent plastic synapses according to the perceptron learning rule Rosenblatt (1962) so that the total synaptic input drives the neuron to the desired activation state it should have at time $t + \Delta t$, after the transition has occurred. To achieve the temporal stationarity of the activity patterns representing the mental states we require that at time t these patterns reproduce themselves at time $t + \Delta t$ (see Fig. 2.5B). In order to additionally guarantee the stability of these patterns, we require that active neurons are driven by a current I that is significantly larger than the minimal threshold value θ (i.e. $I > \theta + d$, where d is known as a "learning margin"). Analogously, inactive neurons should be driven by a current $I < \theta - d$. When the procedure converges for all neurons, the patterns of activity corresponding to the mental states are cooperatively maintained in time through the synaptic interactions and are robust to perturbations.

All conditions corresponding to the mental states and the event-driven transitions can be imposed if there is a sufficient number of RCNs in the network. If it is not possible to satisfy all conditions simultaneously we keep adding RCNs and

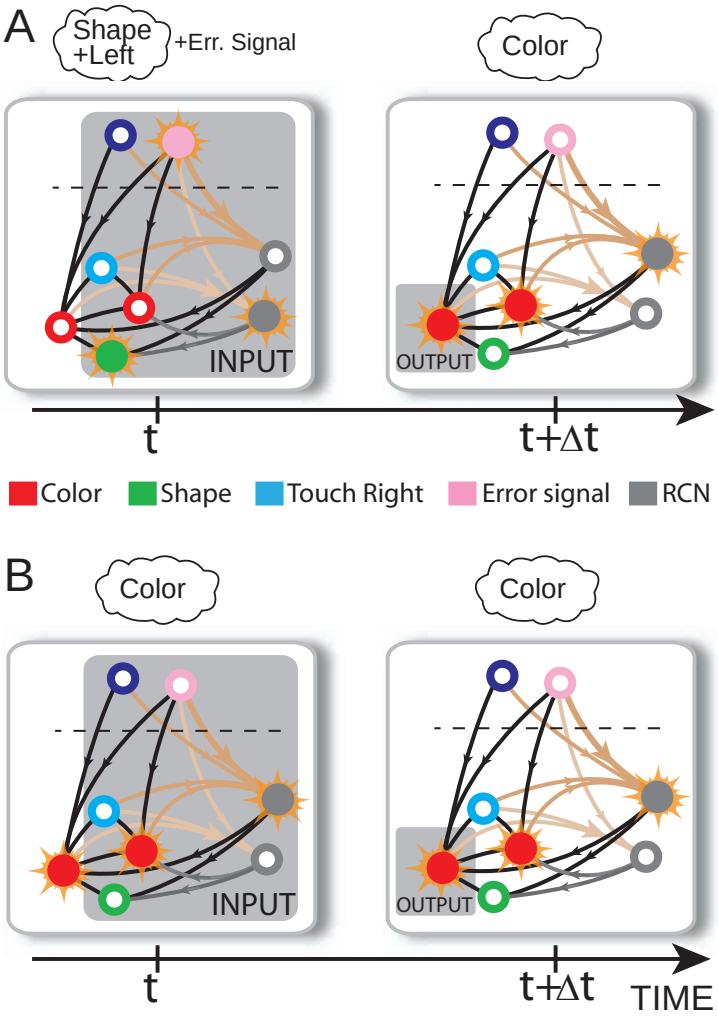

FIG. 2.5: Prescription for determining the plastic synaptic weights. **A**, For the event-driven transitions the synapses are modified as illustrated in the case of the transition from *Shape+Left* to *Color* induced by an *Error Signal*. The pattern of activity corresponding to the initial attractor (*Shape+Left*) is imposed to the network. Each neuron is in turn isolated (leftmost red neuron in this example), and its afferent synapses are modified so that the total synaptic current generated by the initial pattern of activity (time t, denoted by INPUT), drives the neuron to the desired activity in the target attractor (OUTPUT at time $t + \Delta t$). (**B**) For the mental states the initial and the target patterns are the same. The Fig. shows the case of the stable pattern representing the *Color* mental state. The procedure is repeated for every neuron and every condition.

we repeat the learning procedure. In Appendix B.3 we show that such a procedure is guaranteed to converge.

2.2.6 Dense neural representations require a number of neurons that grows only linearly with the number of mental states

We now turn to the question of the capacity of the architecture we sketched in the previous paragraph. That is, if we follow the given prescription to create a neural network, how many RCNs do we need in order to implement a given scheme of mental states and event-driven transitions?

Capacity in the sparse coding limit

The answer to this question strongly depends on the threshold θ for the activation of the RCNs, and hence on the RCNs' coding level f. In particular, in the extreme limit of small f (ultra sparse coding), each RCN responds only to a single, specific input pattern ($f = 1/2^N$, where 2^N is the total number of possible patterns, N being the number of synaptic inputs per RCN). We prove in Appendix B.3 that for the ultra-sparse case, in principel any arbitrarily scheme of attractors and transitions can be implemented, but the number of necessary RCNs grows exponentially with the number of recurrent and external neurons. Such a dependence reflects the combinatorial explosion of possible patterns of neural activity that represent possible conjunctions of events.

Capacity in the dense coding limit

On the other hand, with a larger f, it is more likely that an RCN solves our problem, as it was the case for the mixed selectivity neuron of Fig. 2.2B. What happens when we need to solve multiple correlated context-dependencies? We devised a benchmark to estimate how the number of necessary RCNs scales with f and the complexity of a context-dependent task. We simulated a network in which transitions between randomly selected mental states were all driven by a single event. The representations of the mental state attractors were random uncorrelated patterns, and they can be regarded as different contexts in which the same event can appear. Fig. 2.6A shows the required total number of neurons (recurrent and RCNs) as a function of the coding level f of the RCNs for three different numbers of contexts ($m = 5, 10, 20$). For each choice of m and f, the number of recurrent neurons and RCNs is increased in the same proportion until all transitions are implemented correctly and all attractors are stable and have a basin of attraction of a given size. The minimal number of required neurons is always obtained in correspondence of $f = 1/2$ (dense RCNs patterns of activity), consistently with the results of Fig. 2.3B that showed that the probability that an RCN solves a single context dependence problem is maximal for $f = 1/2$. Such a result is not trivial as we now require not only that multiple context-dependence problems are simultaneously solved, but also that the basins of attraction have at least given finite size.

Scaling in the dense coding limit

With $f = 1/2$, we examined how the minimal number of needed neurons depends on the task complexity, and in particular how it depends on the number of mental

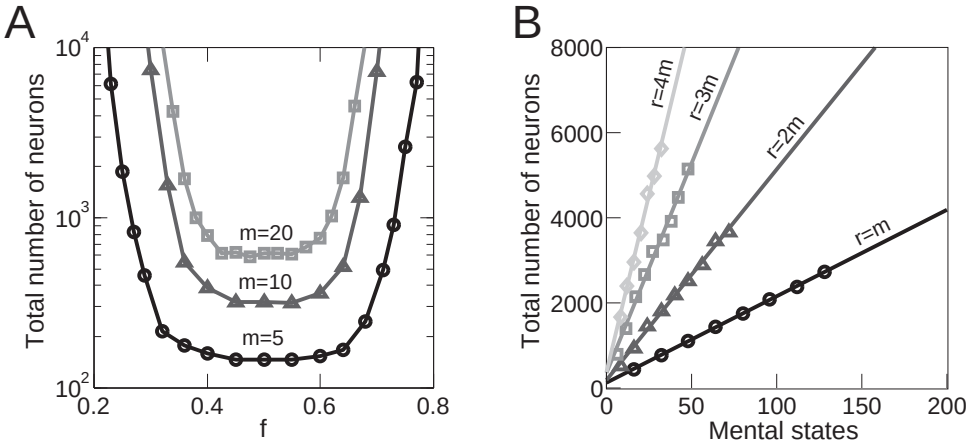

FIG. 2.6: **A**, Distributed/dense representations are the most efficient: total number of neurons (recurrent network neurons+RCNs) needed to implement m transitions between $2m$ random attractor states (internal mental states) as a function of f, the average fraction of inputs that activate each individual RCN. The minimal value is realized with $f = 1/2$. The three curves correspond to three different numbers of mental states m (5,10,20). The number of RCNs is $4/5$ of the total number of neurons. **B**, Total number of needed neurons to implement m random mental states and r transitions which are randomly chosen between mental states, with $f = 1/2$. The number of needed neurons grows linearly with m. Different curves correspond to different ratios between r and m.

states m (which in this analysis equals the number of events) and transitions r (Fig. 2.6B and Appendix B.4). Notice that for $r > m$, the same event drives more than one transition, which is what typically happens in context-dependent tasks. The total number of neurons increases linearly with m for all ratios r/m and the slope turns out to be proportional to the number of contexts in which each event can appear. This favorable scaling relation indicates that highly complicated schemes of attractor states and transitions can be implemented in a biological network with a relatively small number of neurons.

2.2.7 Modeling rule-based behavior observed in a monkey experiment

Simulation of a network performing a rule-based task

The prescription for building neuronal circuits that implement a given scheme of mental states and event-driven transitions is general, and it can be used for arbitrary schemes provided that there is a sufficient number of RCNs. To test our general theory with a concrete example, we applied our approach to a biologically realistic neural network model designed to perform a rule based task which is analogue to the WCST and it is described in Fig. 2.1 (Mansouri et al., 2006; 2007). We implemented a network of realistic rate-based model neurons with excitation mediated by AMPA and slow NMDA receptors, and inhibition mediated by $GABA_A$ receptors. Fig. 2.7A shows the simulated activities in two consecutive trials of two rule selective neurons and two motor response selective neurons. The rule in effect changes from *Color* to *Shape* just before the first trial, causing an erroneous response that is corrected in the second trial, after the switch to the alternative rule. Although the two rule selective neurons in Fig. 2.7A maintain their selectivity throughout the trial, their activity is modulated by other events appearing in the different epochs of the trials. This is due to the interaction with the other neurons in the recurrent network and with the RCNs. Fig. 2.7B shows the activity of three RCNs. These typically have a rich behavior exhibiting mixed selectivity that changes depending on the epoch (and hence on the mental state). Two features of the simulated neurons have already been observed in experiments: 1) neurons show rule selective activity in the inter-trial interval, as observed for significant fraction of cells in PFC (Mansouri et al., 2006). 2) the selectivity to rules is intermittent, or in other words, neurons

are selective to a different extent to the rules depending on the epoch of the trial.

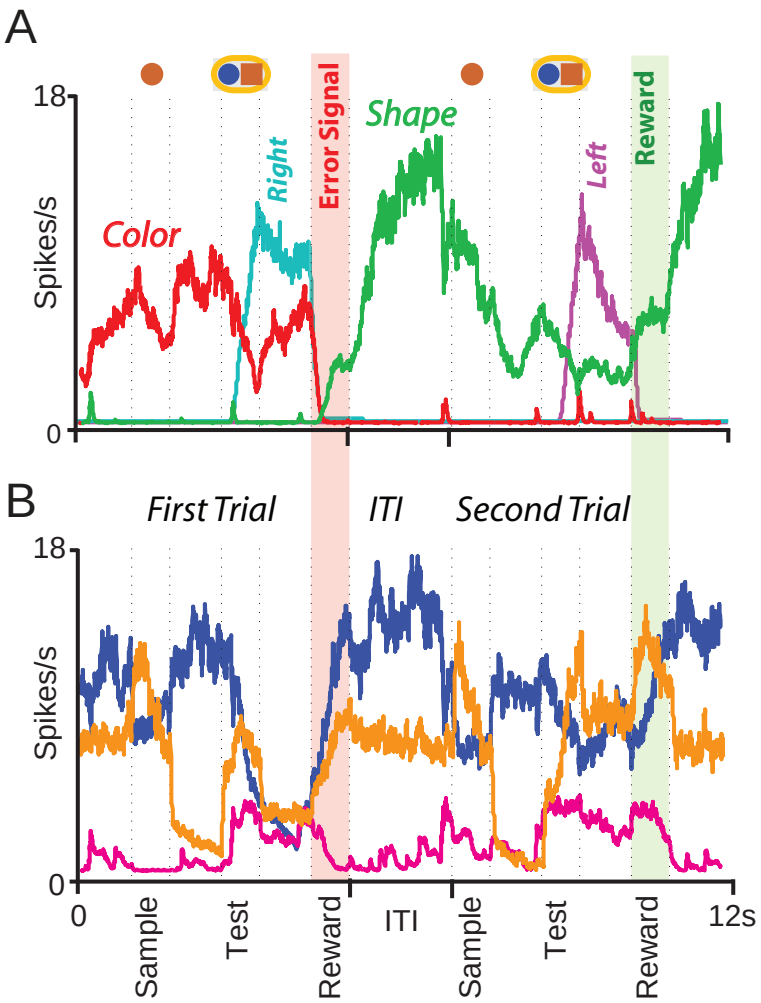

FIG. 2.7: Simulation of a network executing a Wisconsin Card Sorting-type Task. **A**, Simulated activity as a function of time of two sample neurons of the recurrent network which are rule selective, and two which express the motor response. The selectivities of the neurons are: "Color rule" (red trace), "Shape rule" (green trace), Right motor response (sea green trace), Left motor response (pink trace). The events and the mental states for some of the epochs of the two trials are reported above the traces. **B**, Simulated activity as in A, but for three RCNs.

Rule selectivity pattern

To analyze more systematically the selectivity of these cells and to compare it to what is observed in prefrontal cortex, in Fig. 2.8B we plotted for 70 cells whether they are significantly selective to the rule for every epoch of the trial. The cells are sorted according to rule selectivity in different epochs, starting from the neurons

that are rule selective in the inter-trial interval. Whenever a cell is rule selective in a particular epoch, we draw a black bar. In the absence of noise, all cells would be selective to the rule, as every mental state is characterized by a specific collective pattern of activity and the activity of each neuron is unlikely to be exactly the same for two different mental states. However, we consider a cell to be selective to the rule only if there are significant differences between the average activity in *Shape* trials and the average activity in *Color* trials. The results depend on the amount of noise in the simulated network, but the general features of selectivity described below remain the same for a wide range of noise levels.

The selectivity clearly changes over time, as the set of accessible mental states for which the activity is significantly different, changes depending on the epoch of the trial. This intermittent selectivity is also observed in the experimental data (Mansouri et al., 2006) reproduced in Fig. 2.8C.

Possible reasons for the discrepancy between simulations and experiment

There are some discrepancies between the analysis of the simulation (Fig. 2.8B) and the analysis of the experimental recordings (Fig. 2.8C). The simulation reports for instance the presence of neurons which maintain their selectivity to the rule throughout the trial, while the experimental selectivity seems to be more intermittent. One possible reasons for this difference is the higher level of noise to which the experimental preparation in subjected, because of the neural activity being estimated on a limited number of trials from spiking neurons. However, there might be a more profound reason for the discrepancy between experiments and simulations, which is related to the fact that the monkey might be using a strategy that is more complicated than the one represented in Fig. 2.1B. If, indeed, we assume that

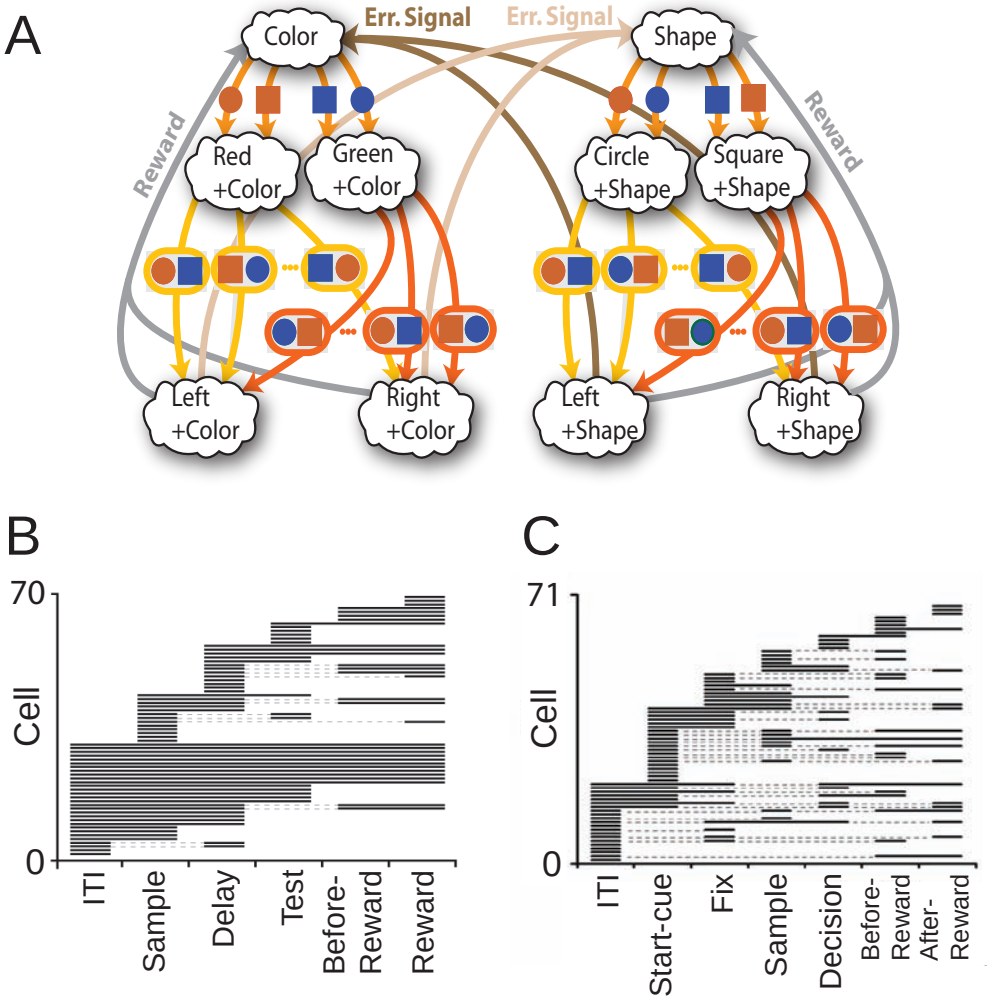

FIG. 2.8: **A**, Scheme of mental states and event-driven transitions for the simplified WCST (same as in Fig. 2.1B). **B**, Rule selectivity pattern for 70 simulated cells: for every trial epoch (abscissa) we plotted a black bar when the neuron had a significantly different activity in shape and in color blocks. The neurons are sorted according to the first trial epoch in which they show rule selectivity. **C**, Same analysis as in B, but for spiking activity of single-units recorded in prefrontal cortex of monkeys performing an analogue of the WCST of Mansouri et al. (2006).

the monkey keeps actively in mind not only the rule in effect, but also some other information about the previous trial that is not strictly essential for performing the task, then the number of accessible states during the inter-trial interval can be significantly larger, and this can strongly affect the selectivity pattern of Fig. 2.8B. This is illustrated in Fig. 2.9 where we assumed that the monkey remembers not only the rule in effect, but also the last correct choice (see e.g. Barraclough et al. (2004) for a task in which the activity recorded in PFC contains information about the last choice). In such a case the activity in the inter-trial interval is more variable from trial to trial and the pattern of selectivity resembles more closely the one observed in the experiment of Mansouri et al. (2006). These considerations simply point out how much of the observed neural variability may be due to the scheme of mental states which is implemented in executing rule-based behavior. When possible a behavioral analysis to establish the strategy employed by the subject is therefore of primary importance (see e.g. Gluck et al. (2002) for a study in which mathematical modeling has been employed to establish the strategy used by subjects to learn a probabilistic category learning task).

It may be also interesting to point out that the statistics of the pattern of selectivity in Fig. 2.8 and 2.9 also depends on the structure of the neural representations of the mental states and on the statistics of the random connections to the RCNs. In particular, the correlations between mental states can generate correlations between patterns of selectivity in different epochs, and across neurons.

Selectivity is not an inherent property of cells

One of the general features of our network which is independent of the implemented strategy or the neural representation of the mental states is the fact that rule

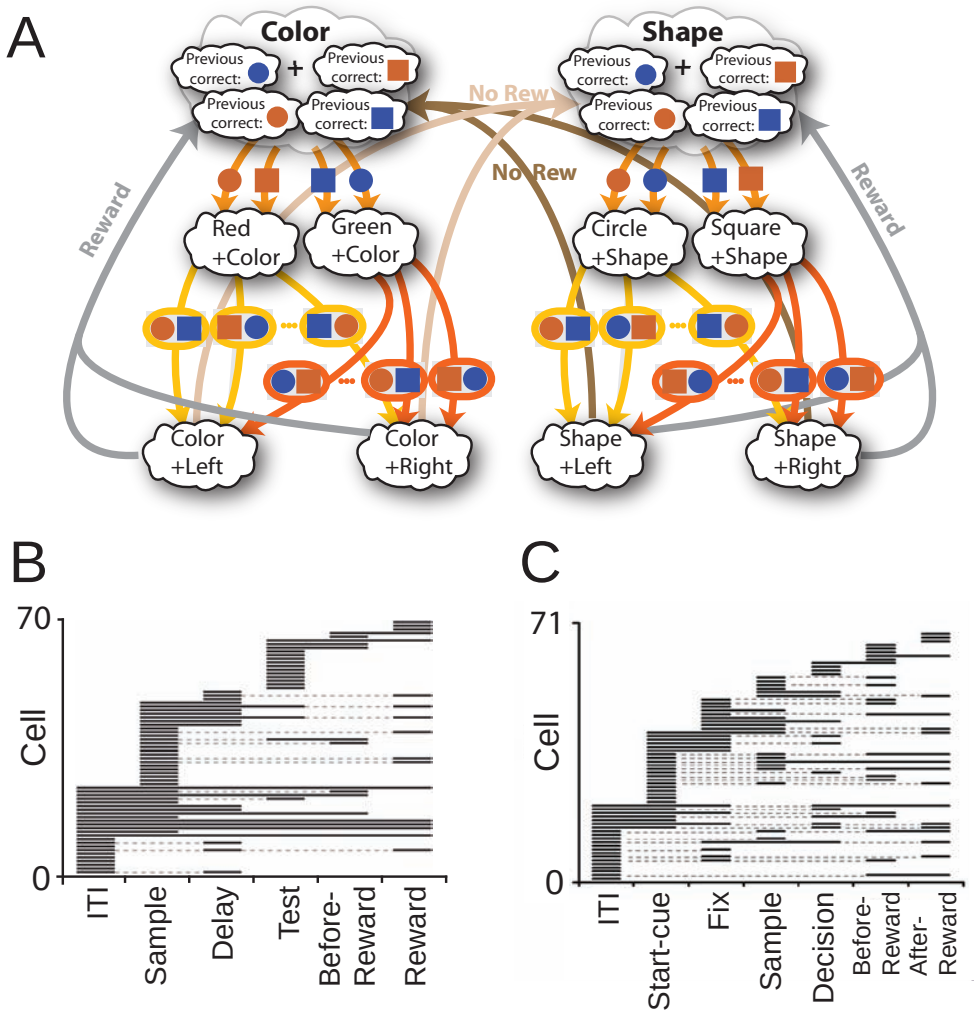

FIG. 2.9: **A**, A second scheme of mental states and event-driven transitions for the simplified WCST. The difference with respect to Fig. 2.1B and Fig. 2.8B is that with this slightly enlarged scheme the network has a memory of the test item which was correct to choose in the previous trial. **B**, Rule selectivity pattern for 70 simulated cells while executing the scheme in A: for every trial epoch (abscissa) we plotted a black bar when the neuron had a significantly different activity in shape and in color blocks. The neurons are sorted according to the first trial epoch in which they show rule selectivity. **C**, Same analysis as in B, but for spiking activity of single-units recorded in prefrontal cortex of monkeys performing an analogue of the WCST of Mansouri et al. (2006).

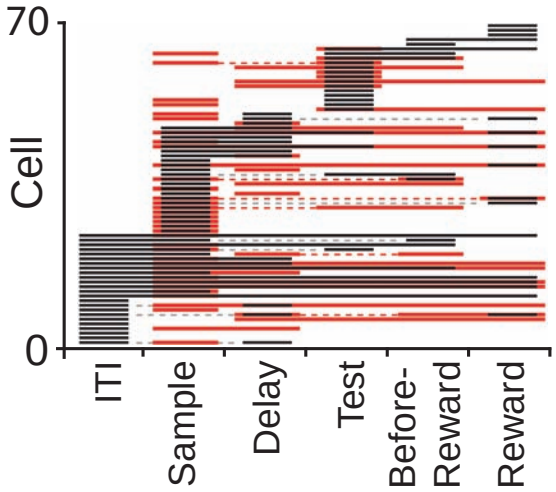

FIG. 2.10: Rule selectivity pattern of Figure2.9B (black bars) overlaid with the the result of the same analysis for selectivity to the color of the samples (red bars). Most of the neurons display multiple mixed selectivity to the rule and the color of the sample in intermittent way.

selectivity is not a property inherent to the cell. These is also the case for other types of selectivity, such as a stimulus feature or reward delivery as observed in the experiment of Mansouri et al. (2006). Particularly striking is the example of the selectivity to a concrete external feature, like the color of an item. Fig. 2.10 shows for the same cells of Fig. 2.9B the selectivity to the color of the sample stimulus (red bars), on top of the bars indicating rule selectivity. Obviously, there is no cell that is selective to the sample stimulus before it is presented (inter-trial interval), but in the remaining part of the trial the pattern of red bars seems to be as complex as the one for rule selectivity. Notice that some cells are selective to both the rule and the color in some epochs. This example illustrates cells commonly found in such a network, which show a complex pattern of mixed selectivity.

2.2.8 Modeling multiple tasks in monkey experiments

We know from our analysis of Fig. 2.6B that the number of mental states and transitions that can be implemented in a network with a biological size is very large as it scales linearly with the total number of neurons. This means that the same network can in principle implement a large number of different tasks, even in the case in which the subject is required to operate on the same sensory stimuli and to make decisions about the same types of actions. Our computational principle based on the mixed selectivity of RCNs is general and is only limited by the number of connected neural cells in the network and not by the complexity of the hierarchy of tasks to be performed.

A network executing a 'match'/'nonmatch' task and a WCST

To illustrate such a generality we show that it is straightforward to extend the network of Fig. 2.7 to perform also a 'match'/'nonmatch' task inspired by another monkey experiment (Wallis et al., 2001). In both tasks, the simulated network reacts to the same sensory stimuli, and responds with the same motor responses. The network is therefore required to face situations in which the information provided by the sole external stimuli is insufficient to determine the proper reaction, and it has therefore to be complemented with a working memory component.

In the 'match'/'nonmatch' task the subject is presented with a sensory cue (say a tone) indicating whether the trial will be performed on the basis of a 'match' or a 'nonmatch' rule. The subject then sees a a sample stimulus and, after a delay period, a test stimulus that is in half of the cases identical to the sample stimulus. If the rule in effect is 'match', the subject has to touch the screen when the test stimulus is the same as the sample. Otherwise the requirement is to withhold any

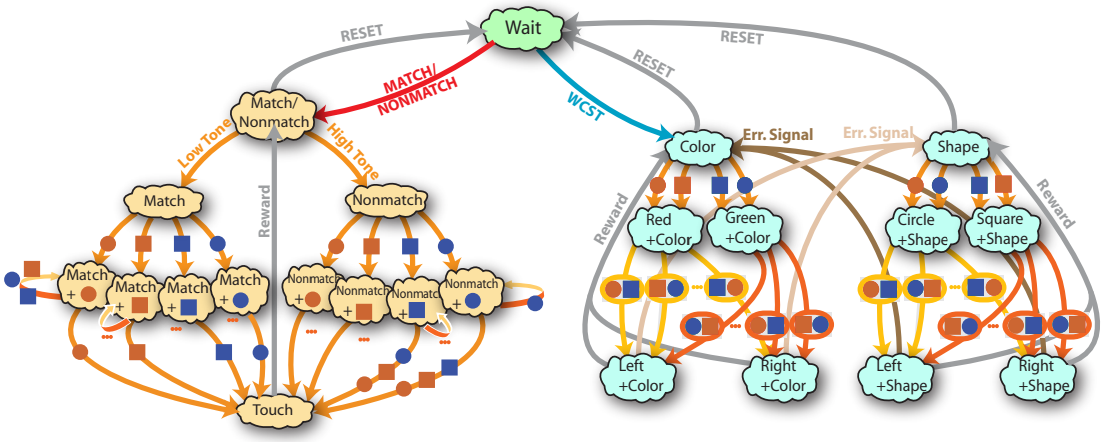

FIG. 2.11: Sets of mental states and transitions for two tasks: Simplified version of the WCST (Fig. 2.7) inspired by the experiment of Mansouri et al. (2006; 2007), and a 'match'/'nonmatch' task similar to an experiment presented in Wallis et al. (2001). Notice that both tasks are performed using the same items as stimuli.

action until the next matching stimulus appears. Analogously, in the 'nonmatch' case the subject has to touch the screen only when the stimulus differs from the sample stimulus. The scheme of mental states and transitions corresponding to a possible strategy to execute this task is reported in the left part of Fig. 2.11.

We implemented the procedure to build a network able to execute both the simplified version of the WCST and the 'match'/ 'nonmatch' task. Such a network was endowed with a recurrent population of neurons selective to the general context indicating which task had to be performed (either the simplified WCST or the 'match'/'nonmatch' task). In Fig. 2.12 we show the results of simulations similar to those presented in Fig. 2.7. The activity of a few recurrent neurons and three RCNs are plotted as a function of time for three consecutive trials. In the first and in the last trial the system performs the simplified WCST, whereas in the intermediate trial the system is instructed by an explicit cue to perform the 'match'/'nonmatch' task. Notice that the activity is again very variable in the different epochs, and

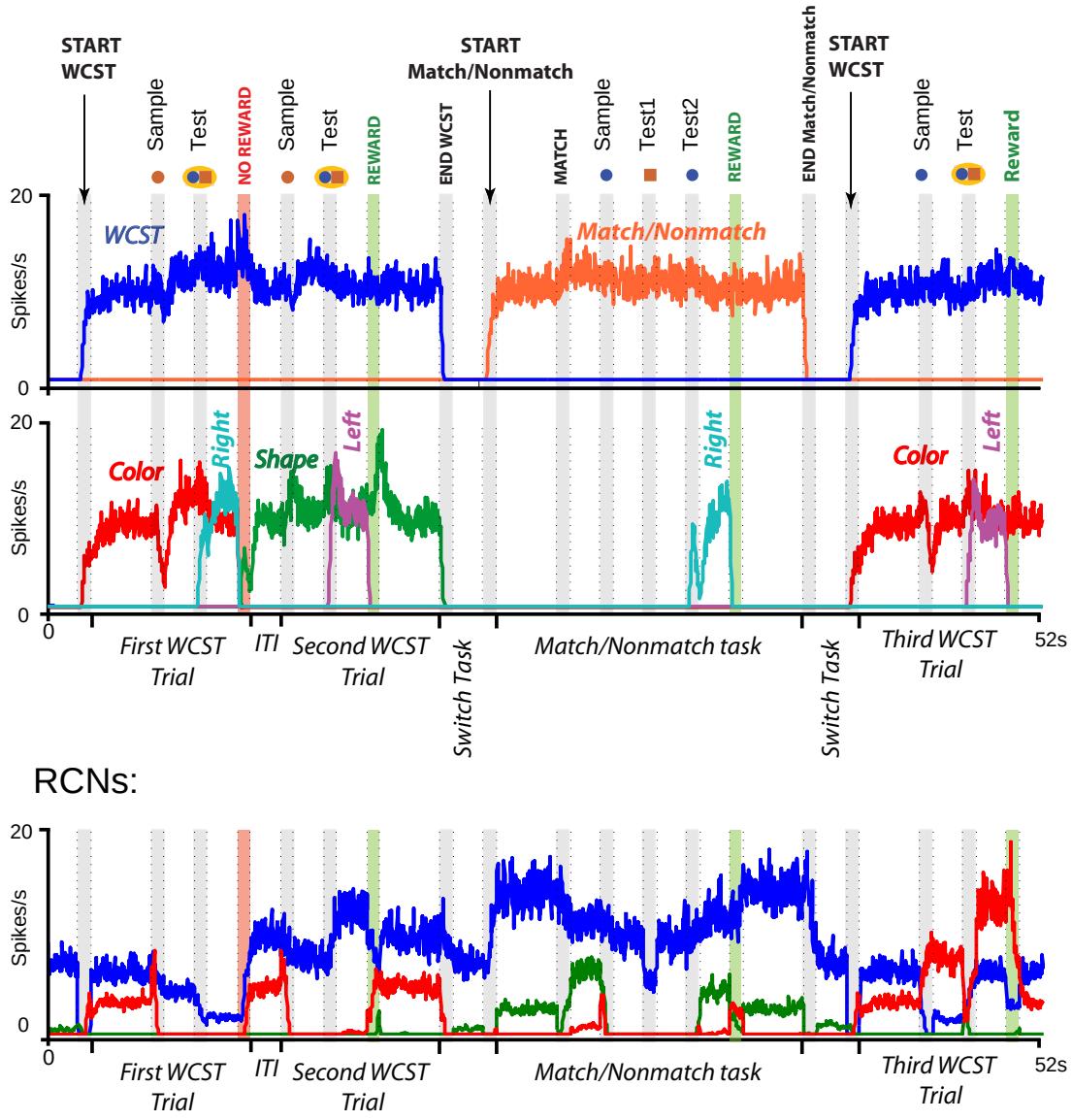

FIG. 2.12: Simulations of a network performing both the simplified version of the WCST and the 'match'/'nonmatch' task. The set of mental states and attractors is illustrated in Fig. 2.11. The activities of recurrent neurons and RCNs are plotted as in Fig.2.7A.

that, interestingly, the neurons encoding the higher order context (the task to be performed), are the most stationary and least variable. This was a general observation, in that neural populations which were required to consistently encode a feature over prolonged time-scales tended also to participate in the most stable attractors. This is due to the fact that such an activity has to be maintained in the face of several possible intervening stimuli, which our training procedure compensate by imposing an appropriate basin of attraction.

2.2.9 Basin of attraction and robustness to cells ablation

Apart from guaranteeing robustness to noise, stable attractors with a large basin of attraction are important to overcome the effect of "lesions" to the network in form of ablation of neurons. The way we're going to show this effect is simply by training a network to perform the WCST task and successively removing an increasing number of neurons from the system to see whether it can still perform at the same level.

We will see that, if the training had converged for a high enough stability parameter (quantified by the learning margin d of Section 2.2.5) the network is still be able to perform the task for mild but still considerable lesions (i.e. even after up to one third of the RCNs are removed).

Beyond this level of severity the performance in executing the task starts to decline. Remarkably, the first performance impairment appearing in the behavior of the network is a difficulty in rule switching, resulting in *perseverative errors* which are strikingly reminiscent of the cognitive symptomatology due to dorso-lateral frontal-lobe lesions in patients (Milner, 1963).

Severe cell ablation

We start by considering a naive instantiation of the network of Fig. 2.12, that is, one which hasn't yet been trained to perform the WCST and the 'match'/'nonmatch' task, although we know from the simulations in the previous section that it can learn both. We next train the network only to perform the WCST. In a second step we "lesion" the network by removing increasing amounts of RCNs, and verify the task performance. Fig. 2.13 shows the activity of the network before the lesions (Fig. 2.13A), after a removal of one third of the RCNs (Fig. 2.13B) and after the more severe removal of one half of the RCNs (Fig. 2.13C).

We will discuss the effect of this last more severe cell ablation (Fig. 2.13C) before the case of the mild cell ablation, as the latter displays a more subtle result, requiring a more extensive reflection.

Figure 2.13C illustrates a difficulty in switching to an alternative rule, the kind of impairment in task execution which is first observed at the behavioral level when ablating cells from the network. At the beginning of the trial in Fig. 2.13C the network is sustaining the attractor corresponding to the *Color rule* and, after the indicated sequence of sample and test stimuli, carries out the corresponding response. However, at the delivery of the *Error signal* signaling a rule change, the network is unable to perform the switch and its behavior results in a *perseverative error* in the following trial.

The reason for the impairment in the transition corresponding to the rule-switch resides in part in the temporal and spatial correlations of the neural representation of the task. The recurrent populations encoding the rule are in fact those whose activity is the most stable, because of the necessity to consistently represent a rule for a prolonged period of time. Additionally, populations encoding

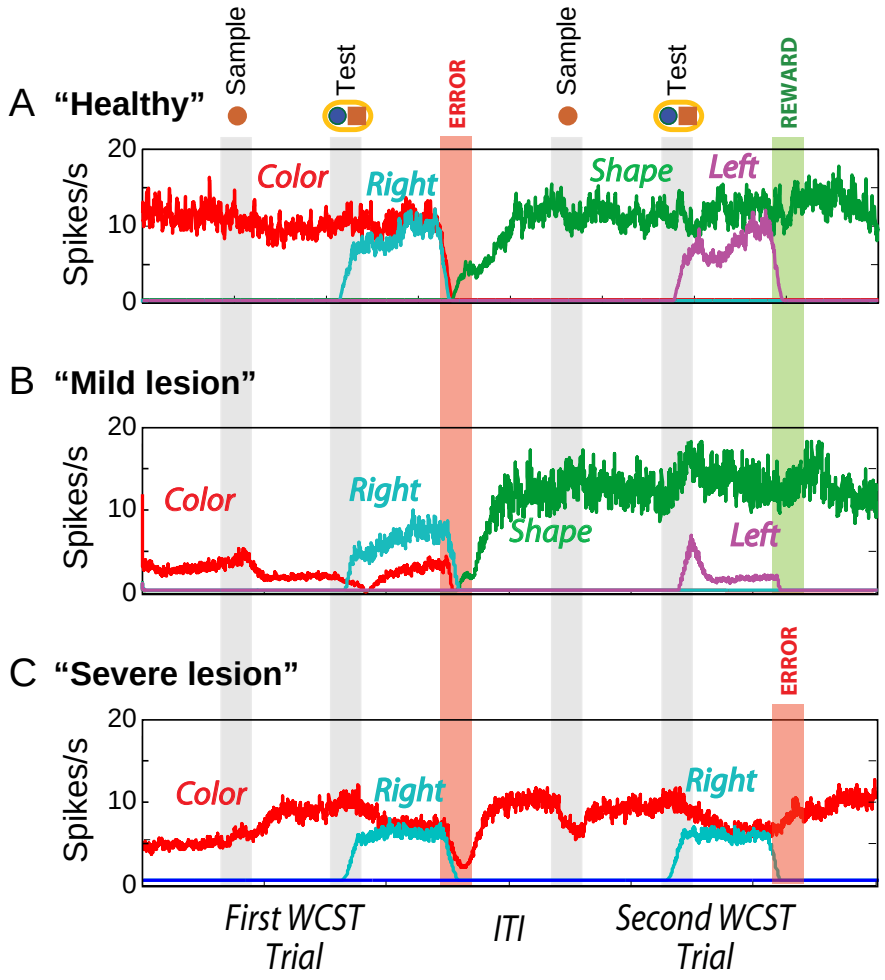

FIG. 2.13: Simulations of the network of Fig. 2.12 after it has been trained to perform only the WCST. We show the activity of neurons selective to the rule (color and shape) and to the motor responses (left and right) before ablation (**A**), after ablating $1/3$ of the RCNs (**B**), and after ablation of half of the RCNs

the opposite rules tend to suppress each other, meaning that the synaptic input implementing the switch has to suppress one rule and simultaneously activate the other. If either one of these mechanisms is weakened, the whole switch is impaired. The loss of RCNs will therefore affect the synaptic input corresponding to a rule switch more strongly than it will affect the stability of the rules themselves.

Another trivial reason of why the rule-switch is the first noticeable performance impairment of the network's behavior is that... it is pretty much the only noticeable performance impairment at all. Removing even more RCNs than the half which was ablated to induce this effect will in fact result in a complete disruption of the network's dynamics. This hints at, loosely speaking, a hierarchical ordering of the level of stability of the attractors and transitions representing the task. The attractors encoding the rules are the most stable, the transitions between this most stable attractors are the most unstable, and all others attractors and transitions are in an intermediate level of stability.

Mild cell ablation

Let us now consider the case of a mild ablation. Removing one third of the RCNs is an operation that the network seems to tolerate very well. The network is in fact still able to perform the task. In particular, it is still capable of switching from one rule to the other, although by comparing Fig. 2.13A and Fig. 2.13B we can see that the activity encoding *Color rule* is slightly suppressed, hinting at a partial destabilization of the corresponding attractor.

If now however we train the lesioned network to additionally learn the "match" / "nonmatch" task, while still retaining the WCST, we incur in some problems. First, the learning algorithm does not converge, unless we diminish the learning

margin d, which was expected. Second, if we test the performance of the network after partial convergence of the training algorithm, we see that it is not able to perform the "match"/"nonmatch" task (see Fig. 2.14). On one hand this is simply a consequence of the reduced capacity of the system (less neurons obviously come down to a smaller learning capacity). On the other hand though, these observations motivate an investigation of the exact properties of the RCNs which determine such a learning capacity, because they highlight that the presence of the specific removed RCNs was essential to the ability to learn a new task.

One of the many advantages in working with simulations is that lesions of simulated networks are completely reversible. So let us take a look at the neurons which were ablated by the lesion to determine what kind of selectivity they display, and whether this tells us something about why they are so important to learn the "match"/"nonmatch" task. What we do is inserting the ablated neurons back in the network which was trained on the WCST, and observe their activity while the network executes the task. Strictly speaking these neurons are superfluous for the execution of the WCST, since the network is able to perform it without them (see Fig. 2.13). However, they are potentially useful during the "match"/"nonmatch" task. In fact, if we probe the network with a stimulus which is relevant in the "match"/"nonmatch" task (the tone cuing the beginning of the task) while it is executing the WCST, we see that some neurons respond to it, even though it is an irrelevant stimulus for the task currently performed (see Fig. 2.15). Panel B of Fig. 2.15 shows the activity of three ablated neurons during the normal execution of the WCST (blue trace), and when the network is probed with the tone during various inter-event intervals (red trace). Some neurons display a differential activity in these two conditions.

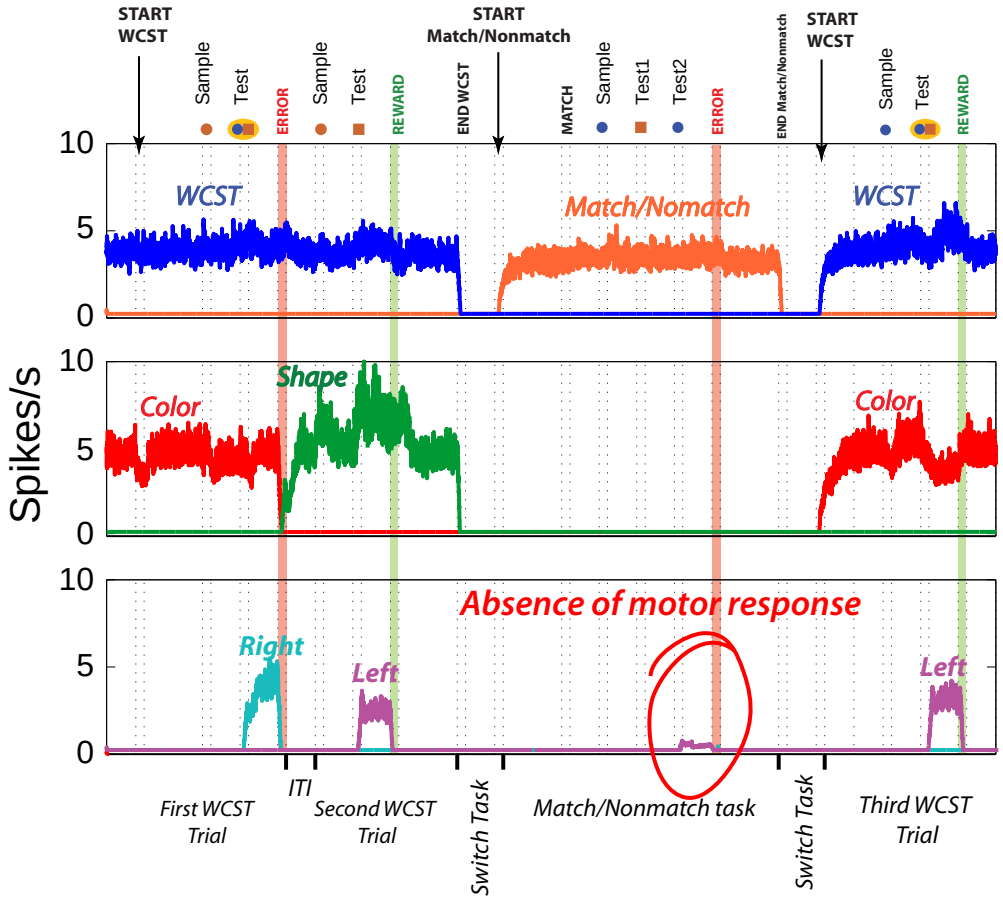

FIG. 2.14: Simulations of the lesioned network of Fig. 2.13B after it has been trained to perform the "match"/"nonmatch" task, in addition to the WCST. The top panel shows the activity of neurons selective to the task to be performed (WCST or "match"/"nonmatch"), the middle panel shows the activity of neuron selective to the rule of the WCST (color or shape), the bottom panel shows neurons selective to the response. The network is not able to execute the correct in the newly learned "match"/"nonmatch" task.

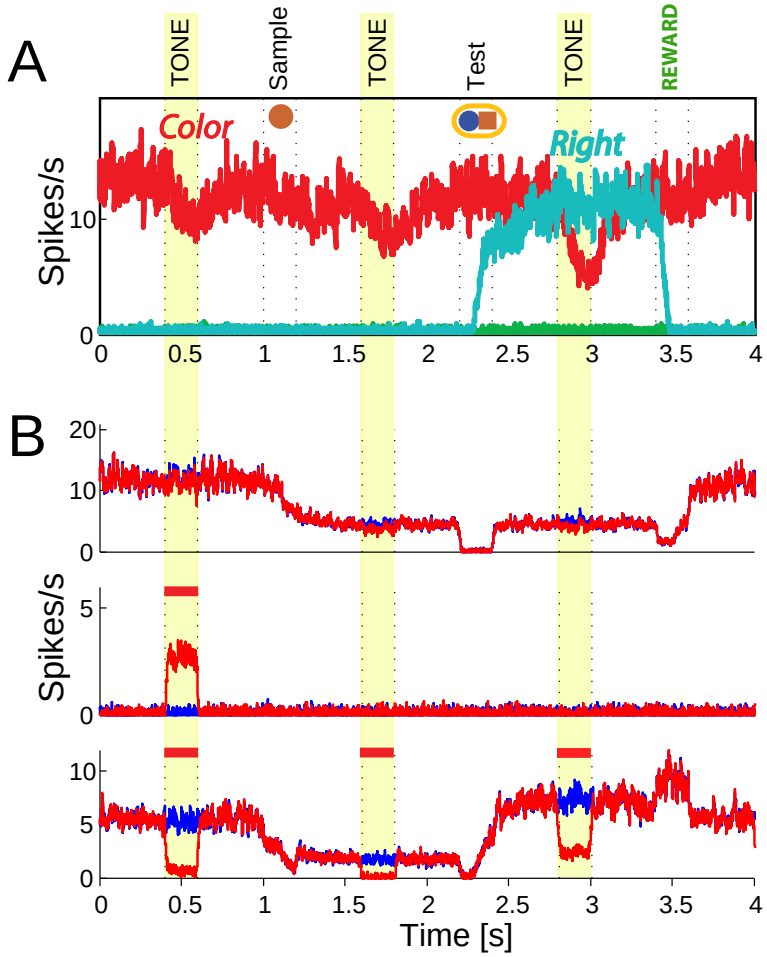

FIG. 2.15: Activity of three "ablated" RCNs. While the network of Fig. 2.13 (not lesioned) is executing the WCST we probe it by presenting stimuli that are not relevant for the task, but are relevant for the "match"/"nonmatch" task. In this case the presented non-relevant stimulus is the tone cuing the beginning of the "match"/"nonmatch" task. The stimulus is presented during the ITI, Delay, Before Reward and After Reward periods. **A**, Activity of some recurrent neurons. They are not perturbed by the presentation of the non-relevant stimulus. **B**, Three RCNs belonging to the population which was ablated in Fig. 2.14. In blue we show the activity during two trials, overlaid by the activity in red in the same conditions except for the presentation in inter-event intervals of an irrelevant stimulus (the tone which cues the "match"/"nonmatch" task). The activity of the neuron in the middle and bottom panels is clearly modulated by the non-relevant stimulus, as it can be inferred by the difference between the blue and the red trace. This indicates that these neurons are likely to be useful for the learning of the future "match"/"nonmatch" task, even though they are superfluous for the execution of the WCST. Red bars indicate selectivity to the tone cuing the "match"/"nonmatch" task.

Basin of attraction through RCNs combines stability and sensitivity

The fact of being selective to irrelevant stimuli is a desirable feature which allows a neural network to establish new response modality to such stimuli, whenever these should become relevant. As we saw in the previous example of the ablated RCNs which are superfluous to execute a task, but are useful to learn a new one, mixed selectivity, combined with the imposition of a finite basin of attraction, offers a mechanism to guarantee stability for the execution of known tasks by preserving at the same time sensitivity to new potentially relevant features. In fact, as we will show in Appendix B.4, the addition of RCNs permits to accommodate an increasing basin of attraction, that is, increasingly stable mental states and task execution. On the other hand, the addition of RCNs is also an effective mechanism to endow the network with a pool of neurons selective to new stimuli which can have a potential relevance for learning future tasks or associations. It may surprise some that these apparently contradicting functions, stability and sensitivity, can be realized through the same mechanism.

2.2.10 Required learning epochs decrease with the number of RCNs

There is a clear way in which the increased overall sensitivity to aspects of a task discussed in the previous section can be exploited, which is to promote learning. Intuitively, an increase of mixed selectivity neurons, that is, of neurons selective to conjunctions of features can be useful to learn associations between these features (cf. discussion in Asaad et al. (1998)).

In general and technically speaking, what happens when RCNs are added to

the network, is that the neural patterns of activity are represented in an increasingly separable way. Indeed, adding RCNs to the network is equivalent to embedding the neural patterns representing the mental states into a higher dimensional space (cf. Appendix B.2). Although the relative distances between different patterns are approximately preserved, the absolute distances increase with the number of RCNs, increasing in turn the separation between the neural patterns that should produce active neurons from those that are supposed to produce inactive neurons. One consequence of this is that it becomes easier to find a hyperplane separating these two classes of patterns, and hence the number of learning epochs required by the perceptron algorithm decreases, as predicted by the perceptron theorem (Block, 1962). In turn, this implies that our training algorithm needs a decreasing amount of time to find a synaptic configuration implementing the desired task.

The phenomenon is illustrated in Fig. 2.16, where we plotted the average number of learning epochs required to satisfy all conditions to realize the attractors and transitions, as a function of the number of RCNs. This was done for three different numbers of attractors and transitions. The number of learning epochs decreases rapidly as RCNs are added to the network. Although this is not the real learning process used by the brain (here we assume that the set of mental states and transitions are already known), it gives strong indication that our network has the highly desirable property that learning becomes simpler and faster as the number of RCNs increases. The convergence time of the perceptron algorithm depends in fact on the separability of the patterns it tries to classify, which is an intrinsic characterization of the difficulty of the task to be learned, independent of the way in which it is learned.

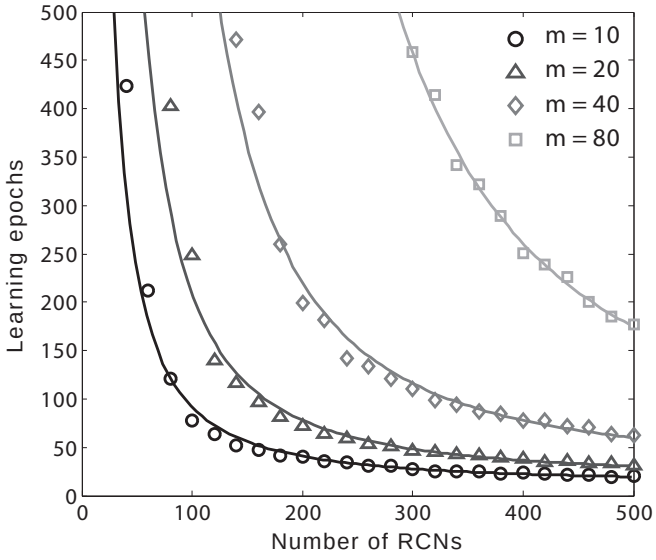

FIG. 2.16: The number of required learning epochs decreases as the number of RCNs increases for a fixed minimal stability parameter $\gamma = 0.5$. The number of epochs is plotted for four different levels of capacity ($m = 10, 20, 40, 80$). The solid lines are the power law curves fitted to the data points (the exponent ranges from approximately -1.5 to -2.2 as m increased). The asymptotic number of learning epochs seems to increase linearly with the number of transitions and the number of attractors m, ranging from approximately 12 to 40 (not visible in the plot), for $m = 10$ and $m = 80$, respectively.

2.2.11 Stochastic transitions to implement probabilistic tasks

In our simulations of the WCST-type and "match"/"nonmatch" tasks, transitions from rule to rule were induced by an *Error Signal* or by an explicit cue indicating the rule in a deterministic manner. However, the parameters of the network and the synaptic couplings can be tuned so as to implement the potentially interesting scenario of stochastic transitions between mental states. In our network the probability for a particular transition to occur depends on the interplay between the noise amplitude σ (cf. equation 2.4 in Section 2.4), and the learning margin d used in the perceptron algorithm to compute the proper synaptic couplings implementing the

attractors and the considered transition (see equation (2.1) in Section 2.4). Let us be more precise on the role of the parameter d by first noticing that a parameter d appears every time the training algorithm is modifying the synaptic matrix to either stabilize an attractor or implement a transition (see Section 2.2.5). In general the higher the d parameter, the more stringent is the condition for the convergence of the learning algorithm (which can anyway always be met by increasing the number of RCNs). However, d plays a different role depending on the case.

In the case in which the training algorithm is stabilizing an attractor, d turns out to be proportional to a stability parameter (Krauth and Mézard, 1987; Abbott, 1990), a quantity which quantifies the average width of the basin of attraction. A high d parameter therefore corresponds to a very stable attractor.

In the case of the implementation of an event-driven transition, the parameter d gives a measure of the strength of the average synaptic current with which every neuron is pushed towards the appropriate firing rate corresponding to the desired activity pattern.

The parameter regime we explored in order to implement probabilistic transitions in one in which the noise is not high enough to disturb the structure of the attractors, but could interfere with event-driven transitions. Transitions are implemented because of the extra synaptic input conveyed to the recurrent network by the RCNs in response to an external stimulation. In order for the transition to occur, such a synaptic input has to be applied consistently for a sufficient amount of time. The effect of the RCNs on the recurrent network is in fact at every moment too low to influence the attractor dynamics. But if this effect is kept constant over a prolonged interval of time the, its temporal sum can overcome the pull of the basin of attraction and bring the activity to the destination of the transition. Noise can

temper with this mechanism by disturbing the temporal coherence of the synaptic input mediating the transitions.

Fig. 2.17 shows the results of a simulation implementing these ideas in the same network considered in Fig. 2.7. The top panel of Fig. 2.17A shows a situation in which the system is maintaining the *Color rule*, until the delivery of the *Error signal* successfully induces a transition to the alternative rule. The bottom panel shows a simulation executed in the same conditions with the same parameters. Here however, because of the different realization of the neural noise, the synaptic input due to the presentation of the *Error signal* is not able to successfully induce a switch to the opposite correct rule. The network commits a *perseverative error*. Fig. 2.17C shows the probability of a successful rule switch as a function of noise amplitude, demonstrating that the network can accommodate a whole range of probabilistic behaviors.

Biological mechanisms to modulate the transition probability

The influence of noise on our neural network could be altered through a gain control mechanism effectively modifying the relative strength of synaptic inputs with respect to neuronal noise. Assuming in fact that the noise is uncoupled from the synaptic input, gain modulation changes the relative importance of the input with respect to the noise. Gain modulation can in turn be due to the production of neuromodulators. The observed differential expression of muscarinic and nicotinic acetylcholine receptors at the thalamic synapses onto excitatory and inhibitory L4 cortical neurons has been already proposed has a mechanism to implement gain control mechanisms (Disney et al., 2007; Disney and Aoki, 2008). Norepinephrine has also been implicated in gain modulation effects (Aston-Jones and Cohen, 2005).

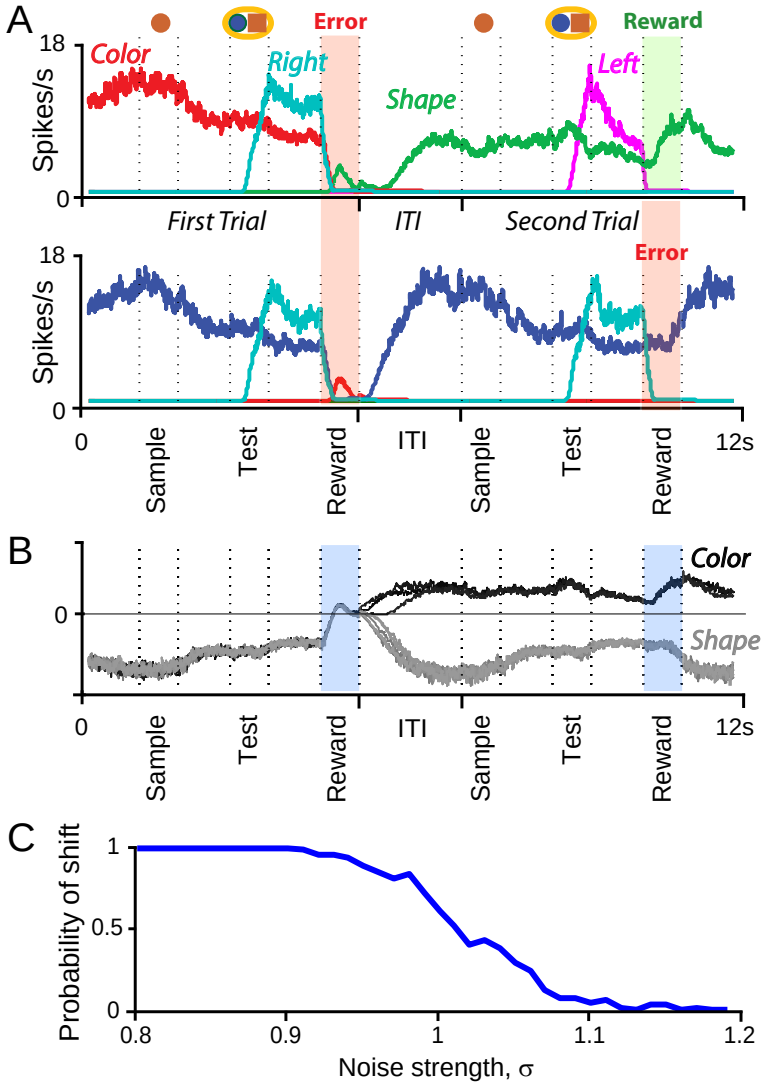

FIG. 2.17: Noise is injected in the simulated neurons of Fig. 2.7. The transitions between the two mental states corresponding to the rules become stochastic. **A**, The neural activity of four highly selective neurons (to color rule, shape rule, touch left and touch right) is plotted as a function of time. In the top panel the *Error signal* induces a transition from one rule to the other, whereas in the bottom panel, under the same conditions, the transition does not occur. **B**, Difference in activity between two neurons selective to *Shape* and *Color rule*, respectively, for several occurrences of the *Error signal* event. In half of the case the transition occurred, and in the other half it did not. **C**, Probability of completing a transition as a function of normalized noise, σ, (noise strength is defined as unitary in correspondence to a $1/2$ transition probability).

Interestingly, both acetylcholine and norepineprhine have been hypothesized to promote exploration by signalling expected and unexpected uncertainty, respectively (Yu and Dayan, 2005).

2.2.12 Predicted features of mixed selectivity: diversity, pre-existence and "universality"

The RCNs and the recurrent neurons show mixed selectivity that is predicted to be characterized by several surprising features. In particular, it is characterized by at least three related properties:

1. Mixed selectivity is highly diverse, in time, as shown in the previous sections, and in space, as different neurons exhibit significantly different patterns of selectivity. Such a diversity is predicted to be significantly higher than in the case of alternative models with hidden units, in which the synaptic connections are carefully chosen to have a minimal number of hidden units. According to our model, neurons with selectivity to behaviorally irrelevant conjunctions of events and mental states are predicted to be observable at any time.

2. Mixed selectivity pre-exists learning: neurons that are selective to conjunctions of mental states and events which are behaviorally relevant to execute a task should be observable even before the task is learned.

3. Mixed selectivity is "universal": mixed selectivity to specific features and conjunctions of features is predictive of the capability of learning a task involving such features. In other words, if a crucial aspect of a task cannot be

represented as a combination of the mixed selectivity neurons in the network, the network will be impaired in the acquisition of such task.

2.3 Discussion

2.3.1 Summary

Heterogeneity is a salient yet puzzling characteristic of neural activity correlated with high level cognitive processes such as decision making, working memory and flexible sensori-motor mapping. Usually models are built to reflect the way we believe the brain solves a certain problem, and neurons with particular functional properties are carefully chosen to make the system work. In some cases these systems are tested to see whether they remain robust in spite of the presence of disorder and the diversity observed in the real brain. Here we showed that heterogeneity actually plays a fundamental computational role in complex, context-dependent tasks. The introduction of cells displaying response diversity in the form of mixed selectivity is sufficient to enable the network to perform complex cognitive tasks and even facilitates the process of learning such tasks. Moreover, mixed selectivity is a property which is computationally very inexpensive to obtain, and therefore does not call for any implausible adhoc supervision mechanism. It is in fact sufficient to introduce neurons that are randomly connected in order to reflect the proper mixture of neural activity encoding the internal mental state and the neural signals representing external events. The number of randomly connected neurons which is typically necessary to implement a task of appreciable complexity is surprisingly small and is generally comparable to the number of cells needed in carefully designed neural circuits. The randomly connected neurons have the advantage that they provide the network with a large variety of mixed selectivity neurons from the very beginning, even before the animal can correctly perform the task. Moreover, when the representations are dense, they are 'universal' as they are likely to solve

multiple tasks.

The properties of randomly connected neurons have been harnessed in neural networks since the 60's (Marr, 1969; Albus, 1971). More recently they have been used to generate complex temporal sequences and time varying input-output relations (Jaeger and Haas, 2004; Maass et al., 2002) and to compress information (Candes and Tao, 2004). They also have been used implicitly in the form of random initial weights in the case of gradient descent learning algorithms like backpropagation (Zipser and Andersen, 1988).

2.3.2 Other approaches based on hidden units

Boltzmann machines (Ackley et al., 1985) have been designed to solve similar problems, in which attractors corresponding to non-linearly separable patterns are stabilized by the activity of hidden units. Recent extensions of the Boltzmann machine algorithm (O'Reilly and Munakata, 2000; Hinton and Salakhutdinov, 2006) can also deal with event-driven transitions from one attractor to another. Our approach is similar because our RCNs are analogous to the hidden units of a Boltzmann machines. However, in our case the synaptic connections to the RCNs are not plastic and we do not need to learn them. The hidden units are pre-wired, similarly to what has been proposed for some feed-forward networks(Rosenblatt, 1962; Marr, 1969; Albus, 1971). Such a procedure is equivalent to embedding the original patterns in a higher dimensional space, a strategy widely used in multi-layer networks and in Support Vector Machines (SVM) (Cortes and Vapnik, 1995). A similar strategy has been also employed in Jaeger and Haas (2004),Maass et al. (2002), and Maass et al. (2007), where the authors realized dynamical systems which utilized a pool of circuits of randomly connected units to generate complex temporal sequences.

In all of these cases the learning procedure is simple and fast. The perceptron algorithm that we use to impose all the conditions for attractors and transitions converges in a few iterations, and the convergence time decreases when the number of RCNs – our hidden units – increases. This is true also when we impose that the basins of attractions must have a given size, or in other words, that the generalization ability of the network remains unchanged for different numbers of RCNs.

We would like to stress that what we propose is not a real learning algorithm, but rather a prescription for finding the synaptic weights. A real, biologically plausible learning algorithm would probably require a significantly more complicated system, with many of the features discussed in O'Reilly and Munakata (2000), like gating. However we believe that it is interesting to show that a network can implement arbitrarily complicated schemes of attractors and event-driven transitions with a very simple prescription to find the desired synaptic configuration. This might greatly simplify and speed up a real learning algorithm.

Recently, it has been shown (Dayan, 2007) that mixed selectivity neurons implemented with multilinear functions can actually play an important role in neural systems that implement both habits and rules during the process of learning of complex cognitive tasks. Multilinearity implements conditional maps between the sensory input, the working memory state, and an output representing the motor response.

2.3.3 How dense should neural representations be?

Although there is no systematic study providing a direct quantitative estimate of the average coding level f, dense representations have been widely reported in

prefrontal cortex (Fuster and Alexander, 1971; Funahashi et al., 1989; Miller et al., 1996; Romo et al., 1999; Tanji and Hoshi, 2008; Wallis et al., 2001; Genovesio et al., 2005; Mansouri et al., 2006; 2007; Nieder and Miller, 2003). These observations seem to contradict basic considerations which would suggest a tendency to reduce the overall activity of a neural network, by employing sparse representations. The reduction of metabolic costs (sparser representations, i.e. small f, require a lower neural activity and hence a lower energy consumption) and of pattern interference (sparse representations minimize the overlap between activity patterns, therefore increasing the capacity of ANNs (Amit, 1989)) would in fact suggest some clear advantages in keeping the coding level f as low as possible.

Our results could offer at least a partial explanation for this apparent contradiction, by indicating a computational advantage in using high coding level f. What we showed in our analysis is in fact that, in order to solve the problems related to context-dependence, the optimal representations for mental states, external inputs and for the patterns of activities of the RCNs should be dense. The optimal fraction f of required coding neurons has been shown to be $1/2$, which is not entirely surprising as such a fraction would maximize the amount of information that can be represented in every moment by the neural patterns of activity of the RCNs. This means however that the majority of the neurons is expected to respond to a large fraction of aspects of the task, and in general to complex conjunctions of events and inner mental states.

We propose that the regime in which higher associative cortical areas operate could be dictated by a trade-off between the minimization of interference and metabolic costs (low f), and a maximization of information content of the activity patterns (high f). Figure 2.3B actually shows that the probability of solving

context-dependence problems is nearly constant around the maximum at $f = 1/2$, implying a large parameter region to accommodate this trade-off.

2.3.4 Other experimentally testable predictions

One of the predictions of our theory is that the mental states are represented by attractors. This has at least two directly testable implications. The first one is that there should be rule selective sustained activity during inter-event intervals. Interestingly, some experiments have already shown rule selective activity, also during the inter-trial interval preceding the beginning of the trial (Mansouri et al., 2006; 2007; Buckley et al., 2009). The second prediction is based on the analysis of the trial-to-trial fluctuations of the sustained activity observed during delay intervals (Amit et al., 1997; Yakovlev et al., 1998; Sugase-Miyamoto et al., 2008). Inter-event persistent activity can be interpreted in at least two ways. The first possibility is that it reflects the collective attractor dynamics of the neuronal circuit, where different events select different patterns of sustained activity corresponding to the mental states. The second possibility is that a form of short term memory is mediated locally, either by the synaptic input or by the inherent cell properties. The activity at the beginning of the inter-event interval is stored in this memory trace, and retained in the form of slowly decaying sustained neural activity. In the case of attractor dynamics, the fluctuations of the neural activity in the initial part of the inter-event interval will not be correlated with the fluctuations at the end of the inter-event interval. Indeed, the fluctuations in the attractor scenario would just reflect the instantaneous noise, as long as the network remains within the basin of attraction. On the contrary, in the second scenario, the inter-event interval neural activity is not the result of a collective dynamics, but it reflects the activity stored

at the beginning of the interval in the short term memory. As a consequence the noise causing trial-to-trial fluctuations will affect the activity as long as the short term memory lasts, and this might extend over the entire inter-event interval. In this second scenario the fluctuations of the activity at the beginning and at the end of the inter-event interval will be correlated.

2.3.5 Trial to trial variability

In Fig. 2.7 we illustrated the typical example of an experiment in which the same stimulus can appear in different contexts and is then represented by different mental states. In most experimental protocols used to study neural activity in behaving animals, the stimuli or other relevant events appear always in the same context and they have invariably the same meaning. We believe that this is not the typical operating mode of the brain. The meaning of the same stimulus can vary significantly depending on the external context and on our present mental state. Even in controlled experiments there might be many uncontrolled factors affecting our perception. This might explain the large variability of single neuron responses across different trials. Indeed, there might be several functionally equivalent mental states corresponding to the same external event, even in the case in which during a trial epoch there is no change in the environment. There is certainly a lot of evidence for trial-to-trial variability, and there is now accumulating evidence for the existence of sequences of sudden transitions between mental states during a particular interval of a trial, even in the absence of task relevant events (Abeles et al., 1995; Jones et al., 2007). Moreover, there is accumulating evidence that the activity preceding a trial is highly variable, significantly more than during the execution of the task (Churchland et al., 2006). This, if interpreted within our frame work, indicates that

the number of accessible mental states might drop drastically from the inter-trial interval to the first epochs of the trial. In our scenario the reduction is predicted to be less evident in the cases in which the subject is required to remember some information about the previous trials.

2.3.6 Why attractors?

The major limitation on the number of implementable transitions in the absence of mixed selectivity units is due to the constraints related to the assumption that initial states are stable patterns of persistent activity, or, in other words, attractors of the neural dynamics. This is based on the assumption that rules are encoded and maintained internally over time as persistent neural activity patterns. Given the price we have to pay, what is the computational advantage of representing mental states with attractors?

Generalization across different timings

One of the greatest advantages resides in the ability to generalize to different event timings, for instance to maintain internally a task rule as long as demanded behaviorally. In most tasks, all animals have a remarkable ability to disregard the information about the exact timing when such an information is irrelevant. For example when they have to remember only the sequence of events, and not the time at which they occur. The proposed attractor neural networks with event-driven transitions can generalize to any timing without the necessity of re-training. Generalizing to different timings is a problem for alternative approaches that encode all the detailed time information (Jaeger and Haas, 2004; Maass et al., 2002; 2007) or for feed-forward models of working memory Goldman (2009). The networks

proposed in Jaeger and Haas (2004); Maass et al. (2002); Goldman (2009) can passively remember a series of past events, in the best case as in a delay line Ganguli et al. (2008). The use of an abstract rule to solve a task requires more than a delay line for at least two reasons: 1) Delay lines can be used to generate an input that encodes the past sequence of recent events and such an input can in principle be used to train a network to respond correctly in multiple contexts. However, the combinatorial explosion of all possible temporal sequences would make training costly and inefficient as the network should be able to recognize the sequences corresponding to all possible instantiations of the rules. 2) Even if it was possible to train the network on all possible instantiations of the rule, it would still be extremely difficult, if not impossible, to train the network on all possible timings. A delay line would consider distinct two temporal sequences of events in which the event timings are different, whereas any attractor based solution would immediately generalize to any timing.

Other models of working memory based on short-term synaptic plasticity Hempel et al. (2000); Mongillo et al. (2008) can operate in a regime that is also insensitive to timing, but they require the presence of persistent activity, similarly to what we proposed in our approach.

Resistance against distractors

The stability requirement implicit in an attractor dynamics offers a simple mechanistic substrate to implement observed features of prefrontal activity like resistance against distractors (Miller et al., 1996; Sakai et al., 2002). This effect was already

modeled within an spiking attractor neural network framework in Brunel and Wang (2001). A similar effect is illustrated in Fig. 2.15, where the network is presented with a distractor (an external stimuli uncorrelated from the other stimuli) during the execution of the task. The distractor does not seem to excessively alter the dynamics of the network (2.15A), since the additional synaptic input due to its presentation is attenuated by the attractor behavior. Remarkably this attenuation is carried out in a collective manner so that the distractor stimulus is not barely ignored or "gated-away" (cf. O'Reilly and Frank (2006)), but its effect is absorbed by the coordinated alteration of the activity of the RCNs. This is a way in which the network stabilizes its own activity without discarding the a sensitivity to irrelevant stimuli. This is a crucial feature which allows our network to be able to respond to new aspects of the environment. Stimuli that are presently irrelevant, can in fact become of importance in subsequent tasks or associations. It would be therefore counterproductive to simply ignore them by simplu suppressing their effect on the network dynamics.

Implementation of probabilistic behavior

The possibility of implementing stochastic transitions is an important property of our network, which derives simply from its underlying attractor dynamics. Stochastic transitions allow to simulation of probabilistic tasks (Gluck et al., 2002), decision making under conditions of uncertainty (Sugrue et al., 2005), and the navigation of uncertain environments (Yu and Dayan, 2005).

In uncertain environments, where reward is not obtained with certainty even when the task is performed correctly, the animal is required accumulate enough evidence before switching to a different strategy. Such a behavior could be im-

plemented in our network by assuming that an independent system keeps track of recent reward history and producing a neuromodulator which controls the probability of making a transition between the mental states corresponding to alternative strategies. Such a mechanism could explain the observed behavior of monkeys in the WCST-type task (Mansouri et al., 2006; 2007), which, when cued with an error signal, switch to a different rule with a probability close to chance level. A detailed analysis of the monkey behavior in the particular experiment that we analyzed is interesting possible future direction of this work.

2.3.7 Conclusion

Mixed selectivity allows the network to encode a large number of facts, memories, events, intentions and, most importantly, various combinations of them without the need of an unrealistically large number of neurons when the representations are dense. The necessary mixed selectivity can be easily obtained by introducing neurons that are connected randomly to other neurons, and they do not require any training procedure. The present work suggests that the commonly observed mixed selectivity of neural activity in the prefrontal cortex is important to enable this cortical area to subserve cognition.

2.4 Details of the implementation of the model

To examine the scaling behavior of our network in the limit of a large number of neurons, we used a simple model of McCulloch- Pitts-like neurons. This model was used to generate Figures 2.6A,B and Figures B.9 and B.10 in the Supplementary Material. We then implemented a more complex, realistic, rate-based neural network model to simulate a version of the Wisconsin Card Sorting Task as well as a "match"/"nonmatch" task (Figures 2.7A,B, 2.8B, 2.9B, 2.10 and 2.12).

2.4.1 The network of McCulloch and Pitts neurons

Network architecture

The architecture of the neural network is illustrated in Fig. 2.3A. There are three populations of cells: 1) the recurrent neurons, whose patterns of activity encode the inner mental state, 2) the external neurons, encoding the events that drive the transitions from one mental state to another, and representing the input neurons that are presumably in different brain areas and 3) the Randomly Connected Neurons (RCN), that provide the network with mixed selectivity neurons. The recurrent neurons receive input from themselves and the other two populations and project back to themselves and the RCNs. The RCNs receive input from both the external neurons and the recurrent network, and project back to the recurrent neurons, but, for simplicity, they are not connected among themselves. The external neurons do not receive any feedback from the other two populations.

All connections to the neurons in the recurrent network are plastic, whereas the connections to the RCNs are fixed, random and uncorrelated. The random connections to an RCN are Gauss distributed with zero mean and with a standard

deviation equal to $1/N$, where N is the number of presynaptic neurons.

Neural dynamics

The recurrent neurons are simplified firing-rate neurons citepDayanAbbott2001 and their dynamics is governed by the equation:

$$\tau \frac{d\nu_i}{dt} = -\nu_i + \phi(I_i - \theta_i), \qquad i = 1, \ldots N,$$

where $\tau = 5ms$, $\phi(x) = \tanh(x)$, θ_i is a threshold, and I_i is the total synaptic current generated by all the afferent neurons (recurrent, RCNs and external):

$$I_i = \sum_j J^r_{ij}\nu_j + \sum_j J^{rcn}_{ij}\nu^{rcn}_j + \sum_j J^x_{ij}\nu^x_j, \qquad i = 1, \ldots N.$$

Here J^r is the matrix of the plastic recurrent connections, J^{rcn} are the plastic connections from the RCNs to the recurrent network, and J^x is the matrix of the plastic synaptic connections from the external neurons to the recurrent neurons. Notice that for these simplified neurons both the neural activity ν_i and the synaptic connections can be positive or negative. The activity of the RCNs and of the external neurons are denoted by ν^{rcn}_j and ν^x_j, respectively. The dynamics of the RCNs is governed by the same differential equation as the recurrent neurons, with the only difference that the total synaptic current is given by $I^{rcn}_i = \sum_j K^r_{ij}\nu_j + \sum_j K^x_{ij}\nu^x_j$, where K^r and K^x are the afferent random connections from the recurrent network and from the external neurons, respectively. The integration time τ plays a role analogous to the transmission delays used in Sompolinsky and Kanter (1986) to implement transitions in temporal sequences of patterns of neural activities.

In the absence of any stimulus, the ν_i^x values are set to a particular pattern of neural activities $\nu_i^x = \nu_i^{x_0}$ chosen at random with the same statistics of the patterns representing an external event. We will name $\nu_i^{x_0}$ "spontaneous" activity pattern. When an external event occurs, the ν_i^x values are instantaneously set to the pattern representing the event for a duration of 2τ, and then are set back to $\nu_i^x = \nu_i^{x_0}$.

2.4.2 The prescription for determining the synaptic weights

The plastic connections J^r, J^{rcn} and J^x are determined by imposing the mathematical conditions that ensure both the stability of the patterns of activity representing the mental states and the correct implementation of the event driven transitions.

The first step is to analyze the task to be performed and construct a scheme of mental states and event driven transitions like the one of Fig. 2.1B. Notice that in general there are multiple schemes corresponding to different strategies for performing the same task. The second step is to choose the patterns of neural activities representing the mental states (for recurrent neurons) and the external events (for the external neurons). The structure of these patterns is normally the result of a complex procedure of learning whose analysis is beyond the scope of this work. However the prescription for constructing the neural network applies to any neural representation. The patterns we chose were all vectors with components $\nu_i = \pm 1$.

The third step is to go iteratively over all mental state attractors and event-driven transitions and modify the weights of the plastic synaptic connections until all mathematical conditions for the stability of the attractors and the event driven transitions are satisfied. The algorithm is illustrated in Figs. 2.5A,B where we show two snapshots of neural activity that are contiguous in time. For each transition

from one initial attractor to a target attractor we set the external input to the pattern of activity that corresponds to the triggering event (see Fig. 2.5A). At the same time we impose the pattern of activity of the initial attractor on the recurrent network. We then compute the activity of the RCNs at fixed external and recurrent neuronal activity. For each neuron in the recurrent network we compute the total synaptic current generated by the activity imposed on the other neurons and we modify the synapses in such a way that the current drives the neuron to the state of activation at time $t + \Delta t$. In particular the synaptic currents at time t will generate an activity pattern, under the assumption that the post-synaptic neurons will fire if and only if the total input currents are above the firing threshold θ. The synaptic weights are updated only if the synaptic currents do not match the output activities in the target attractor (i.e. the pattern of activity at time $t + \Delta t$), as in the perceptron learning algorithm (Rosenblatt, 1962). If they need to be modified, the synaptic weights are increased by a quantity proportional to the product of the pre-synaptic activity at time t and the desired post-synaptic activity (i.e. the pattern of active and inactive neurons at time $t+\Delta t$ in the figure). The stationarity of the patterns of activity corresponding to the mental states is imposed in a similar way, by requiring that the pattern at time t generates itself at time $t + \Delta t$ (see Fig. 2.5B). Such a procedure is iterated until all conditions are simultaneously satisfied, guaranteeing that the patterns of activity of the desired attractors are fixed points of the neural dynamics and that the transitions are implemented in a one-step dynamics.

In order to have attractors, the fixed points should also be stable. This can be achieved by requiring that the total synaptic currents not only satisfy the desired conditions, but also that they are far enough from the threshold θ (Forrest, 1988;

Krauth and Mézard, 1987). In this way, small perturbations of the input modify the total synaptic current, but not the state of activation of the neurons. The distance from the threshold is usually named learning margin, which we will denote by d. The synapses are updated until

$$\nu_i(t + \Delta t)(I_i(t) - \theta_i) > d > 0,$$

where $I_i(t)$ is the total synaptic current to neuron i, θ_i is its firing threshold, and $\nu_i(t + \Delta t)$ is the desired output activity. In other words, synapses are updated as long as $I_i(t)$ does not surpass $\theta_i + d$ when neuron i is required to be active at time $t + \Delta t$. Analogously, the synapses are modified until $I_i(t)$ goes below $\theta_i - d$ when the desired output is inactive ($\nu_i(t + \Delta t) = -1$).

Such a condition can be easily satisfied when all synaptic weights on a dendritic tree are scaled up by the same factor. Unfortunately, this is true also in situations when stability is not guaranteed. To avoid this problem we block synaptic updates only when (Forrest, 1988; Krauth and Mézard, 1987)

$$\nu_i(t + \Delta t)(I_i(t) - \theta_i) > \gamma \sqrt{\sum_j J_{ij}^2}, \tag{2.1}$$

where γ is the stability parameter, and the J_{ij}s are the synapses that are afferent to neurons i. The stability parameter γ is chosen to be maximal, i.e. we progressively increase γ until the algorithm stops converging within a reasonable number of learning epochs (we chose 500). Such a procedure is similar to one of the γ-margin modified perceptron algorithms presented in Korzen and Klesk (2008), and allows us to approximately maximize the size of the basin of attraction of the stable patterns

of activities corresponding to the mental states (Forrest, 1988). Summarizing, the equation for updating a synaptic weight J_{ij} is:

$$J_{ij} \rightarrow J_{ij} + \lambda \nu_i(t + \Delta t)\nu_j(t)\ H\left(-\nu_i(t + \Delta t)(I_i(t) - \theta_i) + \gamma \sqrt{\sum_j J_{ij}^2} \right)$$

where H is the Heaviside function and the learning rate λ is set to 0.01.

Biologically realistic firing rate model

Figures 2.7A,B show simulations of a more complex model in which we used rate models for separate excitatory and inhibitory neurons that integrate NMDA, AMPA and GABA mediated synaptic currents. We started by training the synaptic weights of a simplified neural network of McCulloch-Pitts neurons, as described in the previous section. For the simulations of Fig. 2.7 we implemented the scheme of mental states and transitions of Fig. 2.1B. For the neural representations of mental states and external inputs, we used $N^r = 8$ neurons for the recurrent network, 2 encoding the rule (color, shape), 4 for the identity of the sample stimulus (2 colors and 2 shapes), and 2 for the motor responses (touch left, or touch right). These representations result in highly correlated patterns of mental states. The external stimuli are represented by $N^x = 14$ neurons: 4 indicating the color and the shape of the Sample stimulus, 8 representing the color and shape of the two Test stimuli, and two representing either the Reward or Error. We used 384 RCNs, that is slightly more than a 16 fold amount of the total number of recurrent and external neurons.

The network implementing multiple tasks with the strategy illustrated in Fig. 2.11A is composed of $N^r = 14$ recurrent neurons storing the $m = 39$ mental states

(not all states appear in the figure), and $N^x = 20$ external neurons, selective for the $e = 15$ possible external events, which induce a total number of $r = 120$ possible transitions, thus giving a ratio $r/e = 8$. 510 RCNs were enough to implement both tasks on the same network.

The network implementing the simpler strategy illustrated in Fig. 2.11B, exploiting a same/different information signal coming from sensory areas is slightly simpler. The only difference with the previous network is that, since two more signals have to be distinguished ("same" and "different") the number of external neurons is $N^x = 22$. As the task is simpler, involving a smaller number of mental states ($m = 31$) and transitions ($r = 84$) for about the same number of external events ($e = 17$ for a ratio $r/e \approx 5$), only 432 RCNs were necessary.

After convergence of the learning prescription for the chosen representations of states and scheme of transitions we obtained a matrix J of synaptic weights, which in general can be both positive and negative. We enforced Dale's law and separate excitation and inhibition by introducing a population of inhibitory neurons whose activity is a linear function of the total synaptic input generated by the excitatory neurons. In practice we rewrote the synaptic matrix J as:

$$J_{ij} = J^+_{ij} - J^-,$$

where J^- is the absolute value of the most negative synapse and the J^+_{ij}'s are all positive. J^- can be interpreted as the product of the synaptic strengths from excitatory to inhibitory and from inhibitory to excitatory neurons when the transfer function for the inhibitory neurons is linear. We followed a similar procedure for the RCNs, by replacing each of them with an excitatory neuron, and introducing

a second inhibitory population that allows the connections projecting from the neurons replacing the RCNs to be always positive.

The activity of the excitatory and inhibitory neurons are denoted by the the firing rates ν_i^E and ν^I, respectively. The equations governing the dynamics of these firing rates are:

$$\tau_E \frac{d\nu_i^E}{dt} = -\nu_i^E + F(I_i^{EE} + I_i^{EI} + I_i^{ext}), \qquad \tau_I \frac{d\nu^I}{dt} = -\nu^I + F(I^{IE} + I^{II}), \tag{2.2}$$

where F is a threshold linear function with unitary gain: $F(x) = H(x) \cdot x$, the currents I_i^{ext} are generated by the neurons representing the external events, and the synaptic currents I_i^{xy} are generated by the population of neurons y and injected into population x ($x, y = E, I$ where E and I indicate excitatory and inhibitory neurons respectively). The time development of the synaptic currents is governed by:

$$\tau_{xy} \frac{dI_i^{xy}}{dt} = -I_i^{xy} + \sum_j J_{ij}^{xy} \phi_{xy}(\nu_j^y), \tag{2.3}$$

where J^{xy} is a matrix of synaptic weights. The synaptic currents from excitatory to excitatory neuron ($xy = EE$) are mediated by NMDA receptor channels with a slow timescale $\tau_{EE} = \tau_{NMDA} = 100ms$. They saturate at high frequencies ν_js of the pre-synaptic spikes due to the saturation of the open channels with slow decay rate (Wang, 1999; Brunel and Wang, 2001):

$$\phi_{EE}(\nu_i) = \frac{\nu_i \tau_{EE}}{1 + \nu_i \tau_{EE}}.$$

Currents from excitatory to inhibitory neurons ($xy = IE$) are mediated by fast excitatory AMPA synapses with $\tau_{IE} = \tau_{AMPA} = 5ms$ and $\phi_{IE}(\nu_i) = \nu_i$. Finally, for

$xy = II$ (inhibitory self-couplings) and $xy = IE$ (excitatory to inhibitory) we have GABA synapses with $\tau_{xy} = \tau_{GABA} = 2ms$ and $\phi_{xy}(\nu_i) = \nu_i$. The synaptic matrices J^{IE}, J^{II}, J^{EI} have been chosen so that the total inhibitory current to the excitatory population is proportional to $I_i^{EI} = -J^- \nu_i$, where J^- is the most negative synapse obtained by the learning procedure. This condition can be expressed as:

$$J^- = -J^{EI}(1 + |J^{II}|)^{-1} J^{IE}.$$

Given a set of excitatory synaptic weights J^{EE}, it is always possible to compute a J^{II} large enough so that all fixed points are stable. Then the product $J^{EI}J^{IE}$ is determined by the above expression for a given J^-. We chose without any loss of generality $J^{EI} = 1$ and $J^{IE} = -J^-(1 + |J^{II}|)$.

The network is set to its initial conditions simply by clamping the firing rates ν_i^E of the recurrent and of the external neurons to the pattern of activity representing the desired starting attractor and the "spontaneous activity" stimulus pattern, respectively, and letting all the currents and firing rates variables of the other neurons evolve according to equations (2.3,2.2) until a stationary state is reached.

External events are simulated by changing the activities of the external neurons to the pattern representing the event for a time $\Delta t = 2\tau_{NMDA}$, where τ_{NMDA} is the longest synaptic time scale, and then setting them back to the spontaneous activity pattern.

Additionally, we introduced a multiplicative noise term that modifies the firing rate of the excitatory neurons ν_i^E. This term is meant to capture finite-size fluctuations widely studied in networks of integrate-and-fire neurons (Brunel and Hakim,

1999b). Formally this is expressed by the following change in equation (2.2):

$$\nu_i^E(t) \rightarrow \nu_i^E(t)(1 + \sigma^2\eta(t)), \tag{2.4}$$

where $\eta(t)$ is a Gaussian process with unitary variance and $\sigma^2 = 0.01$.

Chapter 3

Attractor concretion as a mechanism for the formation of context representations [1]

In the previous chapter we presented a theory for the neural representation of mental states as patterns of reverberating activity within the framework of Attractor Neural Networks endowed with mixed selectivity. Our theory gives a prescription to build a neural network able to execute any deterministic or Markovian rule-based task as a series of event-driven transitions between mental states. After having characterized the neural representation of mental states and their role in the execution of rule-based tasks, we now investigate the question of how these representations can be created.

Since an essential component of a mental state is the temporal relationship between the sensory, cognitive and decisional elements which characterize it, we will explore the consequences of the assumption that *temporal contiguity* is the drive of such a creation. We will formalize and exemplify these ideas in the case of a

[1] This chapter is based on the following reference:
Rigotti, M., Ben Dayan Rubin, D., Morrison, S.E., Salzman, C.D., & Fusi, S.,
Attractor concretion as a mechanism for the formation of context representations (NeuroImage, in press)

context-dependent trace conditioning task. Specifically, our hypothesis is that the acquisition of this kind of tasks is subserved by the concomitance of two learning systems: a feedback-based system which rapidly learns stimulus-response associations, and a slow unsupervised system encoding the temporal contingencies of these associations. This last system is assumed to be composed of a pool of mixed selectivity neurons constantly encoding information about the sensory events and the affective value associated to them. The analysis of neurophysiological recordings motivated by this theoretical setup indeed revealed the presence of these mixed selectivity neurons in the Orbitofrontal Cortex (OFC) and the amygdala of monkey engaged in a context-dependent trace conditioning task.

Our framework next assumes that a temporal asymmetric Hebbian-like plasticity mechanism strengthens the connections between mixed selectivity cells which tend to be sequentially activated, and promotes in this way the creation of patterns of self-sustained activity. These patterns will therefore display a correlation structure which encodes the temporal information of the task contingencies. These new self-sustained activity patterns will in turn participate in sequences of activation on longer time-scales and mediate the iterative creation of new patterns of self-sustained activity. We name this process of fusion of self-sustained activity patterns *attractor concretion*, and we will show how it can be used to create mental states effectively representing information about a behavioral context, thereby conveniently biasing decisions in the case of ambiguous external evidence.

3.1 Introduction

When we execute a complex task, we often need to store information about past events in order to decide how to react to a particular stimulus. If the task is familiar, we know what information to store and what to disregard. At the moment we make a decision about our response to a particular event, we are in a specific mental state that contains all the information that we know to be relevant to react to that event. This information is typically about our perception of the environment, our physical position, our memories, our motivation, our intentions, and all the other factors that might be relevant to reach a particular goal. In other words, every mental state is our most general disposition to behavior. In many cases the execution of a task can be considered as a series of transitions from one mental state to the next, each triggered by the occurrence of a particular event.

In order to understand the neural mechanisms underlying the execution of complex tasks we need to answer two important questions: 1) how do we create the neural representations of mental states that contain all the relevant information to execute a task? 2) how do we learn which mental state to select in response to a particular event? Reinforcement learning (RL) algorithms (see e.g. (Sutton and Barto, 1998)) have provided an elegant theoretical framework to answer the second question. In particular they provide prescriptions for the policy of mental state selection that maximize reward and minimize punishment. In RL algorithms, values that represent future cumulative reward are assigned to the mental states. The value increases as the agent moves closer to a pleasant outcome, like the delivery of reward. The table of values of mental states thereby determines the optimal policy. One has simply to select the action that induces a transition to the state

with highest value. However, most RL algorithms presuppose that the set of mental states contain all the relevant information for performing a task, and hence they do not provide an answer to the first question, on how the mental states are created.

In this paper, we make an attempt to answer this question by proposing a mechanism for the creation of mental states in context dependent tasks, in which the optimal policy for maximizing reward is different in different contexts. In particular we will consider all the situations in which the information about temporal context can be used to create mental states that unequivocally determine the state of the environment and the actions to be executed. In other words, if the occurrence of an event in two different contexts requires different policies, we need to react to that event in two different ways that will be encoded in two different sets of mental states. We propose in our paper a mechanism that leads to the formation of different sets of mental states for different contexts.

3.1.1 A paradigmatic experiment

To illustrate the principles behind our proposed mechanism, we will present a neural network model that performs an extended version of an appetitive and aversive trace conditioning task used in recent neurophysiological recording experiments (Paton et al., 2006; Belova et al., 2007; Salzman et al., 2007; Belova et al., 2008; Morrison and Salzman, 2009). In those experiments, monkeys learned whether an abstract fractal image (conditioned stimulus, CS) predicted liquid reward or aversive air-puffs (unconditioned stimulus, US) after a brief time (trace) interval. Single unit recordings in the amygdala and orbitofrontal cortex (OFC) revealed the existence of cells encoding the learned value of the CSs (rewarded or punished). After a variable number of trials, the CS-reinforcement contingencies were reversed and monkeys

had to learn the new contingencies. In the experiments, the CS-US associations were reversed only once. However, in principle, the two contexts defined by the sets of CS-US associations could be alternated multiple times. In this situation, it is possible that the animal at some point creates two representations corresponding to the two contexts and it can switch rapidly from the optimal policy for one context to the optimal policy for the other. This switch is qualitatively different from learning and forgetting the associations as it would not require any synaptic modification. The two independent context representations would be simultaneously encoded in the pattern of synaptic connectivity.

The model of a neural circuit performing the trace conditioning task with multiple reversals will be used to illustrate the mechanism for the formation of context representations. The single unit recordings from experiments (in which there is a single reversal), will be used to support our assumptions about the initial response properties of the model neurons.

3.1.2 The proposed model architecture: the Associative Network (AN) and the Context Network (CN)

In modeling data from this task, we assume that there are two interacting neural circuits. The first one is a neural circuit that we name AN (Associative Network) similar to the one proposed by (Fusi et al., 2007), that learns simple one to one associations between CSs and USs. The second one, the CN (Context Network) observes the activity of the AN, in particular when the AN has already learned the correct associations, and abstracts the representations of temporal contexts. These representations can then be used by the AN to predict more efficiently the value of

the CSs when the context changes and the AN has to learn the new associations.

3.1.3 Learning and forgetting associations: the function of the AN

In every trial the AN starts from a wait state. The CS biases a competition between the populations of neurons representing two different mental states, one that predicts the delivery of reward, and hence has a positive value, and the other that predicts punishment and has a negative value. The delivery of reward or punishment resets the AN to the wait state. The AN encodes the CS-US associations by making CS triggered transitions to the state that represents the value of the predicted US. The CS-US associations are learned by biasing the competition between the positive and the negative state. In particular, the competition bias is learned by modifying the synaptic connections from the neurons that represent each CS and the positive and negative neurons. When the associations are reversed, the learned synaptic strengths are overwritten by the new values. The AN can be in one of the three states (wait, positive, negative), and can implement only one set of CS-US mappings at a time. In the two contexts the AN implements two different "policies", as the same CS induces different transitions. Notice that the CS-US associations are learned independently for each CS. The AN does not store any information about the relations between different CS-US associations, and in particular about the fact that all associations are simultaneously modified when the context changes. This means that for example, when the context changes, the AN cannot infer from the modification of one of the CS-US associations, the value of the other CS. This type of inference requires information about the temporal

statistics of the CS-US associations, which are collected by the CN. We propose that this type of inferential information is stored in the representations of the temporal contexts, which are built from the statistics of the sequence of events and mental states.

3.1.4 The formation of representations of temporal contexts: the main idea

In the trace conditioning task the first context is characterized by the fact that the sequence CS A-Reward is most often followed either by itself or by CS B-Punishment. Analogously the second context is defined by the elevated probability of the transitions between CS A-Punishment and CS B-Reward. The animal observes rarely that CS A-Reward is followed by CS B-Reward or CS A-Punishment. In other words, if we look at the matrix of transitions between these sequences, we can clearly identify two clusters of sequences that are connected by significantly larger transition probabilities. These two clusters define the two relevant contexts. The idea sounds simple, but the detailed implementation turned out to be more difficult than expected because initially, the neural circuit does not know that it needs to consider the CS-US associations as the building blocks of the context representation. The neural circuit observes a series of several events and mental states and it has to abstract autonomously what is relevant for the formation of context representations.

3.1.5 The neural basis of the formation of context representations

In the detailed implementation, the learning process that takes place in the CN iteratively merges the neural representations of temporally contiguous events and mental states. The first compounds that are created represent short temporal sequences, or, more generally, groups of events and mental states that tend to be temporally contiguous. Compounds that often follow each other can also merge into larger compounds. In this sense the process of merging is iterative, and at every iteration the compounds grow to represent larger groups of events that are temporally contiguous to each other. In more technical terms, the temporal statistics of the compounds define a new Markov process that is a coarse grained representation of the Markov process of the previous iteration, similarly to what has been studied in (Li and DiCarlo, 2008). The iterative process stops when all the transition probabilities between the compounds go below a certain threshold. In the specific case of the trace conditioning task, the first representations that merge are those that represent temporally contiguous events like CS A and the mental state predicting a positive value. The parameters are chosen so that the merging process stops when CS A-Reward and CS B-Punishment belong to a single compound representing context 1, and CS B-Reward and CS A-Punishment belong to a second compound that represents context 2. The neural mechanisms that could underlie this iterative learning process are described in detail in the Methods, and the simulations are illustrated in the Results.

The initial conditions: mixed selectivity in theory and experiments

The neurons of the CN are randomly connected to the neurons of the AN and the parameters are chosen such that they respond to conjunctions of external events (CS A or B, reward, punishment) and states of the AN (neutral, positive, negative). The neurons therefore may be described as having mixed selectivity, even before the learning process of the CN starts. We observed these types of neurons both in the amygdala and in orbitofrontal cortex of monkeys performing the trace conditioning task of Paton et al. (2006). We report in this manuscript the statistics of their response properties.

From temporal sequences to context representations: attractor concretion

In the model, the temporal statistics of the mixed selectivity neurons depend on the sequence of patterns of activation of the AN. For example consider a trial in which CS A is followed by reward. The AN would start in a wait state with neutral value and CS A would steer the activity toward a positive state. The US would reset the AN activity back to the neutral state. In the CN we would observe the activation in sequence of the following populations that are selective to conjunctions of states and events: neutral-CS A, CS A-positive, positive-reward, reward-neutral. Initially, each conjunction induces a transient activation of the CN neurons. The synapses between CN neurons that are activated simultaneously are strengthened (Hebbian component of synaptic plasticity), so that the neural representations of individual conjunctions become stable self-sustaining patterns of persistent activity that are attractors of the neural dynamics (see e.g. (Hopfield, 1982; Amit, 1989)). These patterns remain active until the occurrence of the next

event and the activation of a new input from the AN. A second component of the synaptic plasticity, which we call temporal sequence learning (TSL), strengthens the connections between neurons that are activated in succession, similar to what has been proposed for learning of temporal sequences (Sompolinsky and Kanter, 1986) and temporal contexts (Brunel, 1996; Griniasty et al., 1993; O'Reilly and Munakata, 2000; Rougier et al., 2005; Yakovlev et al., 1998). This component causes the merging (concretion) of attractors that are activated in a fixed temporal order, leading to the formation of the representations of temporal contexts.

From context representations to the creation of mental states

At the beginning of the learning process, the CN simply reflects the activity of the AN, and hence the entire AN-CN system has the same number of mental states as the sole AN (neutral, positive, negative). At the end of the learning process, the CN can represent both contexts, with one being the "active context". Hence, the entire AN-CN system has two sets of the three AN states, one for the first context and one for the second. As the AN receives feedback from the CN, it can then easily disambiguate between the CS-US associations of the first context from those of the second. We will show that at the end of the learning process the full AN-CN system can work more efficiently than the AN alone after a context switch. Indeed it can predict the correct value of a CS when the other is already known. We will then discuss quantitative predictions about the behavior and the neural activity that can be recorded.

3.2 Materials and Methods

We first describe the details of the trace conditioning task that has been used in neurophysiological experiments and its extended version, which we used in all model simulations. We then describe the the Associative Network (AN) and the Context Network (CN). The model of the AN has already been introduced in Fusi et al. (2007). Here we summarize briefly its neural and synaptic dynamics and we show simulations of the specific case of the trace conditioning task. We then describe the novel neural and the synaptic dynamics of the CN. The description of the learning behavior and the explanation of the mechanisms are deferred to the Results section. Finally, we describe the details of the analysis of neurophysiological data that we use in the Results to motivate the model assumptions about the initial response properties of the neurons.

3.2.1 The experimental protocol

The appetitive and aversive trace conditioning task described in (Paton et al., 2006; Belova et al., 2007; Salzman et al., 2007; Belova et al., 2008; Morrison and Salzman, 2009) uses a trace conditioning protocol (a type of Pavlovian conditioning) to train an animal to learn the relationship between abstract visual stimuli (conditioned stimuli, CS) and rewards and punishments (unconditioned stimuli, US). While a monkey centers its gaze at a fixation point, CSs are presented followed by a trace interval (a brief temporal gap), and then US delivery. In the experiments (Paton et al., 2006; Belova et al., 2007; Salzman et al., 2007; Belova et al., 2008; Morrison and Salzman, 2009), the animals demonstrated their learning of CS-US contingencies by licking a reward tube in anticipation of reward and blinking in

anticipation of the delivery of the air-puff. After a variable number of trials, the CS-US associations were reversed without any notice and the animals had to learn the new contingencies.

In the version of the task that is simulated in this paper, we will consider two CSs, A and B, and two possible outcomes, reward and punishment, which are delivered after a brief time interval, as in the original task of (Paton et al., 2006). The CS-US associations are reversed multiple times, switching from context 1 in which CS A is paired with reward, and CS B with punishment, to context 2 in which CS A is paired with punishment and CS B with reward. The blocks of context 1 and context 2 trials are approximately 120 trials each, and they are alternated multiple times.

3.2.2 The Associative Network (AN): structure and function

The AN learns the associations between CSs and USs. This Associative Network (AN) receives feed-forward plastic inputs from the neural populations that represent external events (CSs, reward and punishment). The neurons of the AN are grouped in three different populations: two excitatory populations representing positive and negative value compete through a population of inhibitory neurons (see Fig. 3.1). The synaptic connections between these three populations are chosen as in Wang (2002) and Fusi et al. (2007), so that there are only three stable states: a wait state in which the AN is quiet (neutral value state), and the other two corresponding to the activation of one of the two excitatory populations (positive and negative state). The presentation of the CS generates an input that initiates and biases

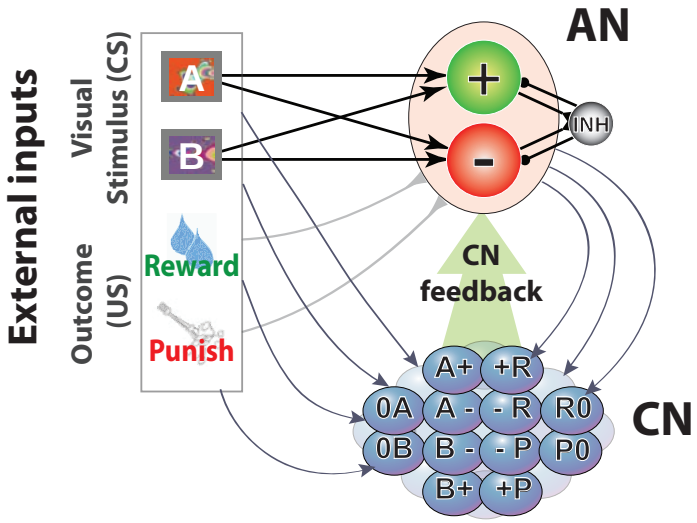

FIG. 3.1: The two networks of the simulated neural circuit: the Associative Network (AN, top right), and the Context Network (CN, bottom right). The AN and the CN receive inputs from the neurons encoding external events (conditioned and unconditioned stimuli). The AN network contains two populations of neurons, +,-, that encode positive and negative values respectively. These neurons are activated by external events (CSs) in anticipation of reward and punishment. The inhibitory population (INH) mediates the competition between the two populations. The connections from the CS neurons to the AN neurons are plastic and encode the associations between the CS and the predicted US. The CN neurons receive fixed random synaptic connections from both the AN and the external neurons. The neurons in the CN respond to conjunctions of external events and AN states and they are labeled accordingly. The recurrent connections within the CN are plastic and they are modified to learn context representations. After learning, the CN neurons encode the context, and they project back to the AN (described later, in Fig. 3.4).

the competition between the positive value and the negative value populations of neurons. The delivery of the US brings the network back to the neutral value state (see Fig. 3.2 for a description of the AN dynamics). Initially, when the CS is novel and the associated US is still unknown, the CS activates an unbiased competition between positive and negative populations and one of the two values is chosen randomly with equal probability. This behavior reflects the fact that the animal is already familiar with the experimental protocol and can predict that the CS will be followed in half of the cases by reward and in the other half by punishment (Fusi et al., 2007). The synaptic weights between the neurons encoding the CS and the AN neurons are modified in every trial depending on whether the prediction of the AN (positive or negative) matches the actual US that is delivered (reward or punishment) or not. When the prediction is correct (i.e. when the CS activates the positive state and it is followed by reward, or when the CS activates the negative states and it is followed by punishment), the synapses from the CS neurons to the correct value encoding population are strengthened (learning rate $q_+^R = 0.042$, see Fusi et al. (2007) for the details of the synaptic dynamics), and the synapses from the CS neurons to the other value population are weakened ($q_-^R = 0.073$). These modifications reinforce the bias in the competition between positive and negative populations towards the correct prediction. When the prediction is incorrect, we assume as in Fusi et al. (2007) that all synapses from the CS neurons to the AN neurons are rapidly depressed ($q_-^{NR} = 0.99$). This reset forces the AN to choose randomly the value of the CS the next time it is presented. The learning rates qs have been chosen to match the learning curves reported in Paton et al. (2006). In particular we chose them so that the simulated AN reaches 90% of the value prediction performance in 10 trials on average.

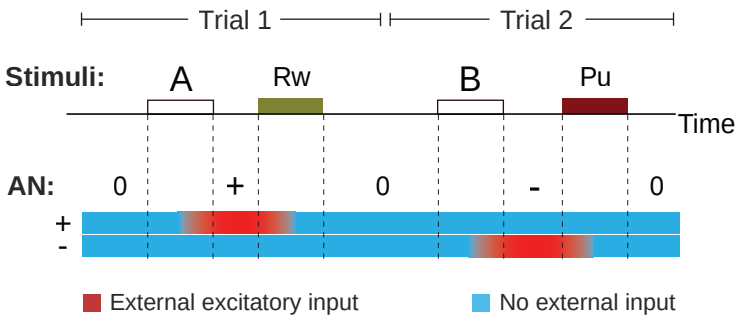

FIG. 3.2: Simulated activity of the AN during two trials of the trace conditioning task of Paton et al. (2006). During the first trial CS A is presented, followed by a reward. The AN network is initially in the neutral state '0' in which all populations are inactive (the activity is color coded: blue means inactive, red means active). The presentation of CS A initiates a competition between the positive coding AN population '+' and the negative coding population '-' which, in this simulation, ends with the activation of population '+'. The delivery of reward resets the AN to the '0' state. In the second trial CS B activates population '-' and punishment resets it.

3.2.3 The Context Network (CN): the architecture

The Context Network is made of neurons that are randomly connected to both the neurons encoding the external events and the excitatory neurons of the AN. The synapses between CN neurons are plastic and they are modified by the learning rule described below in order to create context representations. The random connections are Gauss distributed and the parameters are chosen so that a large number of neurons of the CN neurons respond to conjunctions of external events (CS A,CS B, reward, punishment, denoted by 'A','B','R','P', respectively) and the state of the AN (positive, negative, neutral, denoted by '+', '-', '0' respectively). For simplicity, we will consider in our simulations only the neurons that respond to simple conjunctions of one external event and one state. For example, some neurons would respond to CS B only when the AN switches to a negative state. We will label these neurons with 'B-'. In Fig. 3.3 we show the simulation of a rate model neuron that

behaves like a typical CN neuron with mixed selectivity. These simulations are for illustrative purposes only and to motivate our assumptions about the response properties of the CN neurons. This type of model neurons will not be used in the rest of the paper (see the Methods section about the neural dynamics for the model neurons that will be used). The simulated neuron receives synaptic inputs with equal weights from the neurons that are activated by CS B and from the negative value coding neurons of the AN (simulated as in Fusi et al. (2007)). The firing rate is a sigmoidal function of the total synaptic input. Although the choice of equal synaptic weights might seem special, the behavior illustrated in Fig. 3.3 is actually the typical behavior of neurons with randomly chosen synaptic weights. The probability that a randomly connected neuron exhibits significant mixed selectivity depends on the specific model of the neuron and on the neural representations of the events, but it can be as large as 1/3 (see Appendix B.3).

The neurons with mixed selectivity are the building blocks of context representations and they are assumed to be present in the CN from the very beginning, before any learning process takes place. In the Results we will support this assumption with experimental data.

3.2.4 The neural dynamics of the CN

The activity of the AN drives the CN network, which is composed of N populations of neurons which are either active or inactive. We denote by ξ_i the activity of population i. As we consider only the randomly connected neurons that respond to simple conjunctions of external events and state of the AN, we have in total $N = 12$ different types of populations responding to the following combinations: 0A, 0B, A+, A-, B+, B-, R0, P0, +R, -R, +P, -P. Not all these combinations are

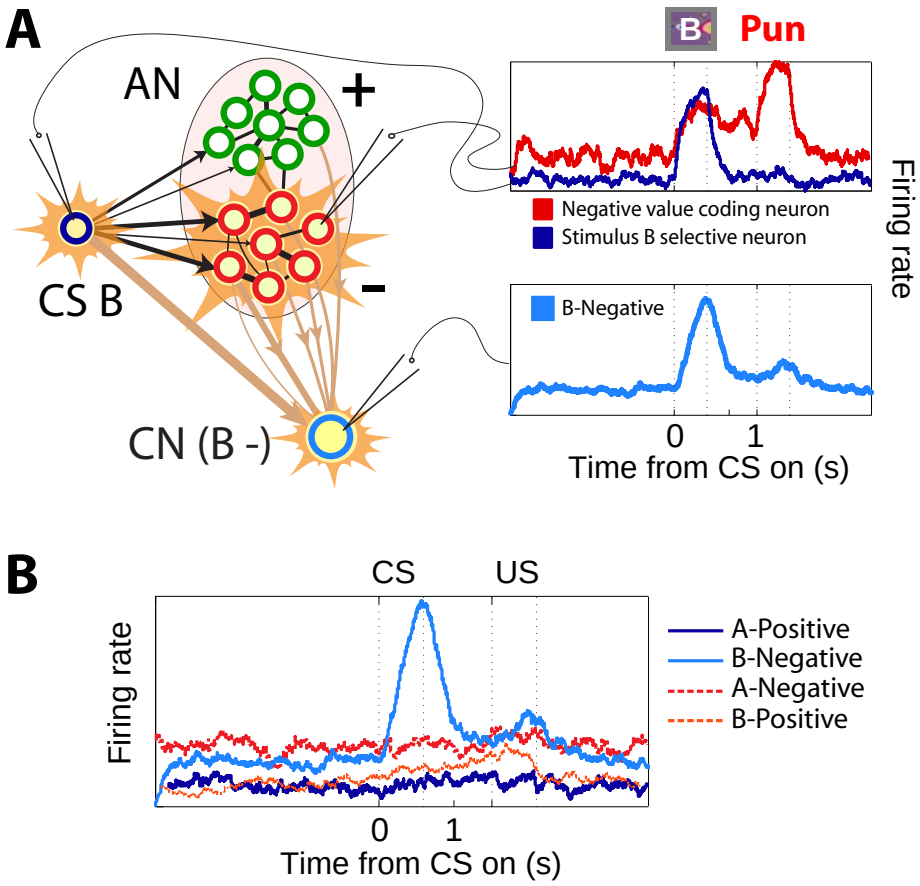

FIG. 3.3: Illustrative firing-rate simulations of a typical CN neuron which exhibits mixed selectivity to the conjunction of an external event (CS B) and an AN value state (negative). **A**, The top plots show firing rate as a function of time for two simulated neurons in response to CS B. The blue trace represents the response of an external neuron encoding CS B, which is flat until the presentation of visual stimulus B. The red trace corresponds to the response of a negative value coding neuron of the AN. CS B is already familiar and its value is correctly predicted by the AN. When CS B is shown, the negative value population is activated, and it remains active until the delivery of the US. In the bottom plot, we show the activity of a CN neuron that, by chance, is strongly connected to CS B external neurons and to the negative value coding AN neurons. The response is significantly different from spontaneous firing rate only when CS B is presented, and the negative value AN state wins the competition. **B**, Mixed selectivity to CS B and negative value. The cell is selective to both the value and the identity of the CS as the neuron responds only to the CS B-Negative combination and not to the other combinations (CS A-Positive, CS A-Negative, CS B-Positive).

necessary for the formation of context representations, but we simulate all of them for completeness, as they are all assumed to be present in the neural circuit. For simplicity we ignore neurons that respond to combinations of three or more events and AN states. We assume full connectivity within the CN. The average strength of the excitatory synaptic connections from neurons of population j to neurons of population i is denoted by J_{ij}, and it can vary from 0 to 1. Every CN population inhibits all the others through a pool of inhibitory neurons. The net effect is to include a constant negative coupling ($g_I = 0.5$) between the excitatory populations. Additionally, every population i receives a constant current θ_i and an external input h_i from the AN neurons. More quantitatively, the activity ξ_i^t of population i at time t is given by the following discrete time evolution equation:

$$\xi_i^t = \Theta\left(\frac{1}{N}\sum_{j=1}^{N}(J_{ij} - g_I)\xi_j^{t-1} + h_i^t + \theta_i\right), \quad i = 1, \ldots, N, \tag{3.1}$$

where Θ is the Heaviside function with $\Theta(x) = 1$ if $x > 0$, and $\Theta(x) = 0$ otherwise.

During the learning process, the single or multiple CN populations can become attractors of the neural dynamics (stable patterns of self-sustaining activity). These patterns will be the building blocks for context representation, as illustrated in the Results. Once activated by the AN and external input, the attractors remain active indefinitely, or at least until the arrival of the next sufficiently strong input. To avoid the simultaneous activity of all CN populations, we need to guarantee that the external input can overcome the recurrent CN input and shut down previously activated stable patterns. This is important also for weak external inputs, that would normally be ignored by the CN. For this reason we complemented the described neural dynamics by a reset signal which inhibits the whole network ev-

ery time an external input h^t targets at least one population which is not already activated by the recurrent input. Such a signal is important not only to reset the activity but also to learn only when the activity pattern of the CN is modified.

3.2.5 The synaptic dynamics of the CN

The CN "observes" the activity of the neurons representing the external stimuli and the neurons of the AN. The context representations are created from the temporal statistics of the patterns of activity of the AN and the external neurons. Here we describe the equations and the details of the synaptic dynamics that lead to the formation of the context representations, but we explain the mechanism and we show simulations only in the Results.

The synapses are modified by two mechanisms: 1) a Hebbian mechanism strengthens the synapses of simultaneously active neurons, and depresses the synapses connecting an active to an inactive neuron. Analogously to the mechanism introduced in Hopfield (1982) and, more recently, in Amit and Brunel (1995)), it stabilizes the CN activity that is initially imposed transiently by the external input and the AN. If the synapses between co-activated neurons become strong enough, the neurons of the activated CN population can excite each other to the point that the transient activity becomes self-sustaining (attractor of the neural dynamics). 2) the TSL (Temporal Sequence Learning) mechanism, which links together patterns of activity that are activated in sequence. This component of the synaptic dynamics is responsible for merging attractors that are often temporally contiguous. It basically strengthens the synapses between two neurons that are activated sequentially one after the other. Moreover it depresses the synapses between active neurons and neurons that are inactivated at the next time step.

Both mechanisms are activated only when the competition between the positive and negative populations in the AN is strongly biased, indicating that the AN has already learned the associations. As the neurons have only two levels of activation, we monitor the total synaptic input to them, and we modify the synapses of the CN only when the current driving the winning population exceeds a threshold $\theta_L = 0.25$ (to be compared to the total synaptic input, i.e. the argument of the Heaviside function in Eq. 3.1). In a more realistic implementation with rate neurons, we could set a threshold for the firing rate of the AN neurons.

The Hebbian mechanism

The modifications of the synapses from population j to population i depend on the current pre and post-synaptic activity $\xi^t_{i,j}$ and on the post-synaptic recurrent synaptic input I^t_i (i.e. the input from the neurons the belong to other populations within the CN):

$$I^t_i = \frac{1}{N}\sum_{j=1}^{N}(J_{ij} - g_I)\xi^t_j + \theta_i. \tag{3.2}$$

In particular we have:

$$\Delta J^s_{ij} = \Theta\left(\gamma^s - \xi^t_i I^t_i\right)\left[q^s_+(1 - J_{ij})\xi^t_i\xi^t_j - q^s_- J_{ij}\left(1 - \xi^t_i\right)\xi^t_j\right], \tag{3.3}$$

Equation (3.3) describes a modified version of the perceptron learning algorithm (Rosenblatt, 1958). The synapse J_{ij} is potentiated when both the pre- and the post-synaptic neuron are simultaneously active ($\xi^t_i\xi^t_j$) and depressed when the pre-synaptic neuron is active and the post-synaptic neuron is inactive (the $\left(1 - \xi^t_i\right)\xi^t_j$ term). The two terms containing J_{ij}, and $(1 - J_{ij})$ impose a soft-bound on the synaptic weights and keep them between zero and one, as in Senn and Fusi (2005).

The synapses are not updated when the post-synaptic neuron is active and the total recurrent input is sufficiently large ($\Theta(\gamma^s - \xi_i^t I_i^t)$, where γ^s is a positive number which is related to the stability parameter (Gardner, 1987)). This term prevents the synapses from being updated when the recurrent input is sufficient to activate the post-synaptic neuron in the absence of the external stimulus. In other words, the synapses are not updated if the pattern of activity can already sustain itself. This term prevents the correlated parts of different attractors to dominate the neural dynamics and hence allows the network to generate attractors that are highly correlated (see e.g. Senn and Fusi (2005)). Moreover, if γ is sufficiently large, it increases the stability of the attractors (Krauth et al., 1988).

This type of learning prescription can be implemented with realistic spike driven synaptic dynamics (Brader et al., 2007). The factors q_+^s and q_-^s are the learning rates for potentiation and depression, respectively. The parameter values are: $\gamma^s = 5 \times 10^{-4}$, $q_+^s = 7.5 \times 10^{-2}/n$ $q_-^s = 15 \times 10^{-2}/n$, where n is the number of time steps in one trial (in our simulations $n = 15$).

Temporal Sequence Learning (TSL)

The second learning component, temporal sequence learning (TSL), is meant to strengthen the synaptic connections between neurons that are repeatedly activated one after the other in a fixed temporal order. At every time step t we calculate how many inactive populations are activated by the new incoming external input h^t and we divide this quantity by the total number of populations N:

$$\Delta^t = \frac{1}{N} \sum_{i=1}^{N} \Theta\left(h_i^t\right) \ \Theta\left(1 - \xi_i^{t-1}\right).$$

This is a measure of the global mismatch between the pattern imposed by the external stimulation at time t and the network activity at the previous time $t-1$. When this quantity is different from zero, it is an indication that the neural activity of the CN has been modified by the external input. In such a case a reset signal is delivered to the CN (see neural dynamics), and the synaptic connections from population j and population i are modified according to:

$$\Delta J_{ij}^{a} = \Delta^{t}\Theta\left(\gamma^{a} - \Theta\left(h_{i}^{t}\right) I_{i}^{t-1}\right) \cdot \left[q_{+}^{a}(1 - J_{ij})\xi_{j}^{t-1}\Theta\left(h_{i}^{t}\right) - q_{-}^{a} J_{ij}\xi_{j}^{t-1}\left(1 - \Theta\left(h_{i}^{t}\right)\right)\right], \tag{3.4}$$

The term in square brackets contains two terms, one potentiates the synapses and the other one depresses them. In particular, the synapses are potentiated when the post-synaptic external current h_i^t is positive and the pre-synaptic neuron was active at the previous time step (ξ_j^{t-1}). The synapses are depressed when the post-synaptic external current h_i^t is negative, and the pre-synaptic neuron was active at the previous time step. The J_{ij} dependent terms implement a soft boundary as in the case of the Hebbian term. The presence of a soft boundary is in general important to estimate probabilities (Rosenthal et al., 2001; Fusi et al., 2007; Soltani and Wang, 2006), and in our specific case to estimate the probability that a particular event is followed by another one. The parameters are: $\gamma^a = 0$, $q_+^a = 1.0$ and $q_-^a = 7.5 \times 10^{-2}$.

At every time step the synaptic weights are updated according to:

$$J_{ij} \leftarrow J_{ij} + \Delta J_{ij}^{a} + \Delta J_{ij}^{s}. \tag{3.5}$$

When the reset signal is delivered, first the weights are updated and then the activity is reset.

3.2.6 The feedback from the CN to the AN

The information about the current context contained in the CN activity after learning, can be used by the AN to predict more efficiently the value of the stimuli. For example, when the CN-AN knows that the CS-US association has changed for CS A, it can predict that the CS-US association for CS B has changed as well. In order to do so, we need to introduce some form of feedback from the CN to the AN. In principle CN neurons could project directly to the AN neurons, as the CN neurons contain all the information that the AN neurons need to know about the current context. However, this feedback input cannot produce the desired context dependent bias on the AN competitive dynamics, unless we introduce an intermediate population of neurons that mixes the external input and the CN activity (see Fig. 3.4). This is a general requirement for many systems in which there is a dependence on context (see Chapter 2). Indeed, without this intermediate layer of neurons, there is no set of synaptic weights that would produce the correct prediction. The general proof is in Appendix B.1, here we give an intuitive argument for this specific case. Consider two input neurons: an external neuron that is active when CS A is presented and inactive for CS B, and a CN neuron that is active for context 1 and inactive for context 2. The AN "output" neuron encoding a negative value should be inactive for CS A+Context 1 (both input neurons active=11), and CS B+Context 2 (00). At the same time it should be inactive for CS A+Context 2 (10), and CS B+Context 1 (01). This mapping of the input to the output is equivalent to the implementation of the logical operation 'exclusive or' (XOR) and it is

known that it is not possible to build a single layer network that implements it (see e.g. Minsky and Papert (1969)). The solution proposed in the previous Chapter 2 is to introduce an intermediate layer of randomly connected neurons. If the number of these neurons is sufficiently large, the problem equivalent to the XOR explained above can be solved. For this reason we introduced an additional population of neurons whose activity depends on the input from the CN populations and the external neurons. Analogously to the CN neurons, which are also connected by random synaptic weights to the AN, most of the neurons of the additional population respond to pairs of CN-external neurons activations. We therefore assume that the feedback population is composed of neurons responding to the $2 \times N$ possible CS-CN population combinations (2 CSs multiplied by the N populations of the CN). These neurons project back to the AN neurons with plastic synapses that are modified with the same synaptic dynamics as the connections from the external neurons to the AN neurons, except that the learning rates are significantly smaller ($q_+^R = 2 \times 10^{-4}$, $q_-^R = 0$, $q_-^{NR} = 4 \times 10^{-4}$). These feedback connections are initialized to zero, and the learning rates are chosen to be small, so that the synaptic input from the CN affects the AN dynamics only at a late learning stage, when the CN context attractors are formed and stable. The AN sees the information about the current context coming from the CN as an additional input, that would operate in the same way as a constantly present contextual cue.

We stop modifying the CN synapses when the feedback input becomes too strong compared to the external input. This prevents the CN from learning from its own activity, with the danger of effects that are difficult to control. For example, if the CN learns rapidly one of the two contexts of the trace conditioning task, and it starts dominating the AN behavior, then it becomes difficult to create the

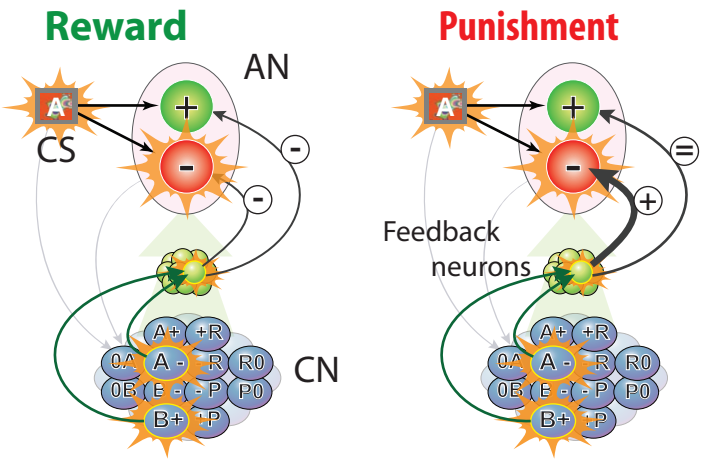

FIG. 3.4: The learning dynamics of the CN to AN feedback. This signal is mediated by a layer of feedback neuron selective to CN and external input activity. The synapses connecting the feedback neurons to the AN are modified with the same learning dynamics as the one used for the AN synapses (see Fusi et al. (2007) and the description of the AN dynamics in the Methods).

representation of the second context because the AN will not have a chance to learn the CS-US associations of the second context. Indeed it will constantly be driven by the CN, which represents and will continue to represent only the first context. In the simulations we block the CN learning dynamics when the synaptic input to the AN coming from the CN feedback is more than 2.5 times larger than the direct feedforward external input.

3.2.7 The analysis of recorded OFC and amygdala cells

Our assumption that the neurons of the CN are initially selective to conjunctions of external events and AN mental states (mixed selectivity), is supported by the analysis of neurophysiological recordings. We analyzed the cells recorded during the trace conditioning task with a single reversal described at the beginning of the

Methods. It is reasonable to assume that this situation (i.e. the single reversal) reflects what happens in the initial or early stages of the learning process that leads to the formation of context representations. Most analyses were performed on spike data from two time intervals during the trial: the CS interval (90-440 ms after image onset for monkey L; 90-390 ms after image onset for monkey R) and the trace interval (90-1500 ms after the image turned off). These time intervals were chosen because more than 90% of visual response latencies exceeded 90 ms as established by an analysis of latencies, described previously (Paton et al., 2006; Belova et al., 2007; 2008; Morrison and Salzman, 2009).

In order to determine the degree to which neural responses are modulated by reinforcement contingencies (image value) or by the sensory characteristics of the CSs themselves, we performed a two-way ANOVA with image value and image identity as main factors. The ANOVA was performed separately on spike counts from the CS and trace intervals for each cell, as cells could encode image value at different times during the trial. If there was a significant effect of image value in either or both intervals ($p < 0.01$), the cell was classified as value-coding. We found a few cells that had opposite image value preferences in the CS and trace intervals, and these were excluded from further analysis. Neurons in OFC and the amygdala that were categorized as "non-value-coding" exhibited a variety of responses to conditioned stimuli; these included neural responses that were similar for all conditioned images, as well as responses that were strongest (or weakest) for the stimulus associated with a weak reward. In addition, a substantial proportion of OFC and amygdala neurons, both value-coding and non-value-coding, showed a significant main effect of image identity in the ANOVA, or an interaction effect of image value and image identity ($p < 0.01$).

We performed an additional analysis to determine whether in the trace conditioning task with a single reversal (presumably the situation preceding the formation of context representations) there are cells that encode the context in the same way as the CN neurons at the end of the learning process. As we will see in the simulation results, the CN neurons that represent the context after learning, are selective to the context in every interval of the trial (in the presence or in the absence of external events like the CS or the US). In particular their activity should be significantly different in the two contexts for all the individual CS-US associations. In practice, in the specific case of the trace conditioning task, there should be a threshold that separates the activity recorded in CS A-Positive and CS B-Negative trials (context 1) from the activity recorded in CS A-Negative and CS B-Positive trials (context 2). Moreover the difference between context 1 and context 2 activities should be significant. We imposed these conditions by considering all pairs of CS-US combinations. In particular, in order to meet the criterion for a "context cell", the average activity of the neuron in the CS A-Positive trials (μ_{A+}) had to be significantly different ($p < 0.05$, t-test) from the average activity in the CS A-Negative trials (μ_{A-}). Moreover, we required that also the differences $\mu_{A+} - \mu_{B+}$, $\mu_{B-} - \mu_{A-}$ and $\mu_{B-} - \mu_{B+}$ are significant. Additionally, we required that all the four differences $\mu_{A+} - \mu_{A-}$, $\mu_{A+} - \mu_{B+}$, $\mu_{B-} - \mu_{A-}$ and $\mu_{B-} - \mu_{B+}$ have the same sign (we always subtract context 2 (reversal) epochs from context 1 (initial) epochs). A cell qualified as a context cell if it satisfied all these criteria at least in one of the two analyzed intervals (the interval during CS presentation and the trace interval).

3.3 Results

We present the results as follows: we first explain the assumptions of the model about the mixed selectivity of the neurons, and we provide experimental evidence to support them. In particular, we show how neurophysiological data recorded in the amygdala and orbitofrontal cortex of monkeys during appetitive and aversive trace conditioning supports the hypothesis that neurons have mixed selectivity to external events like the CSs and inner mental states encoding the predicted value of the stimuli. The model neurons are assumed to exhibit the same response properties before the process of creation of context representations starts. We then illustrate the proposed mechanisms underlying the formation of context representation by describing the simulations of a model neural network performing a trace conditioning task. In particular we show how transient events can generate patterns of sustained activity that bridges the gap between two successive relevant events. We then explain the iterative process of merging of the neural representations of short temporal sequences (attractor concretion) that leads eventually to the temporal contexts. Simulations show that these representations can significantly improve the prediction of the value of a stimulus when the context changes. Finally, we use the model to make specific predictions for the patterns of neural activity that would be observed given new experimental manipulations.

3.3.1 Learning context representations

The initial situation: experimental evidence for neurons with mixed selectivity

In our model, we assume that the neurons of the CN have mixed selectivity to the mental states of the AN (positive, negative, neutral) and external events (CSs, USs). They should exhibit this form of selectivity to the conjunction of mental states and events even before the learning process leading to the formation of context representation starts. The assumption is based on two considerations: 1) these conjunctions contain the basic elements that characterize the contexts. For example, selectivity to CS A would not allow the network to discriminate between the two contexts, as the very same stimulus A is presented in both contexts, in which it would activate the neuron in exactly the same way. However, a neuron that responds to CS A only when the following mental state of the AN is *positive*, would activate only in one context and not in the other. 2) neurons with this form of mixed selectivity can be easily obtained with random connections, and hence without any learning procedure (see Methods and previous Chapter 2). Indeed, neurons that are connected with random synaptic weights to the neurons of the AN, which represent the mental state, and to the neurons representing the external events, are very likely to respond only to the simultaneous activation of these two populations, provided that the threshold for activating the neuron is large enough. Neural representations of patterns of activity across several randomly connected neurons are analogous to random projections and they can encode efficiently the information contained in both the external and internal inputs (e.g. the Johnson and Lindenstrauss lemma (Dasgupta and Gupta, 2002)).

Neurons with the assumed mixed selectivity (in the CN) and with the expected response properties for the AN have been observed in various areas of the brain. For example, the CSs are assumed to evoke value dependent sustained activity in the AN. We observed neurons with these response properties both in the OFC and in the amygdala while the animal was performing the trace conditioning experiment (see Fig. 3.5A,B). The activity can be sustained throughout the trace interval (see the cells of Paton et al. (2006)), for a limited time (Fig. 3.5A), or it can ramp up in anticipation of reinforcement (Fig. 3.5B).

The CN neurons are assumed to respond to conjunctions of events and the internal states of the AN. In the trace conditioning experiment we expect to observe in the first context neural representations of the mixtures, such as CS A-Positive or CS B-Negative, whereas in the second context, the patterns represent CS A-Negative or CS B-Positive. We have often observed neural responses reflecting this type of mixed selectivity in the amygdala and OFC. We recorded and analyzed 216 cells in OFC and 222 from the amygdala (recorded from two monkeys). We used a two way ANOVA to determine whether image value, image identity or an interaction between image value and identity accounted for a significant portion of the variance in neural firing. For this analysis, we excluded the first five trials of the experiment of each type of trial, as well as the first 5 trials of each trial type after reversal. We did this to exclude trials in which neurons were changing their firing rate during learning about reinforcement contingencies. Of particular interest to our proposed model, a substantial number of neurons in both the amygdala and OFC had a significant effect of the interaction between image identity and value (66/216 OFC neurons, and 87/222 amygdala neurons, $p < 0.01$, 2-way ANOVA). Neurons with a significant interaction term indicate that responses to images are

modulated in an unequal manner by reinforcement contingencies, which is the precise type of response profile postulated by the model. Two examples of these types of neurons are depicted in Fig. 3.5C,D. Notice that in each case, neurons represent the conjunction between a particular image and a particular reinforcement. In these two cases, image identity and image value do not have a significant effect on responses, but many of the cells with significant interactive effects also show significant effects of a main factor in the ANOVA.

The first learning phase: from transient conjunctions of events and mental states to attractors

We assumed that the CN initially receives a strong excitatory input from the AN and the neurons representing the external events only when particular events (the CSs A and B, the USs Reward and Punishment) are preceded or followed by specific states of the Associative Network, AN (neutral, positive, negative). For example, consider a trial in which CS A is associated with reward (first trial in Fig. 3.6A). We assume that the AN has already learned the correct association, and the presentation of CS A induces a transition from a state with neutral value (0) to a state with a positive value (+) (see Fig. 3.6B). The neurons encoding a positive value have sustained activity until reward is delivered and the activity of the AN is shut down. The CN neurons observe the following sequence of AN states and events: Neutral-A, A-Positive, Positive-Reward, Reward-Neutral. The corresponding populations of neurons are transiently activated in the same order (0A, A+, +R, R0, see Fig. 3.6C). Analogously, in a trial in which CS B is associated with punishment, we have Neutral-B, B-Negative, Negative-Punishment, Punishment-Neutral (see Fig. 3.6, second trial). For simplicity, we assumed in our simulations that each

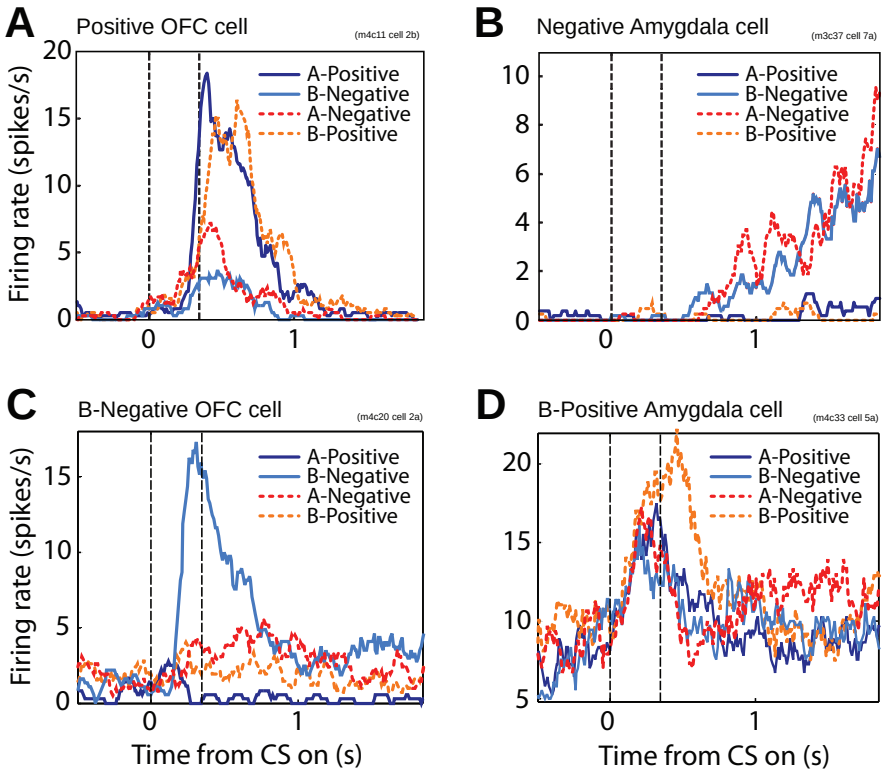

FIG. 3.5: Recorded activity of OFC and amygdala cells that respond as expected in AN **(A,B)** and CN **(C,D)**. The activity has been recorded while the monkey was performing the trace-conditioning task for the four possible CS-US pairings. The continuous traces show the activity after the monkey had learned the associations defining Context 1 (A-Positive, B-Negative). The dotted traces show the activity after learning of Context 2 (A-Negative, B-Positive). The AN cells show sustained activity during the trace interval that encodes the value of the CS. These cells have been observed both in the OFC **(A)** and amygdala **(B)**. The CN cells are selective to specific combinations of CS and value, both in the OFC **(C)** and the amygdala **(D)**.

conjunction like Punishment-Neutral activates a single population of the CN.

The Hebbian component of learning strengthens the synaptic connections between the CN neurons that are repeatedly co-activated. Moreover, it depresses the synapses from active to inactive neurons (see Methods for the detailed equations). As a consequence all patterns of activity of the CN that are activated a sufficient number of times become attractors of the neural dynamics. At the end of the first phase, we have the situation illustrated in Fig. 3.6D. Each conjunction of events and AN states activates a population of the CN, and the activity remains elevated even after the event terminates. However, every time the AN activates a population of CN neurons that was inactive, or deactivates a population of CN neurons that was active, we assumed that a reset signal is delivered and the previous pattern of reverberating activity is shut down by a strong inhibitory input (blue stripes in the figure).

The second learning phase: concretion of temporally contiguous attractors

Now that the representations of the conjunction of events and AN states are attractors of the neural dynamics, the time gap between one event and the next one is bridged by the self-sustained patterns of reverberating activity. Two successive conjunctions of events belonging to the same or different trials, become temporally contiguous. This enables the temporal sequence learning (TSL) mechanism to modify synaptic weights and to link two successive patterns of activity, so that the process of attractor concretion can start. The TSL mechanism operates only when the pattern of activity of the CN changes because it is modified by the AN input, and it strengthens the synapses between an active pre-synaptic neuron and

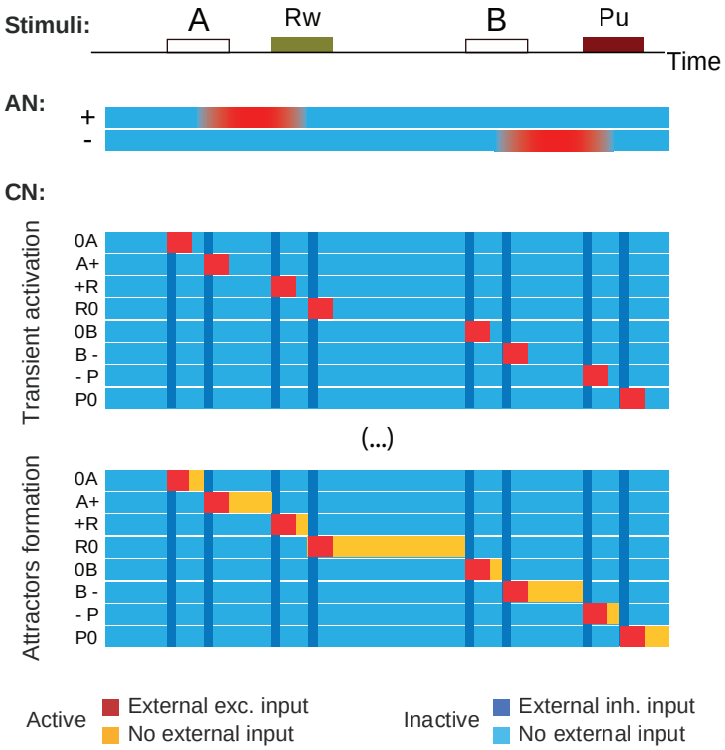

FIG. 3.6: First learning phase: from transient events to attractors. A: The scheme of two consecutive trials. In the first trial the presentation of CS A is followed, after a delay, by the delivery of reward. In the second one, CS B is followed by punishment. B: Color coded activity of the AN (red=active, blue=inactive) as a function of time in response to the events depicted in panel A. The simulation starts in an inactive state with neutral value (0). The presentation of CS A induces a transition to a state in which the neurons encoding positive value (+) have self-sustained activity. The activity is shut down by the delivery of reward. Analogously for the CS B-Punishment case. C: Color coded activity of the CN populations as a function of time (red=active in the presence of external input, yellow=active in the absence of external input, light blue=inactive, blue=inactive because of the strong inhibitory input generated by a reset signal). Each row represents the activity of one population that is labeled according to its selectivity (e.g. 0A is a neuron that responds only when the AN is the neutral state and CS A is presented). The external events together with the activation of positive and negative states of the AN activate the populations of the CN (red bars). Every time a different population is activated a reset signal is delivered (blue stripe). D: First CN attractors: the synapses within each repeatedly activated population are strengthened to the point that the activity self-sustains also after the event terminates (yellow bars).

a neuron that is activated at the successive time step. Moreover it depresses the synapses between active neurons and neurons that are inactivated at the successive time step. If the synapses between two populations, say a and b, are sufficiently potentiated, then the activation of a causes also the activation of b, leading to the merging of the two attractors (attractor concretion).

The process of formation of the context representations requires a few iterations, and the typical phases that we observe in the simulations are illustrated in Fig. 3.7. The iterative process generates representations of progressively longer temporal sequences. To illustrate this process, consider the same two trials considered in Fig. 3.6. The CN dynamics are now simulated at different stages of the learning process (Fig 3.6B-E). A scheme representing the temporal statistics of the activation of the CN populations that is relevant for the concretion process is shown in the right column. Although this scheme does not allow us to make quantitative predictions about the detailed neural dynamics, it is useful to describe the dynamics of the concretion process, and in particular to understand how the temporal statistics of the events and mental states are related to the probability that two attractors merge into a single representation. Each arrow links two CN populations and its thickness represents the propensity of these populations to merge. The propensity depends on both the parameters of the learning dynamics and the temporal statistics of the activation of the two populations. In particular, the arrow connecting two generic populations, say a to b, is proportional to the probability that a is followed by b, multiplied by the number of times that a is activated within a certain time interval, which depends on the parameters of the learning dynamics. This is motivated by the fact that the synapses between a and b are potentiated by the TSL mechanism every time that a is activated, and then it is followed by

b. They are depressed when a is followed by a population different from b. The stronger a synapse becomes, the higher the probability that a activates b, and hence that the two populations merge into a single representation. The propensity to concretion depends also on other details of the TSL mechanism (see the Methods for the description of the full dynamics), but also and more strongly on the effects of the Hebbian mechanism. In particular the Hebbian component of synaptic dynamics strengthens the synapses within population a every time a is activated, and it depresses the synapses from a to all the other populations, including b. This effect of stabilization of a increases with the time that the CN spends in a state in which a is active. This means that the strength of the connections between a and b, and hence the propensity to concretion, should decrease with the time that a is active. This is valid only when a and b are not already co-activated, because in such a case the Hebbian term actually strengthens their connections. For these reasons, the propensity is inversely proportional to the fraction of time that the CN spends in a when b is inactive, multiplied by the ratio between the Hebbian learning rate and the TSL learning rate. Summarizing, the propensity of a to merge with b is high when a is activated repeatedly and it is often followed by b. However it is reduced if the CN spends a large fraction of time in a, or if the CN is often driven to states other than b.

The largest propensities drive the first concretions. For example Fig. 3.7B shows that +R and R0 are the first population to merge into a single attractor. Indeed +R is consistently followed by R0 in both contexts. Notice that also A+ is always followed by +R, but the propensity to merge is smaller because A+ is activated on average half of the times that +R is activated. The result of this first concretion is illustrated in the simulations of Fig. 3.7B. The activation of +R now

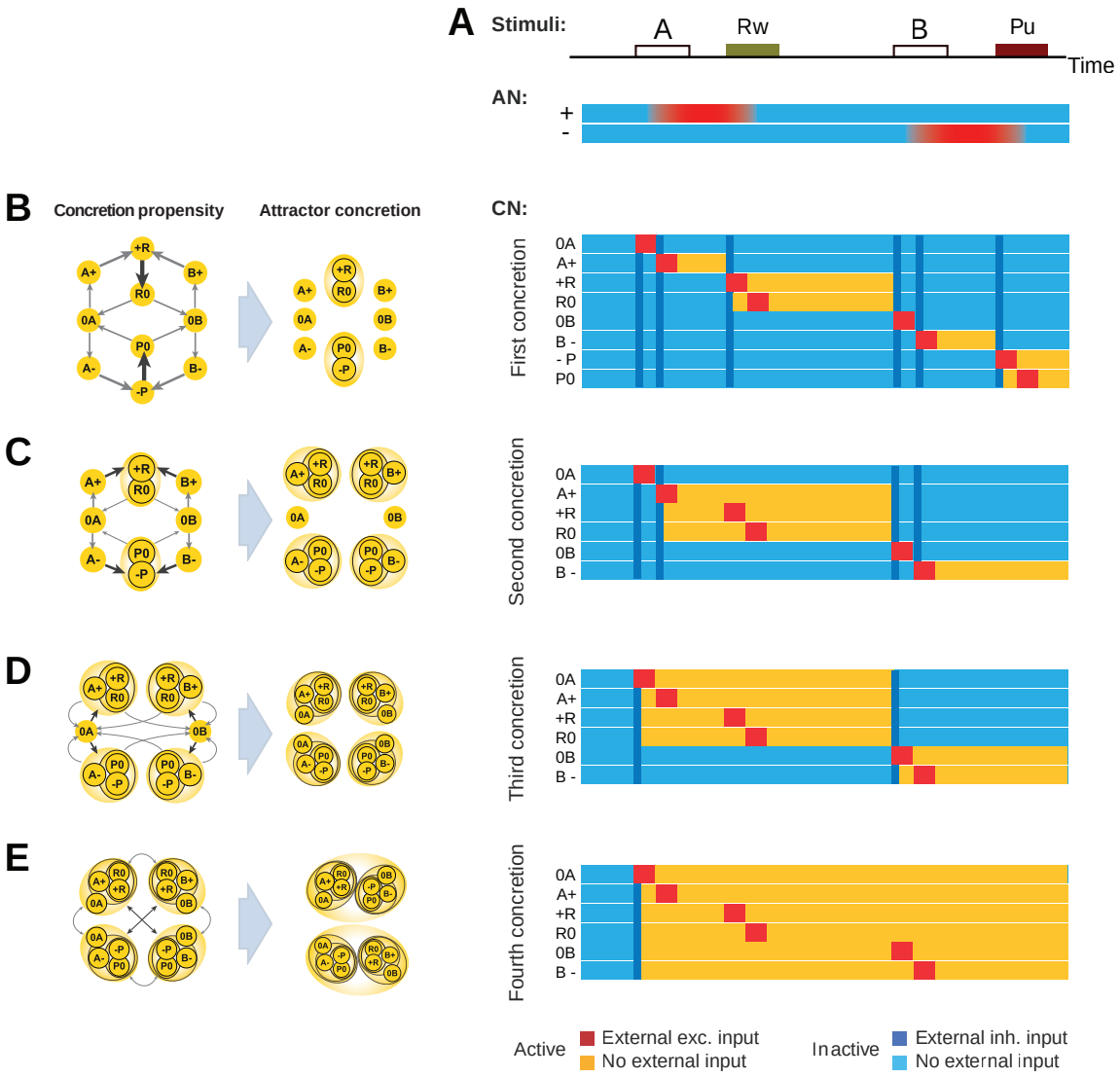

FIG. 3.7: Second learning phase: attractor concretion. **A**, Scheme of two trials and color coded activity of the AN as a function of time as in Fig. 3.6. **B,C,D,E**, From left to right: Scheme of propensities to concretion, scheme of attractors following concretion, Color coded activity of the CN populations as a function of time as in Fig. 3.6 following the concretion. **B,C,D,E** describe different iterations of the concretion process (see the text for a detailed description).

turns on also the R0 population, and from now on, the two populations will always co-activate since they are part of a new compound.

The next iteration is again driven by the concretion propensities, however now there are new attractor states in the CN, and all the propensities must be recalculated. The new scheme of propensities is shown in Fig. 3.7C. The next concretions are again predicted correctly by the propensities. The same process is iterated in Fig. 3.7D and in E, where we finally obtain the representations of the two contexts. Notice that at every iteration the width of the arrows progressively decreases. This is due to the fact that the CN spends more and more time in the new attractors and hence the propensity to concretion with other attractors decreases because of the Hebbian component of learning. At some point the process of concretion stops because the propensity is too small to induce a concretion. We choose the learning rates to stop the process as soon as we have the representations of the two contexts (but see also the Discussion for a different choice of parameters). The full simulations of the learning process are shown in Fig. 3.8.

Predicting context dependent values: the expected behavior

After the learning process described in the previous section, the CN contains a representation of the current context. When the CS-US associations are reversed, the first time a CS is presented, the value is predicted incorrectly by the AN. This resets the synapses from the neurons representing the external events to the AN, and the transitions to the positive and negative states become random with equal probability Fusi et al. (2007). At the same time a 'surprise' signal is generated in the CN and the attractor representing the context is reset. If the AN selects by chance the wrong value state, then it keeps selecting the state randomly. As soon as

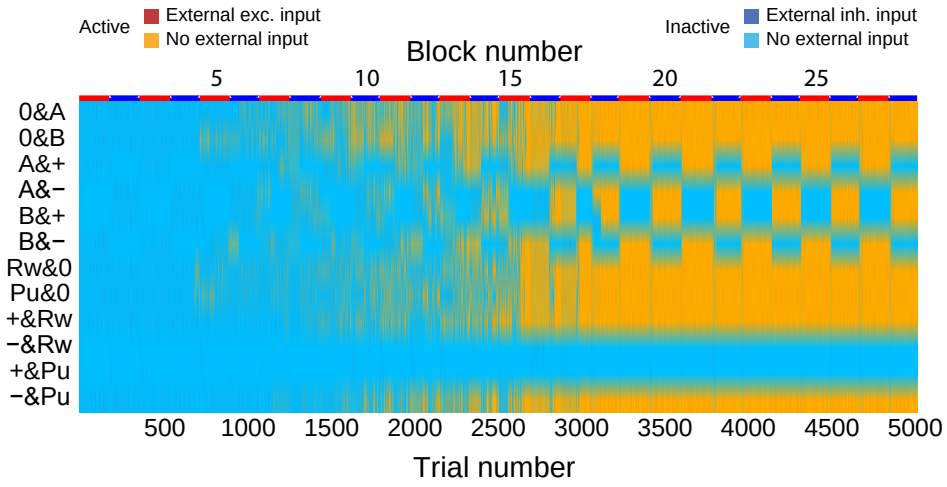

FIG. 3.8: Full learning simulation. Color coded activity of the CN populations as a function of time as in Fig. 3.6. Red and blue bars above the plot indicate context 1 and 2 respectively. Temporally contiguous attractors merge into single representations of short temporal sequences (attractor concretion). Eventually, the context representations emerge, and they are demonstrated by the coactivation of the attractors representing all conjunctions of events and AN states in each context (e.g. 0A, A+, +R, R0, 0B, B-, -P, P0 for context 1).

the AN selects the correct value, then it also activates the correct context in the CN, and the AN-CN system starts predicting the correct values for all CSs. Although it is not possible for our neural circuit to switch with probability one from one context to the other in one trial, it is still possible to harness the information contained in the context representation of the CN in order to improve the prediction of the US. Indeed, as soon as the AN guesses the correct value for one CS, say CS A, the CN also selects the correct context and then it is possible to predict the US that follows CS B with certainty. Summarizing, as soon as the context changes, the AN predicts the wrong value. The surprise signal resets the CN context, and then the AN-CN system selects randomly one of the two possible contexts until it guesses the correct one. This strategy is less efficient than switching to the alternative context

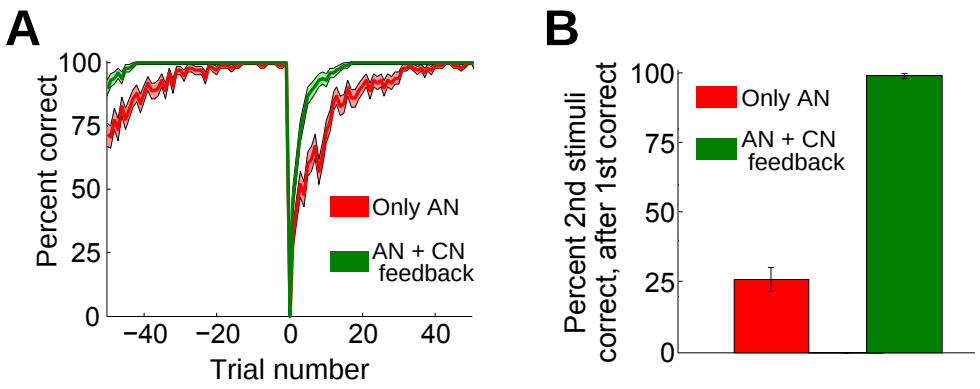

FIG. 3.9: Harnessing the feedback from the CN to the AN. **A**, Performance of the AN around a reversal (at trial number 0) without (red curve) and (green curve) the CN feedback. Performance drops to almost 0% right after a reversal, quickly goes to chance level and recovers exponentially to maximum. Introducing the CN feedback conveying the context information results in an effective learning rate increase. This effect is due to the information gained about one CS while learning the value of the other. **B**, Percent of correct predictions of the value of one CS when the new value of the other CS is already known. The performance is estimated immediately after a context switch. In the absence of the context information provided by the CN, the performance is significantly worse (left) than in the presence of CN feedback, when the performance is close to 100%.

as soon as one knows that the context has changed, but it is still more efficient than learning independently the associations, as it allows the AN-CN system to predict correctly the value of all the CSs once it knows the new value of one CS.

This mechanism is implemented in our neural circuit by the feedback from the CN to the AN, as described in the Materials and Methods. The CN and the external neurons project to a population of randomly connected neurons which represent mixtures of the CN context representations and the external events. These neurons contain the information about the current context and the occurring event. The synapses between these neurons and the AN are plastic, and they are modified in the same way as the synapses from external neurons to the AN, which encode

simple Pavlovian associations. As soon as the context is correctly determined by the CN, it is simple to predict the values of both CSs as the AN sees the CN as an additional input that represents explicitly the current context. Indeed, Fig. 3.9B shows the percentage of correct predictions of the value of one CS when the neural circuit has already guessed correctly the value of the other CS. In other words, we quantify the ability of the neural circuit to use the context information to infer the value of one CS once the value of the other CS is known. In the absence of context information, this percentage is at a significantly lower level, that depends on the specific sequence of events (left). In the presence of the CN feedback, this percentage is close to 100% (right). This behavior is in principle observable in an experiment and the plot of Fig. 3.9B provides us with a behavioral criterion for establishing the existence of context representations.

3.3.2 Experimental predictions about the response properties of recorded cells

As the process of concretion takes place, the neural representations evoked by events like the presentations of the CSs or the delivery of the USs, become progressively more similar in the CN. For example, the activation of CS B& Negative (B-) should eventually evoke the activation of the attractor representation of context 1. Neurons that initially were activated by CS B& Negative only, are predicted to be activated also by CS A & Positive. In particular, as the process of attractor concretion starts from the events that follow each other with the highest probability, the first compounds that form are likely to be [+R,R0], [-P,P0], followed by [0A,A+], [0B,B-], for context 1, and [0B,B+], [0A,B-] for context 2 (see Fig. 3.10).

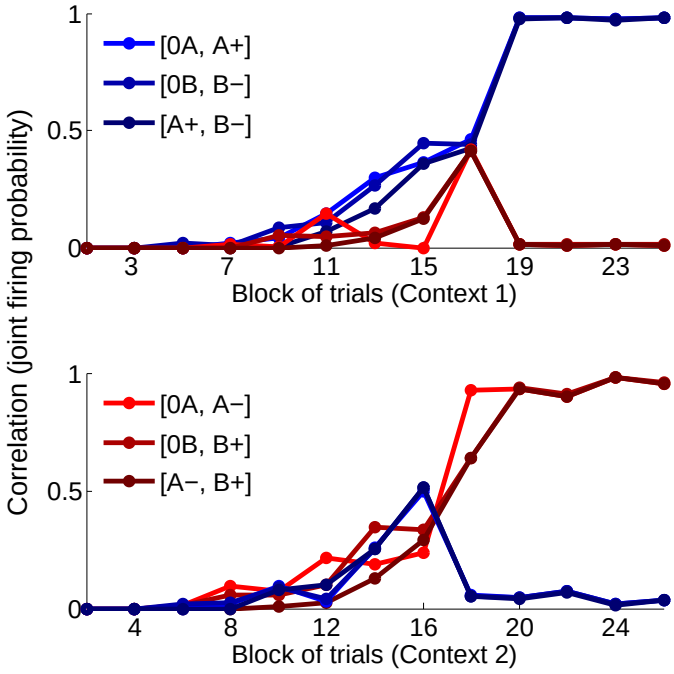

FIG. 3.10: Predictions on the correlations between neurons that respond to conjunctions of events. The probability that two populations of neurons of CN are co-activated is computed by running the simulation several times and it is plotted as a function of the number of blocks of context 1 (top) and context 2 (bottom) trials. Different colors denote different co-activated populations. Initially the probability is zero, as we assumed that in the CN there are only populations that respond to simple conjunctions of events and AN states. As learning progresses, the probability of co-activation of populations that represent events of the context increases. For example, in context 1, neurons that initially respond only to CS A & Positive (A+), after 15 reversals (block 16), respond also to CS B & Negative (B-). The corresponding compound is denoted by [A+,B-].

This prediction can be tested experimentally by single unit recordings. We predict that the fraction of recorded cells that are activated by events that belong to the same context, should increase monotonically with the total number of trials that the animal experiences. Notice that in the simulations of Fig. 3.10 we assumed that there are initially no neurons that already respond to conjunctions of more than two events. In the brain there might be neurons that from the very beginning respond already to those conjunctions of events that represent the contexts in the CN. The probability of finding those neurons is predicted to be significantly smaller than the probability of finding neurons that respond to simple conjunctions of events, like those that we simulated, but in general we cannot exclude that they are already present before the learning process starts. Consistent with our predictions, in the trace conditioning experiment we observed 11/216 cells in OFC and 15/222 cells in the amygdala that are significantly selective to the conjunctions of events that represent the context (see Materials and Methods for the details of the analysis). Our analysis shows that these cells may only be significantly selective by chance, or anyway they may simply be a very small fraction of the recorded cells. In both cases, our assumption that the majority of cells responds to simple conjunctions of events is correct.

3.4 Discussion

We proposed attractor concretion as a possible mechanism for creating representations of contexts. The formation of these representations leads to the generation of new mental states that were not present in the initial AN-CN system. Indeed, the CN, at the end of the learning process operates as an additional input to the AN that contains information about the active context. When the AN-CN system is considered, the number of mental states eventually doubles, as for each AN state there are two states of the CN. In practice, two new sets of mental states are generated as soon as the CN starts encoding the two contexts. In the intermediate stages of the iterative process that leads to the formation of the new states, the CN neurons are activated by short sequences, which generate a number of mental states that is larger than the final one. However the temporal statistics of these sequences is not correlated with any useful information about the contexts. In particular they do not allow the AN to disambiguate between the CS-US associations of the two different contexts. Indeed in the first three stages (in Fig. 3.7B,C,D) there is no evidence of the existence of temporal contexts, both in the transition probabilities and in the propensities. It is only at the end of the concretion iterative process that the CN can detect a clustered structure in the transition probabilities (i.e. large probabilities to make a transition within the context, and small ones to switch to a different one). We believe that the kind of merging behavior we observed in the case we analyzed is very general and it applies to a wide variety of tasks in which the information about temporal contexts are important to improve the performance.

In our manuscript we analyzed a particularly difficult case, because the CSs and the USs appear in the two contexts in a perfectly symmetric way, and the

contexts are solely defined by the temporal statistics of the events. Our model would certainly work and generate quantitative predictions in simpler cases, in which for example a particular CS-US identifies unequivocally one context, or in which an explicit contextual cue appears in all or in some trials to signal which context is active. The concretion rules illustrated in the schemes of Fig. 3.7 allow us to make predictions about the typical behavior of our model network, although it is always important to run a full simulations to analyze the behavior of the concretion process. For example, the representation of a contextual cue that appears in every trial would have the highest propensity to merge with CS-US associations, and it would operate as a kernel around which the context representations are built. The details of this learning process obviously depend on the specific task protocol, on the representations of the CSs and the USs and on the parameters of the model, but the model can certainly generate quantitative predictions in a wide variety of experiments.

3.4.1 A hierarchy of context representations

The final context representations are determined not only by the temporal statistics of the events, but also by the parameters of the neuronal and synaptic dynamics. In a system with multiple neural circuits, each characterized by different dynamical parameters, we expect the creation of a hierarchy of contexts that correspond to different processes of attractor concretion. Some neural circuits can represent only single events, as the TSL component is not strong enough, some others might represent general contexts that correspond to conjunctions of many events, stimuli and internal conditions. For example, there could be cases with no merging, in which the network simply represents the individual events A-Reward, B-Punishment, A-

Punishment and B-Reward. For the parameters used in our simulations the patterns of activity corresponding to A-Reward and B-Punishment merge into the representation of context 1, whereas A-Punishment and B-Reward merge into context 2. For other sets of parameters there could be a unique, large compound comprising all possible events of the context corresponding to the general task. Such a compound would link together A-Reward, B-Punishment, A-Punishment and B-Reward. All these neural circuits with different neural and synaptic parameters are likely to be simultaneously present in the brain, either in one particular area, or spread across different areas. They would provide the brain with a hierarchy of contexts at many different levels that all together determine a general mental state. A population of cells reading out all these context representations could easily encode the value of the current state, and such a value would in general depend on which context pattern is activated in every level of the hierarchy.

The use of a hierarchy of contexts generated by the heterogeneity of the network is a new possibility that will be investigated systematically in future theoretical and experimental studies. In particular, it should be possible to detect to which hierarchical level a recorded neuron belongs. This could be done by manipulating the transition probabilities between events. For example, we considered an experiment in which there are only two CSs and the probability to make a transition from a trial with one CS to a trial with a different CS within the same context is simply $1/2$. In the case of $2p$ CSs, this probability would be $1/2p$. The probability to switch to a different context should be accordingly reduced when all the CSs are presented at least once in each context ($\ll 1/2p$). Hence the clustered structure of transition probabilities that defines the temporal context would still be detectable for any $p > 2$, but the transition probabilities would all be rescaled down when p increases.

The parameters of the neural circuits that determine the propensity to concretion and the maximal number of CSs are always related. In particular, neural circuits with a small propensity to concretion would generate context representations only if p is small enough. $p = 2$ maximizes the probability that we observe what we described in the experimental prediction section. However we also have the additional prediction, that for any neural circuit that shows concretion for $p = 2$, there is always a p that is large enough so that no concretion should be observed. Of course the entire brain would still be able to create context representations because there would be a different neural circuit with the proper parameters to generate the context representations in a different situation.

3.4.2 Previous experimental and theoretical works on temporal context representation

Neural signals that could provide a representation of temporal context have been observed in several experiments. For example, Miyashita investigated the representations of sequences of visual stimuli(Miyashita and Chang, 1988). In these experiments, a monkey was trained to perform a Delayed Match to Sample task in which the sample stimuli were presented in a fixed temporal order. Single neuron activity recorded in infero-temporal cortex revealed that cells activated by one particular stimulus were likely to be activated also by the neighboring stimuli in the temporal sequence. In other words, the spatial patterns of neural activity across multiple cells, induced by temporally contiguous stimuli, were highly correlated, reflecting the order of presentation of the sensory events. This work inspired several theoretical works on the neural mechanisms underlying context representation in

the brain. For example, Griniasty et al (Griniasty et al., 1993) interpreted this data as being an expression of the recurrent dynamics of cortical circuits. In the model the authors proposed, these circuits initially produce different patterns of stable, reverberating activity in response to individual sensory stimuli. If the stimuli are repeatedly presented in a fixed temporal order, the synaptic connections are modified so that a sensory stimulus activates a pattern of activity that is correlated not only to its representation when presented individually, but also the representations of the stimuli that surround it in the temporal sequence. This pattern of reverberating activity may be a neural representation of the context in which the sensory stimulus appears. A more detailed model of the learning process that is responsible for tuning the synaptic weights has been proposed by (Brunel, 1996) and some of the predictions have been verified in experiments (Yakovlev et al., 1998).

In all these models, each stimulus evokes a different representation and the context is encoded in the correlations between the representations. Highly correlated neural patterns correspond to stimuli that belong to the same context. In our model, attractor concretion leads to new inseparable "entities" that represent contexts. Different events, like visual stimuli, activate the very same pattern of activity if they belong to the same context. This is different from the representations in which the contexts are encoded in the correlations because in that case, each event still activates a characteristic pattern of activity that is unique, though similar to the patterns elicited by the events of the same context. Our approach is similar to what has been proposed in (Brunel, 1996) for pair associates. One of the advantages of creating new entities that represent context, is compositionality. Context representations can easily merge to generate new compounds even if they are highly structured and complex. In other words, when a new entity is created,

there is a significant reduction of the dimensionality of the state space. Indeed if two populations are always coactivated, they behave as a single one, reducing the effective number of independent populations, and hence the dimensions of the state space. This reduction greatly simplifies the process of representation of complex contexts. This is not true in the case in which each individual event contains the information about the correlation with all the other population of neurons (i.e. with all the other dimensions). One of the disadvantages of creating new entities is that the information about individual events is lost, unless multiple systems are considered as discussed in the previous section of the Discussion.

3.4.3 Interacting systems learning over different time-scales

An architecture similar to the one we suggested, in which the fast learning AN trains the slower CN network, has been also put forward in McClelland et al. (1995). The authors of this work suggest that fast learning in the hippocampus interacts and "trains" slower cortical connections, in order to combine rapid information acquisition with generalization over longer time-scales.

Neurophysiological studies in monkeys executing a visuomotor association task also reveal the presence of two learning systems following different time courses. Simultaneous recordings in fact reveal rapid changes in the neural activity of the striatum, compared with a slower modification in the prefrontal cortex. These activity in PFC was also more correlated with the slow improvements in behavioral performance (Pasupathy and Miller, 2005). Motivated by these results Miller and Buschman (2008) attribute to the BG the role of a supervised fast learning module "training" the slower and less supervised Prefrontal Cortex in charge of creating more abstract cognitive structures.

Functional magnetic resonance imaging studies (fMRI) (Poldrack et al., 2001) and double dissociation studies of the performance of amnesiac and Parkinson subjects (Shohamy et al., 2008) also suggest the presence of two learning systems differentially engaged in different phases of the acquisition of a probabilistic task. What these studies suggest is that the mediotemporal lobes are involved in rapid associative learning, while Basal Ganglia (BG) incrementally integrate experience over longer time-scales in a feedback-based manner.

3.4.4 Alternative approaches to the creation of context representations

As we briefly discussed in the Introduction, one of the limitations of Reinforcement Learning (RL) algorithms is that a representation of mental states is assumed, yet the algorithms do not provide for how these representations are created. This is a major limitation when the environment is only partially observable, as in the case of limited sensory data, or when agents have limited computational resources to process all the details of the sensory stimuli. In all of these cases, we often might be induced to think that we are in the same situation when we actually are not and we would need to select different actions. For example, when we drive in a forest, we often arrive at two similar crossings, which can lead to confusion if we don't take into account other information, such as where we have been recently. Thus in many circumstances it is possible to decide our actions if we remember some of our previous experiences. For example, one crossing might be preceded by a pump station and the other might be preceded by a level crossing. In this situation, we need to create two distinct mental states, each corresponding to a

different temporal context.

We proposed a simple mechanism and a biologically plausible neural network model that autonomously generates context representations. Alternative and complementary approaches have been proposed to solve analogous problems in different fields. For example, Hidden Markov Models (HMM) have been widely used to predict complex temporal series where the next event might depend on a long sequence of previous observations (see e.g. (Rabiner, 1989)). In these models and in some of their extensions to decision processes (e.g. POMDP, Partially Observable Markov Decision Processes (Kaelbling et al., 1998)), the number and the meaning of the states of the agent are not known a priori. The algorithms usually start from a large number of hidden states that are randomly linked to the observed states of the environment. The states acquire a meaning by iterating a recursive algorithm that estimates both the probabilities that a hidden state is related to an observed state, and that it is followed by another state (see e.g. Viterbi or Baum-Welch algorithms). Although very efficient in many applications like speech recognition, these algorithms suffer from many limitations. They require knowing the number of states a priori. If there are not enough hidden states, then it is possible to add more, but estimating all the transition probabilities between hidden states becomes rapidly an intractable problem as the number of states increases. Moreover, all probabilities should be revaluated every time a new state is introduced, which makes the system rather inflexible. Finally, the convergence of the recursive algorithms to the global optimal solution is not guaranteed and the final scheme of hidden states and transition probabilities depends strongly on the initial condition. Some of these limitations might be related to the fact that the hidden states are initially chosen in a completely random fashion and they are not constructed on

the basis of the temporal statistics of the events. Although we still do not have a general theory and a convergence theorem as in the case of HMMs, we believe that our approach does not suffer from these limitations, and it is closer to the mechanisms that the brain might be using to create context representations.

3.4.5 Where are the cells of the AN and the CN in the brain?

As shown in Fig. 3.5, neurons with mixed selectivity to the relevant conjunction of events for defining the two contexts have been observed both in pre-frontal cortex (PFC), in particular, in OFC, and in the amygdala. As these two areas are strongly interacting (Salzman and Fusi, 2009), it is likely that the neural circuits of the AN and the CN are distributed across these brain regions, and probably, also other areas (e.g. in other subareas of PFC, as well as in the hippocampus and related structures).

3.4.6 When temporal contiguity is broken by intervening distractors

In many realistic situations there are contexts in which the temporal contiguity between relevant events is broken by distractors, e.g by the presentation of a random visual stimulus. In our model, these distractors would disrupt the process of formation of context representations. There are at least two possible solutions to this problem. The first one has been proposed in (O'Reilly and Frank, 2006), and it is based on a gating system that can learn to ignore the irrelevant events. In these models the irrelevant events are simply 'gated', and hence not represented in the

AN and in the CN. The second possibility is based on short term synaptic mechanisms that could preserve the memory of relevant events even when distractors are presented (see Mongillo et al. (2008) for a possible mechanism based on short term synaptic facilitation). For example, the TSL mechanism could be implemented by tagging the synapses at time t, and then modifying them in the next time interval τ, creating links between the event occurring at t and all events occurring the next time interval τ. Such a mechanism would suffer from the introduction of the inherent time constant τ of synaptic tagging, whereas our mechanism can work and generalize with almost any timing between successive events. Notice that synaptic tagging could in principle allow us to eliminate the first phase of learning, in which the attractor representations of individual events are created to bridge the temporal gap between events that are separated in time.

3.4.7 More general mental states and operant tasks

We focused on a trace conditioning task to illustrate the proposed mechanism for the formation of context representations. However the same mechanism can be applied to other experiments and to operant tasks. For example in (Asaad et al., 1998; Pasupathy and Miller, 2005) the monkeys are trained to associate saccadic movements to visual responses. The AN model has already been used to reproduce quantitatively the observed behavior of the monkeys when they learn and forget visuo-motor associations (Fusi et al., 2007). The positive and negative states are equivalent to the decisions of the monkey about the direction of the saccade (left or right). In one context stimulus A is associated with left and B with right, in the second context the associations are reversed. Although in the experiments the monkeys never learned to switch from one context to another immediately, it is

possible that in other conditions (see the Discussion of Fusi et al. (2007)) they would be able to create the representations of the contexts. The model and the predictions would be the same as for the trace conditioning task. Notice that the two contexts correspond to two simple rules that could be expressed as 'whenever A is associated with left, B is associated with right' and 'whenever A is associated with right, B is associated with left'. In this case the representation of the temporal context would be equivalent to the representation of a rule. In recent years, investigators accumulated evidence that the activity of prefrontal neurons can encode abstract rules (Genovesio et al., 2005; Mansouri et al., 2007; 2006; Wallis et al., 2001). These rules allow the animal to respond to the same sensory stimulus in different ways depending on the strategy or on a sequence of sensory cues preceding the stimulus. Hence, they are all analogous to the temporal contexts that we studied here. In one of the cited experiments, Tanaka and colleagues observed sustained activity in the inter-trial intervals that encode the rule in effect when the monkey was performing a simplified version of the Wisconsin Card Sorting Test (Mansouri et al., 2006). This rule was determined by the temporal context of monkey choices, and the rule selective inter-trial activity therefore corresponds to an active representation of context (O'Reilly and Munakata, 2000; Loh and Deco, 2005; Deco and Rolls, 2005; Rigotti et al., 2008; Rougier et al., 2005). In all the models and the experiments that we described, a large proportion of neurons exhibit mixed selectivity. Interestingly, it has been shown (Dayan, 2007) that mixed selectivity neurons implemented with multilinear functions can actually play an important role in neural systems that implement both habits and rules during the process of learning of complex cognitive tasks. Multilinearity implements conditional maps between the sensory input, the working memory state, and an output representing the motor response.

Temporal contiguity can also be important in the creation of invariant representations. Some investigators propose that invariant representations of objects can be generated by linking temporally contiguous views (Rolls and Milward, 2000; Miyashita and Chang, 1988; Li and DiCarlo, 2008). This is yet another example of an abstraction process that relies on temporal contiguity to create the internal representations. As shown in our manuscript, temporal contiguity can also be an important aspect of the statistics of the environment in all the cases in which behavior depends on information about temporal context. We believe that it is particularly important to study the neural mechanisms that allow the animal to encode complex patterns of temporal contiguity as these processes might underlie the neural basis of cognitive functions that range from the creation of invariant representations to rule abstraction.

Chapter 4

Conclusion

4.1 Final remarks

The spiking activity of neurons recorded in behaving animals is typically very heterogeneous, with diverse selective responses encoding different aspects of task events. This diversity is especially bewildering with regard to the prefrontal cortex, a brain structure that has been shown to be critically important for higher cognitive behaviors as it is at the top of the sensory-motor hierarchy (Fuster, 2001). Diversity in the neural responses of PFC has been for instance well characterized during the execution of conditional association tasks. Several studies (Watanabe, 1986; Hasegawa et al., 1998; Asaad et al., 1998) have identified mixed selectivity to conjunctions of cues and the associated delayed motor responses. Bichot et al. (1996) reported neurons in the frontal eye field (FEF) which developed selectivity to the conjunction of position and stimulus color.

Thanks to the valuable insights from neurophysiology and pioneering proposals for the role of PFC in executive control through the representation of context (Cohen and Servan-Schreiber, 1992), there has been steady advancement in computational techniques towards the production of detailed network models of the role

of PFC in rule-based implementation of arbitrary associations. Developments of large-scale architecture in behavioral neuroscience (O'Reilly and Munakata, 2000; Frank et al., 2001), and ideas borrowed from Reinforcement Learning to implement multi-modular system models of decision-making (Daw et al., 2005) are merging to produce networks able to cope with rule-based control (Dayan, 2007; 2008). On the other hand the complex dynamics of realistic networks has been successfully harnessed to reproduce these results in biologically plausible neural substrates (Wang, 2002; Soltani and Wang, 2006; Fusi et al., 2007).

Despite all these progresses, there does not exist a unified theoretical framework about the function of the heterogeneity in the response properties of higher associative areas. The need for the development of new models going in this direction has been stressed very recently in a contribution detailing a sophisticated analysis of neural data recorded in PFC during a delayed somatosensory discrimination task (Jun et al., 2010). This paper explicitly compared the distribution of the experimentally observed selectivity with the predictions made by two paradigmatic models which where proposed for this specific task. The first model was published by Machens et al. (2005) and is based on quasi-continuum attractors, while the second was laid down by Miller and Wang (2006) and uses integral feedback. Both of these models reproduce the basic behavior and the most apparent neural responses. Nevertheless, none of them satisfied more stringent comparison criteria requiring that the cell types predicted by either model dominate the data. What the authors concluded is that the inherent modular structure of the available models are not able to capture the great distribution, heterogeneity and variety in response types.

The type of model we propose in Chapter 2 naturally displays these characteristics. Our theory is based on transitions between mental states represented as

stable patterns of self-sustained activity (attractors), which are implemented by mixed selectivity neurons. These neurons, we showed, find a natural instantiation as Randomly Connected Neurons (RCNs). The kind of selectivity arising from this class of networks is highly heterogeneous, intermittent and distributed as illustrated by the analysis of the models of rule-based behavior detailed in Section 2.2.

It is crucial to observe that random connectivity does not correspond to random responses. Since synaptic randomness is a quenched disorder, random connectivity is rather manifested as quenched heterogeneity, i.e. cells displaying idiosyncratic responses to particular combinations of task variables, but nonetheless reliably spiking in response to such combinations from trial-to-trial.

The creation of new mental states by means of temporal-contiguity-driven learning (Chapter 3), we think, could also account for how heterogeneous activity is modified with training. The selectivity to combinations of position and stimulus color in FEF reported by Bichot et al. (1996) is a property which develops with experience. Analogously, the stimulus-planned motor action mixed selectivity reported in Asaad et al. (1998) seems to arise during learning of the task, because of a progressive shift of the activity related to the response from the end of the trial to the stimulus presentation. This kind of learning effects could be due to the formation of a mental state, similarly to what we proposed in Chapter 3 for the trace conditioning task of (Paton et al., 2006). It would be certainly interesting to model this specific type of task to generate some predications about the exact dynamics of the appearance of such signals with learning, and for instance, contrast it with models where these kinds of temporal shifts would be completely attributed to a teaching signal like a TD-error (Schultz et al., 1997) or the direct unexpected reward (Hazy et al., 2005).

4.2 Future directions

The theoretical ideas exposed in this thesis could be developed in several very interesting directions. The framework of Chapter 2 implementing rule-based behavior through event-driven transitions between mental states could be used directly to model known and well characterized decision tasks. A convenient and timely example would be the task recently analyzed in Jun et al. (2010). It would be interesting to see how well the distribution of selectivity predicted by our model would fit the one reported from the data. Perhaps even more interesting would be modeling the entire learning process, capitalizing on the ideas presented in Chapter 3 about the role of temporal contiguity in the acquisition of the temporal statistics of the ask through the formation of new mental states.

Spiking Q-learning

An ongoing project which goes some first steps in this direction is the unification of methods of Reinforcement Learning (RL) theory (Sutton and Barto, 1998) with realistic spiking attractor neural networks. RL has recently enjoyed a wave of popularity in the neuroscience community because of its successful application in interpreting the functional role of midbrain dopamine neurons (Schultz et al., 1997). The temporal difference (TD) algorithm offer in fact a sound theoretical framework to interpret the phasic dopamine discharge as a reward prediction error, a teaching signal which indicates a surprising outcome and promotes learning of the correct value. The inclusion of these principles in the framework of mental states and event-driven transitions of Chapter 3 would allow us to capitalize on the wealth of methods developed by the machine learning community. Most of all, it would

provide us with a biologically plausible learning mechanism to automatically modify the behavior of the network in response to changing rules in a variable environment.

Some preliminary simulations obtained by including subpopulations selective to reward, value, and prediction error in a spiking implementation of our network developed in Mattia et al. (2007) show some promising preliminary results towards this union of RL and spiking ANN. The way RL is implemented in spiking network is by resorting to a version of TD-learning proposed by Watkins (1989) which is called Q-learning. Q-learning is essentially a technique to learn the value of a given action in a given state (the Q-value), so that the agent can always carry out the most valuable action, i.e. the one which predicts highest expected cumulative reward. It is moreover an algorithm which is known to converge to the optimal solution in stable environments (Watkins and Dayan, 1992). The idea of our spiking implementation of Q-learning is simply to encode the Q-values in the synaptic efficacies of neurons displaying mixed selectivity to mental states and external-events, so that any such a combination triggers a transition to the most valuable mental state.

Although some very recent computational studies have already implemented spiking simulation learning through RL techniques (Potjans et al., 2009), our approach has a different motivation. While the scope of Potjans et al. (2009) (and of most RL-based approaches, for that matter) is to model an agent which is able to faithfully predict the responses of a totally observable environment, our framework centered around the representation of mental states is underlaid by a different viewpoint. The idea is essentially to model a situation in which a mental state selects and activates other mental states on the basis of external evidence, but most of all on the basis of the correlational structure of the mental states themselves. Since this

structure is a repository of the acquired contingencies, and of the relevant affective and motivational aspects, this approach has the advantage of automatically implementing a behavioral outcome which is relevant to the agent and the achievement of current goals. Another important advantage of this approach is the possibility to apply it to navigate partially observable environments (Kaelbling et al., 1998), without the danger of incurring the 'curse of dimensionality' (Bellman, 1957) and the 'computational noise' (Daw et al., 2005) which haunts models trying to implement an optimal policy. An agent conceived within the framework based on mental states would instead start with tentative suboptimal responses and gradually create new mental states iteratively better approximating an optimal response, which incidentally seems to be the strategy employed by healthy subjects when facing complex probabilistic category learning (Shohamy et al., 2008).

Probabilistic tasks

It is interesting to point out that the AN-CN system of Chapter 3 naturally implements simple strategies, like Win-Stay, Lose-Switch (Barraclough et al., 2004). This is simply because an attractor representing a context in the CN stays active until a contingency incompatible with the context comes along; at which point the attractor is reset and a new one is probabilistically selected by the subsequent AN response. The attractor concretion framework could be easily extended to more complex stochastic tasks with probabilistic contingencies, like the probabilistic delivery of reward in a weather prediction task (Gluck et al., 2002). Such an implementation would however need the inclusion of an additional probabilistic component in the network. More specifically, it would require an element able to quantify *uncertainty*. For instance, if a reward is not attributed, it may be because

the response is correct but the delivery of the reinforcement is stochastic, or it may simply be because the response was wrong. Quantifying the degree of how certain one is about the response would therefore help disambiguate the interpretation of this feedback. In particular, we should allow our implemented surprise signal to reset an active attractor only when enough evidence for the need of a switch has been accumulated, in such a way that spurious signals incompatible with the current attractor would be ignored.

A way to implement such an accumulation of evidence has been presented by Yu and Dayan (2005). In that research the authors showed how to utilize uncertainty signals to perform optimal inference based on unreliable observations in changing contexts. Their proposal was to keep track of two uncertainty variables, one standing for *expected* uncertainty, i.e. quantifying how unreliable we know a given outcome can be, and one standing for *unexpected* uncertainty, i.e. telling us how unexpected some observation are, given some prior assumption. The idea is simply that, if the unexpected uncertainty variable goes above the expected uncertainty, then the current prior assumption doesn't account for the observed variability, meaning that it is time to consider some other assumptions. Yu and Dayan (2005) proposed that these two kinds of uncertainty could be represented by acetylcholine and norepinephrine, respectively.

Alternatively, certainty/uncertainty signals have been shown to be directly represented in brain activity. For instance, fMRI experiments reveal signals correlated to uncertainty during perceptual categorization tasks (Grinband et al., 2006). Another example is offered by electrophysiological recordings in monkey LIP that show signals representing the confidence of the animal in its own responses (Kiani and Shadlen, 2009). This kind of information could be used to implement a proba-

bilistic surprise signal, maybe in concomitance with a mechanism gating the neural activity of stimuli which have to be ignored (Vogels and Abbott, 2009).

Experiments to test theory predictions

Another possible future direction of the presented projects, which is actually an undergoing collaborative project with Dr. C. Daniel Salzman's lab, is to test the predictions of the theory about the creation of representations of different mental states. If temporal contiguity underlies the formation of new mental states representing context, then it should be possible to generate new mental states, modify neural selectivity, and change the value of mental states by manipulating the temporal statistics of events. In order to do that, the proposed idea is to train monkeys to perform a context-dependent version of the trace conditioning task (Paton et al., 2006), in which the value of CSs depend on context. As for the data presented in Chapter 3, the target of electrophysiological recordings are the amygdala and orbitorfrontal cortex (OFC), which are reciprocally connected and contribute to different aspects of valuation and decision-making (Padoa-Schioppa and Assad, 2008; Paton et al., 2006) and were recently discovered to encode the positive and negative values of the mental states induced by the visual stimuli and reinforcements (Paton et al., 2006; Belova et al., 2008; Morrison and Salzman, 2009). The idea of the task is to alternate blocks of trials containing two sets of CS-US associations (corresponding to two contexts) until the monkey creates a mental representation of both and can rapidly switch from a mental state corresponding to one context to the mental state corresponding to the other. One of the predictions of the theory is that neural representations of CS-US contingencies appearing in the same temporal context will be highly correlated, meaning that neurons that respond to one par-

ticular conjunction of a CS-US will tend to respond also to the other contingencies appearing in the same context. If such correlations are indeed observable, it should be possible, under the theory of mental states creation, to modify them by driving the formation of new context representing mental states. Preliminary results obtained in a version of this task in which the context is explicitly cued, but only in a random fraction of the trials, already showed that after a context change monkeys can rapidly adapt their differential anticipatory licking to reflect the rewarded and unrewarded CSs in the block. This happens even when the contextual cue does not appear on each trial. In accordance with these behavioral changes, Dr. Alex Saez in Salzman's group observed that a significant fraction of neurons maintain selectivity to the context and switch their response levels depending at a context change, even before the reinforcement contingencies of a new block had been experienced. These signals are reminiscent of the rule-related activity observed by Wallis et al. (2001) and (Mansouri et al., 2006) and motivate further experiments to investigate how they are created.

Appendix A

Attractor Neural Networks

Stability parameter and basins of attraction

Because of its importance, researcher in the field of associative networks were lead to considering the question of ensuring a sufficiently large basin of attraction. Let us formulate the problem in mathematical terms to clearly state the underlying problématique.

We adopt a Hopfield-type model where the activity of neuron i in the network at time t is denoted by $S_i(t)$, a variable which for simplicity we define as being equal to 1 if the neuron is firing, and to -1 if the neuron is inactive. The Hopfield model assumes a dynamics of the type:

$$S_i(t+1) = \text{sign}\left(\sum_{j=1}^{N} J_{ij} S_j(t)\right), \qquad i = 1, \ldots N,$$

which says that the activity of neuron i at time $t+1$ is determined by the sum of the activity of all the N neurons in the network weighted so that a given neuron j gives a contribution proportional to the efficacy J_{ij} of the synapse connecting neuron j to neuron i. If this sum is positive, neuron i fires, otherwise it remains silent.

A given set of patterns of activity ξ_i^μ, $\mu = 1, \ldots, p$ is therefore stable across time (is a set of fix points of the dynamics) if setting $S_i(t) = \xi_i^\mu$ for $i = 1, \ldots N$ and $\mu = 1, \ldots, p$ at time t, implies that the activity pattern does not change at time $t + 1$, i.e. $S_i(t + 1) = \xi_i^\mu$ for $i = 1, \ldots N$ and $\mu = 1, \ldots, p$. The fix-point activity patterns and the synaptic matrix have therefore to satisfy the fix-point equation:

$$\xi_i^\mu = \text{sign}\left(\sum_{j=1}^{N} J_{ij}\xi_j^\mu\right), \qquad i = 1, \ldots N. \tag{A.1}$$

The fix-points of the neural dynamics ξ^μ are then proper *attractors* if, a part from satisfying this last equation, they also allow for their recall when the network is probed with an activity pattern which is sufficiently close to them. In other words we want that if the network activity at time t is $S_i(t) = \xi_i^\mu + \Delta_i$, where Δ_i are the component of a sparse perturbation vector, then at some later time $t + \tau$ the activity be $S_i(t + \tau) = \xi_i^\mu$ (where the index i is always $i = 1, \ldots, N$). This comes down to saying that the dynamics of the network retrieved the pattern ξ_i^μ from the perturbed version $\xi_i^\mu + \Delta_i$ and acted as a content-addressable memory system. In reference to the concept of basin of attraction, it means that the pattern of activity $\xi_i^\mu + \Delta_i$ was within the domain of attraction of the attractor ξ_i, and therefore evolved towards it.

From eq. A.1 we see that a way to guarantee a basin of attraction of a given width r_A is to require that

$$\xi_i^\mu = \text{sign}\left(\sum_{j=1}^{N} J_{ij}\left(\xi_j^\mu + \Delta_j\right)\right), \qquad i = 1, \ldots N,$$

for all perturbation vectors Δ smaller than r_A. This is indeed something which had

been tried at first by Gardner's group. However, this method requires a number of training iterations which is exponential in the number of neurons in the network N, because we need to consider each possible perturbation. It's a method which was therefore quickly abandoned in favor of a more convenient one.

Krauth and Mézard (1987) essentially make the following observation. Equation A.1, the condition for fix-points of the dynamics, can be rewritten in the following way:

$$0 < d \leq \left(\sum_{j=1}^{N} J_{ij} \xi_j^\mu \right) \xi_i^\mu, \qquad i = 1, \ldots N, \tag{A.2}$$

for some positive constant d, defined as a *learning margin*.

Let us consider a perturbation Δ with δ entries equal to 1 or -1, while the others are all zero. Because of eq. A.2 we have:

$$\left(\sum_{j=1}^{N} J_{ij} \left(\xi_j^\mu + \Delta_j \right) \right) \xi_i^\mu \geq d + \left(\sum_{j=1}^{N} J_{ij} \Delta_j \right) \xi_i^\mu \geq d - 2\delta \max_j |J_{ij}| \, .$$

This ensures that the memory ξ^μ will be retrieved when probing the network with the noisy pattern $\xi^\mu + \Delta$, provided that:

$$\delta \leq \frac{d}{2 \max_j |J_{ij}|} \leq \frac{d/2}{\sqrt{\frac{1}{N} \sum_j J_{ij}^2}}, \qquad i = 1, \ldots, N. \tag{A.3}$$

This equation tells us that when we train a network to maximize the basin of attraction of an attractor ξ^μ, we have to maximize the quantity $\frac{d}{\sqrt{\sum_j J_{ij}^2}}$, which is equivalent to maximizing the *stability parameters* (Abbott, 1990).

On the other hand, this tells us that, in order to obtain a basin of attraction of a given width δ, we just need to train the synaptic matrix until eq. A.3 is satisfied. In other words this offer a useful *stop-learning condition*.

It is important to point out that this analysis somehow stresses the advantage of persistent neural activity obtained thanks to circuit-based recurrence rather than to some intrinsic cell property. The latter method alone would in fact not allow any basin of attraction, because it would correspond to having a very low stability parameter, essentially determined only by the diagonal terms J_{ii}. Basically, if a cell displays sustained activity just because of its intrinsic bistability, it cannot count on the rest of the circuit to reactivate it, in case its activity is suppressed by some noisy fluctuations.

Appendix B

Scaling of attractor networks with randomly connected neurons

B.1 Constraints on the number of implementable context-dependent transitions

The conditions corresponding to the attractors and the transitions as explained in Fig. 2.2 cannot be imposed simultaneously when the same event is required to activate a neuron in one context, and to inactivate it in another. Let us formally see why.

Let us consider the example of a set of two transitions from two distinct attractors represented by patterns ξ^1, ξ^3 that are elicited by the same external field, represented by the pattern h^1:

$$\xi^1 \xrightarrow{h^1} \xi^2$$
$$\xi^3 \xrightarrow{h^1} \xi^4.$$

Let us choose the patterns ξ^μ and h to be N_I-dimensional and N_E-dimensional ± 1

vectors, respectively, whose components are independently identically distributed according to:

$$Pr(\xi_i^\mu = 1) = f_I, \quad i = 1, \ldots, N_I,$$
$$Pr(h_i = 1) = f_E, \quad i = 1, \ldots, N_E.$$

The parameter f_I and f_E give the coding level of the recurrent and external input networks.

The described attractors and transitions scheme corresponds to the following four pattern associations:

$$\left[\xi^1, h^0\right] \to \xi^1, \qquad [\xi^2, h^0] \to \xi^2 \quad \text{for the attractors,} \tag{B.1}$$
$$\left[\xi^1, h^1\right] \to \xi^3, \qquad [\xi^2, h^1] \to \xi^4 \quad \text{for the transitions,} \tag{B.2}$$

where the N_E-dimensional pattern h^0 represents the absence of a stimulus, and with $[\xi^1, h^0]$ we represent a column vector obtained by stacking ξ^1 and h^0 on top of each other (that is, $({\xi^1}^\top {h^0}^\top)^\top$ in standard linear algebra notation).

This comes back to finding two matrices J and T satisfying the following equations for $i = 1, 2, \ldots, N$:

$$\xi_i^1 = \text{sign}\left(\sum_{j=1}^{N} J_{ij}\xi_j^1 + \sum_{j=1}^{N_E} T_{ij}h_j^0\right), \qquad \xi_i^3 = \text{sign}\left(\textstyle\sum_{j=1}^{N} J_{ij}\xi_j^1 + \sum_{j=1}^{N_E} T_{ij}h_j^1\right),$$
$$\xi_i^2 = \text{sign}\left(\sum_{j=1}^{N} J_{ij}\xi_j^2 + \sum_{j=1}^{N_E} T_{ij}h_j^0\right), \qquad \xi_i^4 = \text{sign}\left(\textstyle\sum_{j=1}^{N} J_{ij}\xi_j^2 + \sum_{j=1}^{N_E} T_{ij}h_j^1\right)$$

(note that, without loss of generality, we set the threshold to zero). Defining the

following variables:

$$\alpha_i^1 = \sum_{j=1}^{N_I} J_{ij}\xi_j^1 + \sum_{j=1}^{N_E} T_{ij}h_j^0,$$
$$\alpha_i^2 = \sum_{j=1}^{N_I} J_{ij}\xi_j^2 + \sum_{j=1}^{N_E} T_{ij}h_j^0,$$
$$\alpha_i^3 = \sum_{j=1}^{N_I} J_{ij}\xi_j^1 + \sum_{j=1}^{N_E} T_{ij}h_j^1,$$

these can be written as

$$\xi_i^1 = \text{sign}\left(\alpha_i^1\right), \qquad \xi_i^3 = \text{sign}\left(\alpha_i^3\right),$$
$$\xi_i^2 = \text{sign}\left(\alpha_i^2\right), \qquad \xi_i^4 = \text{sign}\left(\alpha_i^2 + \alpha_i^3 - \alpha_i^1\right).$$

Note that this set of equations cannot be solved in general, because fixing the sign of α_i^1, α_i^2 and α_i^3 constrains the sign of $\alpha_i^2 + \alpha_i^3 - \alpha_i^1$. It's straightforward for instance to verify that the case $\xi_i^1 = \xi_i^4 = -\xi_i^2 = -\xi_i^3$ doesn't admit any solution. Note that this is equivalent to the well-known XOR problem for the perceptron in two dimensions (Minsky and Papert, 1969).

A geometrical representation of the problem

What happens in the situation we just illustrated is that we have a set of conditions to satisfy, but not enough unknown to satisfy them. The fact is that the correlations, which are introduced by the structure of the attractors and transitions, constrain the patterns on a two dimensional plane, which causes the classification to be non-linear separable (see Fig. B.1).

Unfortunately, it turns that the probability in incurring in such a non-linear separability when we have the same event triggering a transition from two differ-

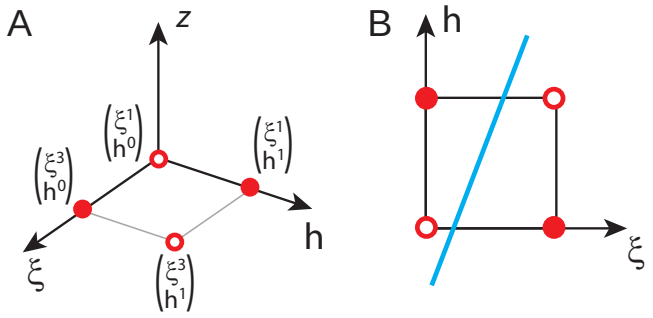

FIG. B.1: Attractors, transitions and non-linear separability. **A** The conditions to be satisfied to generate two attractors and two context-dependent transitions (eq. (B.1,B.2)) correspond to linearly separate four high-dimensional patterns of activity. However, context-dependence introduces correlations which confines them on a two-dimensional plane. **B** It is in general not possible to linearly separate four patterns which are confined on a two-dimensional plane. In fact there is no way to place the blue line on the plane so that it separates both filled red points from the empty ones.

ent mental states is considerably high. For instance, in the case of mental states represented by random and uncorrelated patterns the probability to generate a non-linear separability on one output neuron, is $1/8$. Indeed, there are two possible outputs for each of the 4 input patterns (two attractors and two transitions), for a total of $2^4 = 16$ possible input-output relations. For two of them (the XOR, and its negation) the patterns are non linearly separable. As there are N output neurons, the probability that the patterns are linearly separable for all outputs is

$$\left(1 - \frac{1}{8}\right)^N \sim e^{-N/8},$$

which goes to zero exponentially with N. If the number of contexts C in which the same event occurs is more than 2, then the exponent is proportional to NC. Notice that the probability that the problem is solvable decreases as N increases.

It therefore turns out that the case of random uncorrelated patterns, which requires a simple learning prescription for attractor neural networks (Hopfield, 1982;

Amit, 1989), becomes extremely complicated in the case of attractors and event-driven context-dependent transitions. On the other hand, correlations between patterns might reduce the performance degradation, as they could decrease the probability that the same event modifies in two different directions the activity of a particular neuron.

B.2 The importance of mixed selectivity

B.2.1 Mixed-selectivity and context-dependence

The example considered in Fig. 2.2, where one tries to implement a switch back and forth from two rules with the same *Error signal* represents a particular case of pattern in a non-general low-dimensional position (Fig. B.1). We reformulate that situation from a geometrical point of view as previously illustrated, and interpret the effect of adding mixed selective neurons within this framework. We first illustrate in Fig. B.2C how additional neurons with "pure selectivity" either to the inner mental state or to the external input cannot solve the problem. Then, in Fig. B.2D, we show that there is always a solution when we introduce a mixed selectivity neuron in the network.

Consider a neuron that is selective to the mental states, i.e. when its average response to the inputs containing $\underline{\xi}^1$, (i.e. $[\underline{\xi}^1, \underline{h}^0]$ and $[\underline{\xi}^1, \underline{h}^1]$), is different from the average response to the inputs containing $\underline{\xi}^2$. The left part of Fig. B.2C shows one example of a neuron that is selective to the mental state, but not to the external input. The input space is represented as in Fig. B.2B, and we now consider the output of an additional neuron that activates when in *Shape+Left* mental state, but not in *Color+Left*, regardless of the external input. Active outputs are indicated

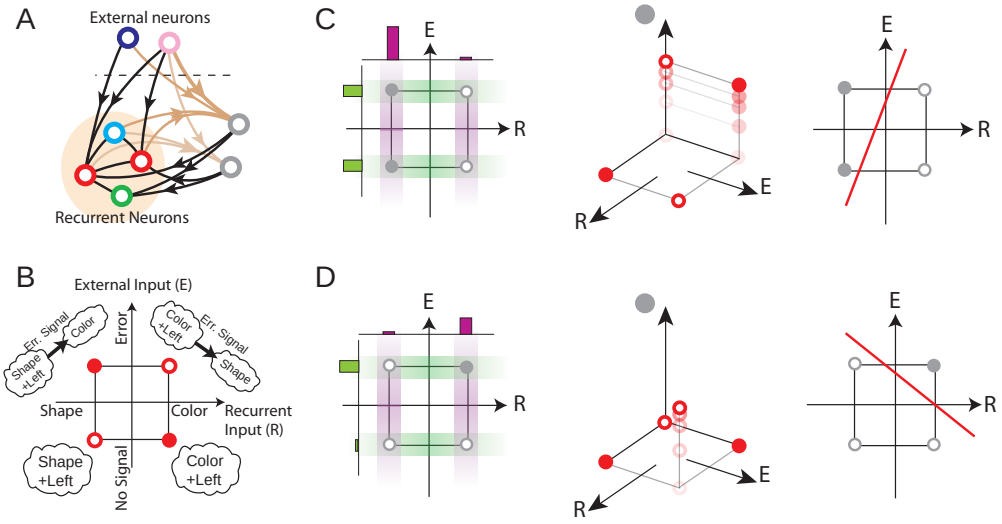

FIG. B.2: **A**, Architecture of the network, reproduced from Fig 2.2 for convenience. **B**, The context dependence problem is equivalent to the XOR (exclusive-OR problem). The N-dimensional space of all possible inputs is projected onto the plane described in the text. The circles represent the desired output activity of a specific neuron (in our case a red, $Color\ Rule$ encoding neuron) in response to the input identified by the location of the circle on the plane. The desired outputs are dictated by the requirements of the conditions corresponding to the attractors (lower quadrants) and the event-driven transitions (upper quadrants). **C**, Effects of an additional neuron with pure selectivity to the inner mental states. Left: the neuron (gray) responds to two of the four possible inputs (leftmost points) and hence it has pure selectivity. The response to the two inner mental states ($Shape{+}Left$, $Color{+}Left$) averaged over the two possible external inputs is represented by two bars above the square. The response to the external inputs averaged over the inner mental states is plotted vertically and it is represented again by two bars. The neuron responds differently and is selective to the inner mental states but not to the external inputs. Center: effects of the introduction of the pure selectivity neuron in the network dynamics. The input space goes from a plane to the third dimension, spanned by the activity of the additional neuron. Two of the circles representing the outputs of the $Color\ Rule$ neurons (see panel B) move up to reflect the activity states of the additional neuron. The axes directions are correct, but their position is arbitrary. Right: an RCN with pure selectivity responding to the same input space represented in panel B. The position and orientation of the red line is determined by the random synaptic weights. For this particular realization separates separates two inputs on the left, which are the input patterns activating the RCN. **D**, Same as in panel C but for a mixed selectivity neuron.

by filled gray circles.

When we introduce such a neuron in the network, the N dimensional input space becomes $N + 1$ dimensional. We can observe the effects on the *Color Rule* neuron of the embedding in a higher dimensionality in the middle part of Fig. B.2C. The extra dimension introduced by the additional neuron is along the z-axis, and the plane of Fig. B.2B is now spanned by the x and y axes. Two of the circles now move up to reflect the activation of the additional neuron when the network is in the *Shape+Left* mental state. Unfortunately, this new placement still does not allow us to draw a plane that separates the inputs activating the *Color Rule* neuron from those that inactivate it. This shows that "pure selectivity neurons" do not solve the non-linear separability problem. The rightmost plot will be explained in the next Section.

Consider now the mixed selectivity neuron of Fig. B.2D. Such a neuron is selective both for the mental states and the external input, as shown by the leftmost plot of Fig. B.2D. Now the embedding in a higher dimensional space can allow us to solve the problem, as only one circle moves up in the central plot of Fig. B.2D. It is easy so see that it is possible to draw a plane that separates the two empty circles from the filled ones. For similar geometrical considerations, we can conclude that the problem of non-linear separability can be solved for all additional neurons that respond to an odd number of the four possible inputs.

B.2.2 The general importance of mixed selectivity

To show the general importance of mixed selectivity we consider, for simplicity, binary neurons that can be either active or inactive. Each neuron can be regarded as a unit that computes a Boolean function $\phi(\cdot)$ of the vector of the N activities

$s_1, ..., s_N$ of the synaptically connected input neurons, which include the recurrent and the external neurons ($s = \{0, 1\}$). The problem of context-dependent tasks is related to the fact that the class of Boolean functions that can be implemented by a neuron is restricted, as it is usually assumed that the neural response is a monotonic function of the weighted sum of the activities of the synaptically connected neurons. More formally, consider a McCulloch-Pitts model neuron that is described by

$$s_i(t + \Delta t) = H\left(\sum_{j=1}^{N} J_{ij} s_j(t) - \theta\right), \tag{B.3}$$

where J_{ij} is the synaptic efficacy of the coupling between neuron j and neuron i, H is the Heaviside function ($H(x) = 0$ if $x \leq 0$, $H(x) = 1$ otherwise), and θ is the activation threshold. Different sets of synaptic efficacies correspond to different Boolean functions. How does the set of functions implementable by a McCulloch-Pitts neuron compare to more general Boolean functions, which would include also the ones that solve context-dependent problems? It is illuminating to expand a general Boolean function in a series of terms containing products of the input variablesWegener (1987):

$$\begin{aligned} s_i(t + \Delta t) = \phi(s_i(t), ..., s_N(t)) = H\left(\sum_{j=1}^{N} C_{ij} s_j(t) + \sum_{j,k=1}^{N} C_{ijk} s_j(t) s_k(t) + \right. \\ \left. + \sum_{j,k,l=1}^{N} C_{ijkl} s_j(t) s_k(t) s_l(t) + \ldots - \theta\right), \end{aligned} \tag{B.4}$$

where the Cs are the coefficients of the expansion. Such an expansion is similar to the Taylor expansion of a function of continuous variables, although in the case of Boolean functions the number of terms is finite and equal at most to 2^N. Every

term is either a single variable, or a product of two or more Boolean variables. This is equivalent to performing the logical OR operation (sum in the expression) of logical ANDs (products) between variables.

A McCulloch-Pitts neuron reads out a weighted sum of the activities $s_1, ..., s_N$, and can therefore only implement Boolean functions that depend on the first order terms of the expansion. The coefficients C_{ij} are equivalent to the synaptic weights J_{ij} of the neuronal inputs. Equation B.4 suggests that, in general, we may need to consider also higher order terms to solve complex problems.

Notice that each term taken singularly, or the sum of terms of the expansion can be considered as the output of an additional neuron that responds to a particular combination of generic events according to Equation B.3. Each C can be then regarded as the synaptic efficacy of the connection from such a neuron to the output neuron s_i. For example, the term $C_{i12}s_1s_2$ can be interpreted as the input to neuron s_i from a neuron that is active only when both s_1 and s_2 are active, with the synaptic strength C_{i12}. The neuron of Fig 2.2B, that solves the problem of context-dependence by responding to the *Error Signal* only when starting *Shape Rule*, actually implements one of these higher order terms.

B.3 Estimating the number of needed RCNs

B.3.1 Single context-dependence: a geometrical analysis

The prescription we use to create Randomly Connected Neurons (RCNs) leads to neurons with mixed selectivity. What is the probability that an RCN solves the problem generated by one particular context dependent transition? In order to solve the problem, we showed in S2 that the neuron should have mixed selectivity,

or in other words, in our paradigmatic example, the neuron has to respond to an odd number of the 4 possible input patterns $[\underline{\xi}^1, \underline{h}^0]$, $[\underline{\xi}^1, \underline{h}^1]$, $[\underline{\xi}^2, \underline{h}^0]$, $[\underline{\xi}^2, \underline{h}^1]$. What is the probability that an RCN has such a response property? The RCN is active if the weighted sum of its inputs ν_j is above some threshold θ:

$$\sum_j K_j \nu_j > \theta, \tag{B.5}$$

where the K_j's are the synaptic weights and the sum extends over both the external inputs and the neurons of the recurrent network. Choosing a specific set of synaptic weights and a threshold is therefore equivalent to drawing an hyperplane in an N-dimensional space (whose equation is $\sum_j K_j \nu_j = \theta$) that separates the input patterns activating the RCN from those that don't activate it. For some of these lines, the RCN implements the mixed selectivity neuron that we need in order to solve the context dependence problem of Section S1. On a plane, the problem amounts to determining a set of synaptic weights and a threshold so that the line $\sum_{j=1}^{2} K_j \nu_j = \theta$ has a particular orientation and displacement with respect to the origin. The rightmost part of Fig. B.2C shows how an RCN responds to the 4 possible input patterns $[\underline{\xi}^1, \underline{h}^0]$, $[\underline{\xi}^1, \underline{h}^1]$, $[\underline{\xi}^2, \underline{h}^0]$, $[\underline{\xi}^2, \underline{h}^1]$, that lie on the same plane introduced in Fig. B.2B. The RCN output is determined by the orientation of the red line that represents one realization of the random synaptic weights. The gray circles on the left of the line indicate that the neuron is activated by *Shape+Left*, no matter what is the value of the external input. Such a neuron has "pure" selectivity and it does not solve the non-linear separability problem. The second example in Fig. B.2D, shows an RCN connected by a different set of synaptic weights. The orientation and placement of the red line isolate only one vertex of the square, and

the RCN shows mixed selectivity. In this second case, the introduction of this RCN solves the non-linear separability problem. What is the probability that an RCN has this kind of mixed selectivity?

Random synaptic weights would imply random orientation and displacement with a distribution that depends on the dimensionality of the original space of input patterns (N), on the statistics of the random weights and on the threshold for neuronal activation. In our case the probability of drawing a particular line depends only on the distance from the center of the square. In particular, it grows to a maximum and then it decays to zero (see Figure B.3B). The only useful RCNs correspond to those that isolate a single vertex. Those lines that are far from the center of the square do not cut any edge joining two of the four input patterns, and they do not solve the non-linear separability. As a consequence, the best distributions are those localized around the center of the square, as in the case of Fig B.3C, i.e. for small thresholds θ. In all these situations the fraction f of all possible patterns of the input space that activate the RCN is close to $1/2$, whereas, when the threshold θ is large, f tends to zero.

We now give a more general and formal explanation for the importance of the kind of mixed selectivity we introduced in our network. We seek to prove that as the number of RCN grows, the probability to be able to implement an arbitrary scheme of attractors and transitions goes to one. We analyze two specific cases, the ultrasparse case in which f is very small and every RCN responds to only one input pattern, and the dense case in which $f = 1/2$.

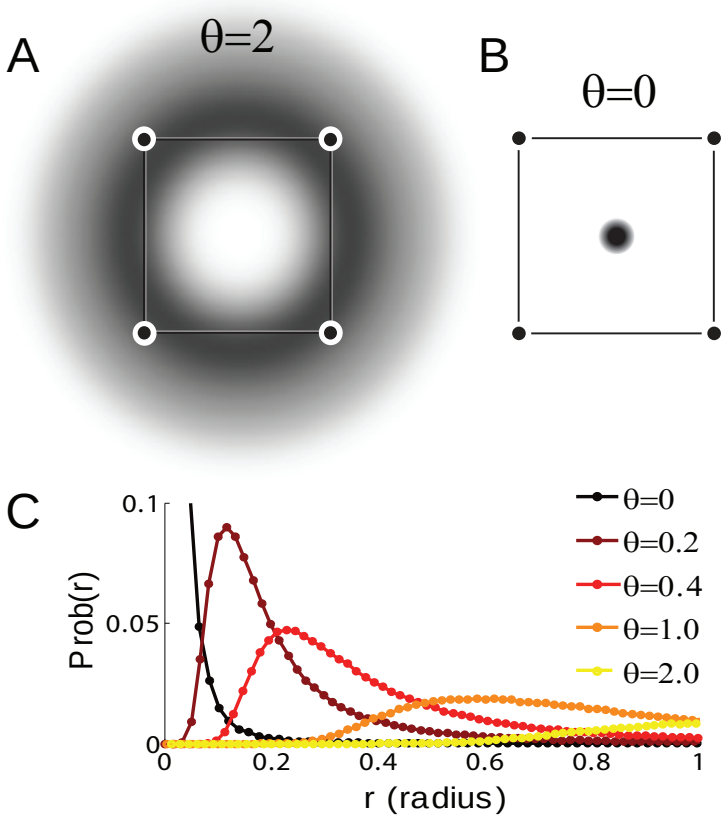

FIG. B.3: **A**, Density of circles tangent to the planes generated by randomly sampling RCNs with Gaussian synapses and $\theta = 2$. **B**, Same as panel A, but with $\theta = 0$. **C**, Distribution of the radii of tangent circles for different values of the firing threshold θ.

B.3.2 The ultrasparse case

In the case in which the RCNs are connected to the neurons of the recurrent network by random binary synapses, we can tune the neuronal threshold such that $f = 1/2^N$, i.e. every RCN is activated by a single input pattern. In such a case every additional unit generates one term of a particular Boolean expansion, known as the Disjunctive Normal Form (Wegener, 1987). Using the same notation as in S2, we can write the activity of a generic neuron s_i of the recurrent network as:

$$s_i(t+\Delta t) = \phi_i(s_1(t), ..., s_N(t)) =$$

$$H(C_{i1}s_1(t)(1-s_2(t))...(1-s_N(t)) + C_{i2}s_1(t)s_2(t)(1-s_3(t))...(1-s_N(t)) + ...),$$

where H is the Heaviside function and the Cs are the coefficients of the expansion. Every term is a product of some Boolean variables and the negation of the others (which is one minus the original variable). If these neurons are part of the recurrent network, then they can also be considered as input neurons and can contribute to the total synaptic current. If we choose the proper synaptic weights and have enough RCNs, we know that we can generate any arbitrarily complex function of the inputs $s_1, ..., s_N$. This is an extreme case in which the number of needed RCNs grows exponentially with the number N of neurons in the recurrent network. However, in such a case, not only we can satisfy all possible conditions for the attractors and the event driven transitions, but in principle we can also shape the basins of attractions arbitrarily.

B.3.3 The dense case: single context-dependence

We consider the paradigmatic case of a single context dependence as the one described in section B.1. Our aim is to compute the probability that an RCN solves the context dependence problem. We will show that such probability depends on the sparseness of the representations of the mental states, the external inputs and the corresponding patterns of activities of the RCNs. The main result of this paragraph will be that the maximum will always be in correspondence of dense representations.

First of all let us start by recalling that, in order to solve the non-linear sepa-

rability due to the context dependence problem, we need an RCN that responds to an odd number of the four possible input patterns $[\underline{\xi}^1, \underline{h}^0]$, $[\underline{\xi}^1, \underline{h}^1]$, $[\underline{\xi}^2, \underline{h}^0]$, $[\underline{\xi}^2, \underline{h}^1]$.

Let us consider one particular randomly connected neuron (RCN) and calculate the probability that it responds in the desired way. Our RCN, whose activity level we will denote by the binary variable η, gets afferences from both pools of internal and external neurons with synapses independently and identically sampled from two normal distributions $K^r \sim \mathcal{N}(0, \sigma_r^2)$ and $K^x \sim \mathcal{N}(0, \sigma_x^2)$, respectively. We assume that the statistics of these synapses is independent from that of the patterns. The activity η depends on the total synaptic input and the firing threshold θ:

$$\eta(\xi, h) = \text{sign}\left(\frac{1}{\sqrt{N_r}}\sum_{j=1}^{N_r} K_j^r \xi_j + \frac{1}{\sqrt{N_x}}\sum_{j=1}^{N_x} K_j^x h_j - \theta\right), \tag{B.6}$$

where the $\frac{1}{\sqrt{N_r}}$ and $\frac{1}{\sqrt{N_x}}$ factors have been introduced to keep the total fields of order one.

Let us now calculate the coding level of the RCN, that is the probability that the η is positive.

Coding level of the RCN network:

The terms contributing to the synaptic input to η in equation (B.6) are distributed according to a normal distribution in the following way:

$$\frac{1}{\sqrt{N_r}}\sum_{j=1}^{N_r} K_j^r \xi_j \sim \mathcal{N}(0, \sigma_r^2),$$
$$\frac{1}{\sqrt{N_x}}\sum_{j=1}^{N_x} K_j^x h_j \sim \mathcal{N}(0, \sigma_x^2).$$

The coding level of one RCN as a function of the firing threshold θ is therefore given by:

$$\begin{aligned} Pr(\eta = 1) &= Pr\left(\frac{1}{\sqrt{N_r}}\sum_{j=1}^{N_r} K_j^r \xi_j + \frac{1}{\sqrt{N_x}}\sum_{j=1}^{N_x} K_j^x h_j > \theta\right) \\ &= \frac{1}{\sqrt{2\pi(\sigma_r^2+\sigma_x^2)}}\int_\theta^\infty \exp\left(-\frac{x^2}{2(\sigma_r^2+\sigma_x^2)}\right)dx \\ &= \frac{1}{2} - \frac{1}{\sqrt{2\pi}\sigma}\int_0^\theta \exp\left(-\frac{x^2}{2\sigma^2}\right)dx \\ &= \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left(\frac{\theta}{\sqrt{2}\sigma}\right) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\theta}{\sqrt{2}\sigma}\right), \end{aligned}$$

with $\sigma^2 = \sigma_r^2 + \sigma_x^2$, and where we used the standard definition of the error function: $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt$ and $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$. The coding level of the RCN network is therefore given by:

$$f = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\theta}{\sqrt{2}\sigma}\right), \qquad \sigma^2 = \sigma_r^2 + \sigma_x^2.$$

Conversely, to obtain RCNs with a given coding level f we can set the firing threshold to be:

$$\theta(f) = \sqrt{2}\sigma\ \mathrm{erfc}^{-1}(2f),$$

where with erfc^{-1} we indicate the inverse function of erfc, i.e. the function for which $\mathrm{erfc}^{-1}(\mathrm{erfc}(x)) = x$.

Randomly Connected Neurons and linear separability:

Let us now calculate the probability p that a particular RCN η responds only for an odd number of cases, that is when all but one of the terms $\eta(\xi^1, h^0)$, $\eta(\xi^2, h^0)$, $\eta(\xi^1, h^1)$, $\eta(\xi^2, h^1)$ have the same sign. To calculate this probability let us explicitely write down the activity of η in the four conditions in the following way:

$$\eta(\xi^1, h^0) = \text{sign}\left(g_+ + g_r + g_x\right), \qquad \eta(\xi^1, h^1) = \text{sign}\left(g_+ + g_r - g_x\right),$$
$$\eta(\xi^2, h^0) = \text{sign}\left(g_+ - g_r + g_x\right), \qquad \eta(\xi^2, h^1) = \text{sign}\left(g_+ - g_r - g_x\right),$$

where we have defined three independent variables:

$$g_r = \frac{1}{\sqrt{N_r}} \sum_{j:\xi_j^1 = -\xi_j^2} K_j^r \xi_j^1, \qquad g_x = \frac{1}{\sqrt{N_x}} \sum_{j:h_j^0 = -h_j^1} K_j^x h_j^0, \tag{B.7}$$

$$g_+ = \frac{1}{\sqrt{N_r}} \sum_{j:\xi_j^1 = \xi_j^2} K_j^r \xi_j^1 + \frac{1}{\sqrt{N_x}} \sum_{j:h_j^0 = h_j^1} K_j^x h_j^0 - \theta, \tag{B.8}$$

the sum $\sum_{j:\xi_j^1 = -\xi_j^2}$ being over all the indices j for which $\xi_j^1 = -\xi_j^2$, etc.

The quantities defined in equations (B.7) and (B.8) are Gaussian variables whose variance depends on the correlations (overlaps) between the patterns ξ, h representing the mental states and the external stimuli. Let us denote with o_r the overlap between ξ^1 and ξ^2, and with o_x the overlap between h^0 and h^1:

$$o_r = \frac{1}{N_r} \sum_{j=1}^{N_r} \xi_j^1 \xi_j^2, \qquad o_x = \frac{1}{N_x} \sum_{j=1}^{N_x} h_j^0 h_j^1.$$

Using the fact that $N_r = \sum_{j=1}^{N_r} 1 = \sum_{j:\xi_j^1 = \xi_j^2} 1 + \sum_{j:\xi_j^1 = -\xi_j^2} 1$ and the analogous

identity for N_x it is simple to verify that $g_{r,x,+}$ are distributed in the following way:

$$\begin{aligned} g_r &\sim \mathcal{N}\left(0,\ (1-\sigma_{o_r}^2)\sigma_r^2\right), \qquad g_x \sim \mathcal{N}\left(0,\ (1-\sigma_{o_x}^2)\sigma_x^2\right), \\ g_+ &\sim \mathcal{N}\left(\theta, \sigma_{o_r}^2\sigma_r^2 + \sigma_{o_x}^2\sigma_x^2\right), \end{aligned} \tag{B.9}$$

where we have used the following definitions

$$\sigma_{o_r}^2 = \frac{1+o_r}{2}, \qquad \sigma_{o_x}^2 = \frac{1+o_x}{2}.$$

Note that $\sigma_{o_r}^2$, $\sigma_{o_x}^2$ are quantities between 0 and 1 quantifying how similar ξ^1 is to ξ^2 and h^0 to h^1, respectively. As a matter of fact $\sigma_{o_r}^2$ is equal to zero if ξ^1 is totally anti-correlated to ξ^2 (that is $\xi^1 = -\xi^2$), $\sigma_{o_r}^2$ is equal to one if ξ^1 is equal to ξ^2, and is equal to one half for the intermediate case of uncorrelated patterns.

We can now calculate the probability p that one of the η's has an opposite sign with respect to all the others. Taking into account the distributions of the variables given in (B.9) this probability is given by:

$$\begin{aligned} p = & \frac{8}{(2\pi)^{3/2}\sqrt{(\sigma_{o_r}^2\sigma_r^2 + \sigma_{o_x}^2\sigma_x^2)\cdot(1-\sigma_{o_r}^2)\sigma_r^2\cdot(1-\sigma_{o_x}^2)\sigma_x^2}} \\ & \cdot \int_0^\infty dg_x \int_0^{g_x} dg_r \int_{g_x-g_r}^{g_x+g_r} dg_+ \ \cosh\left(\frac{g_+\cdot\theta}{\sigma_{o_r}^2\sigma_r^2 + \sigma_{o_x}^2\sigma_x^2}\right) \\ & \cdot \exp\left(-\frac{g_+^2+\theta^2}{2\sigma_{o_r}^2\sigma_r^2 + 2\sigma_{o_x}^2\sigma_x^2}\right) \\ & \cdot \exp\left(-\frac{g_r^2}{2(1-\sigma_{o_r}^2)\sigma_r^2} - \frac{g_x^2}{2(1-\sigma_{o_x}^2)\sigma_x^2}\right) \\ & + (x \leftrightarrow r), \end{aligned} \tag{B.10}$$

where with $(x \leftrightarrow r)$ we indicate a summand equal to the previous term in eq.

(B.10) with the only difference that x and r indices have to be exchanged.

Let us consider the case in which the patterns representing the mental states and the external events have the same statistics. We therefore assume that $o_r = o_x = o$, which implies that $\sigma^2_{o_x} = \sigma^2_{o_r} = \sigma^2_o = \frac{1+o}{2}$. We also assume without loss of generality that $\sigma^2_r = \sigma^2_x = 1$. Equation (B.10) simplifies to

$$p = \frac{16}{(2\pi)^{3/2}} \int_0^\infty dg_x \int_0^{g_x} dg_r \int_{\sqrt{\frac{1-\sigma_o^2}{2\sigma_o^2}}(g_x-g_r)}^{\sqrt{\frac{1-\sigma_o^2}{2\sigma_o^2}}(g_x+g_r)} dg_+ \ \cosh\left(\frac{g_+ \cdot \theta}{\sqrt{2}\sigma_o}\right) \cdot \exp\left(-\frac{g_r^2}{2} - \frac{g_x^2}{2} - \frac{g_+^2}{2} - \frac{\theta^2}{2\sigma_o^2}\right). \tag{B.11}$$

For the special case of random uncorrelated patterns with coding level $f_0 = 1/2$ we have that $o_r = o_x = 0$, which means that $\sigma_o^2 = 1/2$. In this case eq. (B.11) further simplifies to:

$$p|_{o=0} = \frac{2}{\pi} \int_0^\infty dg_x \int_0^{g_x} dg_r \ e^{-g_x^2 - g_r^2} \cdot \sum_{i,j=1}^{2} (-1)^i \ \mathrm{erf}\left(\frac{g_x}{2} + (-1)^i \ \frac{g_r}{2} + (-1)^j \ \frac{\sqrt{2}\ \theta}{2}\right). \tag{B.12}$$

Maximizing the probability of linear separability for random patterns:

We now want to further examine the case in which mental states and external stimuli are represented by uncorrelated random patterns with coding level $f_0 = 1/2$. This is the simplest maximal entropy situation which is also the most commonly investigated in the computational literature. The probability p of linear separability for random uncorrelated $f_0 = 1/2$ patterns is given in (B.12). This expression is clearly symmetric in θ and can be shown to have a maximum at $\theta = 0$. For this case corresponding to dense coding $f = 1/2$ we therefore have a maximal probability

which can be calculated to be

$$\max_{\theta}\left(p|_{o=0}\right)=\frac{1}{3}, \tag{B.13}$$

meaning that on average one additional mixed selective unit out of three will be useful to solve the context dependence problem. This is a surprisingly high fraction, considering that the representations and the synaptic connections to the RCN are completely random.

Figure SB.4 shows the probability p of finding a mixed selective RCN as a function of the RCN's firing threshold for different values of the overlap o. As it can be seen, for positive o the maximum is always at $\theta = 0$ which corresponds to dense coding level $f = 1/2$. Moreover, increasing the overlap o decreases the probability of finding mixed selective RCNs. This can be intuitively understood considering that an increasing value of o corresponds to an increasing similarity between the patterns, and therefore an increasing difficulty to linearly separate them. Notice that the case of positive overlap o can always be led back to a case of random uncorrelated patterns with a coding level f_0 satisfying $o = (2f_0 - 1)^2$. Conversely, the case of random patterns with coding level f_0, corresponds to the case of positive $o = (2f_0 - 1)^2$.

Maximizing the probability of linear separability for anti-correlated patterns:

We now want to consider the case in which we are allowed to manually pick the patterns representing the mental states and the external stimuli. In particular, let us see what happens if we are allowed to choose anti-correlated patterns, that is

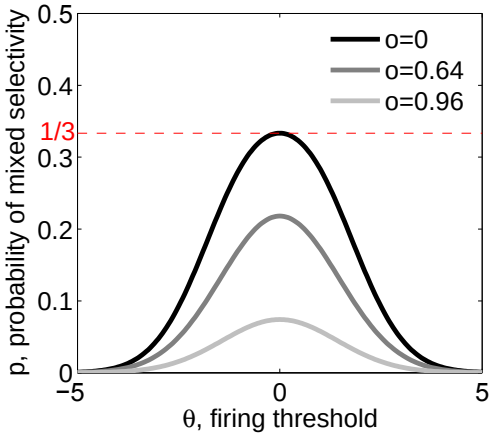

FIG. B.4: Probability of finding an RCN which implements mixed selectivity, therefore allowing to linearly separate the input patterns as a function of the RCN's firing threshold θ. This quantity is calculated in eq. (B.11). Different curves correspond to different positive values of the overlap o of the input patterns representing the mental states and the external events.

pairs of patterns which have a negative overlap o.

Following last paragraph's intuition we would expect that increasingly negative overlaps push the activity patterns further apart, therefore making them easy to linearly separate. From Figure SB.5 we see that this is exactly what happens initially for all values of θ and in particular for $\theta = 0$. When o is decreased below zero the value of p increases for all values of θ, and $\theta = 0$ always corresponds to the maximal value.

This trend crucially stops at a critical value of $o = -1/3$. Below this point, the value of p at $\theta = 0$ starts to decrease and Figure SB.5 shows that the maxima of the value of p shift laterally to $\theta \neq 0$.

It is possible to calculate analitically the critical value $o = -1/3$ of maximal p for $\theta = 0$ maximizing the expression in eq. (B.11). First of all let us compute the

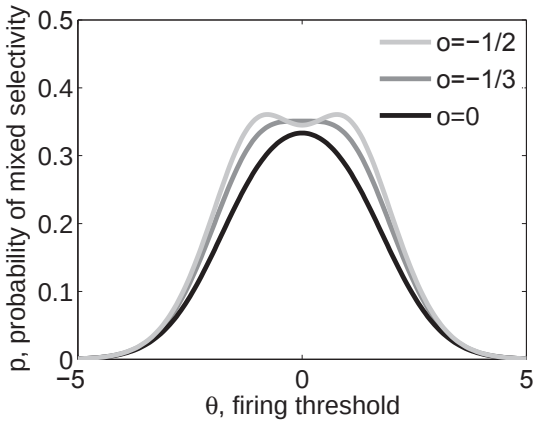

FIG. B.5: Probability of finding an RCN which implements mixed selectivity as a function of the RCN's firing threshold θ. This figure is analogous to Figure SB.4, with the difference that different curves correspond to different *negative* values of the overlap o of the input patterns representing the mental states and the external events.

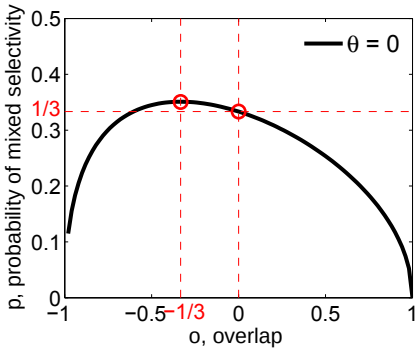

FIG. B.6: Probability p of finding an RCN implementing mixed selectivity as a function of the overlap o between the input patterns for a constant value of the RCN firing threshold θ. We see that by going to negative o we can slighlty increase p until a value $o = -1/3$. At this point $\theta = 0$ stops being a maximum of p.

value of p at $\theta = 0$ from equation (B.11):

$$\begin{aligned} p|_{\theta=0} =& \frac{16}{(2\pi)^{3/2}} \int_0^\infty dg_x \int_0^{g_x} dg_r \int_{\sqrt{\frac{1-\sigma_o^2}{2\sigma_o^2}}(g_x-g_r)}^{\sqrt{\frac{1-\sigma_o^2}{2\sigma_o^2}}(g_x+g_r)} dg_+ \exp\left(-\frac{g_r^2}{2} - \frac{g_x^2}{2} - \frac{g_+^2}{2}\right) \\ =& \frac{4}{\pi} \int_0^\infty dg_x \int_0^{g_x} dg_r \; e^{-\frac{g_x^2+g_r^2}{2}} \left(\operatorname{erf}\left(\frac{g_x+g_r}{2}\;\Sigma\right) - \operatorname{erf}\left(\frac{g_x-g_r}{2}\;\Sigma\right)\right), \end{aligned}$$

where we defined $\Sigma = \sqrt{\frac{1-\sigma_o^2}{\sigma_o^2}}$. The plot of this expression gives the graph in Figure S B.6. To find the maximum we have to calculate the extremal points in o by computing the derivative and setting it to zero. Because of the chain-rule:

$$\left.\frac{\partial p}{\partial o}\right|_{\theta=0} = \frac{\partial \sigma_o}{\partial o}\frac{\partial \Sigma}{\partial \sigma_o}\left.\frac{\partial p}{\partial \Sigma}\right|_{\theta=0}. \tag{B.14}$$

From the definitions of σ_o and Σ the first two factors in equation (B.14) simply give:

$$\frac{\partial \sigma_o}{\partial o} = \frac{1}{4\sigma_o}, \qquad \frac{\partial \Sigma}{\partial \sigma_o} = -\frac{1}{\sigma_o^3\;\Sigma}. \tag{B.15}$$

Because the derivative of erf is just a Gauss function which is easily integrated, also the third term in (B.14) results in a fairly simple expression:

$$\left.\frac{\partial p}{\partial \Sigma}\right|_{\theta=0} = \frac{4}{\pi}\frac{(2-\sqrt{2+\Sigma^2})}{(1+\Sigma^2)\sqrt{2+\Sigma^2}}. \tag{B.16}$$

Putting the last three equations together we obtain:

$$\left.\frac{\partial p}{\partial o}\right|_{\theta=0} = \frac{1}{\pi}\frac{(2-\sqrt{2+\Sigma^2})}{\sigma_o^4\Sigma(1+\Sigma^2)\sqrt{2+\Sigma^2}},$$

which is zero only for $\Sigma^2 = 2$, that is for $\sigma_o^2 = 1/3$, which in turn corresponds to

$o = -1/3$. This in fact is the maximum point which can be graphically inferred from Figure S B.6.

Let us now consider what happens for values of the overlap o which are even more negative than $o < -1/3$. This is illustrated in Figure S B.7.

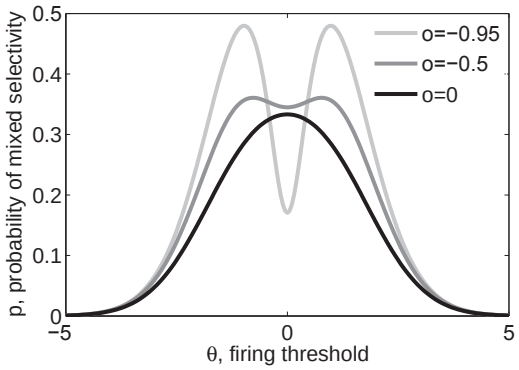

FIG. B.7: Probability of finding an RCN which implements mixed selectivity as a function of the RCN's firing threshold θ. Different curves correspond to different negative values of the overlap o of the input patterns representing the mental states and the external events.

The value of p at $\theta = 0$ goes to zero as $o \to -1$ and the maximum monotonically increases and shifts away from $\theta = 0$. We can therefore ask two questions.

First of all, what is the value of the absolute maximum which is reached at $o = -1$? Derivation and numerical integration of the expression (B.11) for this case shows that this maximum is $p_{max} = 0.5$.

The second question we can ask is, how fast does the value of p go to zero as o approaches -1? To calculate how fast p goes to zero as o goes to -1 let us notice that the quantity $\sigma_o = \sqrt{\frac{1+o}{2}}$ is a measure of how different the pattern ξ^1 is from ξ^2 and h^0 is from h^1, and is exactly equal to zero for totally anti-correlated patterns. We therefore want to Taylor-expand expression (B.11) at $\theta = 0$ around $\sigma_o = 0$, that is for the case of anti-correlated patterns.

To do this we can use equations (B.15) and (B.16) together with the fact that

$$\left.\frac{\partial p}{\partial \sigma_o}\right|_{\theta=0} = \frac{\partial \Sigma}{\partial \sigma_o}\left.\frac{\partial p}{\partial \Sigma}\right|_{\theta=0},$$

which gives:

$$\left.\frac{\partial p}{\partial \sigma_o}\right|_{\theta=0,\ \sigma_o=0} = \frac{4}{\pi}.$$

This means that in the $\theta = 0$ case for very anti-correlated patterns, that is for $\sigma_o \to 0$, the probability of finding a useful RCN goes to zero linearly in σ_o:

$$p|_{\theta=0} = \frac{4}{\pi}\sigma_o + O(\sigma_o^2) = \frac{2\sqrt{2}}{\pi}\sqrt{1+o} + O((1+o)^{3/2}).$$

We can also compute how fast p goes to zero when the input patterns are increasingly similar, that is for the case $o \to 1$, corresponding to $\sigma_o \to 1$. This gives the same type of decay:

$$\begin{aligned} p|_{\theta=0} &= \frac{-4+2\sqrt{2}}{\pi}\sqrt{1-\sigma_o} + O((1-\sigma_o)) \\ &= \frac{-2+\sqrt{2}}{\pi}\sqrt{1-o} + O((1-o)^{3/2}). \end{aligned}$$

In conclusion, we have seen that the case $\theta = 0$ corresponding to a dense RCN coding level $f = 1/2$ always gives the highest probability p to obtain a useful RCN. The only regime for which the case $\theta = 0$ is not the most favorable one is when we are allowed to choose anti-correlated patterns with an overlap below $o = -1/3$. Nonetheless, the probability at $\theta = 0$ decreases relatively slowly when we depart from the random uncorrelated case $o = 0$. Notice that the best possible value of p

which is obtained by chosing ad-hoc the input patterns is $p_{max} = 0.5$, which is a relatively small gain with respect to the value $p_{max} = 1/3$ which we get for purely random input patterns.

Summary of the main results for mental states and events represented by correlated patterns

In the Chapter 2 we illustrated in Figure 2.3 how the number of RCNs varies as a function of the probability f of activating an RCN. The results are valid in the case in which the mental states and external events are represented by random and uncorrelated patterns. In the previous sections we analyzed analytically the more general case of correlated patterns and we summarize here and in Fig. B.8 the main results. As the patterns representing mental states and events become progressively more correlated, the number of needed RCNs increases. In particular, in Fig. B.8A we show the probability of mixed selectivity as a function of f of the RCNs for different correlation levels between the patterns representing mental states and external events. The degree of correlation is expressed as the average overlap o between the two patterns representing the initial mental states (the same overlap is used for the two external events). o varies between -1 and 1, and it is positive and close to 1 for highly similar patterns (Fig. B.8A) or negative (Fig. B.8B), for anti-correlated patterns. The overlap $o = 0$ corresponds to the case of random and uncorrelated patterns. As o increases, it becomes progressively more difficult to find an RCN that can have a differential response to the two initial mental states. This is reflected by a probability that decreases approximately as $\sqrt{1-o}$. For all curves plotted in Fig. B.8A, the maximum is always realized with $f = 1/2$. Interestingly, for anti-correlated patterns, the maximum splits in two maxima that are slightly

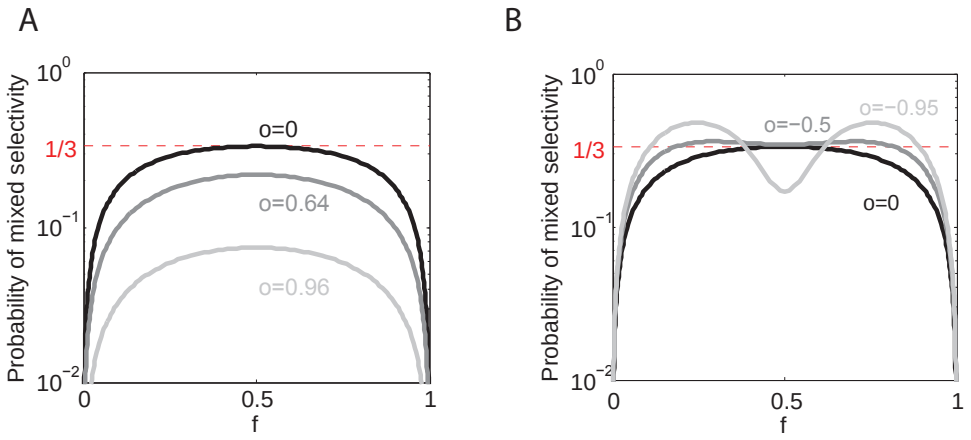

FIG. B.8: **A**, Probability that an RCN has mixed selectivity as a function of f, as in Fig 2.3B, for different positive values of the overlap o between the two initial mental states, and the two external inputs corresponding to the spontaneous activity and the event. Again the peak is always at $f = 1/2$ and the curve decays gently as o goes to 1. **B**, As in A, but for negative values of the overlap o. There are now two peaks that remain close to $f = 1/2$ for all values of o.

above 1/3 (see Fig. B.8B). The maxima initially move away from $f = 1/2$ as the patterns become more anti-correlated, but then, for $o < -5/6$, they stop diverging from the mid point. The optimal value for f remains within the interval $[0.3, 0.7]$ for the whole range of correlations.

B.3.4 The dense case: multiple context-dependencies

Analytical analysis

What is the total number of RCNs needed to satisfy all conditions corresponding to a large number of transitions and stationary patterns of neural activity? Were all context-dependent transitions independent, such a number would be proportional to the logarithm of the number of conditions. This is certainly true for a small number of context-dependent transitions. Unfortunately, the conditions to be imposed for a

large number of context-dependent transitions are not independent, and an analytic calculation turned out to be rather complicated.

The formal generalization of the calculations carried out in the previous section for the effect of an RCN on a single context-dependence can in fact be reformulated as a computation of the rank of a random matrix with correlated entries. Notice in fact that the non-linear separability of the four patterns in Fig. B.1 was due to their relative position in space, which was not general. In practice these four points were confined to a two-dimensional space. Since the set of linear separators in a 2 dimensional space can only separate up to three points in general (a number which equals to the VC-dimension of the set of linear separators (see Scholkopf and Smola (2002)). The role of the additional RCN was therefore to effectively increase the dimensionality of the patterns of activity, that is the rank of the matrix composed by stacking all the patterns together.

It is not clear is this rank can be calculated analytically for more patterns and RCNs considered simultaneously. We therefore decided to carry out this analysis differently in a numerical fashion.

Numerical analysis

We devised a benchmark to characterize numerically the scaling properties of the capacity of the system endowed with RCNs. We considered the case of the simulations where transitions between randomly selected attractors were all driven by a single event. Half of the m mental states were chosen as initial states, i.e. the contexts in which the event can occur. For each initial state we chose randomly a target attractor. The representations of the attractors were random uncorrelated patterns. Fig. B.9A shows the required number of RCNs as a function of the num-

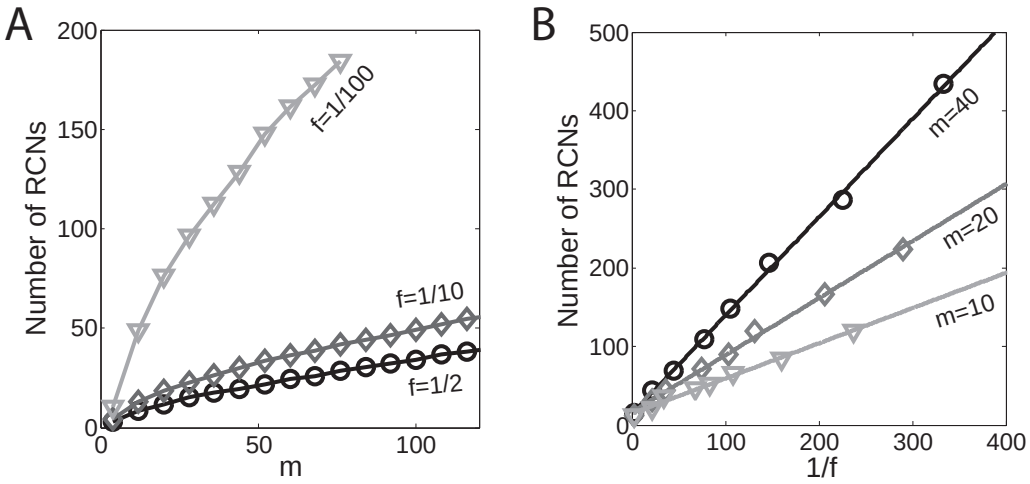

FIG. B.9: **A**, Number of RCNs needed to implement $m/2$ transitions between m random mental states. The number of neurons in the recurrent network is always $N = 200$. Different curves correspond to different choices of the threshold for activating the RCNs, which, in turn, correspond to a different f (average fraction of inputs that activate the RCNs). **B**, Number of needed RCNs as a function of $1/f$ for a different m. $N = 200$ as in panel A.

ber of transitions that are needed in a task. The average number of necessary RCNs scales logarithmically with the number of contexts m for small m values, and then linearly. Moreover, the minimal number of RCNs is achieved for $f = 1/2$, consistently with the full simulations of Fig. 2.6A. The required number of RCNs increases with decreasing f, approximately like $1/f$ when $f \leq 1/2$ (see Fig. B.9B), and like $1/(1-f)$ for $f > 1/2$ (not shown). Notice that in Fig. B.9A,B we plotted the number of needed RCNs for satisfying the mathematical conditions that guarantee the stationarity of the patterns of activities of the mental states and the implementation of the event-driven transitions. When we additionally require that the stationary points are stable and the basin of attraction has a given size, as in Fig. 2.6A,B the situation is significantly worse in the case of $f \neq 1/2$, but the scaling with the number of mental states remains linear (see also next section).

B.4 Scaling properties of the basins of attraction

The size of the basins of attraction of the attractors corresponding to the mental states increases with the number of RCNs. Large basins are important both for the robustness of the system to noise, and for generalizing to novel situations not foreseen when the scheme of states and transitions was decided. For example a novel sensory stimulus, similar enough to a familiar one, might induce the same transitions from one mental state to another. The basin of attraction for a fixed point is estimated in Fig. B.10A. Starting from the fixed point, we perturbed the neurons of the recurrent network, and measured the fraction of perturbed patterns that relaxed back into the correct attractor. The RCNs are also perturbed, as they are connected to the recurrent neurons, but along the abscissa we show only the fraction of neurons of the recurrent network whose activity has been modified. The fraction of correct relaxations stays at 1 when the initial patterns are close to the attractor and then it decreases with the fraction of perturbed neurons. As long as the fraction of correct relaxations is near 1, most of the patterns are within the basin of attraction. The different curves correspond to a different number of RCNs (at fixed number of recurrent neurons) and it is clear that the introduction of RCNs expands the basin of attraction. In Fig. B.10B we plotted the fraction of correct transitions when the external input is perturbed. This fraction also increases with the number of RCNs but it is systematically smaller than in the case of the attractors. This is expected as the external input is perturbed during the entire duration of the transition.

In Fig. B.10C,D we show how the number of required neurons (recurrent neurons+RCNs) scales with the number of mental states for the benchmark of Fig.

2.6B. The two figures differ in the required sizes for the basins of attraction. For Fig. B.10C the basin of attraction had to be large enough to guarantee that initial patterns with a perturbation as high as 3% would all relax back in the attractor. Fig. B.10D is a reproduction of Fig. 2.6B. Here the requirement about the basin of attraction was that initial patterns with a 10% perturbation would all relax back in the attractor. In Fig. B.10C we plotted the minimal number of neurons as a function of the number of mental states for basins of attraction of 3%. The number of required neurons is significantly smaller for smaller basins of attraction when the number of contexts in which an event can appear is small (curves $r = m$ and $r = 2m$). However, as the number of contexts increases, the difference in the required number of neurons decreases.

Notice that the RCNs implement something similar to what is known as a 'random projection'. These projections are obtained by multiplying by a random normalized matrix the vectors representing some patterns in the original space, which in our case are the patterns of activity of the recurrent network and the external neurons. A random projection is known to approximately preserve the similarities between the neural patterns of activity of the recurrent network (Johnson and Lindenstrauss, 1984). This is an important property because it allows the RCNs to enlarge the basins of attraction without altering the relative distances between the patterns of activity of the recurrent network. In particular, initial patterns that are similar enough to let the recurrent network relax into the same attractor will most likely have the same behavior in the complete network that includes the RCNs. The only difference in the complete network is that all distances between patterns of neural activities are stretched on average by the same factor, making learning and separability of the patterns significantly easier.

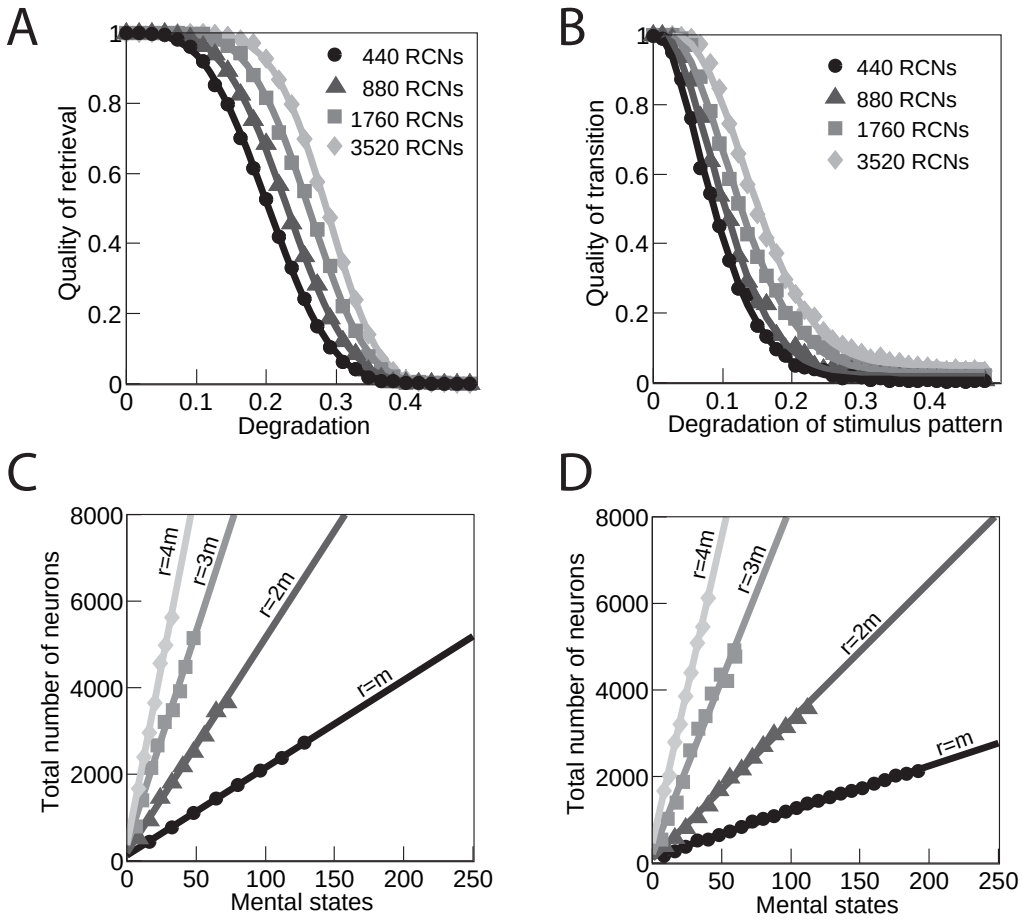

FIG. B.10: Scaling properties of the simulated attractor neural networks. **A**, The size of basins of attraction increases with the number of RCNs. The quality of retrieval (fraction of cases in which the network dynamics flow in the correct attractor) is plotted against the distance between the the initial pattern of activity and the attractor. The four curves correspond to four different number of RCNs. **B**, Fraction of successful transitions in the case in which the representation of the triggering external input is perturbed as a function of the increasing perturbation. **C**, Total number of needed neurons (recurrent network neurons+RCNs) to implement m random mental states and r transitions randomly chosen between the mental states. Different curves correspond to different ratios between r and m. Initial patterns with a perturbation of at least 3% are all required to relax back in the attractor. **D**, Same as C, but for larger basins of attraction, as in Fig. 2.6B (all initial patterns within 10% distance are required to relax back to the attractor).

Bibliography

Abbott, L. F. (1990). Learning in neural network memories. *Network: Comput. Neural Syst*, 1(1):105–122.

Abbott, L. F. (2008). Theoretical neuroscience rising. *Neuron*, 60(3):489–495.

Abbott, L. F. & Kepler, T. B. (1989). Universality in the space of interactions for network models. *J. Phys. A: Math. Gen*, 22(12):2031–2038.

Abeles, M., Bergman, H., Gat, I., Meilijson, I., Seidemann, E., Tishby, N., & Vaadia, E. (1995). Cortical activity flips among quasi-stationary states. *Proc Natl Acad Sci U S A*, 92(19):8616–8620.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, 9:147–169.

Albus, J. (1971). A theory of cerebellar function. *Mathematical Biosciences*, 10:25–61.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol Cybern*, 27(2):77–87.

Amit, D. & Brunel, N. (1995). Learning internal representations in an attractor neural network with analogue neurons. *Network: Computation in Neural Systems*, 6(3):359–388.

Amit, D. & Fusi, S. (1994). Learning in neural networks with material synapses. *Neural Computation*, 6(5):957–982.

Amit, D., Gutfreund, H., & Sompolinsky, H. (1985). Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018.

Amit, D. J. (1988). Neural networks counting chimes. *Proc Natl Acad Sci U S A*, 85(7):2141–2145.

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

Amit, D. J. & Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb Cortex*, 7(3):237–252.

Amit, D. J., Fusi, S., & Yakovlev, V. (1997). Paradigmatic working memory (attractor) cell in it cortex. *Neural Comput*, 9(5):1071–1092.

Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, 21(6):1399–1407.

Aston-Jones, G. & Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *J Comp Neurol*, 493(1):99–110.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nat Neurosci*, 7(4):404–410.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Belova, M. A., Paton, J. J., Morrison, S. E., & Salzman, C. D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron*, 55(6):970–984.

Belova, M. A., Paton, J. J., & Salzman, C. D. (2008). Moment-to-moment tracking of state value in the amygdala. *J Neurosci*, 28(40):10023–10030.

Bernasconi, J. (1988). Analysis and comparison of different learning algorithms for pattern association problems. *Neural information processing systems: Denver, CO, 1987*, page 72.

Bichot, N. P., Schall, J. D., & Thompson, K. G. (1996). Visual feature selectivity in frontal eye fields induced by experience in mature macaques. *Nature*, 381(6584):697–699.

Block, H. (1962). The perceptron: a model for brain functioning. *Reviews of Modern Physics*, 34:123–135. Reprinted in: Anderson and Rosenfeld (eds.), *Neurocomputing: Foundations of Research*.

Boettiger, C. A. & D'Esposito, M. (2005). Frontal networks for learning and executing arbitrary stimulus-response associations. *J Neurosci*, 25(10):2723–2732.

Brader, J. M., Senn, W., & Fusi, S. (2007). Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural Comput*, 19(11):2881–2912.

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

Brunel, N. & Hakim, V. (1999a). Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Comput*, 11(7):1621–1671.

Brunel, N. & Hakim, V. (1999b). Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Computation*, 11:1621–1671.

Brunel, N. & Wang, X. J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J Comput Neurosci*, 11(1):63–85.

Buckley, M. J., Mansouri, F. A., Hoda, H., Mahboubi, M., Browning, P. G. F., Kwok, S. C., Phillips, A., & Tanaka, K. (2009). Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science*, 325(5936):52–58.

Buhmann, J. & Schulten, K. (1987a). Influence of noise on the function of a "physiological" neural network. *Biological Cybernetics*, 56(5-6):313–327.

Buhmann, J. & Schulten, K. (1987b). Noise-Driven temporal association in neural networks. *EPL (Europhysics Letters)*, 4(10):1205–1209.

Candes, E. & Tao, T. (2004). Near-optimal signal recovery from random projections: Near optimal signal recovery from random projections:. *IEEE Trans. Inform. Theory*, 52:5406–5425.

Candes, E. & Tao, T. (2005). Decoding by Linear Programming. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 51(12):4203.

Churchland, M. M., Yu, B. M., Ryu, S. I., Santhanam, G., & Shenoy, K. V. (2006). Neural variability in premotor cortex provides a signature of motor preparation. *J Neurosci*, 26(14):3697–3712.

Cohen, J. D. & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychol Rev*, 99(1):45–77.

Compte, A., Brunel, N., Goldman-Rakic, P., & Wang, X. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314.

Dasgupta, S. & Gupta, A. (2002). An elementary proof of the johnson-lindenstrauss lemma. *Random Structures and Algorithms*, 22:60–65.

Daugman, J. (2001). *Brain metaphor and brain theory*, page 23. Blackwell Pub.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12):1704–1711.

Dayan, P. (2007). Bilinearity, rules, and prefrontal cortex. *Front Comput Neurosci*, 1:1.

Dayan, P. (2008). Simple substrates for complex cognition. *Front Neurosci*, 2(2):255–263.

Deco, G. & Rolls, E. T. (2005). Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. *Cereb Cortex*, 15(1):15–30.

Disney, A. A. & Aoki, C. (2008). Muscarinic acetylcholine receptors in macaque v1 are most frequently expressed by parvalbumin-immunoreactive neurons. *J Comp Neurol*, 507(5):1748–1762.

Disney, A. A., Aoki, C., & Hawken, M. J. (2007). Gain modulation by nicotine in macaque v1. *Neuron*, 56(4):701–713.

Doya, K. (1992). Bifurcations in the learning of recurrent neural networks. In: *1992 IEEE International Symposium on Circuits and Systems, 1992. ISCAS'92. Proceedings.*, volume 6.

Forrest, B. M. (1988). Content-addressability and learning in neural networks. *J. Phys. A: Math. Gen.*, 21:245–255.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cogn Affect Behav Neurosci*, 1(2):137–160.

Freund, Y. & Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, 37:277–296.

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

Fusi, S., Asaad, W. F., Miller, E. K., & Wang, X.-J. (2007). A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. *Neuron*, 54(2):319–333.

Fusi, S., Drew, P. J., & Abbott, L. F. (2005). Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611.

Fusi, S. & Mattia, M. (1999). Collective behavior of networks with linear (vlsi) integrate-and-fire neurons. *Neural Comput*, 11(3):633–652.

Fuster, J. M. (2001). The prefrontal cortex–an update: time is of the essence. *Neuron*, 30(2):319–333.

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

Ganguli, S., Huh, D., & Sompolinsky, H. (2008). Memory traces in dynamical systems. *Proc Natl Acad Sci U S A*, 105(48):18970–18975.

Gardner, E. (1987). Maximum storage capacity in neural networks. *Europhys. Lett.*, 4:481–485.

Gardner, E. (1988). The space of interactions in neural network models. *J. Phys. A*, 21:257–270.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.

Genovesio, A., Brasted, P. J., Mitz, A. R., & Wise, S. P. (2005). Prefrontal cortex activity related to abstract response strategies. *Neuron*, 47(2):307–320.

Gluck, M. A., Oliver, L. M., & Myers, C. E. (1996). Late-training amnesic deficits in probabilistic category learning: a neurocomputational analysis. *Learn Mem*, 3(4):326–340.

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: individual variability in strategies for probabilistic category learning. *Learn Mem*, 9(6):408–418.

Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron*, 61(4):621–634.

Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, 49(5):757–763.

Griniasty, M., Tsodyks, M. V., & Amit, D. J. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation*, 5:1–17.

Hasegawa, R., Sawaguchi, T., & Kubota, K. (1998). Monkey prefrontal neuronal activity coding the forthcoming saccade in an oculomotor delayed matching-to-sample task. *Journal of neurophysiology*, 79(1):322.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2005). Banishing the homunculus: making working memory work. *Neuroscience*.

Hempel, C. M., Hartman, K. H., Wang, X. J., Turrigiano, G. G., & Nelson, S. B. (2000). Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J Neurophysiol*, 83(5):3031–3041.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*.

Hinton, G. (1981). Implementing semantic networks in parallel hardware. *In Hinton, G. E. and Anderson, J. A., editors, Parallel Models of Associative Memory, Erlbaum, Hillsdale, NJ.*

Hinton, G. (1989). Connectionist learning procedures. *Artificial intelligence*, 40(1-3):185–234.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends Cogn Sci*, 11(10):428–434.

Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A*, 79:2554–2558.

Jaeger, H. & Haas, H. (2004). Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80.

Johnson, W. & Lindenstrauss, J. (1984). Extensions of lipshitz maps into a hilber space. *Contemporary Math.*, 26:189–206.

Jones, L. M., Fontanini, A., Sadacca, B. F., Miller, P., & Katz, D. B. (2007). Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proc Natl Acad Sci U S A*, 104(47):18772–18777.

Jun, J. K., Miller, P., Hernndez, A., Zainos, A., Lemus, L., Brody, C. D., & Romo, R. (2010). Heterogenous population coding of a short-term memory and decision task. *J Neurosci*, 30(3):916–929.

Kaelbling, L., Littman, M., & Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence Journal*, 101:99–134.

Kiani, R. & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324(5928):759–764.

Kleene, S. (1956). *Automata Studies*, chapter Representation of events in nerve nets and finite automata., pages 3–42. Princeton University Press, Princeton, N.J.

Kobatake, E., Wang, G., & Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J Neurophysiol*, 80(1):324–330.

Korzen, M. & Klesk, P. (2008). *Artificial Intelligence and Soft Computing ICAISC 2008*, volume 5097/2008 of *Lecture Notes in Computer Science*, chapter Maximal Margin Estimation with Perceptron-Like Algorithm, pages 597–608. Springer Berlin / Heidelberg.

Krauth, W. & Mézard, M. (1987). Learning algorithms with optimal stability in neural networks. *Journal of Physics A: Mathematical and General*, 20(11):L745–L752.

Krauth, W., Nadal, J.-P., & Mzard, M. (1988). The roles of stability and symmetry in the dynamics of neural networks. *J. Phys. A: Math. Gen.*, 21:2995–3011.

Li, N. & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321(5895):1502–1507.

Little, W. & Shaw, G. (1978). Analytic study of the memory storage capacity of a neural network. *Math. Biosci*, 39(3-4):281–290.

Loh, M. & Deco, G. (2005). Cognitive flexibility and decision-making in a model of conditional visuomotor associations. *Eur. J. Neurosci.*, 22:2927–2936.

Maass, W., Joshi, P., & Sontag, E. D. (2007). Computational aspects of feedback in neural circuits. *PLoS Comput Biol*, 3(1):e165.

Maass, W., Natschlger, T., & Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput*, 14(11):2531–2560.

Machens, C. K., Romo, R., & Brody, C. D. (2005). Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science*, 307(5712):1121–1124.

Mansouri, F. A., Buckley, M. J., & Tanaka, K. (2007). Mnemonic function of the dorsolateral prefrontal cortex in conflict-induced behavioral adjustment. *Science*, 318(5852):987–990.

Mansouri, F. A., Matsumoto, K., & Tanaka, K. (2006). Prefrontal cell activities related to monkeys success and failure in adapting to rule changes in a wisconsin card sorting test analog. *Journal of Neuroscience*, 26:2745–2756.

Marder, E. & Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nat Rev Neurosci*, 7(7):563–574.

Marr, D. (1969). A theory for cerebellar cortex. *J. Physiol.*, 202:437–470.

Mattia, M., Rigotti, M., & Fusi, S. (2007). Event driven transitions between attractors in spiking networks. In: *Cosyne: Computational and Systems Neuroscience Conference*.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev*, 102(3):419–457.

McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5(4):115–133.

Miller, E. & Buschman, T. (2008). *Neuroscience of rule-guided behavior*, chapter Rules through Recursion: How Interactions between the Frontal Cortex and Basal Ganglia May Build Abstract, Complex Rules from Concrete, Simple Ones, page 419. Oxford University Press, USA.

Miller, E. K. & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci*, 24:167–202.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J Neurosci*, 16(16):5154–5167.

Miller, P. & Wang, X.-J. (2006). Inhibitory control by an integral feedback signal in prefrontal cortex: a model of discrimination between sequential stimuli. *Proc Natl Acad Sci U S A*, 103(1):201–206.

Milner, B. (1963). Effect of different brain lesions on card sorting. *Archives of Neurology*, 9:90–100.

Minsky, M. & Papert, S. (1969). *Perceptrons*. Cambridge: MIT Press.

Miyashita, Y. & Chang, H. S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, 331:68–70.

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, 319(5869):1543–1546.

Mongillo, G., Curti, E., Romani, S., & Amit, D. J. (2005). Learning in realistic networks of spiking neurons and spike-driven plastic synapses. *Eur J Neurosci*, 21(11):3143–3160.

Morrison, S. & Salzman, C. (2009). The convergence of information about rewarding and aversive stimuli in single neurons. *J Neurosci*, 29(37):11471–11483.

Murray, E. A., Bussey, T. J., & Wise, S. P. (2000). Role of prefrontal cortex in a network for arbitrary visuomotor mapping. *Exp Brain Res*, 133(1):114–129.

Nieder, A. & Miller, E. K. (2003). Coding of cognitive magnitude: compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, 37(1):149–157.

Opris, I., Barborica, A., & Ferrera, V. P. (2005). Microstimulation of the dorsolateral prefrontal cortex biases saccade target selection. *J Cogn Neurosci*, 17(6):893–904.

O'Reilly, R. & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience*. MIT Press.

O'Reilly, R. C. & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18:283–328.

Padoa-Schioppa, C. & Assad, J. A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat Neurosci*, 11(1):95–102.

Passingham, R. (1993). *The Frontal Lobes and Voluntary Action*. Oxford University Press, Oxford.

Pasupathy, A. & Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 433(7028):873–876.

Paton, J. J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, 439(7078):865–870.

Petrides, M. (1982). Motor conditional associative-learning after selective prefrontal lesions in the monkey. *Behav Brain Res*, 5(4):407–413.

Petrides, M. (1985). Deficits on conditional associative-learning tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, 23(5):601–614.

Poggio, T. & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266.

Poldrack, R. A., Clark, J., Par-Blagoev, E. J., Shohamy, D., Moyano, J. C., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, 414(6863):546–550.

Potjans, W., Morrison, A., & Diesmann, M. (2009). A spiking neural network model of an actor-critic learning agent. *Neural Comput*, 21(2):301–339.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rigotti, M., Ben Dayan Rubin, D., Wang, X.-J., & Fusi, S. (2008). Mixed neuronal selectivity is important in recurrent neural networks implementing context dependent tasks. In: *Society for Neuroscience Annual Meeting (Washington, DC, Society for Neuroscience)*, page pp. 929.3/TT11.

Rolls, E. T. & Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput*, 12(11):2547–2572.

Romo, R., Brody, C. D., Hernndez, A., & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399(6735):470–473.

Rosenblatt, F. (1958). *The perceptron: A probabilistic model for information storage and organization in the brain*, volume 65.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan Books.

Rosenthal, O., Fusi, S., & Hochstein, S. (2001). Forming classes by stimulus frequency: behavior and theory. *Proc Natl Acad Sci U S A*, 98(7):4265–4270.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *PNAS*, 102(20):7338–7343.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Sakai, K., Rowe, J. B., & Passingham, R. E. (2002). Active maintenance in prefrontal area 46 creates distractor-resistant memory. *Nat Neurosci*, 5(5):479–484.

Salzman, C. D. & Fusi, S. (2009). Emotion, cognition and mental state representation in amygdala and prefrontal cortex. *Submitted to Annual Review of Neuroscience*.

Salzman, C. D., Paton, J. J., Belova, M. A., & Morrison, S. E. (2007). Flexible neural representations of value in the primate brain. *Ann N Y Acad Sci*, 1121:336–354.

Scholkopf, B. & Smola, A. (2002). Learning with Kernels: Support Vector Machines, Regularization. *Optimization, and Beyond. MIT Press*, 1:2.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Senn, W. & Fusi, S. (2005). Learning only when necessary: better memories of correlated patterns in networks with bounded synapses. *Neural Comput*, 17(10):2106–2138.

Shohamy, D., Myers, C. E., Kalanithi, J., & Gluck, M. A. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neurosci Biobehav Rev*, 32(2):219–236.

Soltani, A. & Wang, X.-J. (2006). A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *J Neurosci*, 26(14):3731–3744.

Soltesz, I. (2005). *Diversity in the Neuronal Machine*. New York: Oxford University Press.

Sompolinsky, H., Crisanti, A., & Sommers, H. (1988). Chaos in random neural networks. *Physical Review Letters*, 61(3):259–262.

Sompolinsky, H. & Kanter, I. (1986). Temporal association in asymmetric neural networks. *Phys. Rev. Lett.*, 57:2861–2864.

Sugase-Miyamoto, Y., Liu, Z., Wiener, M. C., Optican, L. M., & Richmond, B. J. (2008). Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. *PLoS Comput Biol*, 4(5):e1000073.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nat Rev Neurosci*, 6(5):363–375.

Sussillo, D. & Abbott, L. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557.

Sutton, R. (2001). What's wrong with artificial intelligence. http://webdocs.cs.ualberta.ca/ sutton/IncIdeas/WrongWithAI.html.

Sutton, R. & Barto, A. (1998). *Introduction to reinforcement learning*. MIT Press Cambridge, MA, USA.

Tanji, J. & Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiol Rev*, 88(1):37–57.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1142.

Vogels, T. P. & Abbott, L. F. (2009). Gating multiple signals through detailed balance of excitation and inhibition in spiking networks. *Nat Neurosci*, 12(4):483–491.

von Neumann, J. (1956). *Automata Studies*, chapter Probabilistic logics and the synthesis of reliable organisms from unreliable components., pages 3–42. Princeton University Press, Princeton, N.J.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411(6840):953–956.

Wang, X. J. (1999). Synaptic basis of cortical persistent activity: the importance of nmda receptors to working memory. *J Neurosci*, 19(21):9587–9603.

Wang, X. J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci*, 24(8):455–463.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5):955–968.

Watanabe, M. (1986). Prefrontal unit activity during delayed conditional Go/No-Go discrimination in the monkey. II. Relation to Go and No-Go responses. *Brain Res*, 382(1):15–27.

Watkins, C. (1989). *Learning from delayed rewards*. PhD thesis, Cambridge University, Cambridge, England.

Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292.

Wegener, I. (1987). *The Complexity of Boolean Functions*. John Wiley and Sons Ltd, and B. G. Teubner, Stuttgart. ISBN: 3-519-02107-2.

Winograd, S. & Cowan, J. (1963). *Reliable computation in the presence of noise*. MIT Press Cambridge, Mass.

Yakovlev, V., Fusi, S., Berman, E., & Zohary, E. (1998). Inter-trial neuronal activity in inferior temporal cortex: a putative vehicle to generate long-term visual associations. *Nat Neurosci*, 1(4):310–317.

Yu, A. J. & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692.

Zipser, D. & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331(6158):679–684.