Lehrstuhl für Informatik VII der Technischen Universität München

Neurodynamical competition and cooperation mechanisms in brain functions:

Attentional filtering and perceptual learning

Miruna Szabo

Vollständiger Abdruck der von der Fakultät für Informatik
der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Univ.-Prof. Dr. A. Knoll

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. h.c.mult. W. Brauer, em.
2. Univ.-Prof. Dr. H. J. Schmidhuber
3. Prof. Dr. G. Deco, Univ. Pompeu Fabra,
   Barcelona /Spanien

Die Dissertation wurde am 21.09.2007 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 25.04.2008 angenommen.

In memory of my beloved mother, Doina.

Abstract

Cognitive functions of human behavior like selective attention, working memory and decision making require complex information processing techniques in order to select and represent the behavioral relevant information and to create the correct associations. Present-day research addresses these issues by developing neurodynamical computational models inspired from the structure and properties of the nervous system.

Driven by the assumption that the representation of information in the brain is distributed across several cortical areas and that the coherent description of information is achieved at a global level through an intricate inter-areal connectivity, an important approach to the understanding of the neural mechanisms underlying the cognitive functions of the brain assumes that its distinct features might emerge from the mutual interplay of different interconnected brain structures. This implies the necessity of analyzing the biological neural networks at system level.

This work contributes to the understanding of the fundamental brain principles underlying cognitive functions by introducing and systematically analyzing two models that provide fundamental building blocks for the construction of large neurodynamical multi-areal models.

A module inside a cortical area is modeled as a biologically-inspired recurrent network of excitatory and inhibitory spiking neurons described by nonlinear dynamics and characterized by stochastic activity, for which stimuli-related inputs are processed in the context of neuronal reverberation. The modules are structured according to the concept of population coding and to the general framework of Biased-Competition and Cooperation.

Two concrete examples that study specific cognitive phenomena as effects of multi-areal recurrent processing determine a set of underlying working hypotheses and relate these hypotheses to experimental evidence. The first model analyzes the computational principles underlying the neuronal correlates of attentional filtering, showing how an input encoding the attentional state can bias the level of competition and cooperation in a single-area model in order to extract the relevant information for the behavioral task. The second model analyzes the computational principles underlying the neuronal correlates of perceptual learning, showing how learning affects the connectivity between two model areas and how the resulting intrinsic attentional signal affects the level of competition and cooperation in the network in order to express the relevant information for the behavioral task.

Along with correctly describing the experimental findings on awake behaving animals, both models help to extract a number of possible underlying principles, like selective filtering, correlation facilitation and selective tuning. These dynamical features represent fundamental building blocks for large-scale multi-areal neurodynamical networks modeling cognitive brain functions.

Acknowledgments

This research was carried out at the Neural Computation department of Siemens – Corporate Technology where I have met many interesting and inspiring people.

First of all, I am very grateful to my Ph.D. advisor Prof. Wilfried Brauer for his constant support, valuable advice and encouragement throughout my doctoral studies.

Special thanks are addressed to Prof. Gustavo Deco and Dr. Martin Stetter who constantly guided and supported me in all the steps of my research. They helped me to widely broaden my knowledge horizon and provided me with helpful comments and suggestions to improve this work.

Many thanks also to all my colleagues at the Siemens Neural Computation department for the friendly and creative working environment. I especially would like to thank to Dr. Dragan Obradovic who guided my first steps as a Ph.D. student.

Some of the work presented here was done in collaboration with Rita Almeida, Paolo del Giudice, Stefano Fusi and Maurizio Mattia to whom I thank for the fruitful collaboration and interesting discussions.

And last, but not least, I thank my family for their love and support over all these years.

Contents

1 Introduction

Conventional computers use an algorithmic approach to problem solving. They process information in a centralized and predictable way, which makes them unable to solve real world problems that involve adaptation to a changing environment and generalization to new situations. New technological needs like advanced image processing, adaptive pattern recognition, learning autonomous systems, prediction and intelligent control require the development of *intelligent information processing systems*, for which the use of the *biological nervous systems* as inspiration source holds a lot of promise. In tasks like visual recognition, language understanding or motor control, the human brain outperforms today's supercomputers in terms of processing and storing capabilities, reaction times and robustness to novelty. The human brain can be considered as a huge network of interacting neural processing elements (about 10^{11} neurons and 10^{14} connecting synapses) structured in many layered subnetworks with different architectures, properties and functions. The recurrent dynamical nature with nonlinear, parallel and distributed processing capabilities plus the continuous self-adaptivity make our brain a more powerful computational tool than conventional computers are. Therefore studying the way it processes and encodes information represents an important approach for the development of *intelligent* or *thinking computers*.

Understanding how the brain processes information, adapts to complex and novel environments and is able to generalize on one side, and using this knowledge in designing complex adaptive systems that build up their own rules through experience on the other side, are all part of the ***Neural computation*** field of study. The interdisciplinary research goes into distinct directions, from designing tools able to analyze large nonlinear data sets (*artificial neural networks*), trying to understand higher-level brain functions by measuring cortical activity using techniques like EEG, fMRI and PET (*cognitive neuroscience*) to suggesting possible mechanisms underlying brain function using biologically inspired computational models (*computational neuroscience*) and solving engineering problems using cortical inspired architectures (*neuromorphic systems*).

A first step towards biologically-inspired information processing systems was achieved through the development of ***Artificial Neural Networks*** (ANN). Following the general aspects of biological neural systems, they represent an information processing technique that consists of a large number of highly interconnected parallel processing units, whose interactions are modified by simple learning rules. Unlike the biological systems, they are characterized by simple architectures and relatively simple and usually identical processing elements. The first processing units used in ANNs were the McCulloch-Pitts threshold neurons (McCulloch & Pitts, 1943), which are rate model neurons using a step, or threshold activation function. Using these simple units, the multi-layer perceptrons and Hopfield networks were successfully applied in digital computation. Later, rate model neurons using continuous activation functions, like the sigmoid function, used in feed-forward and recurrent networks, were able to approximate also analog computations. Configured through a specific structure and parameter set for a specific application, the ANNs are capable of solving some of the problems that are too complex for conventional technologies based on algorithmic approaches. They are successfully applied to an increasing large number of real world problems like pattern recognition, nonlinear regression, clustering and data classification.

The next step in developing *intelligent* computers, having as inspiration source the biological neural systems, goes in the direction of developing a *general theoretical framework* for studying the fundamental principles underlying the cortical mechanisms of brain function and subsequently understanding the working principles of the brain.

In order to answer fundamental questions like: what functions and computations are performed in our brain and by which mechanisms, the extensive research in ***experimental neuroscience*** investigates the complex nervous system by studying, along with its structure and properties, its responses to specific stimulations under specific conditions. For example neurophysiological experiments using implanted electrodes on awake behaving animals allow measurements of single neuron activities from different cortical areas during various behavioral tasks. Studies of cognitive brain function in humans measure indirectly the neuronal activity through non-invasive techniques such as fMRI. This technique measures the regional changes in blood flow, which is indirectly associated with regional differences in brain activity. Although not as precise as single cell measurements, this non-invasive technique allows the association of different brain regions with particular processing stages of different information types. Detailed introductory descriptions of neuronal physiology, cortical organization and measuring techniques can be found in general neuroscience textbooks, for example Kandel et al. (2000).

In the search to understand and analyze the cortical, neural and synaptic mechanisms underlying the complex functions of the brain, the experimental research is complemented by theoretical models. The ***computational neuroscience*** field of study investigates the dynamics of the neural system at different levels of detail by introducing appropriate mathematical techniques for analyzing and reproducing the rich behavior of the complex neural system from the cortical to the synaptic level. These techniques underlie the development of explicit *mathematical neurodynamical models* of cortical modules, neuronal assemblies, single neurons and individual synapses. The models make explicit assumptions about the underlying neural mechanisms of brain function by integrating different levels of experimental investigation of brain function: behavioral, neuroanatomical, neurophysiological, neuroimaging and single cell studies. Their analysis allows then specific conclusions to be drawn about the studied neural behavior and the assumed underlying neural mechanisms. Different modeling approaches are described in computational neuroscience textbooks, for example Koch (1999); Dayan & Abbott (2001).

This work contributes to the understanding of the fundamental brain principles underlying cognitive functions by introducing and systematically analyzing two concrete models studying specific cognitive phenomena as effects of multi-areal recurrent processing. They provide a number of important basic features of cortical function and represent fundamental building blocks for the construction of powerful computational systems.

Research methodology

Cognitive brain functions, like selective attention, working memory and decision making, require complex information processing techniques in order to select and represent the behavioral relevant information and to create the correct associations. Present-day research addresses these issues by developing neurodynamical computational models inspired from the structure and properties of the nervous system. Important to note here is that distinct features of cognitive brain function seem to emerge from the mutual interplay of interconnected brain structures rather than being generated by individual structures. Searching to explain complex cognitive phenomena as effects of multiareal recurrent processing, *multi-modular systems of interacting* or *coupled recurrent networks* are proposed, which analyze and model the techniques underlying these complex multiareal recurrent processes of the brain.

For this, a cortical module, having a specific functionality, is modeled as a *recurrent network of interconnected excitatory and inhibitory spiking neurons* described by nonlinear dynamics and characterized by stochastic activity, for which stimuli-related inputs are processed in the context of neuronal reverberation (Amit et al., 1994; Amit & Brunel, 1997b,a; Brunel & Wang, 2001). The dynamics of the proposed models are described at the spiking and synaptic activity levels, which provide a quantitative formulation for the temporal evolution of the neural activity at many processing levels: from single neurons, neuron assemblies, recurrent networks of neurons and up to coupled hierarchical modules of networks. This enables psychophysical, neurophysiological and imaging studies to be explicitly simulated and predicted. The structure and properties of these biologically inspired network models, that are able to reproduce experimental measurements studying complex phenomena in the brain, will help to derive relevant biological principles underlying perception and cognition.

Choosing the structure of the cortical modules is a very important aspect, which recent modeling strategies address through the conceptual architectural framework of ***Biased-Competition and Cooperation***. The *Biased-Competition Hypothesis*, formulated for the mechanism of selective attention, assumes that multiple activated populations of neurons engage in competitive interactions that are biased by external interactions in favor of specific groups of neurons (Moran & Desimone, 1985; Spitzer et al., 1988; Motter, 1993; Miller et al., 1993; Chelazzi et al., 1993; Desimone & Duncan, 1995; Reynolds & Desimone, 1999; Chelazzi, 1999). Neurodynamical models constructed within the Biased-Competition framework have been proven to successfully account for different aspects of visual attention (Rolls & Deco, 2002; Corchs et al., 2003) and working memory context dependent tasks (Deco & Rolls, 2003; Deco et al., 2004). The theoretical framework of *Biased-Competition and Cooperation* extends this view: while competition selectively highlights the neural representation of the attended or behavioral-relevant information at the expense of the representation of other present but not relevant information, cooperation promotes the co-activation of the representations of the associated or related information (Szabo et al., 2004; Almeida et al., 2004). Neurodynamical models developed within this conceptual framework have been used to model single neuronal responses, fMRI activation patterns, psychophysical measurements, effects of pharmacological agents and local cortical lesions: Deco & Rolls (2002, 2004); Szabo et al. (2004); Almeida et al. (2004); Szabo et al. (2006).

Thesis overview

This work introduces, by proposing recurrent network models of spiking neurons constructed using the theoretical framework of *Biased-Competition and Cooperation*, a set of underlying working hypotheses for the cognitive brain phenomena, like selective filtering, correlation facilitation and selective tuning, and relates these hypotheses to experimental evidence. Guided by cortical activity measurements from recent neurophysiological experiments on behaving mammals, two examples of functional neurocomputational models are described.

The first study introduces a neurocognitive model of selective attention, and exhaustively analyzes how competition and cooperation biased by behaviorally-relevant information operate within a single model area. The mechanisms of *selective attention* form an important basis of cognitive processing. Through attention, information is selected and filtered out in a context-dependent way, where the context is provided by the internal state of the brain. A remarkable phenomenon of selective attention for human vision known as *inattentional blindness* (for a review see Simons, 2000) refers to the inability of humans to recover any information from the unattended parts of the visual field. *Attentional filtering* represents a particularly strong attentional effect, in which the context gates sensory input in an all-or-none fashion, and might be part of a neural correlate of the inattentional blindness observed in humans. The role of the prefrontal cortex in the attentional filtering mechanism was studied in a recent neurophysiological experiment on awake behaving monkeys engaged in a focused attention task (Everling et al., 2002). Motivated by their results, specifying that only an attended task-relevant stimulus is gated by the context and is allowed to be represented, a neurodynamical computational model of a small part of the prefrontal cortex is proposed to account for the neural mechanisms defining this attentional filtering effect. The model investigates how this strong attentional effect can arise from a weak modulatory bias which mediates the cortical context (Szabo et al., 2004).

The second study introduces a neurocognitive model for learning visual categorization that operates over two different cortical modules. A recent neurophysiological experiment on awake behaving monkeys has shown that learning a visual categorization task shapes the selectivity of inferotemporal cortex (ITC) neurons to the task-relevant features of the presented stimuli (Sigala & Logothetis, 2002). Hypothesizing that the task-dependent shaping of feature-selectivity might emerge as a dynamic effect through the information exchange between ITC and another cortical area, possibly the prefrontal cortex (PFC), where the previously learned stimulus categories could be encoded, a biologically inspired neurodynamical two-layer model is proposed. The model investigates how the selectivity of the ITC model neurons can arise and how it is influenced by learning the categorization task, implemented using a reward-based Hebbian learning algorithm (Szabo et al., 2006). An important feature of the proposed model, which exhaustively analyzes how competition and cooperation operate within a two-area model, is the internal generation of the biases needed for the competition process in the PFC model area.

Along with correctly describing the experimental findings on awake behaving animals, both models help to extract a number of important features underlying cortical mechanisms, like selective filtering, correlation facilitation and selective tuning. These dynamical features represent fundamental building blocks for large-scale multi-areal neurodynamical networks modeling cognitive brain functions and could be an important link to a fundamental brain principle.

The thesis is organized as follows: At first the building blocks of the model networks, i.e. the neurons, are described in chapter 2. The chapter starts with a general introduction in the structure and properties of biological neurons (section 2.1) and follows with a description of different modeling techniques (section 2.2), presenting in detail the modeling strategy adopted in this work: the leaky integrate-and-fire neuron with nonlinear synaptic dynamics (section 2.2.4).

Chapter 3 starts by presenting different assumptions for the encoding mechanisms performed by the nervous system (section 3.1). Afterwards, based on properties of the biological cortical networks, the structure and parameters of the considered network model are presented in section 3.2. The chapter ends with the description and biological motivation for the *Biased-Competition and Cooperation* architectural framework (section 3.3).

Chapter 4 introduces the *Reward-based Hebbian learning mechanism* used for training the proposed biological inspired spiking network model from chapter 7. Chapter 5 describes the Mean-field approximation which simplifies the analysis of the network and allows exhaustive explorations of its parameter space.

Chapters 6 and 7 introduce two examples of functional network models. They study the influence of selective attention on perceptual processing (chapter 6) and the influence of concept formation on the lower-level representations of information in the biological cortical structures (chapter 7). Simulation results of both mean-field analysis for stationary conditions and full spiking-dynamics are presented and discussed. The last chapter discusses the results of the modeling strategies and presents the conclusions of this work.

2 Network processing units

The scientific community in neuroscience has agreed that information processing in the nervous system underlying all sensory, motor and cognitive functions is achieved by the conjoint activity of a large group of specialized cells called ***neurons***. Sparsely and inhomogeneously interconnected, they form an intricate network with remarkable storing and encoding capabilities. Referred to as the *information processing units of the nervous system*, the neurons are complex biophysical and biochemical entities able to perform specific kinds of computations. They are responsible for encoding, integrating and transmitting information in the nervous system.

Anatomical and physiological data accumulated over many years of neurobiological research, in special in-vitro and in-vivo single-cell measurements of neuronal activity, provide significant knowledge about the *neuronal structure* and the biophysical and biochemical mechanisms underlying *neuronal activity*. The generation of action potentials and the synaptic transmission are fundamental mechanisms of information transmission in the biological neural system, at their basis underlying specific ionic mechanisms. Understanding the complex biochemical processes behind these mechanisms creates an extended basis for constructing biological realistic models of neurons.

Detailed ***conductance-based models***, like the Hodgkin and Huxley model (Hodgkin & Huxley, 1952), are able to accurately reproduce experimental neurophysiological single-cell recordings but are too complex and not efficient for constructing and analyzing large neural network models. Therefore simple and efficient neuron models were developed, like the ***leaky integrate-and-fire model***, capturing only the major features and properties of biological neurons. Detailed information about neuron biophysics and the corresponding modeling strategies can be found in many computational neuroscience books, see for example Jack et al. (1983); Koch (1999); Dayan & Abbott (2001).

The first part of this chapter, section 2.1, concentrates on the general characteristics of the structure and communication mechanisms of real neurons. The second part, section 2.2, addresses neuron modeling strategies, in particular the one adopted further in this work.

2.1 Biological neurons

Neurons, on the order of $10^{11} - 10^{12}$ in the human cerebral cortex, are the functional cells of the nervous system specialized in the reception, integration and transmission of information in the form of an electro-chemical process. They have a particular structure and specific electrical and chemical properties that are related to their functionality.

Despite the high functional diversity, neurons have common anatomical structures and physical properties, briefly presented in section 2.1.1. All neurons process and transmit information in the same way using a combination of chemical and electrical signals. Across one neuron the information travels as an electrical signal in the form of modifications of the cell's membrane potential. Between neurons, the information is transmitted as a chemical signal in the form of specialized molecules passing through specialized points of contact. These communication mechanisms are introduced in section 2.1.2.

2.1.1 Structure and properties of biological neurons

Although neurons differ in shape and size from one brain region to another, they have many structural and functional features in common. Distinctive from other cells in the organism, neurons have specialized extensions – the *dendrites* and the *axon*; structures – the *synapses*; chemical elements – the *ion channels*, the *neurotransmitters* and the *receptors*; and electrical properties – the *membrane potential* (Braitenberg & Schütz, 1991; Kandel et al., 2000). A schematic representation of a pyramidal neuron can be seen in Figure 2.1. Named from the pyramidal shape of their body, the cortical pyramidal neurons are the most common neurons in the mammalian cortex (around 80% of all cortical neurons).

The **dendrites** are short ramifying branches (usually no longer than 1 mm) extending from the cell body (also called *soma*). The *dendritic tree*, grouping all dendrites, along with the cell body constitute the *receptive, or input, region of the neuron*, where synaptic contacts are made with other neurons (usually around 5000 and most of them dendritic). The receptive region is characterized by specialized protein molecules embedded in the cell membrane, called **receptors**, that respond to the chemical signals sent by the connecting neurons.

The **axon** is a singular long extension of neuron's body which represents *the output region of the neuron.* The axon ranges in length from a few hundred microns to over a meter (for motor-neurons) and divides in many branches (forming the so called *axonal arborization*). This enables an active neuron to transmit its output, to many neurons in the same and also other remote areas in the nervous system. For the cortical pyramidal neurons, the axon can connect from 10^3 to more than 10^4 other cortical neurons, many of them residing in the same brain area.

The axonal fibers are often wrapped in multiple sheaths of *myelin* – fatty layers acting as an electrical insulator. The myelinization facilitates a high propagation speed of the electrical signal for small diameters of the axon, by increasing its membrane resistance. This is an efficient solution to group numerous connecting fibers in a small limited space. The myelin sheaths surrounding the axon are segmented at regular intervals by small gaps called *nodes of Ranvier* where the impulse can be regenerated. This is important for the transmission of the electrical impulse over long distances.

The **synapses** are specialized junctions that represent the connection points between neurons where information is transmitted in the form of a chemical signal[1]. Synapses are the neural

[1] I refer here only to chemical synapses which are the common type of synapses encountered in the cortical areas

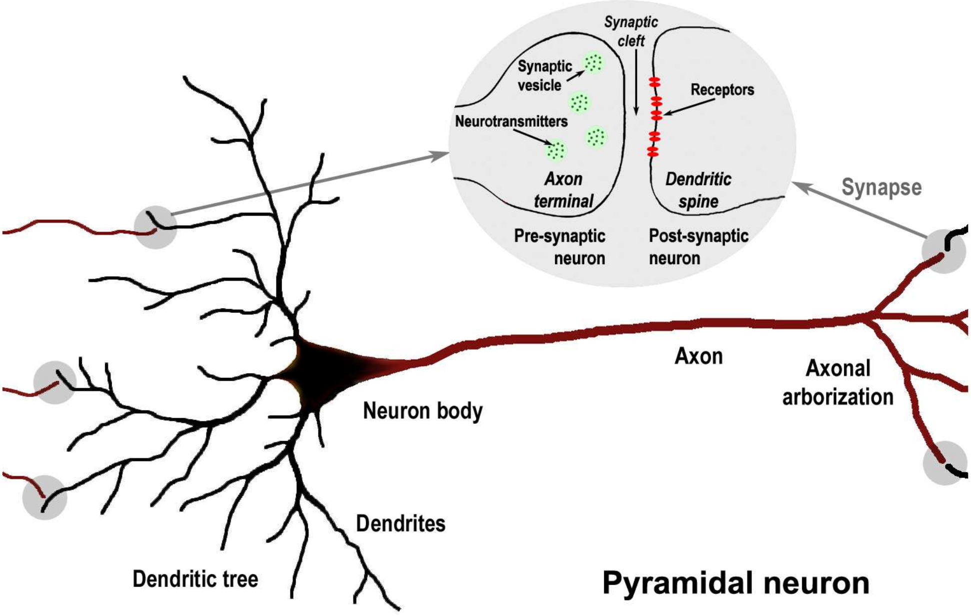


Figure 2.1: Schematic representation of a pyramidal neuron with a detail of the chemical synaptic site. The notations used are explained in the text.

elements believed to be involved in adaptation, learning and memory. A schematic representation of a synapse can be seen in Figure 2.1. The neuron sending the information is referred to as *pre-synaptic* and the receiving neuron as *post-synaptic*. The tips of the axonal branches of the pre-synaptic neuron are called *axon terminals*. They release, in certain conditions, small amounts of special molecules – called ***neurotransmitters*** that diffuse in the gap between the connecting neurons – named *synaptic cleft* – and bind to the *receptors* of the post-synaptic neuron.

For each neurotransmitter type there are more corresponding receptor types, and it is possible that different types of receptors are present in a single synapse. The most common neurotransmitters in the cortex are:

- ***glutamate***[2] – an *excitatory* neurotransmitter that acts most commonly on AMPA[2] and NMDA[2] receptors, and
- ***GABA***[2] – an *inhibitory* neurotransmitter that acts most commonly on GABA_A and GABA_B receptors.

Each synapse is characterized by a specific *synaptic efficacy* determined by the type and amount of released neurotransmitters and the type and number of activated ion channels on the post-synaptic site.

Along with the particular structure and chemical properties, neurons have also particular elec-

[2] Abbreviations: glutamate – amino acid L-glutamate; AMPA – α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid; NMDA – N-methyl-D-aspartate; GABA – gamma-aminobutyric acid

trical properties. They are characterized by temporal and spatial electrical changes that can be accounted by modifications of the electrical potential across neuron's membrane. The **membrane potential** is defined as the voltage difference between the inside of the neuron and the extracellular fluid that surrounds it. Biological values for neuron membrane potentials range from −90 mV to +50 mV.

The neuron is bounded by a selectively-permeable membrane composed of a double layer of lipid (fatty) molecules, 3 to 4 nm thick, embedding specialized proteins referred to as **channels**, that selectively allow the passage of specific charged particles across it (most commonly the Na^+,K^+,Cl^- and Ca^{2+} ions). *Voltage-gated channels* are specific ion channels that open only for some values of the membrane potential. The voltage dependence is important for the generation and conduction of electrical signals in the output region of the neuron. *Ligand-gated channels* open in response to a specific chemical stimulus, and are important for the communication between the neurons. The membrane is an almost perfect electrical insulator being impermeable to most charged particles, acting as a *capacitance*. The embedded ion-conducting channels act as a *conductance* across the membrane.

The resting (quiescent) state of a neuron is the state of dynamical equilibrium in which the intra- and extra- cellular ionic distributions are balanced in such a way that no net ionic flow is present across the membrane. The resting state is a polarized state: there are more negative ions inside and more positive ions outside the neuron, or in other words the inside of the neuron is more negative than the extracellular fluid. For a neuron at rest, the electrical potential across the membrane, known as the **resting membrane potential**, V_{rest}, has a typical value of −70 mV. The changes in the membrane potential are mediated by the flow of sodium Na^+ and potassium K^+ ions, between the intra- and extra- cellular space. Most of the time, the membrane potential has a negative value around its resting value.

A description of the complex biophysical mechanism underlying neural communication is given in the next subsection.

2.1.2 Mechanism of neural communication

One important neural communication mechanism, which will be considered here, consists of three complementary processes: synaptic transmission between neurons, electrical conduction and integration along the neuron membrane and action potential generation in the output region of the neuron.

Synaptic transmission is a complex electro-chemical process that takes place at the synapse. The electrical signal from the pre-synaptic neuron is transformed into a chemical signal that is transmitted through de extracellular space to the post-synaptic neuron where it is transformed back into an electrical signal. A schematic representation of synaptic transmission is presented in Figure 2.2.

The receptors of the post-synaptic neuron are directly or indirectly (using intracellular second messengers) coupled to ligand-gated membrane ion channels, which are induced to open in re-

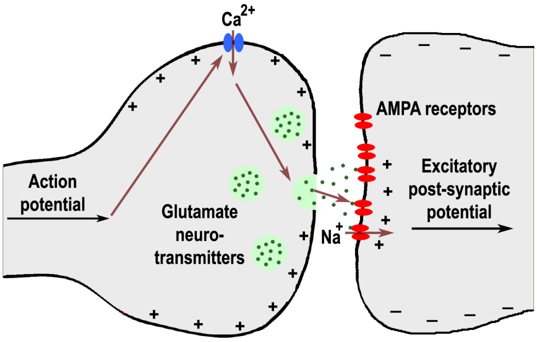

Ca2+
Action potential
Glutamate neuro-transmitters
AMPA receptors
Excitatory post-synaptic potential
Na+

Figure 2.2: Schematic representation of the process underlying synaptic transmission. The electrical impulse reaching the axon terminal of the pre-synaptic neuron causes voltage-sensitive calcium channels to open. The calcium ions trigger the release of neurotransmitters in the synaptic cleft, which bind to the corresponding receptors of the post-synaptic neuron and induce particular ion channels to open. For example: AMPA receptors are Na^+ channels; NMDA receptors are Ca^{2+} channels and GABA$_A$ receptors are Cl^- channels. The resulting ion flow induces specific electrical changes in the post-synaptic neuron.

sponse to the neurotransmitter binding process. This is equivalent to a change in the membrane conductance of the post-synaptic site, called ***synaptic conductance***, that depends on the post-synaptic receptor properties. Different receptors induce different response types: AMPA and GABA$_A$ receptors mediate fast synaptic transmissions, NMDA receptors mediate a slower synaptic transmission, while GABA$_B$ receptors mediate a slow and long lasting synaptic transmission. The NMDA receptors are in addition voltage dependent: they can only be activated if a sufficient level of depolarization is present (which removes the Mg^{2+} ions blocking them). Thus NMDA receptors act as coincidence detectors of pre- and post-synaptic activity, which makes them possible candidates for the biochemical implementation of the Hebbian learning mechanism.

Channel opening induces a small ion flow into the post-synaptic neuron, which generates a transient and local small change in the electrical potential across the membrane – known as **post-synaptic potential**[3].

Depending on the sign of the electrical change, two types of stimulation or input are defined:

- one that *depolarizes* the membrane potential, i.e. the voltage inside the cell becomes more close to the one outside, denoted as **excitatory**, and
- one that *hyperpolarizes* the membrane potential, i.e. the voltage inside the cell becomes even more negative than the voltage outside, denoted as **inhibitory**.

The neurotransmitters, their corresponding receptors and thus the synapse type are also either inhibitory or excitatory, depending on the type of stimulation that they induce. Conform to Dale's principle (Dale, 1935), the synaptic connections made by a pre-synaptic neuron to other

[3] sometimes referred to as graded potential because it represents a small change in the membrane potential that degrades with distance and in time

neurons are either all excitatory or all inhibitory and thus a neuron can be also denoted as excitatory or inhibitory. The pyramidal neurons are believed to be excitatory, acting through the excitatory glutamate neurotransmitter. Most interneurons are believed to be inhibitory, acting through the inhibitory GABA neurotransmitter. They induce *excitatory post-synaptic potentials* (EPSP) and *inhibitory post-synaptic potentials* (IPSP), respectively.

The size of the induced post-synaptic current depends on the deviation of the post-synaptic membrane potential from the reversal potential of the corresponding ionic channels. The ***reversal potential*** is defined as the potential for which the sign of the induced synaptic current reverses and is given by the Nernst potential of the ion channels. The reversal potential depends on the synapse type: for excitatory synapses the reversal potential is much larger than the resting potential, with a typical value of $V_E = 0$ mV, and for inhibitory synapses the reversal potential is close to the resting potential, with a typical value of $V_I = -70$ mV (Connors et al., 1988; McCormick, 1989). The stimulation, either excitatory or inhibitory, drives the membrane potential towards the corresponding reversal potential. In conclusion, the induced synaptic current is strong for membrane potentials far away from the corresponding reversal potential and decreases as membrane potential approaches the reversal potential. If the membrane potential is approximatively equal to the reversal potential of a synapse, an inhibitory input has no direct effect on the membrane potential but it affects the depolarization caused by other excitatory inputs. This effect is called *'shunting inhibition'* because it effectively shunts the inputs received from other excitatory synapses.

The process of **electrical conduction and integration** takes place through passive ion diffusion inside the neuron. The induced post-synaptic potentials are propagated along the dendrites and the cell body in a passive way, with decreasing strength over distance and time. Locally, the tendency of the membrane potential after stimulation is to slowly decay to its resting value due to leakage currents flowing through the membrane. The post-synaptic potentials induced at different locations and times are integrated over the cell body using spatial and temporal summation, respectively.

Action potential generation is conditioned by the depolarization level of the membrane potential at a special region of the cell body connecting the axon, named *axon hillock*. When the accumulated electrical potential exceeds a certain threshold value, called **excitation** or **spiking threshold**, a brief voltage pulse known as **action potential**, **impulse** or **spike** is generated. For cortical neurons, the excitation threshold V_{thr} has a typical value of -50 mV. Excitatory inputs drive the membrane potential towards the spiking threshold, thus making the neuron more likely to fire. Inhibitory inputs drive the membrane potential away from the spiking threshold, thus making the neuron less likely to fire. Cortical excitatory post-synaptic potentials are usually around 0.2 mV, thus more than 100 excitatory inputs have to arrive in a very short time interval, for a spike to be initiated.

The common features of an action potential are the fast depolarization of the membrane above 0 mV, generated by a fast influx of sodium ions, followed by a slightly slower re-polarization of the membrane generated by a slower efflux of potassium ions. Thus for a brief period of time, about 1 ms, the membrane potential suddenly reverses its electrical potential. Afterwards, the membrane repolarizes towards a slightly hyperpolarized state characterized by the ***reset membrane potential***, whose typical value for cortical neurons is $V_{reset} = -75$ mV. During the hyperpolarization period, the sodium channels are inactivated. This inhibits the generation of another action potential for a short time interval called ***refractory period*** and ensures that the action potentials travel in only one direction. The refractory period limits the maximum frequency with which real neurons fire.

Due to the special configuration and structure of the axon, the action potential propagates fast, with constant speed and amplitude, away from the cell body in all its axonal branches. Along the myelinated parts of the axon the impulse is propagated using cable properties, in a similar way how electrical current travels along insulated cables. This type of conduction is fast and efficient, using a small amount of energy, but the strength of the impulse decays in time. The nodes of Ranvier are the places where the action potential is regenerated by permitting ion exchange. Here, the signal regains its initial strength conditioned that it remained above the excitation threshold. This alternating way of conducting the impulses along the axon is called *saltatory conduction*. Because the action potential is generated only in the moment when the membrane potential exceeds the excitation threshold and it maintains a constant amplitude along the axon, it is said to obey an *all-or-none* rule (Adrian, 1914).

The *passive conduction* of the electrical potentials in the dendrites and the cell body, characterized by slow propagation with decaying strength, is only effective for communication on short distances. In contrast, the *active conduction* along the axon, characterized by fast and efficient propagation at constant strength, is effective for communication on long distances.

Important to note here is that across all types of neurons, action potentials have a generally stereotyped shape with a typical duration of the order of one millisecond. Thus, it is reasonable to assume that their actual shape does not contain any information, and that neuron activity can be fully characterized by their time of generation. This implies that the information sent from one neuron to another must be encoded in the *timing* or the *rate* of the transmitted impulses. Different encoding strategies will be discussed in chapter 3.1.

In-vivo single cell recordings show that for real neurons spikes usually occur at irregular intervals and are easily separated one from another. The sequence of spikes emitted by a neuron, i.e. its output, is referred to as ***spike train*** and is usually characterized through its ***spiking rate*** or ***inter-spike intervals***. The spiking rate is usually estimated by averaging over small time intervals and over many instances, i.e. many recordings of the same neuron under the same conditions of stimulation. All these important findings on the structure and functionality of biological neurons that were described in this section, will be used to develop different mathematical models of neurons, from very simple to more complex, as illustrated in the next section.

2.2 Model neurons

A prerequisite of understanding the neurodynamical mechanisms underlying brain function is to construct neuron models extracting the essence from the complex structure and behavior of real neurons. Their level of description should be accurate enough to allow relevant mechanisms at the physiological level to be properly taken into account. Also, their description should be simple and effective, allowing the construction of network models with a large number of neurons and synaptic connections.

As seen in the previous section, the activity of biological neurons is characterized by the emission of discrete electrical impulses, i.e. spikes, which are considered to be the elementary units of signal transmission in the nervous system. Different modeling strategies describe the neuronal activity in terms of:

- a continuous variable representing the average activation rate: ***rate models***;
- a series of discrete impulses representing the spike train: ***spiking models***;
- or more detailed, a continuous variable representing the entire time evolution of the membrane potential: ***conductance-based models***.

Simple models describe the synapses as simple current sources, while complex models include nonlinear synaptic dynamics. Section 2.2.1 captures the tradeoff between accuracy and efficiency in neuronal modeling. Section 2.2.2 introduces the *leaky integrator* rate model. A simple and frequently used spiking neuron model, the *leaky integrate-and-fire* model, is presented in Section 2.2.3. Its extension capturing more complex synaptic dynamics, presented in Section 2.2.4, represents the biological-inspired neuron model used further in this work.

2.2.1 Complex versus simple neuron models

Pioneers in the study of neural mechanisms, ***Hodgkin and Huxley*** analyzed the action potential generation mechanism for the giant axon of the squid (Hodgkin & Huxley, 1952). They developed a complex mathematical model describing the voltage and time dependence of the ionic axonal permeabilities, using curve fitting from experimental data. The voltage dependent conductances were later explained by the voltage-dependent ion channels embedded in the neuron membrane (as described in the previous section). Their detailed model, consisting of a four-dimensional set of coupled nonlinear partial differential equations, generates action potentials extremely accurate in shape and time course as compared with the experimental single-cell recordings.

Following the formalism introduced by Hodgkin and Huxley, the **conductance-based neuron models** describe through membrane conductances the collective behavior of the ion channels of the same type from a region of the cell membrane. The active or nonlinear conductances accounting for channel's dependence on the membrane potential or the presence of a specific chemical particle, enable a more accurate modeling of the specific phenomena observed in experimental measurements. The conductance-based neuron models are complex mathematical models using detailed descriptions of the biophysical mechanisms for synaptic integration and action potential generation. They are continuous-time models that capture the detailed time course of the membrane potential and are able to reproduce with good accuracy the neurophysiological in-vivo single-cell measurements but are difficult to analyze analytically and also computational expensive to be used in large interconnected neural networks.

An important aspect to be taken into account is that real neurons can show substantial differences in the membrane potential across their cell surface. A good modeling strategy would be to divide the neuron into a number of inter-connected simple compartments, characterized by membrane potentials with small spatial variations, and model each compartment separately. *Multi-compartment models* are more realistic but also more complex and difficult to analyze and use in large neural networks. *Single compartment models*, equivalent to point-like neurons, completely ignore the structure of the dendritic tree and consider that the membrane potential is uniform across the entire cell surface. Their internal state can be completely characterized by a single variable representing the average membrane potential across the entire cell.

Considering the behavior of the biological neuron's membrane potential, two distinct regimes can be easily separated: the slow varying *subthreshold regime* and the brief and fast varying *spike generation regime*. Most of the time a neuron finds itself in the subthreshold regime. Here, the membrane potential is continuously and slowly changing by the integration of the incoming post-synaptic potentials. The brief spike generation regime is reached when the membrane potential exceeds the excitability threshold and is characterized by fast stereotyped changes of the membrane potential. It is followed by the short refractory period in which another action potential can not be generated. Neuron models can be considerably simplified without major functionality loss by considering only the time when an impulse is generated without modeling its actual time course. The impulses are reduced to only a point in time, expressing the spike or firing time, which is defined as the moment when the membrane potential crosses the spiking threshold. These models are called ***integrate-and-fire neuron models*** and their activity is fully characterized by the sequence of firing times, called the spike train.

An even more drastic simplification ignores also the exact timing of the impulses and describes the neuron activity only through its average firing or spiking rate. These reduced models, called ***rate neuron models***, represent a good approximation in the case of a large number of uncorrelated pre-synaptic spikes. Because of their simplicity, they make possible the steady state analysis of large network models (as shown in chapter 5).

2.2.2 Leaky integrator rate model

Rate neuron models represent a standard tool in neural network theory that describe the activity of the neurons in terms of their activation or firing rate. The *activation* of a neuron, denoted by $x(t)$, depends on its total afferent input, $I_{tot}(t)$, that is given by the activation of the connecting neurons. This dependence, describing the input-output characteristic of the neuron, is expressed through the *activation* or *transfer function*: $x(t) = f(I_{tot}(t))$. The shape of the transfer function gives the characteristics of the performed computation. For linear transfer functions, the neuron model can compute only linear functions but it has the advantage that its dynamics can be computed analytically. Introducing nonlinearities increases the computational power of the model. A piecewise linear transfer function considers that the activation is zero below some threshold level of the afferent input and linear above it, accounting for the activation threshold of neural excitability. A sigmoidal transfer function accounts for the saturation of firing rates for high inputs, reflecting the refractoriness property of neurons.

The leaky (or forgetful) integrator model accounts for the transient nature of the inputs by introducing leakage currents, which imply temporal decays in the activity level. The activity of neuron i in the network is represented by the variable $x_i(t)$ whose dynamics follows the exponential decay model:

$$\tau \frac{dx_i(t)}{dt} = -x_i(t) + f(I_{i,tot}) \tag{2.1}$$

where τ is the integration time constant that characterizes the response time of the neuron. A change in the synaptic input induces an asymptotical change in the activity level of the neuron towards the fixed point $x_{i,0} = f(I_{i,tot})$.

The total synaptic input driving neuron i integrates by spatial summation all afferent stimulations coming from other neurons in the network and possibly from other sources outside the model network. It can be written as a weighted linear sum of the individual activations:

$$I_{i,tot} = I_{i,ext} + \sum_j w_{ij} x_j(t) \tag{2.2}$$

where $w_{ij}x_j(t)$ denotes the amount of input contributed by neuron j and $I_{i,ext}$ denotes input from other external sources. w_{ij} represents the strength or efficacy of the connection from neuron j to neuron i.

For some models, the activation $x(t)$ and input $I(t)$ are considered to represent the firing rate and the afferent synaptic current, respectively. In this case $f()$ is referred to as the frequency-current $(f - I)$ curve.

2.2.3 Leaky integrate-and-fire neuron model

The **integrate-and-fire** (IF) models, proposed for the first time in 1907 by Lapicque (Lapicque, 1907), represent a class of reduced spiking neuron models characterized by a simple level of description of both integration (subthreshold) and spiking regime, which makes them computationally very efficient. They maintain the essential features of neuronal excitability like the all-or-none behavior, synaptic input integration and refractoriness, but leave out the detailed biophysical descriptions of the synaptic channels and the generation of action potentials. The dynamics of the membrane potential is modeled only for the subthreshold regime. The firing regime is described by simple rules that can include also the refractoriness property.

The standard **leaky integrate-and-fire** (LIF) neuron model is a single-compartment model that accounts for the membrane leakage currents through a passive (constant) conductance and omits all other active synaptic conductances, which makes it easy to implement and simulate. It simply translates the afferent spikes to currents of constant amplitude that are temporally summed into a single variable. A fixed threshold is used for the generation of the discrete identical pulses. Due to its simplicity and efficiency, the LIF neuron has proved to be a practical model for studying the dynamics of neural networks with a large number of interconnected neurons. A detailed description of the model can be found in Gerstner & Kistler (2002).

The equivalent electrical circuit of the LIF neuron model, presented in Figure 2.3, consists of a capacitor, $\mathbf{C_m}$, expressing the total *membrane capacitance* of the cell, in parallel with a resistor, $\mathbf{R_L}$, expressing the total *membrane leak resistance* of the cell. The voltage across the capacitor corresponds to the *membrane potential* of the LIF neuron and is denoted through $\mathbf{V_m}$. The *total afferent input* to the model neuron is modeled through a current source, $\mathbf{I_{tot}}$, that integrates the individual synaptic currents generated by the activations of the pre-synaptic neurons. The RC circuit is characterized by the time constant $\tau_\mathbf{m} = R_L C_m$, referred to as the *membrane time constant*, which determines the response time of the neuron to the afferent stimulations. The *membrane leak conductance* $\mathbf{g_L} = 1/R_L$ characterizes the cellular mechanisms that restore the equilibrium state of the neuron after stimulation and enables the model to take into account the temporal relationships in the inputs.

In absence of stimulation, $I_{tot} = 0$ corresponding to a neuron at rest, the voltage across the capacitor is given by the constant voltage source $\mathbf{V_L}$, the *leakage reversal potential*, which accounts for the polarized equilibrium or resting state of real neurons. Excitatory, $I_{tot} > 0$, or inhibitory, $I_{tot} < 0$, stimulations drive away the membrane potential from its resting value V_L. This results in a potential difference across the resistor, which gives rise to the membrane leakage current, $I_L(t) = g_L(V_m(t) - V_L)$, opposing the voltage change. The membrane potential is thus asymptotically restored to its resting value, with time constant τ_m.

Using simple voltage-current relationships from electrical circuit theory, the ***subthreshold membrane potential dynamics*** of the LIF neuron model can be described by the following integrator equation (Knight, 1972; Ricciardi, 1977; Tuckwell, 1988):

$$\tau_\mathbf{m} \frac{\mathbf{dV_m(t)}}{\mathbf{dt}} = -\mathbf{V_m(t)} + \mathbf{V_L} + \mathbf{R_L I_{tot}(t)} \tag{2.3}$$

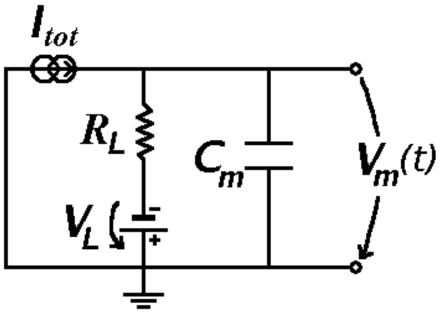


Figure 2.3: Equivalent electrical circuit of the leaky integrate-and-fire neuron model for the subthreshold regime. The parallel $R_L C_m$ circuit models the passive electrical properties of real neurons. The current source I_{tot} accounts for the total synaptic input, and the voltage across the capacitor V_m represents the membrane potential of the model neuron. For details see text.

where $V_m(t)$ is the instantaneous value of the membrane potential that completely describes the internal state of the integrate-and-fire neuron. For the LIF model, the time evolution of the subthreshold membrane potential is described as a simple one dimensional continuous dynamical system.

The LIF model introduces a nonlinearity at the spike generation regime: when $V_m(t)$ is depolarized above a chosen *firing threshold*, θ, a spike is generated and the membrane potential is reset to the *reset potential* $\mathbf{V_{reset}}$. The spikes are described as pulses of infinitely large amplitude and infinitely short duration through the Dirac delta function $\delta(t)$. For a time period corresponding to the *absolute refractory period* $\tau_{\mathbf{ref}}$ the neuron is prevented to emit another spike by keeping $V_m(t) = V_{reset}$. Accordingly, the ***firing regime*** is specified by the rule:

$$\mathbf{if}\ \mathbf{V_m(t^k)} = \theta\ \mathbf{then\,generate}\ \delta(\mathbf{t}-\mathbf{t^k})\ \mathbf{and\,set}\ \mathbf{V_m(t)} = \mathbf{V_{reset}}\ \mathbf{for\ t} \in [\mathbf{t^{k+}}, \mathbf{t^k} + \tau_{\mathbf{ref}}] \quad (2.4)$$

The ***output*** of the LIF neuron is given by the ordered sequence of firing times, i.e. the *spike train*, and is modeled as a series of Dirac delta functions: $\sum_{\mathbf{k}} \delta(\mathbf{t} - \mathbf{t^k})$ with $V_m(t^k) = \theta$.

For constant stimulation, $I_{tot}(t) = I$, the time course of the membrane potential starting from the initial value $V_m(t_0)$ at $t = t_0$, is given by:

$$V_m(t) = (V_L + R_L I) + (V_m(t_0) - (V_L + R_L I))\, e^{-\frac{t-t_0}{\tau_m}}.$$

The tendency of the system is to return from the initial state to an equilibrium state characterized by the asymptotically stable fixed point $V_m(t_\infty) = V_L + R_L I$.

For an input spike train, defined as a discrete series of Dirac delta functions, the total afferent current to the cell can be written as:

$$I_{tot}(t) = \sum_{j=1}^{N} w_j \sum_k \delta(t - t_j^k) \quad (2.5)$$

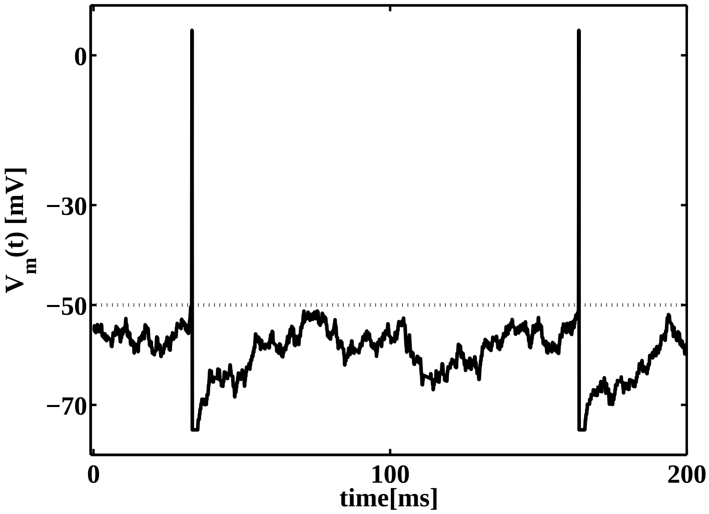


Figure 2.4: Time course of the membrane potential of a standard LIF neuron stimulated by a stochastic current defined as a Poisson-distributed train of Dirac delta impulses with rate $N \cdot \nu$. It was assumed that the neuron is driven by $N = 10^3$ inputs, of which 80% are excitatory with a mean activation rate of $\nu_E = 3$ Hz and 20% are inhibitory with a mean activation rate of $\nu_I = 9$ Hz, and using the following parameters: $V_L = -70$ mV, $\tau_m = 20$ ms, $J_j = +1$ for excitatory inputs and -1 for inhibitory inputs. Once the threshold $\theta = -50$ mV is reached, the neuron emits a spike and stays for the refractory period $\tau_{ref} = 2$ ms at the reset potential $V_{reset} = -75$ mV.

where N is the total number of synaptic connections, w_j represents the efficacy of the synaptic coupling (connection strength) with pre-synaptic neuron j. The sum over k runs over all spikes arriving at at a given site j, characterized by their arrival times t_j^k.

The time course of the membrane potential is determined by solving the ordinary differential equation 2.3 with synaptic input given by 2.5:

$$V_m(t) = V_L + (V_m(t_0) - V_L)\, e^{-(t-t_0)/\tau_m} + \sum_j J_j \sum_k e^{-(t-t_j^k)/\tau_m} H(t - t_j^k) \tag{2.6}$$

where $H(t)$ is the Heaviside step function, $H(t - t_j^k) = 1$ for $t > t_j^k$ and 0 otherwise. Each input spike causes an instantaneous jump of size $J_j = \frac{w_j}{C_m}$ at time t_j^k in the membrane potential (neuron model with instantaneous synapses) that will thereafter decay exponentially with time constant τ_m. For a large time constant as compared to the average inter-spike interval, the decay of the membrane potential is negligible and the output of the LIF neuron is equivalent to a temporal linear integration of the inputs. For a smaller time constant as the inter-spike interval, the leak becomes significant. Hence the excitatory input has to be strong enough to initiate a spike, evidencing the correlations in the input spike trains.

An example of the time course of the membrane potential as described by equation 2.6, considering stochastic excitatory and inhibitory input, is shown in Figure 2.4. For stochastic inputs, the evolution of the membrane potential is nondeterministic. In order to analyze its evolution, the standard trick is to approximate the stochastic input with a continuous diffusion process (Ricciardi, 1977; Tuckwell, 1988). This reduces the spiking model to a rate model describing the average firing rate as a function of the mean and variance of the total synaptic input. For more details see section 5.1 that formulates the Mean-field analysis for LIF model neurons.

Ignoring the synaptic current dynamics of real neurons and approximating the synapses as simple current sources, the standard LIF neuron model can account only for linear synaptic interactions. Also, the model ignores the characteristic reversal potentials of different synapse types. Moreover, because of the saturation effects at the synaptic sites, the total synaptic input coming to a real neuron is not just a linear sum of the independent contributions. All this make the standard LIF neuron model unable to capture the rich spiking dynamics of real cortical neurons.

The standard LIF neuron model represents a framework for modeling the complex behavior of real neurons. It can be extended to include nonlinear synaptic dynamics, to account for different types of synaptic inputs and to include saturation effects at the synaptic sites. Such an extended LIF neuron model using more realistic synaptic dynamics and allowing the use of realistic biophysical constants (like synaptic conductances and delays) is presented in the next section.

2.2.4 Leaky integrate-and-fire neuron with nonlinear synaptic dynamics

The **leaky integrate-and-fire with nonlinear synaptic dynamics** (LIF-NS) neuron model represents a trade-off between the detailed conductance-based models able to accurately reproduce the complex dynamics of neural activity and the computationally efficient LIF models. Recently proposed in Brunel & Wang (2001); Wang (2002) for cortical models of selective working memory, the LIF-NS model extends the standard LIF model presented in the previous section to include nonlinear synaptic current dynamics following a biologically inspired description. The synapses are regarded as membrane ion channels with specific opening dynamics and the afferent inputs are modeled as dynamical changes in the synaptic conductance. The model takes also into account the dependence of the synaptic currents on the reversal potentials for both excitatory and inhibitory inputs. Being characterized by biophysical time constants, latencies and conductances, this description captures more realistic features of neuronal activity as the simple LIF neuron model and allows a better comparison with the dynamics of biological cortical neurons.

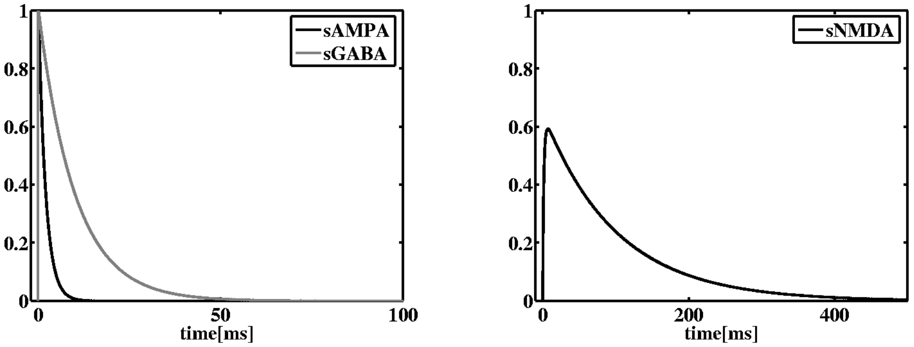
The adopted synaptic model considers a mean-field approach over the contributions of all synaptic ion channels of the same type, i.e. regulated by a specific receptor type R, that are present at a given synaptic site j, assuming that their number is sufficiently large. The synaptic ion channels are assumed to exist in either an open (conductive) state or in a closed state. Their average behavior is described through the gating variable $s_j^R(t)$, expressing the fraction of opened synaptic ion channels of type R at site j.

The LIF-NS neuron model takes into account three types of synaptic input with different temporal characteristics: excitatory input with very fast dynamics mediated by AMPA receptors, excitatory input with slow dynamics mediated by NMDA receptors and inhibitory input with fast dynamics mediated by GABA receptors. For a review of different synaptic current descriptions see Destexhe et al. (1998).

In the case of AMPA and GABA receptors, the responses to the incoming spikes are characterized by very fast activations and fast deactivations. The opening of the channels following a presynaptic spike is smaller than 1 ms and can be considered instantaneous. The general behavior of such fast synaptic transmissions can be described through a steep increase (instantaneous jump) of the gating variables with every afferent spike, followed by exponential decay with time constant τ_{decay}^R (expressing the time constant of the conductance change at the synaptic site). This can be dynamically modeled by a first order differential equation (Destexhe et al., 1998):

$$\frac{ds_j^R(t)}{dt} = -\frac{s_j^R(t)}{\tau_{decay}^R} + \sum_k \delta(t - t_j^k) \tag{2.7}$$

where R is AMPA or GABA. The sum over k runs over all spikes, formulated as Dirac delta functions $\delta(t)$, from pre-synaptic neuron j arriving at times t_j^k. The synaptic decay time constants are set to biophysical realistic values: 2 ms for AMPA (Hestrin et al., 1990; Spruston et al., 1995) and 10 ms for GABA (Salin & Prince, 1996; Xiang et al., 1998). This formulation neglects the saturation of the AMPA and GABA gating variables and is justified only for low



a) fast synaptic dynamics

b) slow synaptic dynamics

Figure 2.5: Time evolution of the gating variables $s_j(t)$ for AMPA, GABA (a) and NMDA (b) receptors in response to one afferent spike at $t = 0$ considering $s_j(0) = 0$.

pre-synaptic firing rates, as in the case of cortical neurons whose mean activation is around the spontaneous firing rate.

In the case of NMDA receptors, the responses are characterized by slower activations and deactivations. The general behavior of the slower synaptic transmission is dynamically modeled by a second order kinetics including both the rise and decay times of the gating variable (Hille, 2001):

$$\frac{ds_j^{NMDA}(t)}{dt} = -\frac{s_j^{NMDA}(t)}{\tau_{decay}^{NMDA}} + \alpha x_j^{NMDA}(t)(1 - s_j^{NMDA}(t)) \tag{2.8}$$

$$\frac{dx_j^{NMDA}(t)}{dt} = -\frac{x_j^{NMDA}(t)}{\tau_{rise}^{NMDA}} + \sum_k \delta(t - t_j^k) \tag{2.9}$$

where the synaptic rise and decay time constants are set to biophysical realistic values: $\tau_{rise}^{NMDA} = 2$ ms and $\tau_{decay}^{NMDA} = 100$ ms (Hestrin et al., 1990; Spruston et al., 1995). The variable $\alpha = 0.5$ ms^{-1} controls the saturation properties of NMDA receptor channels at high pre-synaptic firing frequencies, and $x^{NMDA}(t)$ is an intermediate gating variable.

The time course of the gating variables in the case of one pre-synaptic spike arriving at time $t = 0$ is shown in figure 2.5.a for the fast synaptic dynamics of AMPA and GABA receptors, and in figure 2.5.b for the slow synaptic dynamics of NMDA receptors. This simple and computationally efficient description of the synaptic dynamics, with parameters based on fits to experimental data, captures the essential properties of real synaptic currents. An overview of the synaptic kinetic models is presented in Destexhe et al. (1994).

The effective post-synaptic current mediated by a specific receptor type R at the synaptic site connecting pre-synaptic neuron j, can be expressed, assuming a linear current-voltage relationship, in the following general form:

$$I_j^R(t) = g_{max}^R s_j^R(t) \left(V_m(t) - V_{reversal}^R \right) \tag{2.10}$$

where $g_{max}^R s_j^R(t)$ expresses the time-varying total synaptic conductance at site j mediated by receptor R. g^R_{max} represents the maximum synaptic conductance mediated by the receptor type R and can be thought as channel density multiplied by the maximum conductance of a single channel. The strength of the effective current also depends on the difference between the actual value of the membrane potential and the corresponding reversal potential for that synapse type, $V^R_{reversal}$. Similar to real neuron behavior, when the membrane potential is close to the corresponding reversal potential, the incoming stimulations have almost no effect.

The nonlinear dependence of the NMDA post-synaptic currents on the neuron's depolarization, due to the $[Mg^{2+}]$ blockade, can be expressed through an additional gating factor to the effective current in equation 2.10 (Jahr & Stevens, 1990; Hille, 2001):

$$f_{NMDA}\left(V_m(t), \left[Mg^{2+}\right]_o\right) = \frac{1}{1 + \gamma\, exp(-\beta\, V_m(t))} \tag{2.11}$$

where $\gamma = [Mg^{2+}]_o/3.57$, $\beta = 0.062$ and the extracellular magnesium concentration $[Mg^{2+}]_o = 1$ mM. The constants were determined from single channels studies.

The individual receptor mediated contributions to the synaptic current induced by a pre-synaptic neuron j can be written as:

$$I_j^{AMPA}(t) = g_{max}^{AMPA} \cdot s_j^{AMPA}(t) \cdot (V_m(t) - V_E) \tag{2.12}$$

$$I_j^{NMDA}(t) = g_{max}^{NMDA} \cdot f_{NMDA}\left(V_m(t), \left[Mg^{2+}\right]_o\right) \cdot s_j^{NMDA}(t) \cdot (V_m(t) - V_E) \tag{2.13}$$

$$I_j^{GABA}(t) = g_{max}^{GABA} \cdot s_j^{GABA}(t) \cdot (V_m(t) - V_I) \tag{2.14}$$

where $V_E = 0$ mV is the reversal potential for excitatory synapses ($V_E = V^{AMPA}_{reversal} = V^{NMDA}_{reversal}$); $V_I = -70$ mV is the reversal potential for inhibitory synapses ($V_I = V^{GABA}_{reversal}$); $s_j^{AMPA}(t)$ and $s_j^{GABA}(t)$ are given by the equation 2.7 and $s_j^{NMDA}(t)$ is given by the equations 2.8 and 2.9.

The total synaptic input to a cell is received through recurrent lateral connections coming from neighboring cells located in the same module (area) and external connections coming from distant cells located in other brain areas. The recurrent excitatory post-synaptic currents (EPSCs) are assumed to be mediated by both AMPA and NMDA receptors and the external EPSCs are assumed to be mediated only by AMPA receptors. The inhibitory post-synaptic currents (IPSCs) to both excitatory and inhibitory neurons are assumed to be mediated by GABA receptors (see for example Sheperd (1998)).

The total synaptic current, I_{tot} is calculated as the sum over all synaptic connections and of all receptor-mediated components:

$$I_{tot}(t) = \sum_{j=1}^{N_E} w_j \left(I_{\mathrm{j}}^{AMPA}(t) + I_{\mathrm{j}}^{NMDA}(t)\right) + \sum_{j=1}^{N_I} w_j I_{\mathrm{j}}^{GABA}(t) + \sum_{j=1}^{N_{ext}} I_{\mathrm{j}}^{AMPA}(t) \tag{2.15}$$

where N_E and N_I are the number of excitatory and inhibitory, respectively, recurrent connections with synaptic strengths w_j specified by the network architecture and N_{ext} is the number of external excitatory connections.

For both excitatory and inhibitory LIF-NS model neurons, the subthreshold membrane potential dynamics is described by the integrator equation 2.3, where the total synaptic current $I_{\mathrm{tot}}(t)$ is given by equation 2.15. The same rules describing the spiking regime are used, as for the

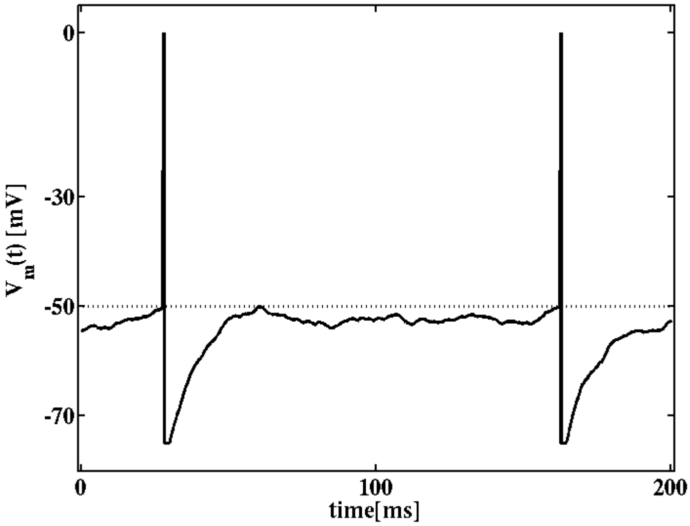


Figure 2.6: Time course of the membrane potential of an excitatory LIF neuron with nonlinear synaptic dynamics stimulated by a stochastic input defined as a Poisson-distributed train of Dirac delta impulses coming from $N_E = 800$ excitatory neurons with a mean firing rate $\nu_E = 3$ Hz and $N_I = 200$ inhibitory neurons with a mean firing rate $\nu_I = 9$ Hz. For more details see the text.

standard LIF neuron (equation 2.4).

For the simulations performed in this work, the biologically inspired model parameters were calibrated based on experimental data from cortical excitatory pyramidal cells and inhibitory interneurons (McCornick et al., 1985; Troyer & Miller, 1997a): resting potential $V_L = -70$ mV, firing threshold $\theta = -50$ mV, reset potential $V_{reset} = -55$ mV[4], membrane capacitance $C_m = 0.5$ nF for excitatory cells and 0.2 nF for inhibitory cells, membrane leak conductance $g_L = 25$ nS for excitatory cells and 20 nS for inhibitory cells, refractory period $\tau_{ref} = 2$ ms for excitatory cells and 1 ms for inhibitory cells, membrane time constant $\tau_m = 20$ ms for excitatory cells and 10 ms for inhibitory cells.

The values for the maximum synaptic conductances can be calibrated such that the spontaneous spiking rates for the excitatory and inhibitory neurons in the model network have the typical values of 3 Hz and 9 Hz, respectively, as observed in the cerebral cortex (Wilson et al., 1994; Koch & Fuster, 1989). In the case of a fully connected network consisting of 800 excitatory and 200 inhibitory neurons, with each neuron receiving 800 external excitatory inputs modeled as independent Poisson processes with an average rate of 3 Hz, the values of the maximum synaptic conductances were calculated by Brunel & Wang (2001): $g_{max}^{AMPA,ext} = 2.08$ nS, $g_{max}^{AMPA} = 0.104$ nS, $g_{max}^{NMDA} = 0.327$ nS and $g_{max}^{GABA} = 1.25$ nS for excitatory neurons; $g_{max}^{AMPA,ext} = 1.62$ nS, $g_{max}^{AMPA} = 0.081$ nS, $g_{max}^{NMDA} = 0.258$ nS and $g_{max}^{GABA} = 0.973$ nS for inhibitory neurons.

[4] In order to explain the high gain of the biological $f - I$ curves of cortical spiking cells and thus explain the high sensitivity of the responses to small changes in the input, Troyer and Miller adjusted the reset potential closer to the threshold potential (i.e. 5 mV under the threshold), Troyer & Miller (1997b). Accordingly, for the simulations performed in this work we use $V_{reset} = -55$ mV instead of -75 mV.

An example of the membrane potential time course for an excitatory LIF-NS neuron receiving a Poisson distributed stochastic input is shown in Figure 2.6. The synaptic current of the standard LIF model consists of instantaneous Dirac impulses, which give rise to discrete jumps in the membrane potential (Figure 2.4). In contrast, the synaptic current of the LIF-NS model is characterized by finite temporal variations giving rise to a continuous modulation of the membrane potential (Figure 2.6). In comparison to the standard LIF neuron for which the membrane potential exhibits large fluctuations between V_{rest} and θ (see Figure 2.4), the membrane potential of the LIF-NS neuron exhibits small fluctuations close to the threshold θ (see Figure 2.6). For balanced excitatory and inhibitory stochastic input, the trajectory of the membrane potential is similar to that of a random walk. An extra excitatory input can move the mean of $V_m(t)$ in the vicinity of the threshold and the stochastic spike arrival can drive $V_m(t)$ over the threshold (firing is driven by the fluctuations in the membrane potential).

Through the detailed synaptic descriptions, the LIF-NS model introduces a nonlinear integration of the afferent impulses. This synaptic filtering gives rise to new dynamical response properties of the LIF neurons, achieving temporal dynamics similar to that of more complex neuron models (Softky & Koch, 1993). This model was also shown to allow a good representation of high frequency neuronal signals (Brunel et al., 2001; Fourcaud & Brunel, 2002).

Assuming to capture a proper level of description of neuronal activity, the LIF-NS neurons represent the building blocks of the biologically-inspired recurrent spiking networks presented in the next chapter.

3 Computational model of cortical networks

In order to properly describe the dynamic aspects of the neural cognitive processes, a *coupled attractor network* view is taken into account: Different cortical modules with specific functionalities are modeled as *recurrent networks of spiking neurons*. Their local attractor dynamics are linked by inter-module connections, corresponding to the long-range axonal fibers in the cortex, in such a way to form a global coherent representation of information. This approach is assumed to model in a conceptual way the neo-cortical operation. The resulting powerful computational networks incorporating the LIF neuron models presented in the previous chapter, exhibit a rich spiking dynamics similar to that of the neurophysiological cortical data (Sima & Orponen, 2003). Consequently, the simulated dynamical processes, that putatively underlie the studied cognitive processes, can be quantitatively contrasted with the neurophysiological experimental measurements.

The first part of this chapter, section 3.1, describes possible neural encoding modalities. Section 3.2 introduces the general structure and properties of the adopted recurrent network model that follows a biologically inspired description. The structure of the recurrent networks follows the general framework of the *Biased-Competition and Cooperation* assuming that model units engage in competitive and cooperative interactions with each other in order to represent their input in a context-dependent way, as described in section 3.3.

3.1 Neural encoding strategies

The scientific community still debates the actual mode in which information is processed and stored throughout the brain, one fundamental unsolved issue in neuroscience being the neural encoding mechanism. As stated in section 2.1.2, it seems reasonable to assume that information is transmitted from one neuron to the other in the form of identical discrete impulses emitted at irregular time intervals (the spike train), but how exactly the information is encoded in these spike trains is still not understood.

Theories, at the basis of ***neural temporal codes***, assume that the precise timing of impulses plays an important role and that the output spike reflects the exact coincidence of several input spikes (neurons are described as coincidence detectors) (Rieke et al., 1997). These theories are supported by different experiments – showing that individual in-vitro neurons produce the same spike train when injected with the same noisy looking input; finding repeating spatio-temporal patterns in a neuron's spike train; or revealing precise temporal correlations of different neurons firing patterns (Abeles, 1994). Experiments on flies (Bialek et al., 1991) show that time dependent stimuli are encoded in the firing times of visual neurons. Temporal codes, where each single spike is thought to carry reliable and precisely timed information, require a detailed description of the neuronal dynamics. They were able to explain the fast transmission of sensory or motor information in small neural systems with few synaptic connections.

In-vivo recordings show that, as opposed to sensory and motor neurons, cortical neurons have a very irregular behavior generating highly stochastic spike trains for both spontaneous low activity and stimulus evoked high activity, that is close to a Poisson point process (Softky & Koch, 1993). The inter-spike interval (ISI) distributions are roughly exponential, with a coefficient of variation close to 1. It is not known if this variability is just noise or is actually given by an efficient encoding technique.

Given the noisy cortical environment, the theories at the basis of ***neural rate codes*** assume that some sort of average measure of the spike trains is the modality by which neurons communicate and that the actual timing of individual spikes is stochastic and gives little information (Rolls et al., 2004; Treves et al., 1999). Cortical neurons are reported to fire irregularly in-vivo although able of regularly firing in-vitro. This suggests that the irregularity comes from synaptic sources interacting with deterministic membrane properties rather than from any intrinsically stochastic spiking mechanism. The spike trains from the same cortical neuron but different trials of the same experimental condition express also high variability, which implies that the cortical processes might not be deterministic. It is assumed that the combined large number of inputs to a cortical neuron results in a fluctuating change of its membrane potential that will stochastically cross the spiking threshold and thus generate an irregular behavior. Rate models are better suited for cortical circuits where the impact of only one spike is relatively small – given that cortical neurons have thousands pre-synaptic connections, many of which coming from neighboring cells with similar receptive fields.

A first encoding modality, ***neuron rate coding***, considers that the neuronal response can be completely described by the average firing rate of single cells, calculated as the mean number of spikes emitted over a small time interval. An event-related response of a single neuron can be expressed by the time average of the traces collected by recording it for many times under the same conditions of stimulation. Many studies showed simple stimuli features (like intensity) to be correlated with the average spiking rate and not with the exact spiking time of the measured sensory neurons (Hubel & Wiesel, 1959). The intuitive assumption that neurons use a rate code for encoding the corresponding information, is subject to an ongoing debate (Rieke et al., 1997). The problem with this simple code is that it can only express constant or slow varying stimuli. Given a neuron's typical time constant on the order of 15-20 ms and the multi-layered structure of the brain, this type of rate coding cannot explain the short reaction times on the order of only $30 - 40$ ms measured in behavioral experiments on animals, and the fast visual recognition in humans on the order of a few hundreds of milliseconds (Thorpe et al., 1996). This problem can be, however, overcome by using population rate coding (see below).

An important characteristic of the neural code is its efficient representation of the sensory world. The important requirement of the neural communication mechanism – to allow fast reactions to fast changes in the input – can be met by another encoding modality that takes into account the structural features of cortical circuits. These consist of a large number of neurons organized into functional groups of neurons with similar properties, like the columns in visual cortex (Hubel & Wiesel, 1962), populations of motor neurons (Kandel et al., 2000) or the modules observed in ITC or PFC (Miyashita & Chang, 1988; Wilson et al., 1993). This leads to the assumption that information could be extracted from the average activity of such assemblies of neurons, and the new measuring techniques using large arrays of electrodes and amplifiers that make possible the simultaneous recording in-vivo of many neighboring neurons, facilitate their study.

Donald Hebb introduced in 1949 the concept of neural assembly as a modality of information representation in the cortical structures, Hebb (1949) The average activity of the population of neurons removes the details of individual spike trains, but emphasizes the common trend of the neuronal assembly. ***Population rate coding*** is sustained by many studies showing that the temporal variations of the average firing rate of a large population of neurons can accurately express the fast temporal variations of the inputs on a time scale much smaller than the integration time constant of single neurons (see for example Gerstner, 1995; Amit & Brunel, 1997a; Fusi & Mattia, 1999). Population rate coding represents a powerful tool for investigating the function of large neural systems and will be used in this work to study the neural mechanisms underlying high-level cognitive brain functions.

3.2 Recurrent network of spiking neurons

The correct operation of the nervous system relies on the coherent flow of information through elaborate neural circuits whose structure and properties are important elements for understanding the cortical mechanisms underlying brain function. At the macroscopic level, the neocortex is hierarchically organized into spatially segregated regions with different functional roles and is characterized by a distinct layered structure. For example, the occipital cortex is specialized for vision and its different macroscopic subregions represent different stages in the processing of visual information. Motivated by the modular architecture of the cortex, it is assumed that small cortical areas with specific functionalities can be treated as recurrent networks, which are interconnected to form complex computational systems able to solve elaborate tasks.

In this work, the model cortical areas are constructed as fully connected recurrent networks of spiking neurons that incorporate biologically inspired cortical features (as earlier introduced in Amit & Brunel, 1997b), which are coupled to each other by excitatory long-range connections. Each module contains some 10^3 excitatory and inhibitory integrate-and-fire neurons that fire spike trains with Poissonian statistics. A recurrent network consisting only of excitatory cells, was shown to be very unstable to small fluctuations in the mean afferent input and thus unable to reproduce the spontaneous firing rates of biological cortical circuits (Amit & Brunel, 1997b). By introducing inhibitory cells, the mean of the stochastic afferent current to a cell is reduced while its standard deviation is increased, and the network becomes able to develop stable attractors at low spontaneous rates. The structure and parameters of the model recurrent network are chosen in agreement to important neuroanatomical and neurophysiological cortical findings that will be specified next.

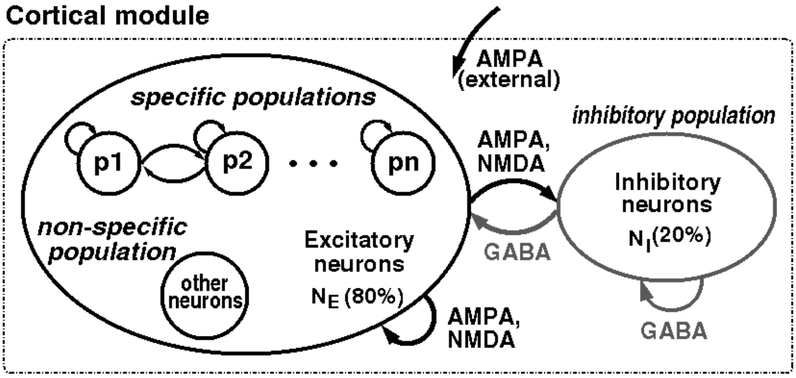


Figure 3.1: Architecture of a cortical module consisting of N_E excitatory neurons, organized into n specific populations and one non-specific population, and N_I inhibitory neurons that form the inhibitory population. Recurrent EPSCs are mediated by AMPA and NMDA receptors and IPSCs by GABA receptors. External EPSCs coming from other cortical modules are assumed to be mediated only by AMPA receptors.

The cortical areas contain a large number of interconnected cells. About 80 % of cells are excitatory pyramidal neurons which communicate via the glutamate neurotransmitter. About 20 % of cells are inhibitory interneurons which communicate via the GABA neurotransmitter (Abeles, 1991). A dominant cortical feature is the dense recurrent intra-areal and inter-areal connectivity, cortical neurons receiving on the order of 10^3 afferent synaptic connections. Typical excitatory pyramidal cells have both short-range collateral connections and long-range connections reaching other remote cortical areas. Inhibitory interneurons are believed to have only short-range connections which keeps their direct influence local. The external long-range connections are assumed to be only of excitatory nature and make up half of the total excitatory inputs to a cell (Braitenberg & Schütz, 1991).

The proposed recurrent network model for a cortical module consists of N_E excitatory and N_I inhibitory model neurons, chosen in the proportion 80% and 20%, respectively, to be consistent with the neurophysiological experimental data (Abeles, 1991). The single neuron dynamics is described by the LIF-NS model capturing detailed synaptic current dynamics, that was presented in section 2.2.4. The recurrent excitatory postsynaptic currents (EPSCs) are modeled to have two components, mediated by AMPA (fast) and NMDA (slow) receptors. The inhibitory postsynaptic currents (IPSCs) into both excitatory and inhibitory cells are mediated by $GABA_A$ receptors. Aside the recurrent excitatory and inhibitory inputs, the model considers that all cells receive, on average, the same number of external excitatory inputs, N_{ext}, originating from outside the module. N_{ext} is chosen to make up half of the total excitatory input to a cell, i.e. $N_{ext} = N_E$ for a fully connected network. The external EPSCs are assumed to be driven only by AMPA receptors. The detailed description level of the LIF-NS synaptic currents allows the use of realistic biophysical time constants, latencies and conductances, which permits realistic time scales and spiking dynamics of the simulated neural activity, that can be afterwards quantitatively contrasted with experimental neurophysiological cortical data.

Motivated by the concept of population coding, assuming that cortical neurons can be organized into groups with similar properties, the model recurrent network is structured into a discrete non-overlapping set of populations defined as groups of excitatory or inhibitory spiking neurons sharing the same inputs and connectivities. Previous studies showed that groups of co-activated neurons form local attractors of the recurrent neural dynamics (Hopfield, 1982; Amit et al., 1994; Amit & Brunel, 1997b). Attractor dynamics for biologically inspired networks of spiking neurons were recently investigated by Amit & Brunel (1997a); Brunel & Wang (2001); Deco & Rolls (2003). The model defines three general types of populations: the *specific populations* gather excitatory neurons selectively encoding information associated to a specific behavioral function; the *non-specific population* groups all other excitatory neurons present in the modeled cortical area that are not involved in the specified behavioral function; and the *inhibitory population* groups all local inhibitory neurons present in the modeled brain area, which helps to regulate the overall activity in the network. The architecture of the proposed cortical module is depicted in Figure 3.1.

A characteristic of the cortical system is its continuous activity. About 99% of the cortical neurons are spontaneously active at a low rate of about 3 Hz. The rest 1% of neurons are active with higher than spontaneous rates, typically some tens of Hz (Wilson et al., 1994; Koch & Fuster, 1989). This characteristic and the dense cortical connectivity suggest that each neuron is mostly driven by a strong background current consisting of spontaneous or unrelated activity from the same or other cortical areas. Few external connections carry specific or related inputs. These inputs can be seen as small perturbations, on the order of a few percents, on top of the background current. It is assumed that the recurrent neural circuits amplify these small inputs in a useful way to achieve a coherent processing of the behavioral relevant information. Another important feature is the stochastic character of the cortical spike trains (as discussed in the previous section).

For the proposed model, all neurons receive an external stochastic background input assumed to originate in other, not explicitly modeled, areas. The background input is modeled as a Poisson process with constant rate. It is assumed that all N_{ext} external connections are activated by independent Poisson processes with a mean rate of 3 Hz (the typical value observed in the cerebral cortex for the spontaneous activity rate of excitatory pyramidal cells, Wilson et al., 1994; Koch & Fuster, 1989). Signals conveying task-related information, like stimulus presentation, attentional state or context knowledge, are also assumed to originate in other cortical areas not explicitly modeled. These relevant inputs are modeled as small, additive excitatory inputs on top of the background input to the specific neuron populations in the model, and are implemented by small increases in the rates of the corresponding Poisson processes.

In the presence of fluctuations, the attractor dynamics of the recurrent network is very unstable, and can respond very differently to small changes in the inputs. The attractor landscape, characterizing the function carried out by the cortical model, is determined by the structured connection strengths of the recurrent network. The connection strengths, or weights, describe the efficacy of the synaptic couplings in the network. The choice for these parameters will be discussed in the next section that introduces the unifying principle of Biased-Competition and Cooperation for neocortical modeling of higher-level cognitive brain functions.

3.3 Architectural *Biased-Competition and Cooperation* framework

An important aspect regarding the information processing and encoding mechanisms of the brain is how the information is represented across the cortical areas. Recent findings lead to the hypothesis that the representation of specific information is distributed across several cortical areas and that one cortical area can hold partial representations of many different views. It is assumed that these *partial representations* represent an incomplete description of the behavioral input and / or internal cortical state, like for example object features, object identities, spatial relationships, behavioral rules or associations. Moreover it is assumed that, a set of different views representing different conflicting or related partial information could be encoded in the neural activities of a particular cortical area.

In order to achieve a coherent global representation, the partial representations are integrated by mutual cross-talk through inter-areal neural connections. Between the cortical areas, *feed-forward connectivity* is usually complemented by *feedback connectivity* between the same neural assemblies, i.e. the neurons feeding back from the higher-stage processing area preferentially address the same neurons in the lower-stage processing area which drive them. The feed-forward input from a lower-stage processing area, which is characterized by less abstract representations, is referred to as ***bottom-up*** **driving input**. The feedback input from a higher-stage processing area, which is characterized by more abstract representations, is referred to as ***top-down*** **biasing input**. Whereas bottom-up driving input is thought to activate a set of concepts consistent with the lower level (e.g., sensory) features, the top-down biasing input is thought to back-propagate higher level (i.e. more abstract) information to the lower-stage processing area, that helps selecting one activation pattern among several possible ones.

The **Biased-Competition and Cooperation** framework assumes that in a cortical area the *conflicting partial representations* ***compete*** with each other in order to be represented, while the *related partial representations* ***cooperate*** with each other, mutually reinforcing their activities.

By *Biased-Competition*, the competitive intra-areal dynamics is assumed to be resolved by a top-down bias coming from a higher-stage processing area that favors a certain representation over the others. In the context of visual attention, for example, a top-down signal encoding the attentional state, could bias the competition in a visual cortical area such that, when multiple stimuli appear in the visual field, only the attended stimulus will end up being represented, thereby suppressing the representation of all other distracting stimuli (Duncan & Humphreys, 1989; Desimone & Duncan, 1995; Duncan, 1996). Neurodynamical models developed within the conceptual framework of the *Biased-Competition Hypothesis* (Moran & Desimone, 1985; Chelazzi et al., 1993; Desimone & Duncan, 1995; Chelazzi, 1999; Reynolds & Desimone, 1999) have been proven to successfully account for different aspects of visual attention (Rolls & Deco, 2002; Corchs et al., 2003) and working memory context dependent tasks (Deco & Rolls, 2003; Deco et al., 2004).

In parallel to the competition view, a *cooperation* view has been formulated, where neural correlates are represented by different co-activated assemblies of neurons (Hebb, 1949). The theoretical framework of *Biased-Competition and Cooperation* for modeling higher-level cognitive

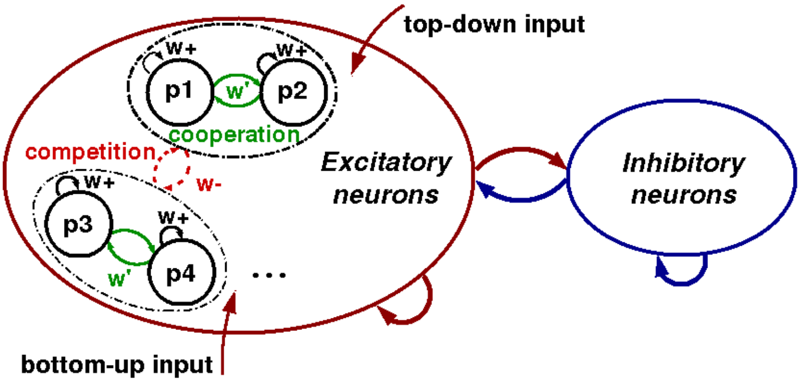


Figure 3.2: Single-layer network architecture using 'Biased Competition and Cooperation' framework. Within the specific populations, the neurons are mutually connected by stronger than average mean synaptic strength $w+$. Cooperation is implemented through stronger than average excitatory lateral interactions ($w_0 \leq w' \leq w+$) between populations representing related information. Competition is implemented through weaker than average excitatory lateral interactions ($w- \leq w_0$) and the global inhibitory signal.

brain functions was introduced by Rolls & Deco (2002); Szabo et al. (2004). This framework has been used to describe different types of experimental data, like single neuronal responses, fMRI activation patterns, psychophysical measurements, effects of pharmacological agents and effects of local cortical lesions (Rolls & Deco, 2002; Deco & Rolls, 2002; Deco et al., 2004; Szabo et al., 2004, 2006).

A multi-areal biased-competition and cooperation model is constructed as a multi-layered recurrent network in which populations of neurons interact with each other in a hierarchical way. By *biased-competition* it is assumed that in a model cortical-area multiple activated populations of neurons encoding conflicting information engage in competitive interactions. The internal competition is assumed to be *biased* by external top-down signals representing attention or behavior in favor of specific groups of neurons encoding the relevant information for behavior (e.g., the attended stimulus). The competition between different populations of neurons selectively highlights the neural responses of the attended or task-relevant stimuli features at the expense of the responses of other non-attended or not relevant features. Cooperation, on the other hand, promotes the co-activation of neuronal populations that represent stimuli features associated with each other and accounts for the correlations in the neural responses (Szabo et al., 2004; Almeida et al., 2004). Using both biased-competition and cooperation, the cortical areas could mutually guide each other's internal dynamics until a maximally coherent state is reached, in which each area's represented partial information is consistent with the represented partial information of the other areas. The result is a coherent global representation of information in the neocortical system (Szabo et al., 2006).

For the adopted multi-layer recurrent network model, biased-competition and cooperation is implemented by structured excitatory connection strengths within and between the model layers and is mediated by the mutual inhibitory signals produced inside each layer. In each layer, all excitatory neurons drive a single population of inhibitory neurons. The inhibitory population regulates the overall activity in that layer by spreading back a common inhibitory signal to all its neurons.

The connection strengths are modulated, starting from an average value w_0 over the entire network, such as to intuitively match a Hebbian learning paradigm. First, it is reasonable to assume that the neurons within a specific population are strongly co-activated and hence the synaptic strength of the connections between them will become stronger than average. Their mean strength is denoted here by $w+$. These strong synaptic couplings implement neuronal reverberation in the population and can, in certain conditions, underlie the formation of working memory[1]. The specific populations representing related information in the context of a certain behavioral task are likely to have correlated activities and therefore are assumed to be linked through stronger-than-average synaptic weights, denoted here by w'. This weight setting implements *cooperation* and underlies the formation of co-activated cell assemblies in the model. The specific populations representing unrelated information are likely to have uncorrelated activities and thus are assumed to be connected through weaker than average synaptic weights, denoted here by $w-$. This weight setting implements *competition* and underlies the formation of competitive dynamics inside the model layer. The specific populations influence each other's activities mainly through the inhibitory neurons, implementing a mechanism of global competition.

Different weight settings correspond to different functionalities of the model network. In order to achieve a certain, desired, operation of the cortical model, the network has to be trained using an appropriate learning algorithm. A biologically-inspired learning mechanism, compatible with the type of network model used and the type of experimental behavioral tasks modeled, will be covered in the next chapter.

[1] working memory can be seen as sustained high cortical activity after the corresponding stimulus was removed

4 Learning mechanism for biologically inspired spiking networks

The most remarkable feature of our nervous system is the ability to modify its internal parameters and even its structure in response to a changing environment, in other words to learn. Learning can be defined as the capability of a system to acquire information and skills in an adaptive and goal-directed way.

The first part of this chapter, section 4.1, captures the important aspects of learning in biological and artificial networks. A biologically inspired learning algorithm is proposed in section 4.2, and its implementation is described in section 4.3.

4.1 Learning in biological and artificial networks

How learning occurs in biological systems is still not exactly understood, but it is clear that they are able to adaptively acquire behaviorally relevant information and relate it to other already stored information. In the vertebrate brain learning is usually associated with the modification of the synaptic efficacies, i.e. chemical changes at the synaptic level, but it is unclear what neural mechanism optimizes these changes in the direction of obtaining the desired associations. These changes can last from a few seconds, being related to *short time memory* or up to many days, being related to *long time memory*. This type of learning is referred to as ***synaptic plasticity***. Another type of learning, still debated in the scientific community, is thought to involve structural changes in the brain. This type of long time learning, referred to as ***structural plasticity***, is associated to the apparition of new synaptic connections and the disappearance of other existing synaptic connections. In this work, only learning in the form of synaptic strength modifications will be regarded.

An important contribution to the study of learning mechanisms was done by Donald O. Hebb (Hebb, 1949). Inspired from the physiology of the nervous system, he developed a theory of how learning could occur in a biological system. Hebb's postulate holds that learning occurs as modifications of the synaptic efficacies between neurons and that these modifications are driven by correlations in the neuronal activity, i.e. synapses are strengthened between neurons that fire at the same time. Thus ***Hebbian learning*** refers to a *correlation based synaptic plasticity*: learning is driven by the multiplicative correlation of pre- and post- synaptic activity.

In the neuroscience research, a persistent increase in synaptic efficacy which follows high-frequency stimulation of afferent fibers is called *Long-Term Potentiation* (LTP) and a decrease in the synaptic efficacy is called *Long-Term Depression* (LTD). Experimentally observed LTP often occurs when there is a coincidence of pre- and post- synaptic activity and it is thus explained through the putative mechanism of Hebbian learning. Experimentally observed LTD can occur in two cases: when there is presynaptic activity in the absence of postsynaptic activity (Homosynaptic LTD) or when there is postsynaptic activity in the absence of presynaptic activity (Heterosynaptic LTD). LTP and LTD are considered to be the neuronal correlates of *learning* and *memory* in the biological systems.

In artificial neural networks, learning or training is a dynamical process that modifies the connection weights between the individual elements according to a learning algorithm or rule. Learning is seen as an optimization process. An artificial neural network usually consists of three layers: an input layer, a hidden layer and an output layer, and the functionality of the network is given by the weights connecting the hidden units. Three major learning strategies were developed:

- **Unsupervised learning** schemes are used to identify patterns (statistical regularities) in data sets and perform data clustering based on the correlations in the inputs. For example the Kohonen self-organizing feature map is an autonomous, bottom-up and data-driven learning strategy. Unsupervised Hebbian learning rules modify the connection weights based on activities of the pre- and post-synaptic neurons, and are used for clustering the input data.

- **Supervised learning** schemes use a teacher specifying the desired output for each specific input vector. The corresponding learning algorithms for multilayer networks (for example the LMS algorithm and the error-backpropagation algorithm) are based on propagating error signals from the output layer to the input layer and modifying the weights in the direction of minimizing some error functions. As an example, the error-backpropagation algorithm computes the gradient of a cost function that quantifies the performance of the network with respect to a desired input-output relationship and with respect to the parameters that should be optimized, i.e. the connection weights. Although not realistic (there is not any known neurobiological equivalent for the supervisor), they support nonlinear input-output mappings which makes them effective in training artificial neural networks.

- **Reinforcement learning** schemes (for example the Monte Carlo prediction and control algorithm (Metropolis & Ulam, 1949; Fishman, 1995) and the temporal differences (TD) learning algorithm (Sutton, 1988)) use an external observer (critic) that evaluates qualitatively the network's performance (when the actual desired output is not given or known) and decides to punish or reward its behavior in order to optimize some reward or cost function. It is applied in the cases when learning takes place through trial-and-error interactions with a dynamical environment. ***Reward based*** (goal directed) learning algorithms are based on rewarding or punishing the system given its response to a certain stimulation. The system is trained using a global reinforcer that modifies the internal parameters in such a way as to maximize the expected cumulative reward. For more details see Sutton & Barto (1998).

4.2 Reward-based Hebbian learning mechanism

Learning in a biological system usually takes place by evaluating the response of the environment to the performed action and modifying accordingly the system's internal associations in order to follow a desired goal. Behavioral experiments on operant conditioning show that the probability of performing one action in relation to a given input (referring to voluntary actions) changes with the behavioral consequences of that action. Actions followed by a positive consequence, called *reward*, become more frequent, implying that the corresponding cortical stimulus-response associations are *strengthened* or reinforced. Similarly, actions followed by a negative consequence, or *punishment* (often regarded as the consequence of not receiving reward), become more infrequent thus the corresponding associations are supposed to be *weakened*. In other words, the biological system determines the behavioral relevance of different sensorial information through the consequences - receiving reward or not - of the selected action and tries to maximize the future reward by remodeling its internal associations.

To model this learning strategy, we construct a ***reward-based Hebbian learning*** algorithm that modifies the synaptic efficacies according to the resulting network activities and a reward signal, using a simple regulatory mechanism. This type of learning algorithm was shown to be compatible with a learning procedure that quantitatively reproduces the behavior of the recorded cortical activity of monkeys trained to learn visuo-motor associations in a continuously changing environment (Asaad et al., 1998; Fusi et al., 2007).

On the network level, assuming that learning takes place in an interactive way using a trial-and-error strategy, the *reward-based model of learning* modifies the future behavior based on the success or failure of previous trials such as to increase the expected reward. This type of reinforcement learning is the biologically plausible choice from the well established learning strategies in the artificial neural networks field (Sutton & Barto, 1998). This kind of interactive learning modifies the future behavior, based on the success or failure of previous trials, such as to increase the desired reward.

On the single synapse level, we consider a *biologically inspired Hebbian learning scheme*, as described in Figure 4.1. Following reward, a synapse is potentiated if the presynaptic and post-synaptic neurons are simultaneously active; depressed if the presynaptic neuron is active but the post-synaptic neuron is inactive; and not modified otherwise. Following non-reward, the synapse is depressed if both presynaptic and post-synaptic neurons are simultaneously active; and not modified otherwise.

A pure Hebbian learning scheme, acting locally at each synapse and for each afferent spike, would be very difficult to control. Thus, the learning dynamics is modeled in this work using a *synaptic mean-field approximation*, which captures, for computational convenience, the average synaptic dynamics between two given populations of neurons. Using this approximation, the average effects of a supposedly underlying single-synapse dynamics are described through a single variable characterizing the synaptic population dynamics. A synaptic population gathers all connecting synapses between two neural populations that originate in one of the two populations and end in the other, i.e. between each pair of neural populations there are well-defined two

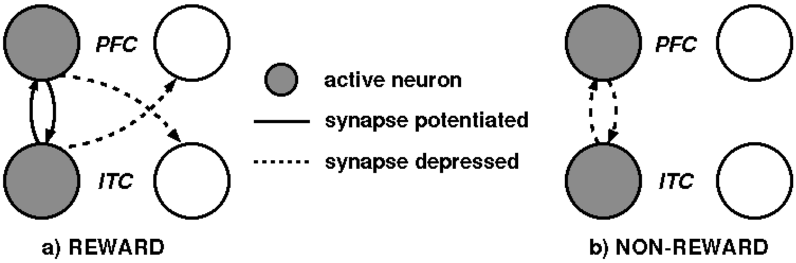


Figure 4.1: Modified Hebbian learning scheme applied for the reward (a) and non-reward (b) cases. The filled blobs represent *active* neurons and the empty blobs represent *inactive* neurons. The potentiated synapses are represented by bold lines and the depressed synapses by dotted lines. The connections that are not marked are not modified. In the reward case, a synapse is potentiated if the pre-synaptic and post-synaptic neurons are simultaneously active, depressed if the pre-synaptic neuron is active but the post-synaptic neuron is inactive. Feedback and feed-forward weights are not always changed in the same way: the learning rule is asymmetric in the reward case. In the non-reward case, the synapse is depressed if the pre-synaptic and post-synaptic neurons are simultaneously active.

synaptic populations.

For the hidden single-synapse dynamics, the individual synapses are considered binary (one *potentiated* and one *depressed* state), as suggested by experimental data from Petersen et al. (1998). Learning dynamics in the case of binary synapses are best described by a stochastic process for which the probability of transition depends on the pre- and post- synaptic activities. Models of learning dynamics using synapses that undergo stochastic transitions between two available states depending on the pre- and post-synaptic neural activity were developed in earlier studies (Fusi et al., 2000; Fusi, 2002; Amit & Fusi, 1994).

At the mean-field level adopted here, all the connections belonging to the same pair of pre- and post-synaptic populations are forced to have the same weight; the latter is updated by first computing the fraction of synapses that would get potentiated or depressed on the basis of the estimated firing rate distributions, as detailed in the next section. Then, the resulting average weight is computed and, finally, that value is assigned as the common new weight for all the synapses of the population. In order for this to be a good description of what would result from the detailed single-synapse dynamics, the non-trivial distribution of firing rates inside each neural population plays an important role. This is why it is important to keep the detailed spiking dynamics of the neurons in the face of the more abstract description of the learning dynamics. The efficacies for the various synaptic populations are thus changed in a way consistent with what would result on average from the spike-driven stochastic changes of single synapses under the same firing conditions. Implemented for a fast approximate evaluation of the learning behavior, this approximate description of the learning dynamics was shown to be a good estimate of the probabilistic model's dynamics (for details see Amit & Fusi, 1994; Fusi, 2002).

4.3 Implementation of the learning algorithm

During learning, the network configuration changes with the modification of the synaptic weights from an initial, chosen, configuration to a final, learned, configuration. For each trial, the network is stimulated with a randomly chosen input pattern and its simulated activity is recorded for a determined time period characterizing its response. The global reward signal estimating the overall performance of the network is not explicitly modeled. Its value is calculated based on the network response. Using the reward variable and the neuronal activations, the network configuration is changed using the following scheme:

Consider a presynaptic neuronal population with a total number of n_i neurons from which n_i^a neurons are considered active and a post-synaptic neuronal population with a total number of n_j neurons from which n_j^a neurons are considered active. In case of reward, the learning dynamics has to reflect for each synaptic population, that all synapses between pairs of active neurons are potentiated and all synapses from an active neuron to an inactive neuron are depressed (as described in Figure 4.1.a). This can be expressed through the fraction of synapses *to be* potentiated, N^p, and depressed, N^d, for each synaptic population:

$$N_{ij}^p = n_i^a \cdot n_j^a / (n_i \cdot n_j) \tag{4.1}$$
$$N_{ij}^d = n_i^a \cdot (n_j - n_j^a) / (n_i \cdot n_j) \tag{4.2}$$

In case of non-reward, the fraction of synapses to be depressed is given by the pairs of active neurons and there are no synapses to be potentiated (as described in Figure 4.1.b):

$$N_{ij}^p = 0 \tag{4.3}$$
$$N_{ij}^d = n_i^a \cdot n_j^a / (n_i \cdot n_j) \tag{4.4}$$

The N_{ij}^p and N_{ij}^d variables embody the firing distributions of the neuronal populations in the network, i.e. the variability in the neuron spiking rates inside the populations after each stimulus presentation, and thus require the full simulation of the neural dynamics. Consequently, the efficacies of the synaptic populations will be changed in a consistent way with what would result, on average, from the spike-driven stochastic changes of single synapses under the same firing conditions.

Considering the variable C_{ij} as being the current fraction of potentiated synapses from the synaptic population ij, its value is updated after each trial as follows:

$$C_{ij}(t+1) = C_{ij}(t) + (1 - C_{ij}(t)) N_{ij}^p q_+ - C_{ij}(t) N_{ij}^d q_- \tag{4.5}$$

where i and j generally denote the pre- and post-synaptic neuronal populations, respectively; q_+ and q_- are the transition probabilities for potentiation and depression, respectively (i.e. the learning rates); $(1 - C_{ij}(t))$ and $C_{ij}(t)$ are the fractions of depressed and potentiated synapses, respectively; and t is the trial number. The same equation (4.5) applies both in the reward and non-reward case but different learning rates can be used.

Accordingly, the average synaptic weight between each pair (ij) of neuronal populations is determined by the following relation:

$$w_{ij} = w_+ C_{ij} + w_-(1 - C_{ij}), \tag{4.6}$$

where w_+ and w_- are the values corresponding to the connection strength between two populations when all synapses are potentiated or depressed, respectively. Different values of w_+ and w_- can be used for the feed-forward and feedback connections in the network.

It should be remarked that the wide firing rate distributions of the neuronal populations in the network can provoke unwanted drifts in the learning history of some of the synaptic populations. Several regulatory mechanisms might in principle help to keep under control the effects of fluctuations in the synaptic dynamics (Miller, 1994; Stetter et al., 1994, 1998). The solution adopted here, is to keep the average synaptic efficacy in the network constant.

For this, a subtractive normalization of the total afferent synaptic connectivity is applied, which is calculated over all synaptic connections to each neuron in the network (Miller, 1994). The average synaptic weight for all connections between the pre-synaptic population i and post-synaptic population j is normalized as follows:

$$w_{ij}^{norm}(t) = w_{ij}(t) - \frac{1}{N}\left(\sum_{k=1}^{N} w_{kj}(t) - \sum_{k=1}^{N} w_{kj}(t-1)\right), \tag{4.7}$$

where N is the number of pre-synaptic populations connected to the post-synaptic population j.

The values for the C_{ij} variables are then recomputed based on the new w_{ij} normalized values in order to keep valid the equality in equation 7. For the next stimulus presentation, the connection weights between the neuronal populations are set to the calculated normalized values w_{ij}. The algorithm is repeated until convergence to a stable network configuration is reached. This configuration associates different neuronal assemblies in the model network such as to maximize the expected reward for the specified input-response relationship.

5 Network analysis using the Mean-field approximation

The recurrent network of integrate-and-fire neurons is a system with many non-linear interacting units whose dynamics is difficult to describe analytically. The standard trick, originating from statistical mechanics and referred to as *the Mean-field approximation*, is to replace all individual interactions from any one unit to any other unit with an average or effective interaction. The analysis is thus simplified to a system of effective units, each of them being influenced by a mean-field interaction. Mean-field models represent a well-established means for efficiently analyzing the approximate network behavior (Tuckwell, 1988; Stetter, 2002), at least for the stationary conditions (i.e. after the dynamical transients).

The Mean-field analysis of the recurrent network of integrate-and-fire neurons ignores the dynamics of individual neurons and calculates in an efficient manner, for the stationary regimes, the mean firing rates of the populations inside the network. This approximation allows an exhaustive analysis of the network dynamics: using different initial conditions, the parameter space of the recurrent network can be systematically explored in order to identify its qualitatively different functional regimes. Subsequently, bifurcation diagrams showing the possible dynamical states of the system as a function of the model-parameters can be constructed. For the parameter region showing the desired behavior, simulations of the full non-stationary spiking dynamics can be performed using the set of coupled differential equations describing the explicit neuronal dynamics of the recurrent network.

A recent derivation of the Mean-field approximation, consistent with the model neurons and network structure used in this work, will be considered for the systematic parameter explorations of the proposed models. This chapter is organized as follows: section 5.1 describes the assumptions and approximations used to predict the average firing rate of a LIF-NS neuron characterized by stochastic inputs. The derivation of the average firing rate of a population of identical and asynchronous firing LIF-NS neurons is presented in section 5.2.

5.1 Approximation of neuronal dynamics

The state of the LIF-NS neuron is fully characterized by the membrane potential dynamics, which is given by the coupled differential equations describing the integration (2.3) and the synaptic currents (2.7, 2.8 and 2.9) (see chapter 2.2). Fixed, regular inputs generate a deterministic behavior of neuronal activity. Stochastic spike arrival generates fluctuations in the membrane potential and thus a non-deterministic behavior of neuronal activity.

In the presence of noise, the exact value of the subthreshold membrane potential and thus the next firing time of a neuron cannot be predicted in a deterministic fashion. The state of the neuron can only be described statistically by estimating the probability that its membrane potential will be in a certain interval after a certain time. A large noise level will lead to a broad distribution of the membrane potential.

Starting from the membrane potential dynamics of the spiking neuron, the standard Mean-field approach performs a stochastic analysis of the mean-first passage time of the membrane potential above the spiking threshold, i.e. computes the average firing rate as a function of the input statistics and model parameters. The approximation provides a simplified description of neuron stochastic dynamics for stationary regimes and represents a well-established way of efficiently analyzing the approximate network behavior (Wilson & Cowan, 1972, 1973; Ben-Yishai et al., 1995; Gerstner, 1995, 2000; Eggert & van Hemmen, 2000; Stetter, 2002). An extended Mean-field formulation consistent with the LIF-NS model neurons considering both the fast and slow glutamatergic excitatory synaptic dynamics (AMPA and NMDA) and the GABAergic inhibitory synaptic dynamics, was derived by Brunel & Wang (2001). The main assumptions and estimations will be captured in the remaining part of this section:

First, it is assumed that the model neurons are characterized by *stochastic spike arrival*, i.e. that the individual spike trains arriving at different synaptic sites of the neuron are independent and that for each individual synapse each spike arrives independently of the previous one (the Markov assumption). The assumption of stochastic spike arrival is sustained by in-vivo recordings of cortical neuronal activity that show a high irregularity in the afferent spike trains (as mentioned in section 3.2). The individual spike trains are described in the model as independent Poisson processes with mean activation rate ν. A Poisson process with rate ν can be characterized by its mean $\mu = \nu$ and variance $\sigma^2 = \nu$. The coupled differential equations describing a neuron's state contain thus random terms which make it difficult to calculate the subthreshold membrane potential evolution and to find the neuron firing times in a deterministic fashion.

Second, it is assumed that the model neurons receive a large number of uncorrelated inputs (spikes) in small time intervals compared to their integration time constant (τ_m) and that individual spikes generate on average very small post-synaptic potentials as compared to the spiking threshold. Hence, the total afferent stochastic current to each neuron can be approximated by a Gaussian process: *the Diffusion approximation* (Tuckwell, 1988). These assumptions are sustained by experimental cortical findings showing dense recurrent connectivity and continuous and highly stochastic cortical activity. Even if all presynaptic neurons fire at low spontaneous rates, the high number of pre-synaptic contacts to a neuron, on the order of 10^3, produce in total a high presynaptic activity. This implies also that each presynaptic spike induces a very small change in neuron's membrane potential.

In the framework of the diffusion approximation, the equations of the subthreshold membrane potential dynamics can be simplified by replacing the pre-synaptic inputs through an average DC component and a random component. The latter can be treated as Gaussian white noise (in the case of instantaneous synapses) or Gaussian colored noise (denoting a finite correlation time of the random components in the case of nonlinear synaptic current dynamics) (Tuckwell, 1988; Amit & Tsodyks, 1991a,b). Consequently, the dynamics of the subthreshold membrane potential follows a Brownian motion that can be described by an Ornstein-Uhlenbeck process in the form of a Langevin stochastic differential equation (Uhlenbeck & Ornstein, 1930).

For the LIF-NS neuron dynamics, the Langevin equation is given by (Brunel & Wang, 2001):

$$\tau_j \frac{dV_m(t)}{dt} = -(V_m(t) - V_L) + \mu_j + \sigma_j \sqrt{\tau_j}\eta(t) \tag{5.1}$$

where

- τ_j is the effective membrane time constant that considers the impact of the various modeled synaptic currents on the membrane conductance,
- μ_j is the mean value that the membrane potential would have in the absence of spiking and fluctuations,
- σ_j measures the magnitude of the fluctuations due to the variance of the stochastic synaptic inputs,
- η is a Gaussian stochastic process with exponentially decaying correlation function with time constant τ_f that is given by the fluctuating terms of the synaptic input.

For the considered model, the fluctuations of the recurrent synaptic inputs, i.e. from other neurons in the modeled area, can be neglected when compared to the fluctuations of the external AMPA inputs. Thus σ_j can be calculated considering only the external stochastic inputs and τ_f is considered to be equal to τ_{AMPA} (for details see Brunel & Wang, 2001). The expressions for these variables depend on the network structure. They were derived by Brunel & Wang (2001) and are given in the second part of this chapter (section 5.2).

Because of the random component of the synaptic current, the solution of the Langevin equation (i.e. $V_m(t)$, the time-course of membrane potential), can be given in the form of a probability density function $p(V_m, t)$: the probability that the membrane potential has the value $V_m(t)$ at time t, starting from V_0 at time t_0. Thus, the probability that the membrane potential at time t will be in the interval (V_1, V_2), given the initial condition $V_m(t_0) = V_0$, will then be:

$$Pr\left\{V_1 < V_m(t) < V_2 \quad | \quad V_m(t_0) = V_0\right\} = \int_{V_1}^{V_2} p(V_m, t)dV_m \tag{5.2}$$

The solution of the Langevin equation (5.1) can be estimated using an equivalent description of the diffusion of the membrane potential, given by the Fokker Planck equation (Risken, 1984):

$$\tau_j \frac{\partial p(V_m, t)}{\partial t} = -\frac{\partial}{\partial V_m}\left(-(V_m(t) - V_L) + \mu_j\right) p(V_m, t) + \frac{\sigma_j^2}{2\tau_j}\frac{\partial^2}{\partial V_m^2} p(V_m, t) \tag{5.3}$$

This equation describes, in the diffusion limit, the temporal evolution of the probability density function of the membrane potential $\frac{\partial p(V_m, t)}{\partial t}$. The first term represents the systematic drift of the membrane potential due to the leakage and mean afferent input. The second term represents the diffusion term, with diffusion constant $\frac{\sigma_j}{2\tau_j}$, and accounts for the fluctuations of the membrane potential. The solution of the Fokker Planck equation gives the probability density of the solution of the original Langevin equation, i.e. the probability density function of the membrane potential.

$$p(V_m, t) = \tau_j \frac{\sqrt{\pi}}{\sigma_j} \int_{V_m}^{\theta} du \exp\left(\frac{(u-\nu_j)^2}{\sigma_j^2}\right) \left[erf\left(\frac{u-\nu_j}{\sigma_j}\right) + 1\right] \tag{5.4}$$

With this, the mean first passage time of the membrane potential can be calculated as the probability of the membrane potential first crossing the threshold θ in the time interval $[t, t+dt]$ starting from V_0 at t_0:

$$P_{firing}(t) = \int_{\theta}^{\infty} p(V_m, t) dV_m \tag{5.5}$$

For the LIF-NS neuron model, an analytical solution can be calculated for the stationary regime, i.e. the regime where the mean input current is constant over time for any given neuron in the network, taking into account the normalization and boundary conditions that restrict the depolarization values to the interval $[V_{rest}, \theta]$ (Treves, 1993). The normalization condition states that the probability that the membrane potential has a value in this interval equals to 1. The first boundary condition describes the firing threshold as an absorbing barrier: $p(\theta, t) = 0$, $\forall t$ (all depolarizations crossing the threshold θ are absorbed and reset to V_{reset}). Another boundary condition describes the reflecting barrier at V_{rest} (no depolarization goes below this value).

With these conditions, the mean firing (emission or discharge) rate, ν_j, calculated as the inverse of the mean inter-spike interval (i.e. the mean time interval between two consecutive spikes) can be approximated by the solution of the mean first passage time of the membrane potential (Ricciardi, 1977; Tuckwell, 1988), and was derived for the considered neuron model in the limit $\tau_{AMPA} << \tau_m$ by Brunel & Sergi (1998); Brunel & Wang (2001):

$$\nu_j = \phi(\mu_j, \sigma_j) = \left(\tau_{ref} + \tau_j \sqrt{\pi} \int_{\beta(\mu_j,\sigma_j)}^{\alpha(\mu_j,\sigma_j)} du \varphi(u)\right)^{-1} \tag{5.6}$$

where

$$\varphi(u) = \exp(u^2)[1 + \text{erf}(u)] \tag{5.7}$$

$$\alpha(\mu_j, \sigma_j) = \frac{(V_{thr} - \mu_j)}{\sigma_j}\left(1 + 0.5\frac{\tau_{AMPA}}{\tau_j}\right) + 1.03\sqrt{\frac{\tau_{AMPA}}{\tau_j}} - 0.5\frac{\tau_{AMPA}}{\tau_j} \tag{5.8}$$

$$\beta(\mu_j, \sigma_j) = \frac{(V_{reset} - \mu_j)}{\sigma_j} \tag{5.9}$$

and erf() is the error function.

$\phi(\mu_j, \sigma_j)$ is the transfer function of the LIF-NS neuron, also known as the frequency - current or response function. It depends on the mean μ_j and standard deviation σ_j of the Gaussian afferent current in the stationary conditions.

5.2 Approximation of population dynamics

The recurrent network of integrate-and-fire neurons is divided into populations of neurons and it is assumed that inside a population i, the neurons are driven by stochastic recurrent and external currents with the same mean μ_i and variance σ_i. Sharing the same statistical properties of the total afferent current, the neurons inside a population fire independently at the same mean firing rate ν_i. The population activity can then be described by only one equation representing the average firing rate of the neurons inside that population, which equals, given the identical input statistics, the mean firing rate of the individual neurons. The population dynamics is described through the evolution of membrane potential densities, denoting the probability that an arbitrary neuron in the population has a specific internal state (at a given time, each neuron may be in a different internal state, given by its membrane potential). This is the same as the probability density function of the membrane potential of an individual neuron inside the population. Thus the dynamics of the population is reduced to the dynamics of a single reference neuron.

Mean-field analysis is a well known approach used to approximate the full spiking dynamics of a population of identical and asynchronously firing neurons (Abbott & van Vreeswijk, 1993; Brunel & Hakim, 1999; Fusi & Mattia, 1999; Nykamp & Tranchina, 2000; Omurtag et al., 2000; Brunel, 2000). For a similar spiking recurrent network as the one presented in chapter 3, the analytical formulas describing the firing rate of a population of LIF-NS neurons for stationary regimes were derived by Brunel & Wang (2001). The derivation assumes that the dynamics of the network will converge to a stationary attractor which is consistent with the asymptotic behavior of an asynchronous firing network of IF neurons (Brunel & Wang, 2001; Del Giudice et al., 2003; Fusi & Mattia, 1999).

The stationary dynamics of each population can be described by the *population transfer function* ϕ, which provides the average population rate as a function of the afferent current statistics and is given, as explained in the beginning of this section, by the single neuron transfer function (equation 5.6). The set of stationary, self-reproducing rates ν_i for all populations $i = 1 \ldots N$ in the network can be found by solving the set of coupled self-consistency relations, using standard numerical integration techniques:

$$\nu_i = \phi(\mu_i(\nu_1, ..., \nu_N), \sigma_i(\nu_1, ..., \nu_N)) \tag{5.10}$$

where $\mu_i()$ and $\sigma_i()$ are the mean and standard deviation of the afferent input to poplation i that depend on the mean firing rates of all other populations in the network. In order to find the stationary rates, a set of first-order differential equations, describing a *fake dynamics* (in contrast to the 'true' underlying spiking dynamics) of the system is used, whose fixed point solutions correspond to the solutions of equation 5.10:

$$\tau_i \frac{d\nu_i}{dt} = -\nu_i + \phi(\mu_i(\nu_1, ..., \nu_N), \sigma_i(\nu_1, ..., \nu_N)) \tag{5.11}$$

For a fully-connected recurrent network consisting of N_E excitatory LIF-NS neurons grouped into P populations and N_I inhibitory LIF-NS neurons, and considering that each neuron receives N_{ext} excitatory external connections, the corresponding parameters are (see Brunel & Wang (2001)):

$$\mu_i = \frac{(S_i^{AMPA,ext} + S_i^{AMPA} + \rho_1 S_i^{NMDA})V_E}{S_i} + \frac{\rho_2 S_i^{NMDA}\langle V\rangle + S_i^{GABA}V_I + V_L}{S_i} \tag{5.12}$$

$$\sigma_i^2 = \frac{g^2_{AMPA,ext}(\langle V\rangle - V_E)^2 N_{ext}\nu_{ext}\tau^2_{AMPA}\tau_i}{g_m^2\tau_m^2} \tag{5.13}$$

where:

$$\tau_i = \frac{C_m}{g_m S_i} \tag{5.14}$$

is the effective membrane time constant;

$$S_i = 1 + S_i^{AMPA,ext} + S_i^{AMPA} + (\rho_1 + \rho_2)S_i^{NMDA} + S_i^{GABA} \tag{5.15}$$

is a shunting factor;

$$S_i^{AMPA,ext} = \frac{g_{AMPA,ext}}{g_m} N_{ext}\tau_{AMPA}\nu_{ext} \tag{5.16}$$

$$S_i^{AMPA} = \frac{g_{AMPA,rec}}{g_m} N_E \tau_{AMPA}\sum_{p=1}^{P} f_p w_{p,i}\nu_p \tag{5.17}$$

$$S_i^{NMDA} = \frac{g_{NMDA}}{g_m} N_E \sum_{p=1}^{P} f_p w_{p,i}\psi(\nu_p) \tag{5.18}$$

$$S_i^{GABA} = \frac{g_{GABA}}{g_m} N_I \tau_{GABA} w_{I,i}\nu_I \tag{5.19}$$

are the average DC components of the synaptic variables. Here f_p is the fraction of the neurons in the p^{th} excitatory population; $w_{p,i}$ is the incoming connection weight from excitatory population p; ν_p is the discharge rate of the p^{th} excitatory population; $w_{I,i}$ is the incoming connection weight from the inhibitory population; ν_I is the discharge rate of the inhibitory population and ν_{ext} is the discharge rate of the external excitatory stimulation.

$$\rho_1 = \frac{1}{J} \tag{5.20}$$

$$\rho_2 = \beta\frac{(\langle V\rangle - V_E)(J-1)}{J^2} \tag{5.21}$$

$$J = 1 + \gamma\exp(-\beta\langle V\rangle) \tag{5.22}$$

are parameters derived from linearizing the voltage dependence of the NMDA conductance around the mean value of the voltage $\langle V\rangle$ (Brunel & Wang, 2001);

The average membrane potential $\langle V\rangle$ was calculated in Brunel & Hakim (1999):

$$\langle V\rangle = \mu_i - (V_{thr} - V_{reset})\nu_i\tau_i, \tag{5.23}$$

The static component of the gating variable of NMDA channels, in the case of Poissonian input spike trains is approximated by a function $\psi(\nu)$ depending on the presynaptic rates ν:

$$\psi(\nu) = \frac{\nu\tau_{NMDA}}{1+\nu\tau_{NMDA}}\left(1+\frac{1}{1+\nu\tau_{NMDA}}\sum_{n=1}^{\infty}\frac{(-\alpha\tau_{NMDA,rise})^{n}T_n(\nu)}{(n+1)!}\right) \tag{5.24}$$

$$T_n(\nu) = \sum_{k=0}^{n}(-1)^k\binom{n}{k}\left(\frac{\tau_{NMDA,rise}(1+\nu\tau_{NMDA})}{\tau_{NMDA,rise}(1+\nu\tau_{NMDA})+k\tau_{NMDA,decay}}\right) \tag{5.25}$$

$$\tau_{NMDA} = \alpha\tau_{NMDA,rise}\tau_{NMDA,decay} \tag{5.26}$$

For the models developed in this work, both mean-field simulations using the formulation presented in this chapter and full spiking-dynamics simulations using the formulations presented in sections 2.2.3 and 2.2.4 will be performed.

6 One-layer neurodynamical model for attentional filtering

This chapter analyzes the computational principles underlying the neuronal correlates of attentional filtering, showing how an input encoding the attentional state can bias the level of competition and cooperation in a single-area model in order to extract the relevant information for the behavioral task.

6.1 Selective attention and inattentional blindness

The overwhelming amount of sensory information coming in a wide variety from the environment at any moment to our nervous system can be dramatically reduced if only the relevant information is considered. In spite of its parallel processing capabilities and its huge amount of processing elements, the brain seems to employ a selection strategy even at early processing stages, for managing this enormous amount of sensory information. This selection process is referred to as *attention* and denotes the ability to concentrate on a particular thing while ignoring others. Attention represents an important cognitive process of perception which involves an organized processing of all sensory information in such a way as to produce a coherent experience of the surrounding environment. The concept of attention implies that only certain behaviorally relevant information from the sensory input or internal representations are processed at a given time. It is assumed that the focus of attention modulated by the behavioral context can be shifted from one sensory information to another or from one internal representation to another in a serial fashion.

Given the limited amount of information that can be processed in the brain at any time, attention represents an important basis of cognitive processing by selecting and filtering the information in a context-dependent way. The context is provided by the internal state of the brain, reflecting the subject's current hypotheses about its surrounding (external) environment. ***Selective*** or ***focused attention*** may be defined as the cognitive process by which the perception of certain relevant stimuli in the environment is favored in preference to other concurrent stimuli of less importance. The attention is said to be focused on selected parts of the environment. For example, certain parts of the visual input can be selectively attended to, depending on their relevance for the current task to be subserved or the current goal to be achieved. The neural process of *visual selective attention* enhances the signals representing visual information relevant for behavior, while suppressing the representation of the non-relevant visual information.

A remarkable phenomenon of selective attention, known as ***inattentional blindness***, has been described for human vision (for a review see Simons (2000)). It refers to the unawareness of a certain visual event when attention is focused on another event, and is thought to be part of an important cognitive mechanism, namely that of focusing or concentrating on a task to be performed. The neural mechanisms underlying selective attention are subject to ongoing debate. Hence, it is interesting to investigate which mechanism could produce the signals generating attention, how they can be flexibly controlled by the current brain state and what are their effects on tuning the neural activity.

Recently, a neurophysiological study performed by Everling, Tinsley, Gaffan & Duncan (2002), investigated the possible underlying mechanisms of visual selective attention by monitoring the activity level of single neurons in the prefrontal cortex (PFC) of awake behaving monkeys which were engaged in a focused attention task. In this experiment, a monkey, after being cued to attend one of the two visual hemifields (i.e. the left or the right visual eye-field), had to watch a series of bilateral stimuli that consisted of different pairs of objects, and to react with a saccade[1] if and only if a predefined target (previously learned object) appeared in the cued hemifield. In order to correctly perform this cognitive task, the monkey had to ignore any presented object in the uncued hemifield and to concentrate (focus his attention) on the cued location.

At first, using unilateral stimuli only, Everling and coworkers observed that some of the measured neurons from PFC were selective for target or non-target stimuli (stimuli requiring or not a behavioral response). These neurons were also found to have a preference for the stimulus location in one of the two hemifields (see figure 6.1). Next, during the focused attention task with bilateral stimulus presentation, the PFC neurons again discriminated between target and non-target stimuli, but only for the attended location, (figure 6.6, column 1). The target / non-target discrimination disappeared if the objects were presented in the unattended visual hemifield, i.e. the presented stimulus in the non-attended location had no influence on the neuronal response (cf. figure 6.6-1b red line). The experimental results showed that only a task-relevant stimulus (i.e., target in the cued hemifield) is gated by the context and is allowed to be represented. Thus, attention acts not only in a modulatory way but imposes a multiplicative effect upon the sensory driven neuronal response. Consequently these neurons seem to code for the behavioral relevance of a stimulus rather than for its identity. This effect is referred to as ***attentional filtering***.

When humans perform similar tasks, they cannot recover any information from unattended sensory stimuli – the inattentional blindness effect as mentioned before. The attentional filtering of object's representation for the unattended hemifield strongly resembles this behavior, possibly explaining the blindness to ignored inputs, implying that the reported properties of the neuronal response might be part of the neural correlate of cognitive inattentional blindness.

Motivated by these observations, we developed a minimal neurodynamical computational model of a small part of the monkey's PFC, which preserves the biological relevance and follows the Biased-Competition and Cooperation architectural framework, in order to investigate how this strong attentional effect can arise from a weak modulatory bias which mediates the cortical context (Szabo et al., 2004). This chapter is organized as follows: Section 6.2 describes the

[1] a saccade is a rapid intermittent eye movement occurring when eyes fix on one point after another

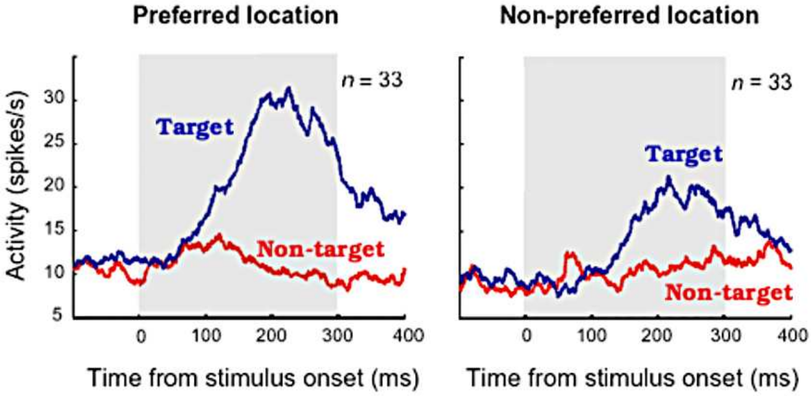


Figure 6.1: Experimental results of the focused attention task in the case of unilateral stimuli presentations – adapted from Everling et al. (2002). The two graphs depict the population activity of all selective PFC neurons when a target (blue lines) or a non-target (red lines) was presented in the preferred location (left graph) or non-preferred location (right graph).

chosen network structure and parameters. Section 6.3 presents results of exploring the parameter space using the mean-field approximation and section 6.4 presents explicit simulations of the network spiking dynamics that are compared to the experimental results. At last, section 6.5 discusses the obtained results.

6.2 Network structure and parameters

In the search to understand the underlying neural substrate of selective attention, a neurodynamical computational model of a small part of the monkey PFC is proposed. The model simulates the conditions of the visual attentional experiment performed by Everling et al. (2002) and tries to reproduce the attentional filtering effect observed in the experimental results. The model was initially constructed using the conceptual framework of *Biased-Competition*. We observed that the mechanism of biased-competition alone could not account for the experimental results and show that biased-competition and cooperation between stimulus selective neurons are, in combination, required conditions for reproducing the referred effect. The theoretical framework of *Biased-competition and cooperation* was described in section 3.3.

Similar to previous studies (Brunel & Wang, 2001; Deco & Rolls, 2003; Deco et al., 2004), a biologically inspired minimal model is set up as a single-layer recurrent network of spiking neurons, whose generic structure and parameters were described in section 3.2. The detailed level of description of the spiking dynamics and the biological inspired network parameters allow thorough studies in realistic time scales of the firing rates involved in the evolution of the modeled neural activity (as shown in section 6.4).

The network is constructed of $N_E = 800$ excitatory neurons and $N_I = 200$ inhibitory neurons,

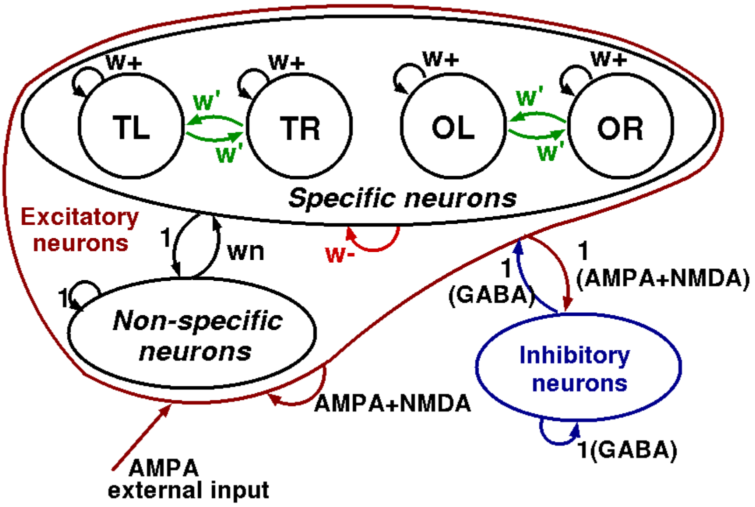


Figure 6.2: Schematic network architecture of the prefrontal cortical module. The four specific populations correspond to target and non-target selective neurons with preferred location left or right. The inhibitory population sends a global inhibitory signal to all neurons and the non-specific population stabilizes the activity inside the network. There are two types of synapses coming from the excitatory neurons: AMPA and NMDA, and one type from the interneurons: GABA.

which are grouped, by setting common inputs and connectivities, into distinct populations, as depicted in figure 6.2. The specific populations, are chosen to show the same selectivities as found in the experimental results (Everling et al., 2002). Accordingly, under a non-attentive control task they encode information about the object identity ('T' for target stimulus, 'O' for other non-target stimulus) and spatial location ('L' for left, 'R' for right visual hemifield). Thus there are four interconnected specific populations, that encode for target with preferred location left (TL), target with preferred location right (TR), non-target (other) left (OL) and non-target (other) right (OR). For simplicity, all specific populations were chosen to have the same number of excitatory neurons: $f \cdot N_E$, with $f = 0.1$. The non-specific population contains all other excitatory neurons in the area, $(1-4f)N_E$, which are not involved in the current task. The inhibitory population, grouping the N_I inhibitory neurons, sends a global inhibitory signal to all the neurons and thus balances the overall activity in the network.

The individual populations are driven by four different kinds of inputs. First, all neurons in the model receive *spontaneous background activity* from outside the module through $N_{ext} = 800$ external excitatory connections carrying Poisson spike trains with a rate of 3 Hz. Second, the neurons in the specific populations receive, in addition to the background noise, bottom-up sensory inputs and two kinds of top-down biasing inputs, as summarized in figure 6.3:

- the ***bottom-up sensory inputs*** selectively encode whether there was a target or a non-target in the left or the right visual hemifield and drive the corresponding specific populations. It is assumed that a lower-level visual cortical area processes the visual scene such as to provide these signals. The four possible combinations of sensory inputs are depicted

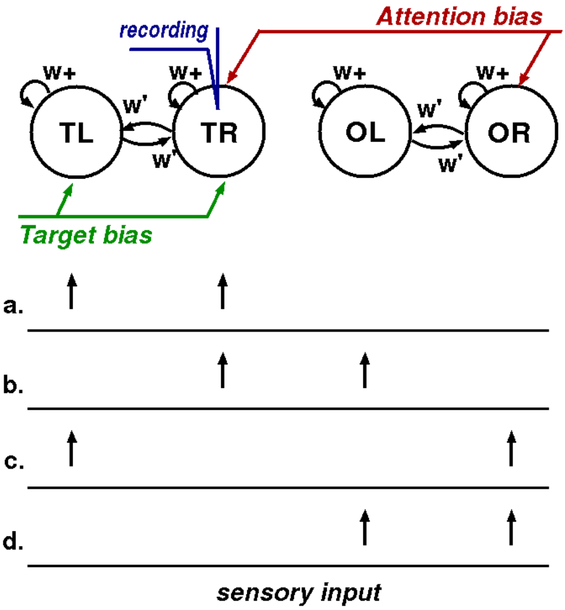


Figure 6.3: Simple scheme illustrating the inputs to the model network, given by the conditions of the visual attentional experiment. The arrows in the lower part of the figure represent the four possible sensory stimulus-combinations of the focused attention task to be modeled: (a) target in both locations; (b) target right and other object left; (c) target left and other object right; (d) non-targets in both locations. The model dynamics is further guided by a target bias and an attentional bias. The *recording* pin on the TR population (the group of target-specific neurons with preferred location right) specifies the neurons whose activity will be considered for the simulation results presented in section 6.4 (in the same way, the experimental results refer also only to the target selective PFC neurons).

in figure 6.3.

- the ***top-down target bias*** accentuates the representation of the relevant behavioral information, by biasing the neurons that are selective for the target object (i.e. from populations TL and TR). The origin of this signal is not explicitly modeled, but it is assumed to originate from a working-memory module that encodes and memorizes context in terms of rules.

- the ***top-down attention bias*** facilitates neurons which have as preferred location the currently cued location. Also the origin of this bias, which might be sent from a spatial working memory area, is not explicitly modeled here.

Both biases together can provide sufficient information about the task to be solved. They are found to guide the competition and cooperation processes. All sensory inputs and biases are set on top of the spontaneous background activity imposed onto the network from outside. The non-specific excitatory neurons do not have sensory or biasing inputs. They are thought to be involved in other cognitive tasks and to be only spontaneously and non-selectively active in the present model and under the present task.

It is considered that the model corresponds to the situation of an already trained monkey, thus the learning process itself is not explicitly modeled. Therefore, the connection weights between the neurons are chosen (not learned or optimized in any manner) such as to intuitively match a Hebbian learning paradigm. The model network is chosen to be fully connected, and for simplicity reasons the neurons between any two populations are connected with the same average synaptic weight. Within the same specific population, neurons are strongly co-activated and are therefore, in agreement with the Hebbian learning rule, connected with a stronger than average weight, $w_+ > 1$ (the average connection weight in the network is chosen $w_0 = 1$).

The activity of the neurons residing in populations selective for different objects is likely to be anticorrelated resulting in weaker than average connections between them, $w_- < 1$. These weak synaptic connections were found to mediate competition between the populations selective for different objects, and are referred to as ***competitive weight setting***. Further, it is not unreasonable to hypothesize that neurons with the same object selectivity are often co-activated and consequently linked strongly, by Hebbian learning mechanisms, yet weaker than neurons which share both object selectivity and location preference. Therefore, the weights between the populations that encode for the same object are set in the range $1 < w' \leq w_+$. This weight setting is found to mediate cooperation between the populations selective for the same object and is referred to as ***cooperative weight setting***.

Activities between specific and non-specific populations are likely to be close to uncorrelated. Therefore, the weights from specific to non-specific populations are set to the average value $w_0 = 1$ and the corresponding feedback connections to $w_n \leq 1$. Finally, all connections from and to the inhibitory population are set to the average value $w_0 = 1$. The resulting connectivity setting is illustrated in figure 6.2. In order to determine the system's performance, the next section investigates different operational modes of the model network by systematically exploring the absolute strengths of the connections between different populations.

6.3 Exploration of network connectivity

Explicit simulations of network dynamics accurately capture the full temporal evolution of network activity. However, they are computationally expensive and thus not efficient to use for systematic parameter explorations. For stationary conditions, the parameter space will be explored using the mean-field formulation consistent with the type of neurons and network structure used in this work, that was presented in chapter 5. The mean-field approximation formulates the average firing rates of the populations as a function of the model parameters. By performing a systematic exploration of the structured connection weights, different parameter regimes corresponding to qualitatively different modes of operation (responses) of the proposed recurrent network model can be investigated.

The values for the sensory inputs and biases have been chosen after exploring the effect of different signal strengths on the results. It was found that the network behavior does not considerably depend on their exact strengths, as long as they stay within certain values: First, the target bias should not be too strong, because otherwise it causes unattended target stimuli to win the competition, but it is needed to stabilize the behavior that target stimuli generally win over non-target stimuli. A suitable target-bias can be thought of breaking the symmetry between target and non-target but weakly affecting the network behavior otherwise. Second, the attentional bias should be not too weak compared to the input. The stronger the attentional bias is, the more complete the attentional filtering effect becomes. For the present network, the target bias should be in the range of 30 Hz or less, and the attentional bias in the range of 100 Hz or more, but it should be stressed that these numbers are only coarse estimates and because of their dependency on network size, structure and parameters cannot be taken as a quantitative predictions but only as qualitative trends.

Unless otherwise stated, the mean-field simulations were carried out as follows: The neuron average firing rates were initialized to 3 Hz for the excitatory populations and 9 Hz for the inhibitory population (typical values for the cortical spontaneous activity rates). The external spontaneous background input received by each neuron was set to $800 \cdot 3$ Hz $= 2.4$ kHz. On top of it, an attentional-bias of 100 Hz and a target-bias of 30 Hz were applied to the corresponding populations throughout the whole simulation. A sensory input was encoded as a small increase of 200 Hz in the external background input of the stimulated specific populations. For each parameter set consisting of different values for the parameters w_-, w_+, w' and w_n, the mean-field equations (see chapter 5) using the neuron model parameters from section 2.2.4 were integrated over 1000 iterations.

Searching to reproduce the attentional filtering effect captured by the neurophysiological experiment, we observed, as indicated by the comparison of the experimental results with explicit simulations of the network spiking dynamics, that the network requires the following combination of ***response properties***:

i). the network must show responses at all,

ii). the network should not show persistent post-stimulus activity (because it was not present in the experimental results),

iii). the network should carry out competition between a target and a non-target population, to implement the suppression of the non-attended object representation, and

iv). populations selective for the same object need to cooperate, in order to facilitate the activation of the neurons selective for the attended object and non-attended location.

We tested the presence or absence of these four properties of network activity over the explored parameter space as described next:

i). First, the *responsiveness* of the network was tested by applying the most effective stimulus (target in both locations) and checking if the resulting activity of the target selective populations exceeded a threshold of 10 Hz.

ii). Second, the *persistent post-stimulus activity* was tested by simulating the immediate post-stimulus phase of the network, i.e. the time after the stimulus – target in both locations – offset, and checking if the populations previously driven by input were still active. The average firing rates of the neurons in the network were initiated to 50 Hz for the two cooperating target selective populations, 3 Hz for the other excitatory populations, and 9 Hz for the inhibitory population. Persistent activity was assumed, if the activity of the target selective populations still exceeded 3 Hz after 1000 iterations.

iii). The presence of *competition* was tested by applying one target and one non-target stimulus and checking if the cooperating non-target populations could win against the target populations when the stimulated non-target population was biased by the attentional signal. Winning was tested by requiring that the activity of the non-attended target selective population does not exceed 3 Hz.

iv). Finally, it was examined under which conditions *cooperation* between neurons selective for the same object became prominent. Cooperation was tested by providing one sensory input and attentional bias to a target specific population, and the other sensory input to a non-target specific population, and measuring whether the other target-specific population was co-activated even without receiving any afferent input or bias.

At first, we set $w_- = 0$, denoting complete competition, and explored how the network's behavior changes when w_+, w' and w_n are varied. Figure 6.4 summarizes the borders of the parameter regimes given by testing the four response properties, which are plotted as a function of w_n and a number of w_+ and w' combinations: *i).* right to the dotted line the network shows responsiveness; *ii).* right to the dashed line the network shows persistent activity; *iii).* left to the dash-dotted line the network shows competition between target and non-target populations; *iv).* right to the solid line the network shows cooperation between the target specific populations.

As it can be seen from figure 6.4, all regimes are relatively robust, for a variety of weight values. Changing $w_+ = w'$ in concert results only in a shift of the regimes along the w_n axis. The regime where the network shows competition, cooperation and no persistent activity is associated to the attentional filtering effect. It can be observed that when the weights w_+ and w' become too small, the attentional filtering regime destabilizes. Likewise, if w' becomes too small, attentional filtering vanishes.

Next, we examined how different values for the weights connecting the target and non-target populations, w_-, and the specific and non-specific populations, w_n, affect the collective behavior of the network. The other two parameters are fixed to: $w_+ = 1.6$ and $w' = 1.6$ (the same values for w_+ and w' will be used for the explicit network dynamics simulations in the next section). The borders of the parameter regimes are summarized in figure 6.5 and correspond to the four response properties of the network:

i). The dotted line marks the border – almost parallel to the w_- axis – above which the network is effectively driven by the weak sensory inputs, i.e. for $w_n > 0.5$ the network shows responsiveness;

ii). The dashed line separates the regimes of persistent activity ($w_n > 0.71$) from pure stimulus coding ($w_n < 0.71$);

iii). Below the dash-dotted line the common input provided by the fibers from the non-specific neurons added to the input coming from the other specific neurons is balanced with the inhibitory input in such a way that it allows for efficient competition in the network;

iv). Right to the solid line ($w_n > 0.61$) the system shows cooperation between the target

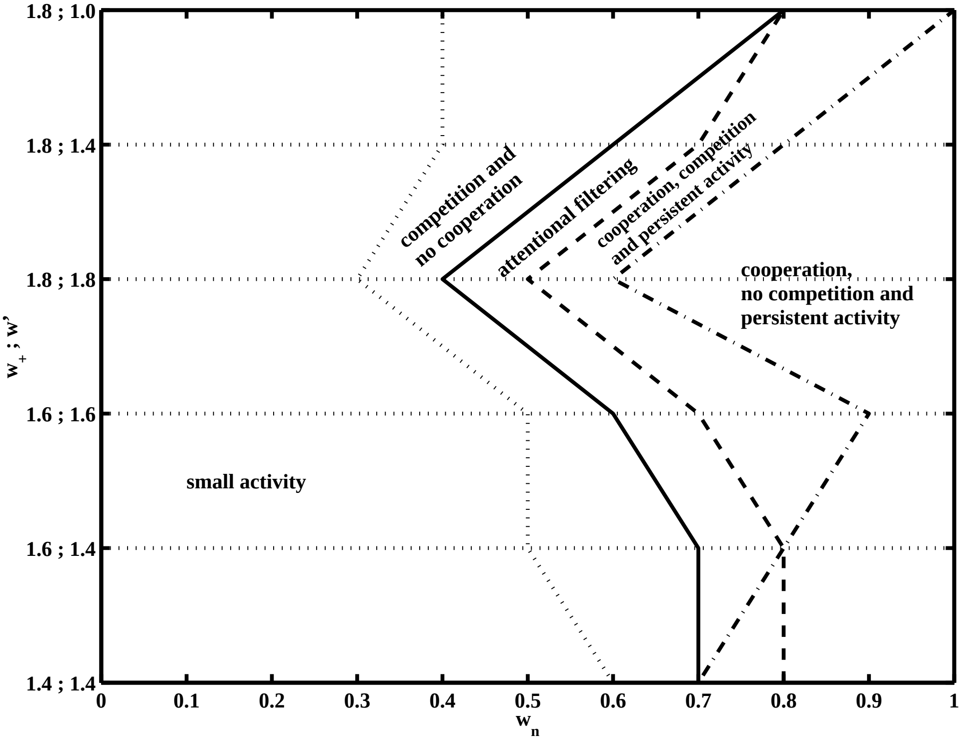


Figure 6.4: Parameter space exploration illustrating the influence of w_+ and w' on the parameter regimes as a function of w_n, for $w_- = 0$. The values on the y-axis denote the two parameters w_+ and w'. The parameter regimes are separated by border lines denoting *responsiveness* (dotted line), *persistent activity* (dashed line), *competition* (dash-dotted line) and *cooperation* (solid line). For more details see text.

specific populations.

It turns out that for stabilizing the attentional filtering regime w_- should be weak, but otherwise its actual value does not strongly influence the attentional filtering behavior. As it can be seen in figure 6.5, the border lines for responsiveness, cooperative setting and persistent activity are given by fixed values of w_n and are independent of the w_- value. For w_n above 0.51 the total excitation in the network becomes sufficient to amplify small external inputs such that the network becomes responsive. Further on, increasing w_n above 0.61 the total excitation in the network becomes sufficient to implement cooperation. And for w_n above 0.71 the reverberations in the network become strong enough to autonomously stabilize the post-stimulus activity.

Most of the excitatory neurons considered in the model belong to a population of non-specific neurons. In the real brain, these neurons would probably contribute to the implementation of some other functions, not related to the task modeled here. In particular some of these neurons could encode stimuli that are irrelevant for the present task. In both explorations, all four tested properties of the network activity were influenced by the parameter w_n expressing the weights from the non-specific to the specific neurons. This indicates that the activity of the non-specific

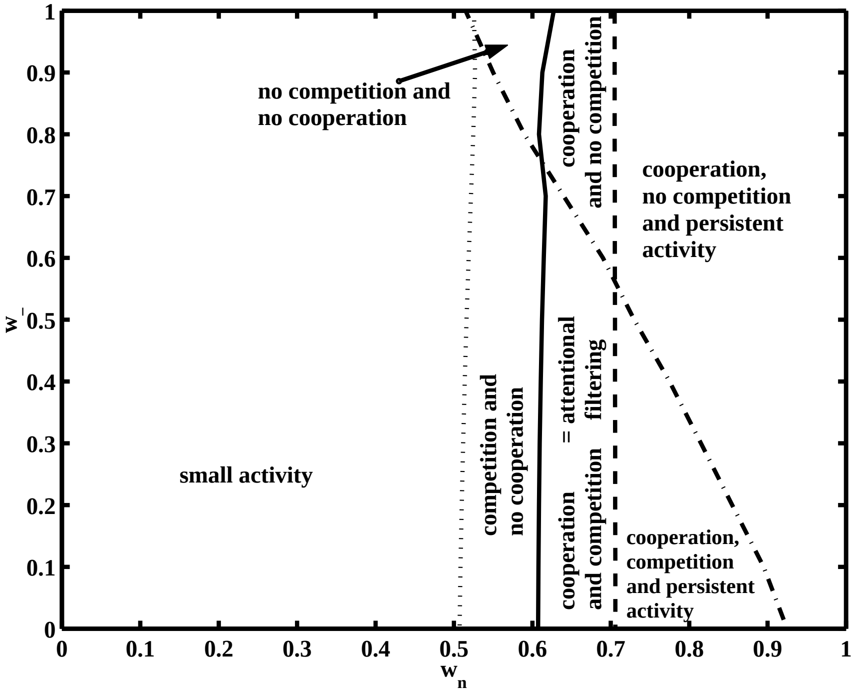


Figure 6.5: Parameter space exploration over different values of w_- and w_n, for $w_+ = 1.6$ and $w' = 1.6$. The parameter regimes are separated by border lines denoting *responsiveness* (dotted line), *persistent activity* (dashed line), *competition* (dash-dotted line) and *cooperation* (solid line).

neurons affects the total level of excitation in the network and can be thought of as a background current whose strength influences the functional regime of the network (as described in Stetter (2006)).

From the presented mean-field simulation results, a number of conclusions can be drawn. First, they show the existence of different parameter regimes, for which the model spiking network shows qualitatively different kinds of behavior. The following operational modes of the simple one-layer network can be specified:

- *Input amplification mode* - for w_- around 1 and small values of w_n, the network shows neither competition nor cooperation;
- *Selective mode* - for small values of both w_- and w_n, the network shows pure competition;
- *Attentional filtering mode* - for small values of w_- and intermediate w_n, the network shows both competition and cooperation, and no persistent activity;
- *Correlation facilitation mode* - for w_- around 1 and intermediate w_n, the network shows pure cooperation;
- *Non-selective working memory mode* - for large values of both w_- and w_n, the network

shows cooperation and persistent activity;

- *Selective working memory mode* - for small values of w_- and large values of w_n, the network shows competition, cooperation, and persistent activity.

Second, the population of non-specific neurons proves important for setting a certain activity level in the network, which corresponds to a certain operational mode. One thing to note here is that the connectivity to and from the non-specific population is non-symmetric. This was mainly chosen in order to remain close to the architectural choice of previous models and thus permit the comparison of these results with other modeling results. Changing the weights of the non-specific population to be symmetric and equal to w_n gives qualitatively the same results as with the non-symmetric setting.

In summary, cooperation, $w' > 1$, in combination with competition, w_- small and w_n intermediate, appears to be important for the generation of the measured attentional filtering effect. Using these results, full non-stationary simulations of the spiking network model were performed, and their results will be presented in the next section.

6.4 Explicit temporal simulations of spiking dynamics

Explicit simulations of the network spiking dynamics were carried out by using the coupled differential equations given in section 2.2.4. All four different bilateral stimuli combinations that were used in the visual attentional experiment of Everling et al. (2002) were applied to the model network, as presented in figure 6.3 a-d. Each sensory stimulus was applied for a time period of 300 ms, with a total strength of 400 Hz per stimulus-input, distributed over the 800 afferent fibers. The target and attentional biases were set to 16 and 160 Hz, respectively, and were left constant throughout the entire simulation run. As it can be observed, these values are different from the ones used in the mean-field analysis. The spiking network is more unstable given the random firing times, and it needs a stronger sensory input and attentional bias and also a weaker target bias for a consistent response. Using this setting, the 1000 coupled equations (2.3) were integrated numerically using the second order Runge-Kutta method with a step size of 0.1 ms. After 800 ms from stimulus onset, an excitatory flush of strength 5 Hz per afferent fiber was given to all neurons. This simulates a strong generic brain activity and causes the reset of the network activity that is mediated over the inhibitory interneurons.

Figure 6.6 demonstrates the necessary ingredients for the attentional filtering effect to occur. The left column, figure 6.6-1, displays the experimental results from Everling et al. (2002) in the case of the four bilateral stimulus combinations (illustrated as insets). The blue lines correspond to attention directed to the preferred location and the red lines correspond to attention directed to the non-preferred location. For the model network simulations (figure 6.6.2-4), the results represent the population-averaged responses of the model 'target right selective' (TR) neurons for the same stimulus conditions and attentional states as in the experimental results.

In the second column, figure 6.6-2, competition and cooperation are combined, using the parameters $w_+ = 1.6$, $w' = 1.6$, $w_n = 0.62$ and $w_- = 0.3$. It can be observed, that the simple network architecture using both conditions of competition and cooperation shows the attentional filtering of the information present in the unattended hemifield, similar to the experimental results measured in the monkey PFC. It can be concluded that **attentional filtering** consists of four different phenomena which can be assigned to the four stimulus conditions:

1. *Location preference*: When both hemifields contain target stimuli, the strength of the response reflects the preference for a specific location of the respective target selective neurons (figure 6.6-1a, 6.6-2a).

2. *Attentional suppression*: Although the target appears in the preferred location of the measured neurons, the response is completely shut down, as soon as attention is shifted away from the target-stimulated side (figures 6.6-1b and 6.6-2b, red lines).

3. *Attentional facilitation*: In contrast, when a target appears in the non-preferred location of the measured neurons, the neural response is increased, as soon as attention is shifted towards it (figures 6.6-1c, 6.6-2c, red lines).

4. Finally, when both hemifields are stimulated with non-target stimuli, the response stays low, reflecting the *target-selectivity* of the measured neurons (figures 6.6-1d, 6.6-2d).

Combining these effects, both biological and model neurons encode only the information presented in the attended hemifield (compare blue lines in figure 6.6-1, 6.6-2 a and b with c and d, compare the red lines in figure 6.6-1, 6.6-2 a and c with b and d), and ignore the content of the non-attended hemifield (compare blue lines in figure 6.6-1, 6.6-2 a with b and c with d, compare the red lines in figure 6.6-1, 6.6-2 a with c and b with d). The content of the non-attended hemifield is not encoded in the responses of the measured neurons.

It can be seen that the model traces from figure 6.6-2 are in good agreement with the experimental results. This demonstrates that a weak attentional bias can be strongly and selectively amplified by cooperation and competition that leads to an all-or-none attentional filtering effect. By choosing different parameters, it was shown that in the framework of the presented model both cooperation and competition are needed together in order to reproduce the referred effect. Competition, mediated by a small weight $w-$, implements attentional suppression, and cooperation, mediated by a strong weight w', implements attentional facilitation. When both mechanisms act together, this simple model shows a strong 'all-or-none' attentional filtering effect, which is mediated by weak top-down biases.

Two further simulations, for which the network was equipped only with competition (figure 6.6-3) or only with cooperation (figure 6.6-4) were carried out in order to examine the roles of cooperation or competition, respectively, for producing the referred attentional filtering effect. We say that a network exhibits cooperation when, due to recurrent processing, the activities of different populations end up being equalized. In contrast, a network exhibits competition, when differences between activities become amplified. In order to observe cooperation, the network needs a stronger connection weight w' between populations coding for the same category. This accounts for the equalization of the two target or two non-target population activities. In contrast, in order to observe competition, a weak connection weight w_- is required between

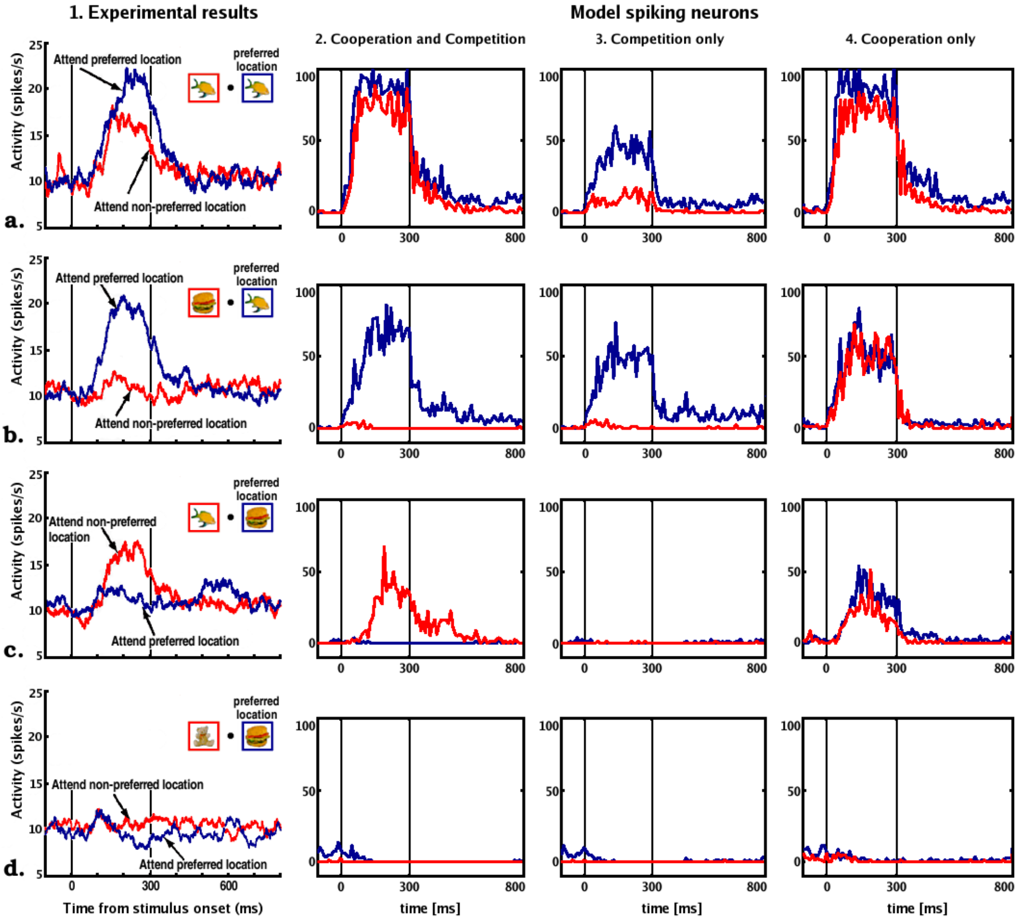


Figure 6.6: Experimental results of the focused attention task (column 1, adapted from Everling et al. (2002)) and model simulation results for three different parameter settings: cooperation and competition setting: $w_+ = 1.6, w' = 1.6, w_n = 0.62, w_- = 0.3$ (column 2); competition only setting: $w_+ = 1.6, w' = w_n = 0.76, w_- = 0.3$ (column 3); cooperation only setting: $w_+ = 1.6, w' = 1.6, w_- = 1, w_n - = 0.65$ (column 4). The results depict the mean neuronal responses of the target selective neurons to the four bilateral stimuli combinations shown in the inset: (a) target in both locations, (b) target in preferred location only, (c) target in non-preferred location, (d) non-target stimuli in both locations. Blue lines correspond to attention focused to the preferred location and the red lines to attention focused to the non-preferred location of the measured PFC neurons and model-neurons.

the populations encoding for different objects. Mediated by global inhibition, this succeeds in amplifying differences between the corresponding population activities.

When the network is dominated by competition (figure 6.6-3), the responses in the case of a target stimulus presented in the preferred location decrease (figure 6.6-3 a and b), and there is no attentional facilitation in the network (see the zero activity in figure 6.6-3c, red line). This is

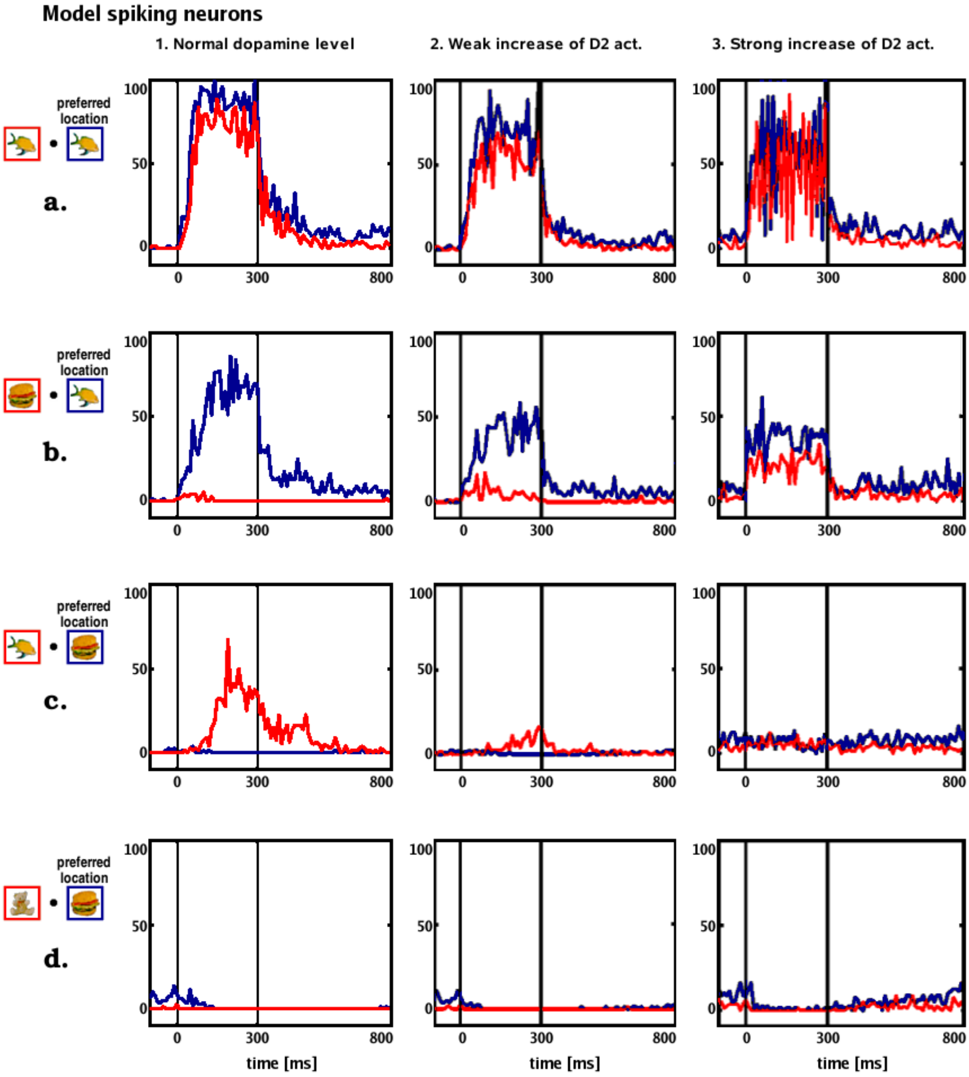


Figure 6.7: Simulation results of the model network for the competition and cooperation setting in the case of three different dopamine levels: normal dopamine level (column 1); weak increase of dopamine D2 receptor activity (column 2); strong increase of D2 receptor activity (column 3). The results depict the mean neuronal responses of the target selective neurons, to the four bilateral stimuli combinations shown in the inset: (a) target stimuli in both locations, (b) target in preferred location only, (c) target in non-preferred location, (d) non-target stimuli in both locations. For each stimulus combinations the two traces reflect the attended location. Blue lines correspond to attention focused to the preferred location and the red lines to attention focused to the non-preferred location, of the target selective model-neurons.

the case, because in the present model the facilitation effect is caused by a lateral propagation of activity from the stimulated TL population to the non-stimulated TR population over the recurrent connections. Because these connections are too weak in the competition-only setting (i.e., w' is too small), facilitation does not occur.

When the network is dominated by cooperation (figure 6.6-4), the responses between attended and non-attended conditions are equalized, and as a consequence the attentional effects dissappear (compare blue with red lines in figure 6.6-4). In particular, attentional suppression is no longer observed. In summary, competition, mediated by a small weight $w-$, implements attentional suppression, and cooperation, mediated by a strong weight w', implements attentional facilitation. When both mechanisms act together, the model shows the strong all-or-none attentional filtering effect, which is mediated by the weak top-down biases.

Departing from the parameter setting which correctly reproduces the experiment, one can also formulate concrete quantitative predictions, suitable for experimental testing, by observing the effect of neurotransmitters or pharmacological treatment. In particular, the effects of manipulating the dopamine level are studied, by evaluating the effect of an increase in dopamine concentration on the attentional filtering mechanism for the proposed network. An increase in dopamine concentration, accompanied by an increase in D2 receptor activation, is known to decrease both NMDA and GABA conductances (Law-Tho et al., 1994; Zheng et al., 1999). Hence, the increase in dopamine concentration is modeled here by a decrease in both NMDA and GABA conductances. This was achieved by multiplying both NMDA and GABA conductances by a factor of 0.7 or 0.2 to obtain a weak or strong, respectively, increase in D2 receptor activation.

The results are presented in figure 6.7. The graphs show the population-averaged responses of the model TR neurons for the same stimulus conditions and attentional states as for the experimental results. For all three cases, the network was set up to show both cooperation and competition: $w_+ = 1.6, w' = 1.6, w_- = 0.3, w_n = 0.62$. The first column presents the simulation results of the model network with a normal level of dopamine. The model neuronal responses in the case of a weak increase in D2 receptor activation, modeled by multiplying both NMDA and GABA conductances by a factor of 0.7, are presented in the second column. The model neuronal responses in the case of a strong increase in D2 receptor activation, modeled by multiplying the NMDA and GABA conductances by a factor of 0.2, are presented in the third column.

It can be seen that when the dopamine level increases slightly, the attention facilitation effect previously defined becomes impaired (figure 6.7, column 2-c). This might be related to a degraded ability to shift selective attention to a new, non-prominent stimulus. As the level of dopamine is further increased, the attentional suppression effect also becomes impaired, and the attentional effect becomes, in general, more impaired (figure 6.7, column 3). Hence, the results suggest that an increase in dopamine concentration will lead to a progressive weakening of the attentional filtering effect.

6.5 Discussion of the results

In this chapter, a network of integrate and fire neurons characterized by biophysically realistic spiking and synaptic dynamics, was used to model possible mechanisms underlying the visual attention filtering effect measured in Everling et al. (2002). To our knowledge, this is the first computational model proposed to explain the referred attentional effect. The experimental results show that the presence of a target in an unattended hemifield is not at all signaled in the target-selective neurons of the PFC. The information is totally filtered out by an attentional mechanism. However, presenting a target in the attended visual field will always exert a strong excitation in the target-selective neurons even for those that are preferentially activated by stimuli in the other visual field. This strong effect was explained in this work through competition and cooperation between the units of a simple one-layer network model, where competition was mediated by a weak attentional bias defining the relevant information to the current behavior.

Earlier models of attentional (Rolls & Deco, 2002; Corchs et al., 2003) and working memory phenomena (Brunel & Wang, 2001) were developed without considering cooperative weight settings. However, when working memory becomes selective, the currently stored information becomes dependent on the current brain state. Hence, context-dependent working memory requires a mechanism of association, which in general allows context to modulate working memory formation. Recent neurodynamical models specify that a more complex structure of the recurrent weights than the one originally proposed by in Brunel and Wang is needed in order to create selective working memory or rule-dependent working memory (Deco & Rolls, 2003; Deco et al., 2004).

In this study, the effects of such a complex weight setting on the network operation were systematically explored. It was found that cooperation between populations is an important structural feature of the network. Roughly speaking, cooperation allows that the activation of a population can be propagated to other populations encoding behaviorally associated information. The first population serves as the context, which enables or facilitates the response of the other populations. In general, cooperation can be hypothesized as the basis of categorization (binding information together referring to the same category), of different associations (for example between sensations) or even of mental manipulation. Here we suggest cooperation - besides competition - as a second fundamental principle for the neural basis of cognitive processes in the prefrontal cortex leading to what we name the *Extended Biased Cooperation-Competition Hypothesis*.

The attentional facilitation effect was successfully implemented through the *cooperative weight setting*, i.e. strong weights between populations encoding associated information. It might be hypothesized that the same effect could be obtained by divergent inputs, for example the neurons in the TL (target-left) population would also receive some external sensory input when the target was presented in the right hemifield. This divergence of the inputs was not modeled in our network since preliminary results showed that it could not account so well for the attentional filtering effect. The facilitation effect seems to need the reverberation between the pairs of specific populations.

The network model used in this work represents a small area of the PFC. The PFC is a neocortical structure connecting with all sensory and motor areas and with other cortical and sub-cortical systems. This broad connectivity structure makes the PFC suitable to coordinate different types of information converging from many brain regions. In fact, neurophysiology, imaging and computational studies have suggested that PFC plays an important role in cognitive and behavioral control, and it is thought that the wide connectivity that characterizes PFC might determine its crucial role in cognition. However, the focus of this work involves the investigation of the possible neural mechanisms within the PFC that underlie complex behaviors and hence the model is restricted to a minimal network of spiking neurons. The emphasis is not put in modeling in detail the whole hierarchy of cortical regions involved in the attentional filtering effect. Therefore, in the presented network, the two biases (for attention and identity of the target) are hypothesized to come from cortical areas not explicitly modeled. We also do not explicitly model the cortical processing of visual information before it reaches the prefrontal cortex. Instead, we consider that the inputs to the network are already signaling a specific pair of object and location. At a later stage one could model the biases explicitly by having specific populations operating in a bistability regime (i.e., able of persistent activity over a delay period) coding for the present context of the task. In principle, the processing of the visual input by early cortical areas could also be added to the present model, using for example models similar to those presented by Rolls & Deco (2002).

Given the complexity of the PFC and its associated projections, it is remarkable that we could explain the experimental results using this very simple network model. Important to note however, is that although we reproduced qualitatively the attentional filtering effect, the quantitative aspects of the results could not be explained by this simple model. In particular, the results of the simulations could not account for the values of the baseline activity and the strength of the responses. We tried to reproduce these features of the data by modifying the values considered for the conductances. The high firing rates observed in figure 6.6-2 a and b could be decreased. However this could not be achieved without reducing also the attentional facilitation effect (figure 6.6-2 c), which tended to disappear. In order to modify in a disproportional way the peak activations, a more detailed network model, containing several processing layers, would be required. The interplay between such layers might contribute to the reduction of the peak activity found when targets are presented to both hemifields, by weakening the input at an earlier stage through a top-down attentional bias. The attentional weakening would lead to a smaller input to the target encoding population with preference for the non-attended location and hence induce less activity for the cooperating pair of target encoding populations.

The number of neurons used in the network (1000) is relatively small, and was chosen for computational feasibility. The network can be scaled to have larger number of neurons (see Brunel & Wang, 2001), thus reducing the finite-size effects while preserving the qualitative behavior of the system. The system is robust to changes in the relative sizes of the neuronal populations, provided that a large part of the excitatory neurons are non-specific. This characteristic is important to assure the stability of the activity in the network, as well as the stability of several important operational regimes identified in the present work. In fact, in biology, for any possible state of the system there is always a large population of non-specific neurons, corresponding to all other neurons not involved in coding the present particular state. Stability of operational modes might then be an important functional role for distributed representation and sparse coding in the brain.

A putative aspect of attentional filtering – inattentional blindness – has been behaviorally studied in humans. The neuronal mechanisms underlying this effect have been hypothesized to be the ones experimentally measured by Everling et al. (2002). However, single neuron measurements cannot be directly compared with behavioral performance. The presented neurodynamical computational model with biologically inspired spiking dynamics captures the neuronal behavior underlying the mechanisms of visual attentional filtering, as measured in Everling et al. (2002), and hence can in principle be used to make predictions concerning human psychophysical experimental results. It can also be extended to allow comparison with neuroimaging results (see Deco et al., 2004). The model can thus provide an ideal theoretical framework to link the electrophysiological measurements with the results from human studies, both measuring performance and brain activity through imaging methods.

In this work two experimentally testable predictions of the model were formulated. First prediction relates an increase in the level of dopamine with a progressive impairment of the attentional filtering effect. In particular, as the dopamine concentration increases, the model predicts that attentional facilitation is first affected, followed by the impairment of attentional suppression. According to these results, in terms of behavior, an increase in the level of dopamine is expected to impair performance, in the sense that the presence of distracting stimuli will interfere with the processing of the task-relevant information. The other prediction is that the presentation of an increasing number of task irrelevant stimuli will eventually lead to the disappearance of the attentional filtering effect. The observed behavior of the model suggests that the ability to filter out stimuli with basis on attention will degrade as the number of distracting stimuli increases.

Further, the mean-field explorations allowed us to characterize the model network for a number of different overlapping working regimes: competition, cooperation, persistent activity and non-competitive amplification, which in combination form quite different modes of operation. For example non-persistent activity and competition yield an attentional selective mode, while non-persistent and non-competition result in a non-competitive amplification mode. The grouping of specific populations, effect we name cooperation, together with competition for both bistability and single stability regimes might be a more general mechanism used in the brain to implement computation. The contribution of this study is then twofold: we reveal different modes of operation of the simple one-layer network, which can be used to perform distinct operations, and model the neuronal mechanisms underlying visual attentional filtering.

7 Two-layer spiking neural network modeling selectivity tuning in ITC

This chapter analyzes the computational principles underlying the neuronal correlates of perceptual learning, showing how learning affects the connectivity between two model areas and how the resulting intrinsic attentional signal affects the level of competition and cooperation in the network in order to express the relevant information for the behavioral task. The study shows that higher-level cognitive feedback encoding the learned categories can explain the enhancement of selectivity in ITC neurons to the stimulus features which are relevant for a learned visual categorization task and that the referred tuning effect can be correctly reproduced by a biologically-inspired learning algorithm that robustly converges to a stable fixed point of the learning dynamics.

7.1 Neural selective tuning in ITC and concept formation in PFC

Perceptual learning represents an important cognitive process that involves structural and functional modifications of the brain following sensorial experience, and leads to improvements in task performance with training or practice (Goldstone, 1998). Different studies show that neurons from cortical areas involved in higher-stages of visual processing become tuned to some particular patterns of the visual input. These changes in the response properties of the cortical neurons, supposed to be mediated by higher-level top-down inputs resulting from cognitive mechanisms like *concept formation* and *attention*, are associated with perceptual learning (Fine & Jacobs, 2002).

Trying to understand the neural mechanisms of perceptual learning represents a challenging task that aims, along with other studies, at a better understanding of brain functionality. *Inferotemporal cortex (ITC)* and *prefrontal cortex (PFC)* are two interconnected cortical areas thought to be involved in the performance of visual tasks, such as visual recognition, categorization and memory, although the contribution of each of these two areas in visual processing is not fully understood. In this context, recent studies have suggested that *PFC* is mainly associated with *cognitive processing* (such as categorization), while *ITC* is more associated with *feature processing*, (Freedman et al., 2003). *Categorization* is an important cognitive mechanism for information processing, involved in concept formation. Further studies suggest that, top-down signals from PFC could partially determine ITC neuronal responses (Freedman et al., 2003; Tomita et al., 1999).

In a recent neurophysiological experiment Sigala & Logothetis (2002) have studied how the representation in ITC of different visual stimulus features was affected by their behavioral relevance, by monitoring the activity level of single ITC neurons of awake behaving monkeys engaged in a visual categorization task. For this experiment, the monkeys learned to categorize a set of schematic images, representing faces or fish (the visual stimuli), into two categories. Each category was associated to one lever that the monkey had to pull when the presented stimulus belonged to that category. The stimuli were characterized by a fixed set of four varying features, each of them having a discrete small set of values (high, medium or low) as shown in figure 7.1.a. Only two of the varying features, referred to as diagnostic, were relevant for solving the categorization task. The two categories could be linearly separated along the two diagnostic features in the stimulus space, as depicted in figure 7.1.a bottom. The other two features, referred to as non-diagnostic, gave no information about the stimulus associated category and were irrelevant for the ongoing task. After training, the activity level of the visual responsive ITC neurons was measured. The experimental results, presented in Sigala & Logothetis (2002) and reproduced in figure 7.1.b show that after training the selectivity for the different levels of the task-relevant features, the diagnostic features, was enhanced (figure 7.1.b - top panel) in comparison to the selectivity for the levels of the other non-diagnostic features (figure 7.1.b - lower panel). Their results suggest that ITC not only encodes objects and features, but their representation is tuned by their relevance to behavior.

Taking into account all these findings on perceptual learning, higher visual processing and the tuning of ITC neurons during the categorization task (as reported by Sigala & Logothetis, 2002), we hypothesize that the enhancement of selectivity to the diagnostic features (see figure 7.1.b-top panel) in ITC might emerge, in the behavioral context, through a higher-level cognitive feedback, originating from category encoding neurons, possibly residing in the PFC. According to this hypothesis, the selectivity for the diagnostic features, which is acquired during training, is formed as a consequence of the top-down signals coming from an area where information about the learned categories is stored. This could explain the underlying neural substrate of the referred perceptual tuning effect.

In order to test the above mentioned hypothesis and account for the presented experimental results, a neurodynamical two-layer cortical model that simulates two small interconnected areas from ITC and PFC is proposed (Szabo et al., 2006). The model is constructed in the framework of *Biased-Competition and Cooperation* (Szabo et al., 2004; Almeida et al., 2004) and the choice of its structure and parameters are presented in section 7.2. The next section, 7.3, characterizes network's modes of operation by exploring different parameter regimes using the mean-field approximation. In section 7.4 the non-stationary dynamics of the model network is simulated using a parameter set where the neurons show selectivity tuning and the results are compared with the experimental results. Choosing a learning prescription that robustly modifies the network free parameters to reach a configuration where the desired associations of the categorization task are correctly performed, section 7.5 presents the evolution of selectivity tuning of the ITC model neurons during learning. The last section, 7.6, contains the discussion of the results and conclusions.

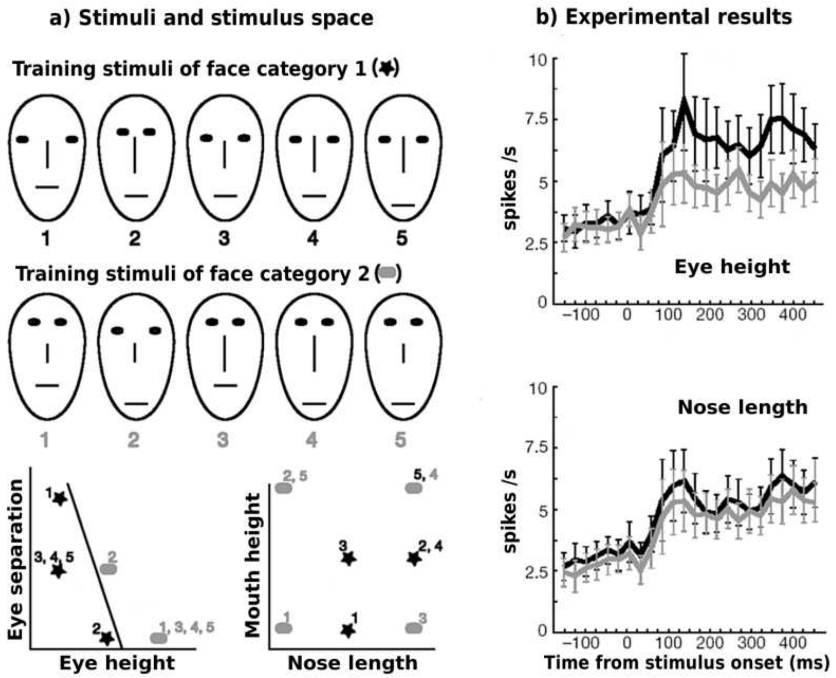


Figure 7.1: a) Stimuli and stimulus space for the visual categorization task adapted from Sigala & Logothetis (2002). The stimuli have four varying features: *Eye-height*, *Eye-separation*, *Nose-length* and *Mouth-height*, and can be linearly separated in two categories along two of the four dimensions: *Eye-height* and *Eye-separation*. b) Experimental results adapted from Sigala & Logothetis (2002). The traces represent the average activation of the recorded visual responsive ITC neurons (a total of 96 units) after training (i.e., after the monkeys have learned to categorize the presented stimuli). For each neuron, the responses were sorted by the presented features and averaged over many trials. The resulting average neuronal activity levels reflect which feature value excite a given neuron most and least, respectively. The population average activation was calculated by grouping these average neuronal activity levels according to their best (black lines) and worst (gray lines) responses to the levels of the diagnostic feature *Eye-height* (top panel) and the non-diagnostic feature *Nose-length* (bottom panel).

7.2 Network structure and parameters

Extending previously introduced biologically-inspired neurocomputational models that were shown to capture many aspects of the dynamics shown by neurophysiological measurements (Brunel & Wang, 2001; Deco & Rolls, 2003; Szabo et al., 2004), a minimal model accounting for the response enhancement to the relevant features of the ITC neurons is proposed. The model network, presented in figure 7.2, is structured in two interconnected layers of integrate-and-fire neurons that follow the general structure and properties described in chapter 3. The first layer corresponds to a small area in ITC that receives external bottom-up information about the presented stimulus and is organized into populations of neurons which receive feature specific inputs. The second layer corresponds to a small area in PFC that contains neuronal populations connected to the populations in ITC in a way which allows them to encode for the corresponding learned categories.

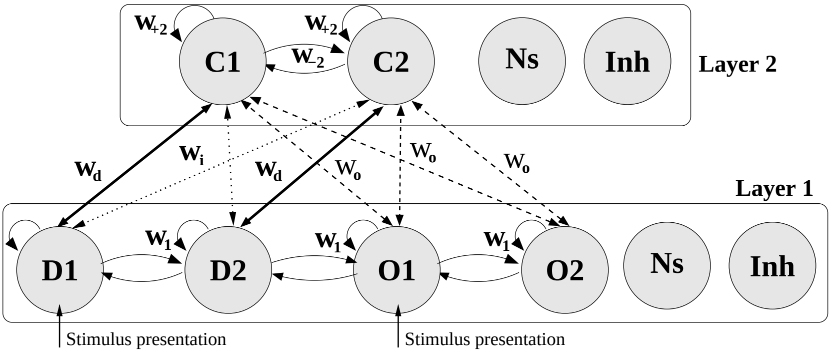


Figure 7.2: Schematic representation of the two-layer model architecture. Each layer consists of a fixed number of specific populations, a non-specific population (Ns) and an inhibitory population (Inh). First layer (ITC model layer) receives external inputs from lower cortical areas encoding the presented stimulus features: D1 receives input when the *diagnostic* feature is *high*, D2 when the *diagnostic* feature is *low*, O1 when the *other* feature is *high*, and O2 when the *other* feature is *low*. The second layer (PFC model layer) encodes the two categories to be learned (C1 and C2), and interacts with the first layer through the connections w_d, w_i and w_o. (for more details see text)

For this minimal model, we consider that the presented stimuli are characterized by only two features: Eye-height and Nose-length, each with two discrete values (high and low), and that the two categories are determined exclusively by one of the features, the diagnostic feature: Eye-height. This results in four specific populations in the ITC layer, denoted according to the specific input that they receive: one population receives input when the stimulus is characterized by the diagnostic feature being high (population D1), one when the diagnostic feature is low (population D2), one when the other feature is high (O1) and one when the other feature is low (O2). The specific populations in the PFC model layer are selective for the first category (C1) and second category (C2), respectively (see figure 7.2). The stimuli having the value high for the diagnostic feature will be classified as belonging to category C1 and the stimuli having the value low for the diagnostic feature as belonging to category C2.

Each layer is constructed of a large number of LIF-NS cells (presented in section 2.2.4), fully connected. First layer, corresponding to ITC, is composed of $N_{E1} = 800$ excitatory model neurons organized into groups of $f \cdot N_{E1}$ neurons for each specific population, $(1 - 4f) \cdot N_{E1}$ neurons for the non-specific population and $N_{I1} = 200$ inhibitory model neurons forming the inhibitory population. The second layer, corresponding to PFC, is composed of $N_{E2} = 520$ excitatory model neurons (with $f \cdot N_{E2}$ neurons for each specific population, $(1 - 2f) \cdot N_{E2}$ neurons for the non-specific population) and $N_{I2} = 130$ inhibitory model neurons grouped into the inhibitory population. For simplicity we used the same number of neurons for the specific populations in each layer and set $f = 0.1$. The chosen values respect the proportion of 80% excitatory neurons and 20% inhibitory neurons, in order to be consistent with the neurophysiological data (Abeles, 1991).

The individual populations are driven by two different kinds of input. First, all neurons in the model network receive spontaneous background activity from outside the module through $N_{ext} = 800$ external excitatory connections, each carrying Poisson spike trains at a spontaneous rate of 3 Hz (the typical value observed in the cerebral cortex Wilson et al., 1994; Koch & Fuster, 1989), which amounts to a background external input of 2.4 kHz for each neuron. And second, the neurons in the four specific populations from the first (ITC) layer also receive, in addition to the background input, external inputs encoding stimulus specific information, which are assumed to originate from lower areas which process the visual scene such as to provide these signals. For the simple case of having only two features, each with two values, results four different combinations of inputs that can be presented to the network. For the corresponding neurons in the specific populations, the rate of the external Poisson train is increased by $\lambda_{stim} = 150$ Hz. The non-specific excitatory neurons receive only the common background input. They are thought to be involved in other cognitive tasks and to be only spontaneously and non-selectively active in the present framework.

The conductance values for the synapses between pairs of neurons are modulated by connection weights, which can deviate from their default value of 1. The structure and function of the network is achieved by differentially modulating these weights within and between populations of neurons. The structure is set so that the sum of all connection weights to each neuron is 1, to assure stability (see Brunel & Wang, 2001). The two layers are fully connected, but the interlayer connectivity is restricted to the specific populations only. The labeling of the weights is defined in figure 7.2.

Inside each layer the weights are considered fixed and chosen as follows: According to preliminary simulations it has proven necessary to chose no structure in the ITC model layer in order to achieve weak selectivity for the case when the feedback was not present. In this way, the enhancement of selectivity in the ITC layer is obtained only as a result of categorization encoding in the PFC layer. All weights are set equal to the default value $w_1 = 1$, thus implementing cooperation between all ITC specific populations (Szabo et al., 2004). The cooperating setting in the ITC layer can be argumented by the fact that different features of a face always co-occur in natural images, e.g. nose (long or short) will always come together with eyes (high or low). In the PFC layer, neurons encoding different categories are likely to have anti-correlated activity resulting, following a Hebbian learning paradigm, in weaker than average connections between the populations C1 and C2, implementing thus competition between them. For this we choose the extreme case $w_{-2} = 0$. The weights within the same category population are set to the default value $w_{+2} = 1$, as we are not interested in persistent activity (achieved for strong recurrent connections, Almeida et al. (2004)) and look for a regime with small firing rates, as reported in the experimental data. For both layers, the weights from and to the non-specific populations were computed so that the sum of all excitatory connection weights to each excitatory neuron sums up to 1 and all connections from and to the inhibitory population are set to the default value of 1.

Finally, the connections between the two layers, restricted to the specific populations only, are set as follows: the neurons selective for the non-diagnostic feature and the category neurons are assumed to have uncorrelated activities, and the same weight w_o is chosen for all the connections between them. The activities of the diagnostic selective neurons and the corresponding category neurons are likely to be correlated, so we hypothesize that the connection strengths between D1 and C1 (and D2 and C2, respectively) will increase by training so that $w_d > w_o$. Likewise the neurons in D1 and C2 (D2 and C1, respectively) probably have anti correlated activities, resulting in a decrease of the strengths by training so that $w_i < w_o$. Unless specified otherwise, throughout this chapter, the weights for the feedback connections (from PFC to ITC) are taken to be half of the weight values for the corresponding feedforward connections (from ITC to PFC) and the values used in the next sections for the parameters w_d, w_i and w_o will represent the feedforward connection values. The choice for the feed-forward connections to have, on average, the double strength of the feedback connections is inspired from the idea that between the cortical areas of the brain the feed-forward projections have a strong driving role, while the feedback projections have a weaker modulatory role. The absolute strengths of the connection weights between the two layers are explored in the next section in order to analyze different operational modes of the model network.

7.3 Exploration of network connectivity

The behavior of the model network whose architecture and parameters were presented in section 7.2 is analyzed by exploring the structure of the excitatory weight setting between the two model cortical layers, using the mean-field approximation, presented in chapter 5 that is fully consistent with the model used. The mean-field approximation allows an exhaustive analysis of the network regimes as a function of the parameter space. For each simulated point in the parameter space, corresponding to a fixed set of inter-layer weights, the Mean-field simulation was performed for all four possible stimuli presentations and the results were grouped according to the best and worst responses for each feature, in the same way as for the experimental results. The simulation started by initializing the frequencies to 3 Hz and 9 Hz for the excitatory and inhibitory neurons, respectively, and setting the rate of the external input to each model neuron to 2.4 kHz. The stimulus presentation was modeled as an extra input rate of 150 Hz to the corresponding specific populations in the ITC layer. After setting all the parameters, the Mean-field equations (see chapter 5), were integrated using the Euler algorithm with a step size 0.1 for 3000 iterations,

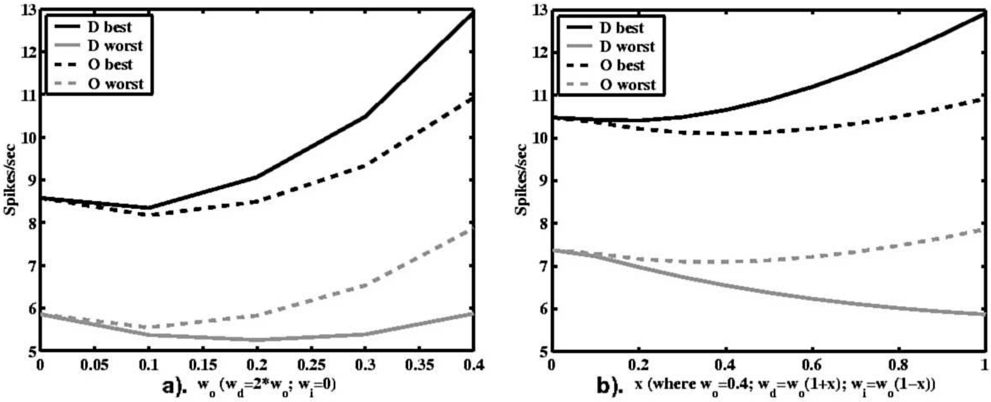


Figure 7.3: Parameter space exploration over different values of the weights between the two layers. Both graphs show the average spiking rates of feature selective neurons according to their best (black lines) and worst (gray lines) responses to the levels of diagnostic (solid lines) and other non-diagnostic (dashed lines) features, exploring (a.) the necessary connection strength between the two layers in order to obtain the desired effect (w_d takes values from 0 to 0.8, $w_o = w_d/2$ and $w_i = 0$) and (b.) the relative difference of these connections (starting from equal connectivity $w_d = w_o = w_i = 0.4$, w_d is increased and in the same time w_i is decreased, until $w_d = 0.8$, $w_o = 0.4$, $w_i = 0$).

enough to achieve convergence.

Figure 7.3 shows how the selectivity for the diagnostic feature, when compared to that of the other non-diagnostic feature, behaves as a function of the connection strengths between the two layers. Figure 7.3.a explores the necessary connection strength between the two model cortical areas needed to influence one another. It was obtained keeping $w_i = 0$ and changing w_d and w_o, subject to the constraint $w_d = 2 \cdot w_o$. Figure 7.3.b explores the relative difference of the connection strengths between the category populations in PFC and each of the specific populations in ITC. It was obtained keeping fixed $w_o = 0.4$ and increasing the difference between w_d and w_i subject to the constraint $w_d + w_i = 2 \cdot w_o = 0.8$. Both graphs show the average spiking rates of all selective neurons in the ITC model layer according to their best (black lines) and worst (gray lines) responses to the levels of diagnostic (solid lines) and other non-diagnostic (dashed lines) features.

The results show that the selectivities for the diagnostic and non-diagnostic features are equal in the case when there is no connectivity between the two layers (figure 7.3.a-left side) or when all the connection weights are equal (figure 7.3.b-left side). But when w_d increases in comparison to w_i, both changing w_o (figure 7.3.a) or keeping it fixed (figure 7.3.b), the selectivity for the level of the diagnostic feature increases while the selectivity for the level of the non-diagnostic feature remains approximately constant. Thus the increase in selectivity is mediated by category specific top-down input from the PFC layer, and the connectivity between the two layers should reflect the following categorization rule: neurons in ITC receiving inputs encoding for the diagnostic feature (for example D1 receive input when presented stimulus has diagnostic feature being

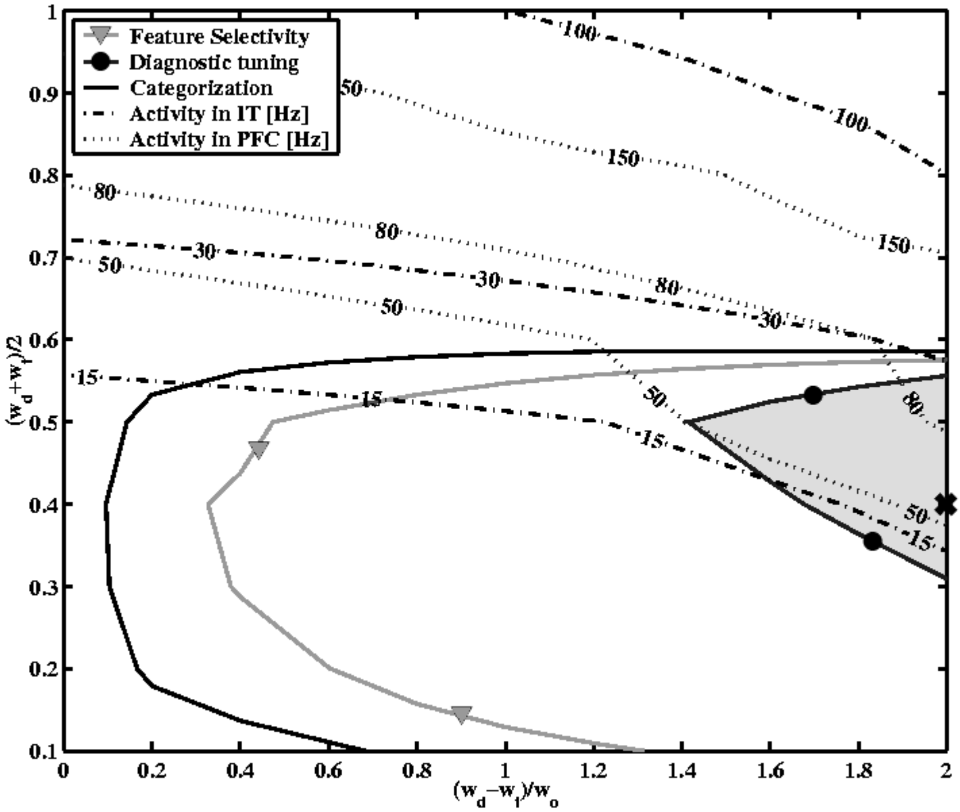


Figure 7.4: Parameter space exploration for different connection weights between the layers. The x axis represents the relative difference in the connectivity of the diagnostic specific populations to their corresponding (w_d) or non-corresponding (w_i) category, and the y axis represents the average connection strength between the layers. The weights connecting the category populations with O1 and O2 in the ITC layer, were subject to the constraint $w_o = w_d/2$. The activity of the four specific populations in the ITC layer was evaluated, and two parameters were calculated in order to measure the *Feature selectivity* and *Diagnostic tuning* of the ITC specific neurons. Inside the black solid line without markers the network performed correct categorization. The dashed (dotted) lines represent the maximum activities in the ITC layer (PFC layer, respectively). The gray area corresponds to the region in the explored parameter space where all the conditions were fulfilled. The black **x** in the graph corresponds to the parameters used in the spiking simulation presented in figure 7.6.

high) should be strongly connected to the corresponding category (C1 through w_d) and weekly connected with the other category (C2 through w_i). The increased difference for the spiking rates associated with the diagnostic feature shows that there is an enhancement of selectivity for the level of this feature in the ITC layer that it is mediated by the category specific top-down inputs from the PFC layer.

The two-dimensional parameter explorations in figures 7.4 and 7.5 further analyze the behavior of the network for a large value range of the excitatory weights connecting the two model cortical layers. The relation between w_d and w_i, relative to the total connection strength is explored in figure 7.4, and the relation between the feedforward and feedback connectivities in figure 7.5.

For both graphs each point was simulated for all four possible stimuli presentations, and the activities of all specific populations were evaluated in order to conclude about the network's operational modes. From the activity of the four specific populations in the ITC layer, two parameters that measured the *Feature selectivity* (gray solid line with ▿ markers) and *Diagnostic tuning* (black solid line with • markers) of the ITC selective neurons were calculated. Feature selectivity measures, for a specified set of parameters, the selectivity for the diagnostic feature through a selectivity index calculated as the difference between the best diagnostic feature value and worst diagnostic feature value activities divided by their sum. Diagnostic tuning measures the difference between the selectivity for the diagnostic feature and the selectivity for the non-diagnostic feature calculated also as a selectivity index. From the activity of the two category populations in PFC, another parameter was calculated as the difference between the activity of the population encoding the correct category of the presented stimulus and the activity of the population encoding the other category divided by their sum. It measures the correct association of the presented stimulus with the corresponding category, i.e. *Correct categorization* (black solid line without markers).

For each of these parameters we chose a threshold that marked the limit where the requirements of having the respective selectivity, tuning or categorization were still satisfied. The figures 7.4 and 7.5 plot the border lines of these limits that separate the different operational regimes of the network. Because the network should work in biologically relevant activation regimes, the maximum activities that the network reached during the simulations were also checked. The dashed-dotted (dotted) lines represent the maximum activities in the ITC layer (PFC layer, respectively). With all this information, we could separate the area in the explored parameter space where all the conditions were fulfilled (the gray area in the graphs).

The results presented in figure 7.4, indicate that as the average connection strength between the layers increases, the activity in both layers increases, and when it exceeds some value (here for $(w_d + w_i)/2 > 0.6$), the functionality of the network becomes impaired (both categories show high activity and the neurons in ITC don't show selectivity anymore). In the same way, for small differences between w_d and w_i ($(w_d - w_i)/w_o < 0.1$) the network shows categorization impairment. As this difference increases, the network starts performing correct categorization. But this difference has to increase even more to achieve feature selectivity ($(w_d - w_i)/w_o > 0.5$). The gray area shows the region for which also the diagnostic tuning requirement is met. We conclude that the average connection strength should be not too high so that the network keeps its functionality with reasonable values for the population activities, but also high enough so that the PFC layer can influence the ITC layer and achieve in this way diagnostic tuning. Only a strong average connection strength is not enough to achieve feature selectivity and diagnostic tuning, another requirement is that w_i should be a very small fraction of w_d ($w_i << w_d$).

Figure 7.5 studies the influence of feedforward versus feedback weight strength on network's behavior. The values for the feedback weights are taken relative to the values for the feedforward weights from $0 \cdot w_{ff}$ to $2 \cdot w_{ff}$. As can be seen from the figure, for high feedforward values,

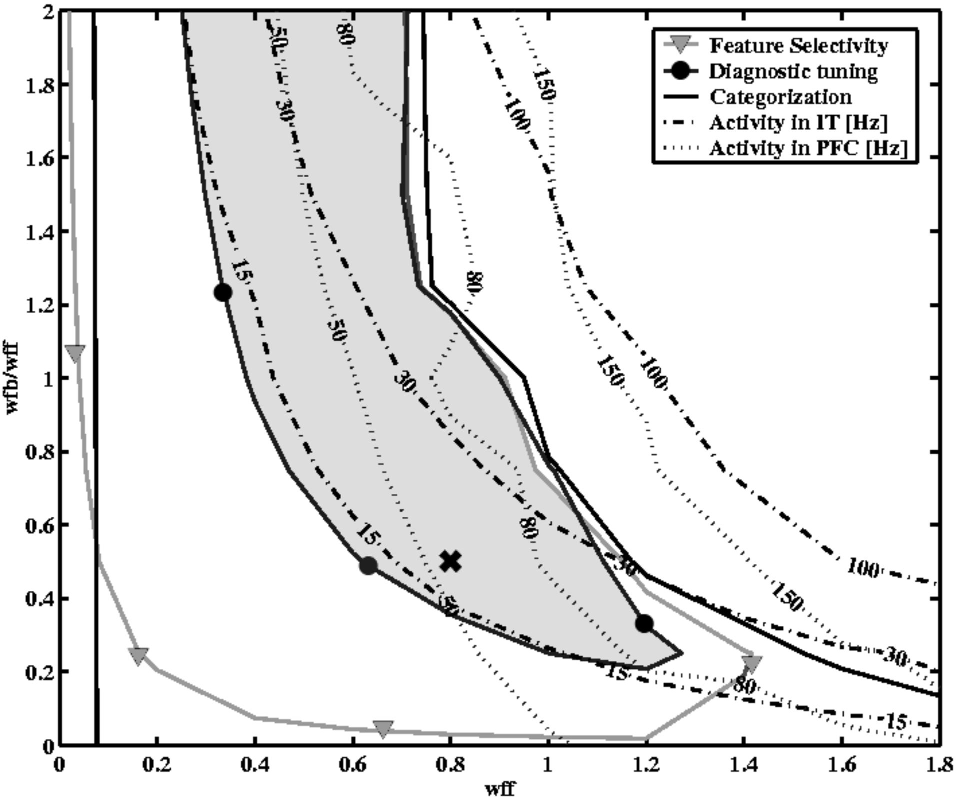


Figure 7.5: Parameter exploration that studies the influence of feedforward - feedback strengths on the network regimes. The x axis represents the strength of the feed-forward value of w_d (the other weights were chosen: $w_o = w_d/2$ and $w_i = 0$). The y axis represents the ratio between the feed-forward and feedback values for all interlayer connections. Same evaluation parameters and border lines were used as in figure 7.4.

$w_{ff} > 1$, the second layer receives significant input which will result in high activities. Also, if we use high feedback values (upper right corner of the figure), the network will end up in a high amplification regime with categorization impairment and no feature selectivity. Similar if small feedforward and feedback values are chosen (lower left corner of the figure), the activities in the PFC will be too small and unable to drive the specific populations in ITC. The gray area in the figure shows the regime where all the conditions: feature selectivity, diagnostic tuning and correct categorization are fulfilled.

7.4 Selectivity shaping in the spiking network

Explicit simulations of the spiking dynamics for the model network were carried out for a specific connection weight setting between the two layers and for all possible combinations of stimulus inputs. For each simulation, a different combination of stimulus features was applied, for a time period of 500 ms, with a total strength of $\lambda_{stim} = 150$ Hz per stimulus-feature input. The 1650 coupled differential equations (2.3) with parameters given in section 7.2 were integrated numerically using the second order Runge-Kutta method with step size 0.1 ms.

Figure 7.6 shows in parallel the experimental results from Sigala & Logothetis (2002) (figure 7.6.1) and the results from the spiking model simulations (figure 7.6.2), for a fixed parameter set: $w_d = 0.8$, $w_o = 0.4$ and $w_i = 0$. The experimental results show the population average activity for all recorded visual responsive neurons, when different combinations of features were presented. Four responses were selected: the highest (black line) and lowest (gray line) responses sorted according to one diagnostic feature (figure 7.6.1.a) and the highest (black line) and lowest (gray line) responses sorted according to one non-diagnostic feature (figure 7.6.1.b).

The simulation results were obtained by doing the corresponding calculations in the model network. In order to reproduce the experimental data, we also took into account all the neurons responding to the presented stimuli, so the average firing rate over all specific populations in the ITC model layer is considered. The model network is simulated for all possible combinations of the input values (D1+O1, D2+O1, D1+O2, D2+O2), and each combination was simulated 10 times. After this, we check for each specific population, which value of the diagnostic feature (non-diagnostic feature, respectively) produces a higher response and we use the corresponding activity to compute the average rate representing best value for the diagnostic feature (non-diagnostic feature, respectively). Similarly the lower responses was used to compute the average rates representing the worst values. This average activities over all specific populations for the best and worst values of the diagnostic (non-diagnostic, respectively) feature are presented in figure 7.6.2 (left column). The right column of figure 7.6.2 presents the firing rates of the specific populations from both layers (D1, D2, O1, O2 for ITC and C1, C2 for PFC) in the case of a presented stimulus with diagnostic feature high (D1) and other feature high (O1) (i.e. external input to D1 and O1 in our model).

As it can be observed from figure 7.6, the simulations results are, qualitatively, in good agreement with the experimental results and show that there is an enhancement of the selectivity for the level of the diagnostic feature, as compared to the non-diagnostic feature (the lines in the first two columns in figure 7.6.a are more separated than those in 7.6.b). And since all weights in the ITC layer were chosen equal, results that the enhancement of selectivity emerges due to the top-down inputs from the PFC layer, which encodes the previously learned stimulus categories. From the time when the stimulus is presented to the network (time = 0 ms in figure 7.6), the selectivity of the category specific populations (figure 7.6b, right most column) emerges through the feed-forward connections (ITC -> PFC) from the activation of the specific populations in the ITC layer. Through the feedback connections (PFC -> ITC), this selectivity is transmitted afterwards to the feature-selective populations in ITC.

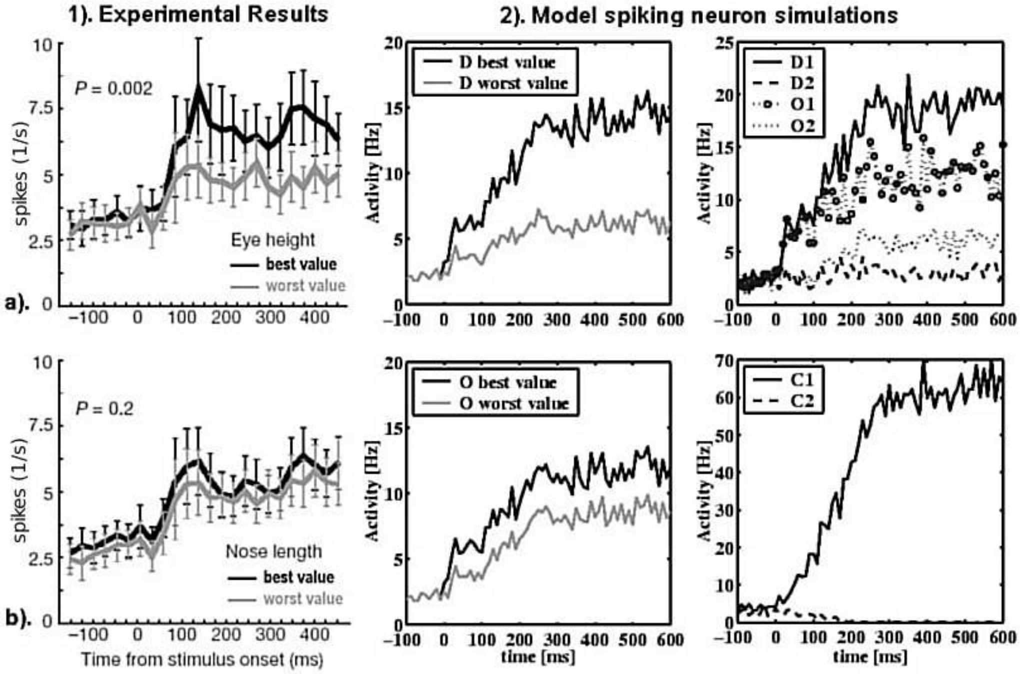


Figure 7.6: Experimental results adapted from Sigala & Logothetis (2002) (1) and model spiking neurons simulation results (2). The first two columns show the average spiking rates of stimulus responsive neurons, grouped according to their best (black lines) and worst (gray lines) responses to the levels of diagnostic (a) and non-diagnostic (b) features. The right most column shows the average spiking rates of the four specific populations (D1, D2, O1, O2) in ITC layer (a) and the two specific populations (C1, C2) in the PFC layer (b) in the case of the stimulus presentation having the diagnostic feature high and the other non-diagnostic feature high (external input to D1and O1). The model simulations corespond to the parameter set: $w_d = 0.8$, $w_o = 0.4$ and $w_i = 0$.

The results show that the enhancement of selectivity for the task-relevant features in ITC can result from top-down information signaling the category. The model ITC neurons respond correctly to the complete stimuli, if they are strongly linked and hence cooperate with each other. Further, the categories can be represented, if the category populations are weakly linked and hence compete with each other. In addition, the network categorizes correctly if the recurrent interareal connections are differentiated for the ITC populations driven by different relevant (diagnostic) features and non-differentiated for ITC populations driven by category-irrelevant (non-diagnostic) features. Under these conditions, only the model ITC populations driven by the diagnostic features develop feature selectivity, whereas the other populations remain non-specific. Category dependent feature selectivity arises in an emergent way from the network dynamics and the weight setting.

7.5 Learning to categorize enhances the neuronal selectivity

In order to enable the network to associate in a biologically plausible manner a set of stimuli, characterized by different combinations of the feature values, with a certain category, the model network is trained using the *reward-based Hebbian learning algorithm*, that was presented in chapter 4. This section is organized as follows: Section 7.5.1 outlines the learning procedure chosen to train the model network. Section 7.5.2 discusses the capability of the model network to start learning from an unbiased weight configuration. The sample learning histories, presented in section 7.5.3, show that the model network successfully develops both a forward ITC → PFC synaptic structure, able to support correct classification, and a backward PFC → ITC synaptic structure producing a task-dependent modulation of the ITC responses. At last, section 7.5.4 analyzes the evolution of network performance during learning.

7.5.1 The learning procedure

Before presenting the learning procedure, we specify the choice for the parameters of the learning algorithm: For all simulations, the learning rates were fixed to $q_+^{reward} = q_-^{reward} = 0.01$ and $q_-^{non-reward} = 0.05$. The learning rate in the non-reward case is chosen to be greater than the learning rate in the reward case. The difference is mostly motivated by previous experimental studies on the learning and forgetting rates of a monkey performing a visuo-motor task (Asaad et al., 1998; Fusi et al., 2007). In these studies, non rewarded trials led to a quick reset of the previously memorized associations, as opposed to learning new associations that required 20-30 trials. In order to reproduce this behavior the modifications in the case of no reward had to be significantly larger than in the case of reward.

To ensure network's stability for all points in the learning process, we have to chose the connection weights between the two layers not too small so that there is information exchange between the two modeled areas and not too high so that the network does not evolve into an amplification regime where neurons lose their selectivity. Also the biological constraint of achieving realistic neuronal activities for the modeled neurons needs to be considered (see Szabo et al., 2005).

In the following simulations, the values for the synaptic strengths in the potentiated and depressed states, respectively, were chosen to $w_+^{ff} = 0.8$ and $w_-^{ff} = 0$ for the feed-forward synapses connecting the populations from ITC to PFC and $w_+^{fb} = 0.4$ and $w_-^{fb} = 0$ for the synapses in the feedback direction. The feed-forward connections are chosen double in strength on average than the feedback connections. The above choice is inspired by the idea that between cortical areas, feed-forward projections have a strong driving role, while feedback projections have a weaker modulatory role.

Using the *reward-based Hebbian learning algorithm*, presented in chapter 4, the synaptic efficacies between the ITC and PFC layers are modified after each trial (i.e. stimulus presentation) according to the network outcome (i.e. resulting network activities) in the PFC layer and a reward variable, using a simple regulatory mechanism. The learning procedure runs as follows:

Starting from a chosen initial configuration of the weight setting, different stimuli are presented to the network in a random order. For each stimulus presentation, all network internal variables are reset and the network configuration, given by the synaptic weight variables, is set to the latest *learned* configuration. Using this setting, the spiking dynamics, modeled by the 1650 coupled differential equations 2.3, is simulated for 500 ms under spontaneous activity followed by 800 ms under specific input encoding the presented stimulus. The simulation running time was chosen such that convergence to a stable configuration is reached. For the period of time when the stimulus is presented to the network, the first 300 ms are regarded as transient time, and only the last 500 ms are used to acquire the time-averaged spiking rates of each simulated neuron. These rates are used to calculate the population firing distributions, the reward variable and finally the synaptic modifications as presented in more detail in the following.

For the typical average firing rates in our simulations, the 500 ms time window used to estimate single neuron rates implies non-negligible fluctuations in the estimated values. As a consequence, despite the full synaptic connectivity and the common value of the synaptic efficacies for each synaptic population, a wide distribution of estimated firing rates in each neuron population arises for each trial. This brings non-trivial consequences in the mean-field learning dynamics, in that superimposing tails of the rate distributions for different pairs of populations can induce unwanted potentiations or depressions, thereby pushing the learning trajectory to wrong directions. We decided to keep this feature to show the robustness of the model to the finite-size effects of various kinds that would affect the dynamics in a less constrained and more realistic setting.

The firing distribution of each population i is expressed through the fraction of active neurons, n_i^a, calculated by comparing the previously computed time-averaged spiking rate of each neuron inside this population with a chosen threshold: above $8Hz$ for the ITC model layer and $14Hz$ for the PFC model layer a neuron is considered to be active. When the population encoding the correct category has more than half of the neurons active and also there are more than double as many neurons active in this population than in the other category population, the trial is assigned a reward, otherwise no reward is given. Next, for each ITC-PFC pair of specific populations, the fraction of synapses *to be* potentiated, N_{ij}^p, and *to be* depressed, N_{ij}^d, are evaluated using equations 4.1-4.4 (where $(i;j)$ or $(j;i) \in (\{D1, D2, O1, O2\}, \{C1, C2\})$). Then the presynaptic efficacies are modified using equations 4.5-4.7.

The network configuration is thus changed, through the modification of the inter-layer synaptic weights, from the initial, chosen, configuration to a final, learned, configuration. In the next subsections, it will be shown that this learning rule robustly modifies the network's free parameters to reach a final configuration where the desired associations are correctly performed.

7.5.2 Network capability to learn from an unbiased starting configuration

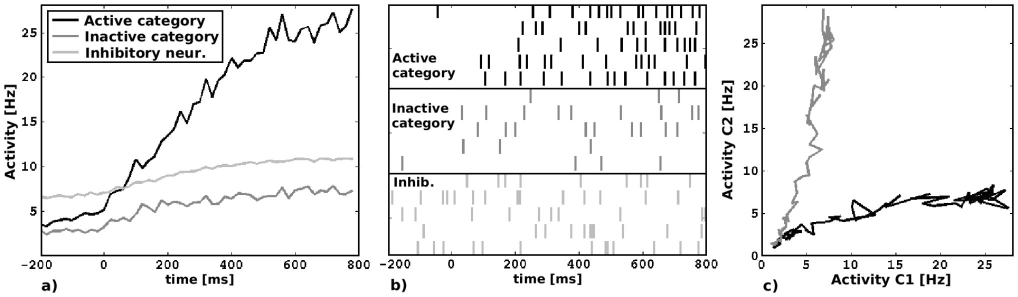


Figure 7.7: Dynamics of the PFC model layer for the initial network configuration, when all the connecting weights between the two layers are set to be equal to $(w_+ + w_-)/2$. Driven by the fluctuations in the inputs, one of the two category populations will end up being represented irrespective of the stimulus presented to the network. The first two graphs show average spiking rates over 50 trials for specific and inhibitory neurons in the model PFC layer. The responses of the specific populations were grouped based on their activity level: the higher responses were averaged into the active category response (black lines) and the lower responses were averaged into the inactive category response (dark gray lines). Population averages are depicted in (a) and spike raster plots for five neurons from each population in (b). The light gray lines represent the averaged activity over all 50 trials of the inhibitory neurons. The right most graph (c) plots the PFC layer dynamics in the phase space.

Before illustrating how learning proceeds in the system, we start by analyzing the properties of the network characterized by an uniform connectivity between the two layers. It is important that for this configuration corresponding to an untrained system, the network exhibits the capability to decide stochastically with 50% probability in response to a stimulation. Figure 7.7 presents the activities of the specific and inhibitory neurons in the PFC model area, averaged over 50 trials, for the untrained network configuration where all inter-layer connections are chosen equal. The strong competition between the populations encoding for the two categories and the stochastic fluctuations present in the network, ensure that even in the beginning of the learning process, when both categories are identically connected with the ITC layer, one of them randomly wins the competition. Hence, even the untrained network always reaches a clear decision.

The learning process will develop as follows: in the untrained system, where all weights between ITC and PFC have equal strengths, when a stimulus is presented to the ITC layer one of the populations in the category layer will, by chance, receive a stronger input, mediated through network's fluctuations. These fluctuations, which are a finite-size effect, are another needed dynamic element of the model that requires the explicit description of neural dynamics at the spiking level. Because there is strong competition between the category populations, mediated through the recurrent connectivity inside the category layer, the category being driven slightly stronger will win this competition and thus will be more strongly activated (figure 7.7). If this category happens to be the correct one, the network is rewarded. Synaptic populations that contributed most to this category's input, are hence potentiated, the ones that were driving the wrong category are weakened.

As a consequence, the next presentation of that stimulus will be more likely to activate the correct category in the future. If the chosen category was wrong, the system is not rewarded, and the synaptic populations that were driving the wrong category are weakened. As a consequence, this category is less likely to be activated by that stimulus in the future. By repeated stimuli presentations, the ITC populations representing the diagnostic features will be consistently associated with the correct category population. In contrast, the ITC populations representing non-diagnostic features will be associated with each output population with the same probability. Consequently, the network will learn to perform better and better the categorization task.

7.5.3 Learning to categorize in the model network

We start training the network from an unbiased initial configuration, where all connections between the two model layers are equal to the average synaptic strength: $w_{ij}^{ff} = (w_{+}^{ff} + w_{-}^{ff})/2 = 0.4$ for the feed-forward connections and $w_{ji}^{fb} = (w_{+}^{fb} + w_{-}^{fb})/2 = 0.2$ for the feedback connections, where $(i; j) \in (\{D1, D2, O1, O2\}, \{C1, C2\})$. Thus the initial network configuration corresponds to half of the synapses being potentiated between each pair ITC-PFC of specific populations. To this end, we will show both the time course of the average synaptic efficacies for the weights of interest, and the manifestation of the plastic synaptic rearrangement in the ITC and PFC neural activities during learning, providing evidence of a qualitative agreement with the findings of Sigala and Logothetis.

Figure 7.8 presents average network activities (over 50 consecutive trials) in three moments of the learning process: at the beginning of learning, at an intermediate point (after 200 trials) and after the convergence of the synaptic parameters (after 1500 trials). The plots in the first row were obtained by performing the same calculations as for the experimental data (figure 1.b). For each specific neuron in the ITC model layer the spiking rates for all 50 consecutive trials were grouped based on the presented stimulus values and were averaged. Each specific neuron has a different response level to the two values of each feature. The highest responses for the diagnostic feature of all specific neurons in ITC model area were averaged producing the *best Diagnostic* response. The lowest responses for the diagnostic feature of all specific neurons in ITC model area were averaged to generate the *worst Diagnostic* response. Similar calculations were done for the non-diagnostic feature.

These average activities over all ITC specific neurons are presented for three points in time in figure 7.8 top row. In the beginning of learning, there is no bias in the input to the PFC layer, the C1 and C2 populations are activated randomly with the same probability (figure 7.8.a, bottom). Thus there is no difference between the tuning of the diagnostic and non-diagnostic features (figure 7.8.a, top). As learning progresses and the synaptic weights evolve, the network now correctly solves the categorization task (figure 7.8.b, bottom). At the same time we notice the beginning of the tuning process that will be enhanced in time (figure 7.8.b, top). After convergence, the selectivity for the level of the diagnostic feature is enhanced, as compared to

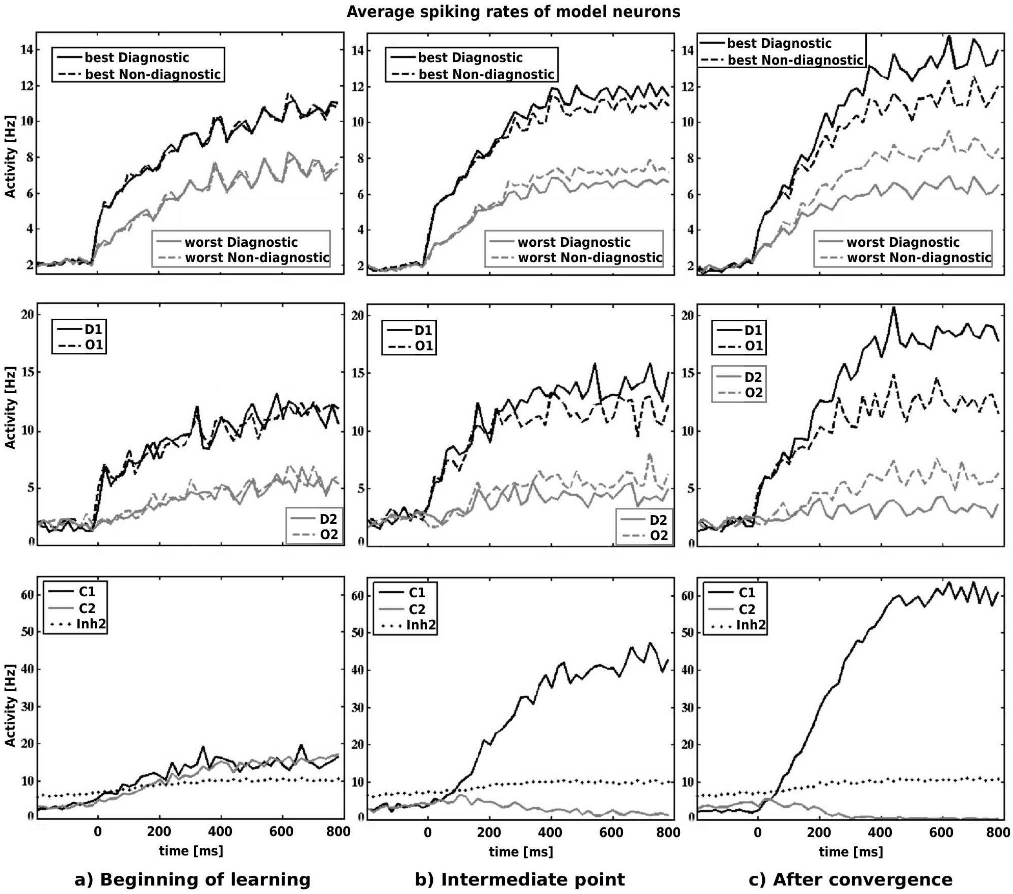


Figure 7.8: Training the spiking network: Simulation results of population activities averaged over 50 successive trials for three points in the learning process: a) in the beginning of learning; b) an intermediate point during learning (after 200 steps); c) after the weights converged to a stable configuration (1500 steps). The top row shows the average spiking rates of stimulus responsive neurons, grouped according to their best and worst responses to the levels of diagnostic and non-diagnostic features. The middle and bottom rows show the average spiking rates of the specific populations in the ITC layer (D1, D2, O1, O2) and the PFC layer (C1, C2), respectively, for the trials among the 50 successive trials where the presented stimulus was characterized by diagnostic feature high and other feature high (external input to the populations D1 and O1).

the non-diagnostic feature (figure 7.8.c, top). The activities for the best and worst diagnostic feature values are more separated than those for the best and worst non-diagnostic feature values. This result is in good qualitative agreement with the experimental results, (figure 1.b), that reflect the ITC activity after the monkeys had learned to categorize the stimuli.

The middle and bottom rows in figure 7.8 show the average spiking rates of the specific populations in the two layers for the selected trials among the 50 successive trials where the presented

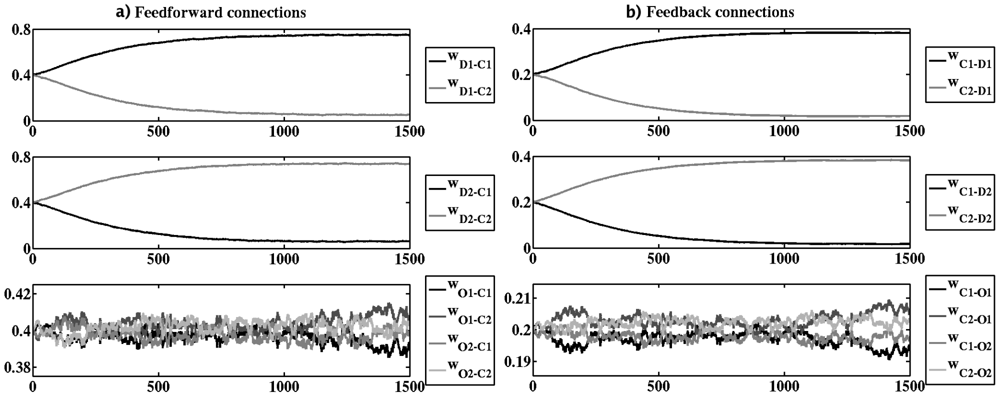


Figure 7.9: Evolution of the average synaptic weights between the selective populations of the two model layers in the process of learning. The graphs present the evolution for both feed-forward (a) and feedback (b) connections, in the case of an unbiased initial condition denoted by equal connectivity between the two layers.

stimulus was characterized by diagnostic feature high and other non-diagnostic feature high (populations D1 and O1 stimulated). Since there is no structure in the model ITC layer, the enhancement of selectivity emerges due to the top-down input from the PFC layer, which encodes the previously learned stimulus categories. The right-most column, figure 7.8.c, corresponds to the point in the learning process where the weights converged to a stable configuration.

From the time when the stimulus is presented to the network (time = 0ms in figure 7.8), the selectivity of the category specific populations (figure 7.8.c, bottom row) emerges through the competition biased by feed-forward inputs (ITC → PFC) from the specific populations of the ITC layer. Through the feedback modulatory inputs (PFC → ITC), this selectivity is transmitted afterwards to the feature-specific populations in ITC (figure 7.8.c, middle). It can be seen that in the first 100 ms after the stimulus onset the D1 and O1 (stimulated) or D2 and O2 (non-stimulated) populations do not differ in activity. Hence there is no diagnostic tuning. Only after the correct category population becomes active, the diagnostic tuning builds up.

The evolution of the synaptic weights between the two layers is presented in figure 7.9. For both feed-forward (figure 7.9.a) and feedback connections (figure 7.9.b) the links between the diagnostic features and the visual object categories are selectively modified. Weights between a diagnostic feature population and the correct category population are increased, those connecting the wrong category population are weakened. The connections between non-diagnostic feature populations and category populations remain around the starting point, corresponding to half of the synapses being potentiated. This learning case corresponds to the network learning the task from scratch, and the initial condition is referred to as **unbiased**.

Figure 7.10 shows learning trajectories for two other initial network configurations. In Figure 7.10.a the network was previously tuned for the feature which is non-diagnostic in the present task protocol, i.e. the Nose-length feature. In figure 7.10.b the network was previously tuned for both features (Eye-height and Nose-length), and also the Eye-height feature was differently

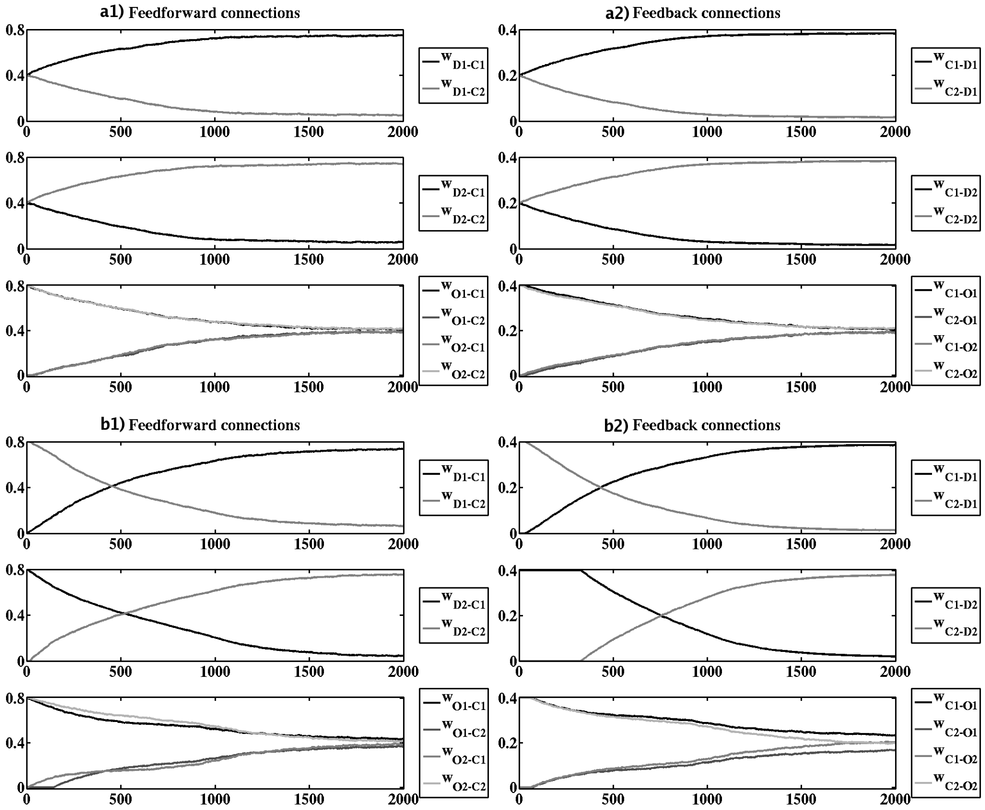


Figure 7.10: Learning trajectories for the average synaptic populations between the two layers, starting from two other initial network configurations which imply a switch in the behavioral task: In the first case the diagnostic feature becomes non-diagnostic and the non-diagnostic feature becomes diagnostic (a). In the second case, both features were important for categorization before the task switch (b). The weights for both feed-forward (a1, b1) and feedback connections (a2, b2) are presented.

associated to the two categories. Both cases correspond to a modification in the task protocol.

The results show that the connection weights with the diagnostic feature are selectively modified in the direction of increasing the synaptic strength with the corresponding category population and decreasing the synaptic strength with the other category population. The connection weights transmitting signals from the non-diagnostic feature converge to the average synaptic strength between the two layers corresponding to the unbiased situation of equal connectivity with both category populations. Because here the network needs to react to a task switch, the corresponding initial conditions are called ***biased***.

7.5.4 Evolution of network performance during learning

For the three simulation runs whose learning trajectories are presented in figures 7.9, 7.10.a and 7.10.b, we define and calculate specific parameters describing key features of the network performance along the learning process. The results are presented in figure 7.11. For each trial, e.g. point in the learning trajectory, the time-averaged activities of the model neurons, calculated over the last 500 ms of the stimulus presentation, are used to compute, in the same manner as described for the results in figure 7.8.a, the best and worst responses for the diagnostic and non-diagnostic features of all specific neurons in the ITC model layer. The evolution during learning of these responses is presented in figure 7.11-top row.

The tuning of the two features Eye-height (diagnostic) and Nose-length (non-diagnostic), is evidenced through a ***feature selectivity index*** calculated as the difference between the best and worst activities for the corresponding feature divided by their sum. The time evolution of the tuning for both features is presented in figure 7.11-middle row. The classification performance during learning, depicted in figure 7.11-bottom row, was estimated through a ***category selectivity index*** calculated as the difference between the average activities of the population encoding the presented category and the population encoding for the other category, divided by their sum.

It can be seen that for the run where the network was initially unbiased, the selectivity of the diagnostic feature starts building over time, whereas the selectivity for the non-diagnostic feature remains constant at a low value (figure 7.11.a-middle row). In the case of a switch in the behavioral task, where the network was previously tuned for the Nose-length feature, the selectivity for the non-diagnostic feature (Nose-length in the present task protocol) decreases while the selectivity for the diagnostic feature (Eye-height in the present task protocol) builds up (figure 7.11.b-middle row). In the last run, the network was previously tuned for both features and also the Eye-height feature was differently associated to the two categories. As it can be seen from figure 7.11.c-middle row, the tuning of the diagnostic feature initially goes down, as for the present task protocol the feature was previously erroneously associated to the two categories.

The learning traces show that on average, after 500 stimulus presentations, the tuning of the diagnostic feature starts building up in accord to the chosen task protocol. Also the tuning of the non-diagnostic feature goes down as it becomes irrelevant for behavior. From the bottom row in figure 7.11 we remark that the network performance in classification reaches a high value after 300 stimulus presentations for the case when the network was initially unbiased, after 600 stimulus presentations in the case of a modification in the behavioral task of only one variable (non-diagnostic previously tuned) and after 1000 stimulus presentations in the case of a modification in the behavioral task of two variables (non-diagnostic initially tuned and diagnostic erroneously tuned). In the latter case, the diagnostic feature of the task to be learned at present was diagnostic before as well, but with the opposite mapping to the categories. This is reflected by a negative category selectivity index (the network is worse than guessing) in the initial phase of learning. It is remarkable that even this severe re-orientation towards a completely new task is robustly achieved by the learning network.

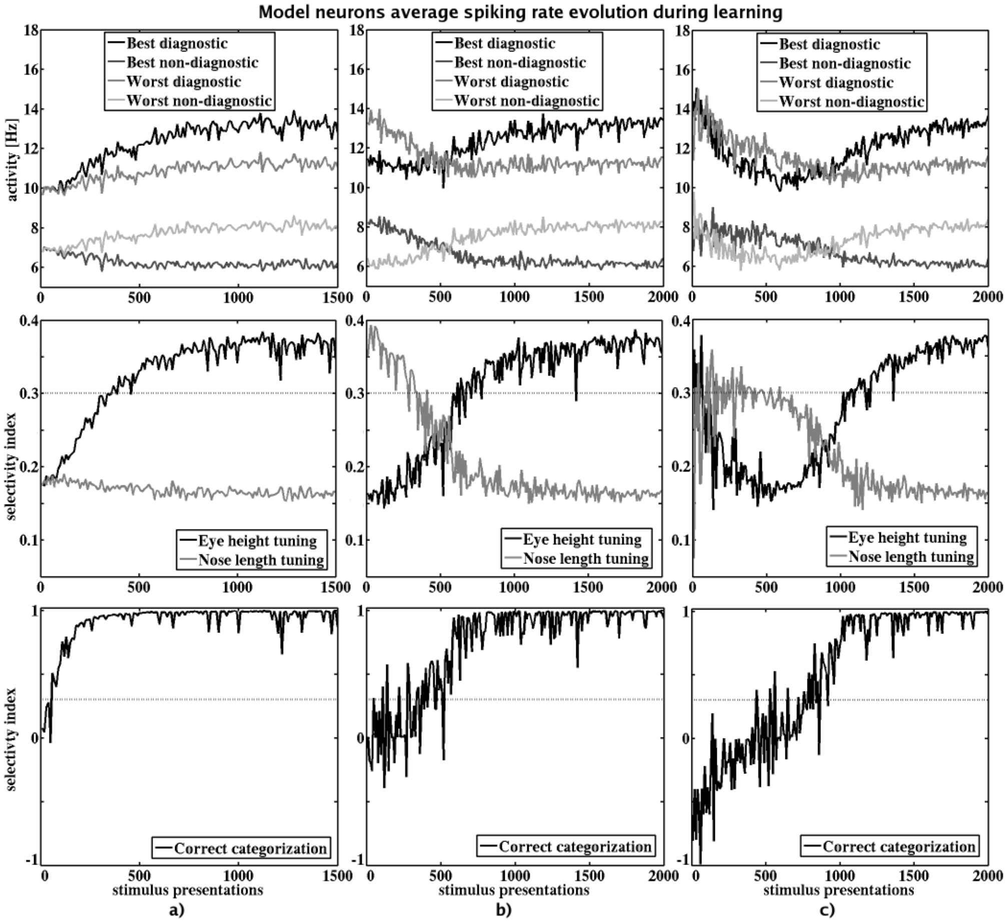


Figure 7.11: Task performance evolution during learning for the three learning histories of figures 7.9, 7.10.a and 7.10.b: (a) learning from scratch (figure 7.9) and (b, c) learning after task switches (figure 7.10.a and b). The top row shows the time evolution of the best and worst responses for the diagnostic and non-diagnostic features of all specific neurons in the ITC model area. The middle row shows the time evolution of the tuning for both features. The bottom row shows the classification performance during learning. For details see text.

As it can be seen in figures 7.9 and 7.10, some weights change in a similar manner. We conclude that we do not need all 16 free parameters (describing the feed-forward and feedback synaptic populations between the two layers' specific populations) to describe network's behavior and reduce the parameter space to the important dimensions only. We define four effective weights: w_d, w_i, w_{o1} and w_{o2}. w_d relates to the average connection weight between D1-C1 and D2-C2 populations, which for our task protocol corresponds to the connections between the diagnostic feature values and the corresponding categories. w_i relates to the average connection weight between D2-C1 and D1-C2 populations, corresponding to the connections between the diagnostic feature values and the non-corresponding categories. Similar w_{o1} and w_{o2} are defined

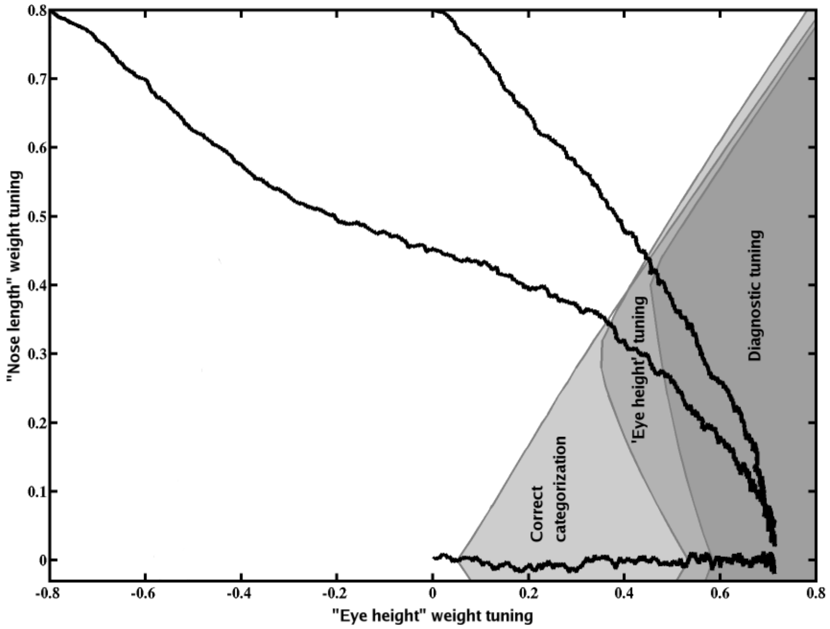


Figure 7.12: Evolution of Eye-height and Nose-length tuning during learning, for three different initial network configurations. They all converge to the same regime corresponding to high selectivity of the diagnostic feature Eye-height and low selectivity for the non-diagnostic feature Nose-length (bottom right corner of the graph). The network's performance is characterized through an extensive exploration of network's effective parameters using the mean-field formulation. The network performed *Correct categorization* in the light gray area, presented *Eye-height tuning* in the medium gray area and showed *Diagnostic tuning* in the dark gray area. For details see text.

for the connections of the non-diagnostic feature with the two categories. Because the feed-forward weights are on average twice as big as the feedback ones, we weighted the feedback connection strengths with a factor of 2. This simplifies the representation by making a simple correspondence between the effective weights and the feed-forward or feedback ones.

$$
\begin{aligned}
w_d &= \frac{1}{4}(w_{D1-C1} + w_{D2-C2} + 2 \cdot w_{C1-D1} + 2 \cdot w_{C2-D2}) \\
w_i &= \frac{1}{4}(w_{D1-C2} + w_{D2-C1} + 2 \cdot w_{C2-D1} + 2 \cdot w_{C1-D2}) \\
w_{o1} &= \frac{1}{4}(w_{O1-C1} + w_{O2-C2} + 2 \cdot w_{C1-O1} + 2 \cdot w_{C2-O2}) \\
w_{o2} &= \frac{1}{4}(w_{O1-C2} + w_{O2-C1} + 2 \cdot w_{C2-O1} + 2 \cdot w_{C1-O2})
\end{aligned}
$$

Using this simplification, we can capture an important result, namely that the network shows robustness to the starting point in the parameter space. All three learning trajectories converge to the same final network configuration as illustrated in figure 7.12 by the black lines. The two axes reflect the tuning of the two features Eye-height and Nose-length expressed through the variables $w_d - w_i$, denoting *Eye-height tuning*, and $w_{o1} - w_{o2}$, denoting *Nose-length tuning*, calculated using the formulas above. The zero values correspond to equal connectivity of the feature values to the two categories, which is equivalent to no tuning for that feature. High values correspond to different connectivities between the two values of the feature with the two categories, which is equivalent to feature tuning. All traces converge to the area where there is no selectivity for the Nose-length feature and high selectivity for the Eye-height diagnostic feature.

In order to illustrate the regimes that the network crosses during training, the effective parameter set describing the excitatory weight setting between the model cortical layers is explored using the mean-field formulation (chapter 5). Similar to the analysis from section 7.3, each point in the parameter space, was simulated for all four possible stimulus presentations. The resulting mean firing rates of the specific populations were evaluated by calculating three suggestive parameters, as described below. From the activity of the four specific populations in the ITC layer, two parameters that measured the *Eye-height tuning* and *Diagnostic tuning* were calculated.

Eye-height tuning measures, for a specified set of effective weights, the selectivity for the Eye-height feature through a selectivity index calculated as the difference between the mean firing rates for the best Eye-height feature value and worst Eye-height feature value, divided by their sum. Note that the best and worst values are calculated in the same way as for the results presented in figure 7.8 and 7.11, with the only difference that the population average activities are used instead of single neuron activities.

Diagnostic tuning measures the difference between the selectivity index for the diagnostic feature (which corresponds to Eye-height tuning) and the selectivity index for the non-diagnostic feature (which corresponds to Nose-length tuning). The Eye-height and Nose-length tuning were also calculated for the single neuron activities, as presented in figure 7.11 middle-row.

From the activity of the two category populations in PFC, another parameter *Correct categorization* was calculated as the difference between the mean firing rates of the population encoding the correct category of the presented stimulus and the population encoding the other category, divided by their sum. This parameter was also calculated for the single neuron activities, as presented in figure 7.11 bottom-row. It measures the level of association of the presented stimulus and corresponding category.

For each of these parameters we chose a threshold that marked the limit where the requirements of having the respective selectivity or categorization are still satisfied, as shown in figure 7.11 middle and bottom rows by the horizontal dotted lines. The network was defined to show *Eye-height tuning* when the best Eye-height value response is twice or greater than the worst Eye-height value response. We say that the network shows *Diagnostic tuning* when the selectivity for the diagnostic feature is twice or greater than the selectivity for the non-diagnostic feature. Also a correct categorization corresponds to an activation of the correct category more than twice greater than the activation of the other category. In figure 7.12 we plotted the areas where these three performance criteria were satisfied. We notice that the learning trajectories converge to the area in the explored parameter space where all three conditions were fulfilled (the darkest gray area in the graph).

7.6 Discussion of the results

The study presented in this chapter, formulates a biologically-inspired two-layer spiking neural network that accounts for the enhancement of selectivity in ITC neurons for stimulus features which are relevant for a learned visual categorization task (as described in Sigala & Logothetis, 2002). The behavioral task consisted of categorizing a set of schematic images based on their varying features, from which only some of them, named diagnostic, were relevant to solve the task. The experimental measurements from ITC (Sigala & Logothetis, 2002) showed that, after training, the neuronal selectivity to the diagnostic features was enhanced as compared to the selectivity to the other, non-diagnostic, features (figure 1.b). Other studies of perceptual learning and visual encoding suggest that the tuning of sensory neurons can be mediated by top-down information and that prefrontal cortex can be associated with processing of category related information. Trying to explain the neural substrate of the experimentally observed phenomena, the present study tests the assumption that the enhancement of selectivity to the behavioral relevant features in ITC might be determined by higher-level cognitive feedback from category encoding neurons possibly residing in PFC and demonstrates a learning scenario which robustly produces such a selective enhancement.

The model is constructed as a biologically-inspired two-layer network of integrate and fire neurons that is able to capture the temporal dynamics of experimentally measured neural spiking rates and thus represents an appropriate choice in the search to explain the neural substrate of the observed cortical effects. Such networks, using the conceptual framework of Biased-competition and cooperation, successfully accounted for different aspects of visual attention and working memory in context dependent tasks. A key new feature of the model proposed in this study is that the biases needed to guide the competition between different neuronal populations are internally generated using recurrent signals produced inside the network. As a stimulus is presented to the network, the sensory inputs (coming from lower visual processing areas) activate the neurons in the ITC model layer and are propagated through feed-forward connections to the PFC model layer. This bottom up input from ITC biases the competition between the category encoding populations. The winning category expresses the monkey's decision (as in Wang, 2002) and influences the activity of the neurons in the ITC model layer such that, after a successful learning, they become selective for the behaviorally relevant features.

Based on an extended parameter exploration, we found that the described effect could be achieved for a specific structure of the connections between the feature encoding layer (ITC) and the category encoding layer (PFC). Namely, the ITC model neurons activated by a feature value determinant for categorization are strongly connected to the associated category and weakly connected to the other category. ITC model neurons activated by a task-irrelevant feature are connected to the category neurons with an average weight value, not significantly changed during training. For this particular structure of the interlayer connectivity, the network was able to reproduce the experimental results, by achieving a high selectivity of the ITC model neurons for the diagnostic feature and a low selectivity for the non-diagnostic feature. By setting all weights equal, so that no structure exists in the interlayer connectivity, the network could not reproduce the enhancement in selectivity for the diagnostic features.

By construction, having identical inputs from the lower sensory areas encoding for the presented diagnostic and non-diagnostic feature values and no structure in the connectivity of ITC specific model neurons, a single layer model (only ITC) would induce identical tuning for the diagnostic and non-diagnostic features. The selective tuning of the ITC model area emerges then only through top-down modulatory signals from the PFC model area where the learned categories are encoded. A side effect of the identical inputs to the ITC model neurons, as can be seen from the results in figure 7.8, is that the model shows some selectivity also for the non-diagnostic feature as compared to the experimental results that show almost no selectivity for the different values of the non-diagnostic feature (figure 7.1.b).

The network is trained using a biologically inspired reward-based Hebbian algorithm, which robustly modifies the connections between the feature encoding layer (ITC) and the category encoding layer (PFC) and ensures convergence for different initial network configurations. This simple model is constructed with a small number of excitatory and inhibitory neurons for each cortical area. The finite size effect creating random fluctuations in the population firing rates enables the spontaneous transition, in the beginning of learning when the two categories are equally connected to the ITC specific populations, of one category to win the competition. Increasing the network size reduces the probability of these spontaneous transitions.

The results show that the learning dynamics converges to a stable fixed point denoting the setting where the ITC model neurons activated by a diagnostic feature value are strongly connected to the associated category and weakly connected to the other category, and the ones activated by task-irrelevant features, are equally connected to the category neurons with an average weight. This structure of the interlayer connectivity was shown to be able to reproduce the experimental data, by achieving a high selectivity of the ITC neurons for the diagnostic feature and low selectivity for the non-diagnostic feature.

The modeling approach from the present work can be used to generate some experimentally testable predictions: Learning to provide the correct categorization, by modification of the ITC → PFC synapses, occurs before the backward plastic reorganization of the PFC → ITC synapses which produces the modulation effect on the selectivity in ITC neurons response. We infer this from the results presented in Figure 7.8: for the intermediate point in the learning process, the network categorizes correctly but the responses in ITC are not yet tuned to the behavioral relevant features. Also the other two cases of contextual task change show that the ITC tuning occurs after correct categorization is achieved. Such a scenario would be consistent with the prediction that the tuning effect is an epiphenomenon of the primary synaptic process that allows to achieve the correct categorization.

Another interesting experimental scenario is suggested by the analysis presented in Figs 7.10 and 7.11, where different initial conditions corresponding to switches in the behavioral task, are chosen: in one case exchanging the role of the diagnostic feature and in the other case switching from both features being diagnostic to only one. We can infer that the number of stimulus presentations needed until convergence to the final learned network configuration, i.e. the time needed by the network to learn the new association, increases with the number of modifications made in the task protocol. Figure 7.11 shows that for convergence in the case of unbiased starting configuration around 500 trials are needed (figure 7.11.a), in the case of a simple change where the non-diagnostic feature was previously associated to the two categories around 900 trials are needed (figure 7.11.b), and for a more complex change where in addition to the change in the latter experiment the diagnostic feature was also differently associated to the two categories around 1300 trials are needed (figure 7.11.c).

8 Conclusions

Aiming to understand the way in which our brain perceives, reacts, reasons and takes decisions, powerful methods are needed to investigate the mechanisms and principles underlying the higher-level cognitive functions of human behavior, like selective attention, working memory and concept formation. An important approach is provided by the analysis of the biological neural networks, characterized by a large number of dense interconnected cells, at the *system level*. This idea is driven by the assumption that the representation of information in the brain is distributed across several cortical areas and that the coherent description of information is achieved at a global level through the intricate inter-areal connectivity. For this, *multi-areal neurodynamical models* based on knowledge from cognitive neuroscience and techniques from computational neuroscience and artificial neural networks are being developed.

The neurocognitive modeling approach presented in this thesis considers neurodynamical computational models inspired from the structure and properties of the nervous system. Neural modules from specific cortical areas having specific functionalities are modeled as *recurrent networks of spiking neurons* that include important features of the biological cortical structures. The neurons are implemented as *Leaky Integrate-and-fire neurons* including *nonlinear synaptic current dynamics* and following a biologically inspired description – given by the use of realistic biophysical time constants, latencies, and conductances. At this level of detail the model allows to perform a thorough study of the realistic time scales and firing rates involved in the evolution of the modeled neural activity. Consequently, the simulated neuronal dynamics, that putatively underlies cognitive processes, can be quantitatively contrasted with experimental data from a wide variety of sources: single cell measurements, fMRI imaging measurements, psychophysical results, effects of pharmacological agents and effects of damage to the neural system. This kind of powerful computational networks was shown to exhibit a rich spiking dynamics similar to that of the neurophysiological cortical data (Sima & Orponen, 2003) and was successfully applied in this work to study attentional and cognitive phenomena as effects of multiareal recurrent processing.

The architecture of the model networks is set up using the concept of population coding and follows the theoretical framework of *Biased-competition and cooperation*, which assumes that in a cortical area the conflicting partial representations *compete* with each other in order to be represented, while the *related partial representations cooperate* with each other, mutually reinforcing their activities. The model neurons encoding the same information are gathered in populations receiving common bottom-up stimulus-related inputs and common top-down signals encoding the attentional condition or task-relevance of the presented information. The populations are interconnected with stronger or weaker synaptic weights in order to implement competition or cooperation between them.

Different weight settings give rise to different overlapping operational regimes of the recurrent networks, like for example: pure amplification mode, selective (competition) mode, correlation facilitation (cooperation) mode and working memory (persistent activity) mode. Thus, the grouping first of neurons into populations and then of different specific populations, encoding related information, into correlated structures (through cooperative weight settings), together with implementing competition between the populations representing anti-correlated information, for both bistability and single stability regimes, is assumed to represent a general mechanism used in the brain to manage complex computation. In general, cooperation can be hypothesized as the basis of categorization (binding information together referring to the same category), of different associations (for example between sensations) and even of mental manipulation.

The Biased-Competition and Cooperation mechanism can be seen as an extension of the early competitive-cooperative mechanism implemented in the Kohonen self organizing networks (maps). They implement a competitive learning process using a winner-take-all strategy between distant neighbors and a cooperative learning process using a neighborhood function between close neighbors (Kohonen, 2001). The competitive-cooperative mechanism can explain experimental findings in early visual system and its extension is recently applied, as described in this work, to cognitive modeling. The Biased-Competition and Cooperation mechanism can then represent a fundamental principle of brain operation.

The contribution of this work is twofold: First, different operational regimes corresponding to different functional modes of first one-layer and then two-layer recurrent network are revealed. Second, two functional neurocomputational models are proposed that, guided by cortical activity measurements from recent neurophysiological experiments on behaving animals, specify possible neuronal mechanisms underlying cognitive phenomena like selective visual attention and selective neuronal tuning.

The first study introduces a one-layer neurodynamical model that investigates how the effect of attentional filtering could arise from a weak modulatory bias which mediates the cortical context (Szabo et al., 2004). In the context of the proposed biologically inspired minimal model, the study shows how competition and cooperation, biased by behaviorally relevant information, operate within a single model area. Relevant parameter explorations reveal regimes where the network shows different modes of operation: selective working memory, attentional filtering, pure competition and non-competitive amplification.

Attentional filtering represents a particularly strong attentional effect, in which the context gates sensory input in an all-or-none fashion. The simulation results showed that both cooperation and competition are needed for reproducing the referred attentional filtering effect, and suggest them as fundamental principles for the neural basis of cognitive processes in the higher-level cortical areas. Also, the presence of non-specific neurons in the model area was important to assure the stability of the activity in the network, as well as the stability of several important operational regimes identified in the present work and can be related to biological evidence, where for any possible state of the nervous system there is always a large number of neurons not involved in coding the present particular state. A recent study, Stetter (2006), showed analytically that the level of the global spontaneous background current dynamically tunes the functional mode of the neural circuit. The stability of operational modes might then be an important functional role for distributed representation and sparse coding in the brain.

The study presents also two experimentally testable predictions of the model: First, an increase in the level of dopamine is related with a progressive impairment of the attentional filtering effect and thus of the task performance. And second, the ability to filter out stimuli with basis on attention is assumed to degrade as the number of distracting stimuli increases.

An early model investigating the mechanisms of visual attentional that follows the Biased-Competition hypothesis, was performed by Reynolds & Desimone (1999). Their implementation, which is not biologically motivated, consisted of a feed-forward model in which the top-down biases – encoding the attentional state – modulated the strength of the connections coming from neurons selective for the attended stimulus. Another quantitative neural model of Reynolds et al. (1999) for visual attention in areas V2 and V4 studied the effects of biased-competition using a simple feed-forward competitive neural network. An alternative biologically motivated implementation of Spratling and Johnson uses a rate-based approach for the neural activity and considers that the top-down signals modulate in a multiplicative way the bottom-up sensory-driven inputs to the model neurons, which compete by sending lateral inhibitory signals to their neighbors (Spratling & Johnson, 2004b,a). In comparison, the biologically motivated model presented in this work addresses the neural mechanisms of attentional filtering using a recurrent network consisting of a large number of spiking neurons, where the top-down biases effect on the activations of the selective excitatory neurons, which compete through local inhibitory neurons. The study presented here suggests cooperation - besides competition - as a second fundamental principle for the neural basis of cognitive processes in the prefrontal cortex leading to the *Extended Biased Cooperation-Competition Hypothesis*. The present study provided also detailed parameter space explorations of the different dynamical attractors of the model network in order to characterize its working regimes. The simple one-area network could show quantitatively different kinds of operational modes given by the overlapping of a small set of fundamental mechanisms: competition, cooperation, persistent activity and input amplification.

The second study introduces a two-layer neurodynamical model for learning visual categorization that investigates how the task-dependent shaping of neuronal selectivity could arise from top-down biases encoding the presented category information, and how this shaping evolves during learning the categorization task (Szabo et al., 2006). In the context of the proposed biologically-relevant minimal bi-areal model, the study shows how competition and cooperation operate over two different cortical modules. A key new feature of the proposed model is that the biases needed to guide the competition between different neuronal populations are internally generated using recurrent signals produced inside the network.

The model assumes that the neurons in ITC, modeled as receiving feature specific sensory inputs, will develop during learning stronger or weaker connections to the category-encoding model neurons from PFC to which they are consistently associated or not. Using a reward-based Hebbian learning mechanism, the proposed model shows a robust change in the bi-areal attractor dynamics towards increased performance of the entire system, even in the case of an unexpected switch in the behavioral task. The results show that the learning dynamics converges to a stable fixed point characterized by the setting where the ITC diagnostic selective neurons are strongly connected to the associated category and weakly connected to the other category, and the non-diagnostic selective neurons are equally connected to the category neurons with an average weight value.

After learning, the results show that the top-down influence of the model category encoding neurons will enhance the selectivity of the responses in the ITC model layer for the behavioral relevant (diagnostic) features of the presented stimuli, explaining the experimentally observed effect. This suggests that task-dependent feature tuning might be a neuronal correlate of top-down hypothesis-driven visual perception and that the perceptual representation in visual areas can be strongly affected by the interaction with other areas which are devoted to higher cognitive functions.

The study also presents two experimentally testable predictions of the model: First, learning to provide the correct categorization, by modification of the ITC → PFC synapses, occurs before the backward plastic reorganization of the PFC → ITC synapses, which produces the modulation effect on the selectivity of the ITC neurons response. Such a scenario would be consistent with the prediction that the tuning effect is an epiphenomenon of the primary synaptic process that allows to achieve the correct categorization. And second prediction is that the time needed by the network to learn a new association, increases with the number of modifications made in the task protocol.

An alternative attentional-gated reinforcement learning paradigm was recently introduced for the Sigala and Logothetis perceptual learning task by Roelfsema & Van Ooyen (2005). They suggest that the selective tuning of ITC neurons could arise from learning the feedforward connections coming to ITC from lower visual processing areas. They use a feed-forward model receiving a generic attentional feedback signal that modulates the learning of the feedforward synaptic weights. As opposed to this implementation, the model presented in this thesis considers a recurrent network in which the activities of the category encoding neurons directly affect the ITC neuronal activities, thus explaining the ITC tuning effect. An experimental scenario that could distinguish between the two network predictions is PFC cooling. After learning the categorization task, the influence of the top-down signals from the category encoding neurons could be measured: in case of a feedforward learning scenario the effect would reside with the same strength, while in the case of a recurrent network the effect it is predicted to decrease or even vanish.

Recent work in the field of cognition of Spratling and Johnson (Spratling & Johnson, 2004b,a, 2006) use computational models of interacting cortical regions to study the possible effects of top-down signals in visual information processing, like visual attention and perceptual learning. They show that the competition between neural representations and the neural activity modulation are common features of cortical information processing that could result from common mechanisms, like the top-down biasing from higher cortical regions. Although using a biologically motivated architecture, their connectionist model is based on simple forms of rate-based response functions and a dendritic inhibition model that allows negative weight values. Other studies from O'Reilly, Cohen, Braver and colleagues (Frank et al., 2001; O'Reilly et al., 2002; Rougier & O'Reilly, 2002; Cohen et al., 2004) explore the biological properties of the PFC underlying working memory and cognitive control by implementing several connectionist models of the PFC based on the Leabra framework[1] (O'Reilly & Munakata, 2000) in which cortical areas are organized according to different levels of abstraction. The models involve competition implemented by lateral inhibition and biased by top-down inputs from PFC, and take into account the role of the dopamine system believed to regulate the learning process and control the update of the PFC representations. Although biologically motivated, this framework uses a rate-based approach for the neural activity which is only valid under stationary conditions, and does not consider an explicit neuronal spiking mechanism and thus is not able to capture the non-stationary temporal dynamics of neural activity.

Compared to the high-level modeling of the above mentioned connectionist type models that implement artificial dynamics and learning schemes (Roelfsema & Van Ooyen, 2005; Miller & Cohen, 2001; O'Reilly et al., 2002), the model presented in this work implements biologically inspired neuronal spiking and synaptic mechanisms and a biologically plausible Hebbian learning algorithm, which make it a realistic model of the actual dynamical processes occurring in the brain. An important feature of the proposed model, which exhaustively analyzes how competition and cooperation operate within a two-area model, is the internal generation of the biases needed for the competition process in the PFC model area.

In order to facilitate a consistent theoretical analysis of the underlying mechanisms of neural computation, one has to consider in the future also discussing the effects and implications of several choices taken in this modeling approach, for example: full versus partial connectivity, local versus global inhibition, the activation and influence of the external neurons, the number and connectivity of the non-specific neurons, different types of neurotransmitters.

Also it should be noted that present spiking neuron models focus on the fast chemical interactions and assume that all synapses coming from one neuron can be either excitatory or inhibitory. Future models, that will take into account also longer time-scale chemical interactions, will enable extra functionalities – like a switch in the activity mode of a neuron, and consequently a switch in the functionality of an entire module, like for example from an attentional module to an selective working memory module.

In conclusion, the presented neurodynamical models describe and analyze a small number of mechanisms, like biased-competition and cooperation, that combined give rise to a number of basic features of cortical processing, namely input amplification, selective filtering, correlation facilitation, attentional suppresion and selective tuning. These features allow relevant insights about the neurodynamical mechanisms underlying higher-level cognitive functions of the brain and represent fundamental building blocks of large neurodynamical models. We belive that these fundamental mechanisms can be used to form very powerful computational systems and represent an important contribution to the modern cognitive neuroscience.

[1] the Leabra framework is a coherent set of basic neural processing and learning mechanisms

A Appendices

A.1 List of common used abbreviations

AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ANN	artificial neural network
EEG	electroencephalography
EPSC	excitatory post-synaptic current
EPSP	excitatory post-synaptic potential
fMRI	functional magnetic resonance imaging
GABA	gamma-aminobutyric acid
IPSC	inhibitory post-synaptic current
IPSP	inhibitory post-synaptic potential
ISI	inter spike interval
ITC	inferotemporal cortex
LIF	leaky integrate-and-fire
LIF-NS	leaky integrate-and-fire with nonlinear synaptic dynamics
LTD	long-term depression
LTP	long-term potentiation
mM	millimolar (concentration)
ms	millisecond
mV	millivolt
nF	nanofarad
NMDA	N-methyl-D-aspartate
PET	positron emission spectroscopy
PFC	prefrontal cortex

A.2 List of common used symbols

V_{rest}	resting membrane potential
V_E	reversal potential for excitatory synapses
V_I	reversal potential for inhibitory synapses
V_{thr}	excitation or spiking threshold
V_{reset}	reset membrane potential
$I_{tot}(t)$	total afferent input to a neuron
$I_{ext}(t)$	total external input to a neuron
$f(I_{tot})$	activation or transfer function
τ	integration time constant
w_{ij}	synaptic strength from neuron j to neuron i
C_m	total membrane capacitance (LIF neuron)
R_L	total membrane leak resistance (LIF neuron)
V_m	membrane potential (LIF neuron)
τ_m	membrane time constant (LIF neuron)
g_L	membrane leak conductance (LIF neuron)
V_L	leakage reversal potential (LIF neuron)
θ	firing threshold (LIF neuron)
$\delta(t)$	Dirac delta function
τ_{ref}	absolute refractory period (LIF neuron)
$H(t)$	Heaviside step function
$s_j^R(t)$	fraction of opened synaptic ion channels of type R at site j (LIF-NS neuron)
τ_{decay}^R	synaptic decay time constant for receptor type R (LIF-NS neuron)
τ_{rise}^R	synaptic rise time constant for receptor type R (LIF-NS neuron)
g_{max}^R	maximum synaptic conductance mediated by receptor type R (LIF-NS neuron)
N_E	number of excitatory recurrent connections (LIF neuron)
N_I	number of inhibitory recurrent connections (LIF neuron)
N_{ext}	number of external excitatory connections (LIF neuron)
ν	firing rate
$w+$	weight setting between neurons within a specific population
w'	weight setting implementing cooperation
$w-$	weight setting implementing competition
q_+	learning rate for potentiation
q_-	learning rate for depression
w_+	total connection strength between two populations when all synapses are potentiated
w_-	total connection strength between two populations when all synapses are depressed
N^p	fraction of synapses to be potentiated
N^d	fraction of synapses to be depressed
C_{ij}	current fraction of potentiated synapses from the synaptic population ij

Bibliography

Abbott, L. & van Vreeswijk, C. (1993). Asynchronous states in networks of pulse-coupled neuron, *Phys Rev E* **48**: 1483–88.

Abeles, A. (1991). *Corticonics*, Cambridge University Press, New York.

Abeles, A. (1994). Firing rates and well-timed events, *in* E. Domany, K. Schulten & J. L. van Hemmen (eds), *Models of Neural Networks II*, Springer, New York.

Adrian, E. (1914). The all-or-none principle in nerve, *J Physiol (London)* **47**: 460–74.

Almeida, R., Deco, G. & Stetter, M. (2004). Modular biased-competition and cooperation: a candidate mechanism for selective working memory, *Eur J Neurosci* **20**: 2789–803.

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

Amit, D. & Brunel, N. (1997b). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex, *Cereb Cortex* **7**(3): 237–52.

Amit, D., Brunel, N. & Tsodyks, M. (1994). Correlations of cortical Hebbian reverberations: experiment versus theory, *J Neurosci* **14**: 6435–45.

Amit, D. & Tsodyks, M. (1991a). Quantitative study of attractor neural networks retrieving at low spike rates I: Substrate - spikes, rates and neuronal gain, *Network* **2**: 259–73.

Amit, D. & Tsodyks, M. (1991b). Quantitative study of attractor neural networks retrieving at low spike rates II: Low rate retrieval in symmetric networks, *Network* **2**: 275–94.

Amit, J. & Fusi, S. (1994). Dynamic learning in neural networks with material synapses, *Neural Comput* **6**: 957.

Asaad, W., Rainer, G. & Miller, E. (1998). Neural activity in the primate prefrontal cortex during associative learning, *Neuron* **21**: 1399–407.

Ben-Yishai, R., Lev Bar-Or, R. & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex, *Proc Natl Acad Sci USA* **92**(9): 3844–48.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. & Warland, D. (1991). Reading a neural code, *Science* **252**(5014): 1854–57.

Braitenberg, V. & Schütz, A. (1991). *Anatomy of the Cortex*, Springer Verlag, Berlin.

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

Brunel, N., Chance, F., Fourcaud, N. & Abbott, L. (2001). Effects of synaptic noise and filtering on the frequency response of spiking neurons, *Phys Rev Lett* **86**: 2186–89.

Brunel, N. & Hakim, V. (1999). Fast global oscillations in networks of integrate-and-fire neurons with now firing rates, *Neural Comput* **11**: 1621–71.

Brunel, N. & Sergi, S. (1998). Firing frequency of leaky integrate-and-fire neurons with synaptic currents dynamics, *J Theor Biol* **195**: 87–95.

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

Chelazzi, L. (1999). Serial attention mechanisms in visual search: a critical look at the evidence, *Psychol Res* **62**: 195–219.

Chelazzi, L., Miller, E., Duncan, J. & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex, *Nature* **363**(6427): 345–47.

Cohen, J., Aston-Jones, G. & Glizenrat, M. (2004). A systems-level perspective on attention and cognitive control: guided activation, adaptive gating, conflict monitoring, and exploitation versus exploration, *in* M. Posner (ed.), *Cognitive Neuroscience of Attention*, Guilford Press, pp. 71–90.

Connors, B., Malenka, R. & Silva, L. (1988). Two inhibitory postsynaptic potentials, and GABA-A and GABA-B receptor-mediated responses in neocortex of rat and cat, *J Physiol* **406**: 443–68.

Corchs, S., Stetter, M. & Deco, G. (2003). System-level neuronal modeling of visual attentional mechanisms, *Neuroimage* **20**: 143–60.

Dale, H. (1935). Pharmacology and nerve endings, *Proc R Soc Med* **28**: 319–32.

Dayan, P. & Abbott, L. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, MIT Press, Cambridge, MA.

Deco, G. & Rolls, E. (2002). A neurodynamical theory of visual attention: comparisons with fMRI and single-neuron data, *in* J. Dorronsoro (ed.), *Proceedings of the Artificial Neural Networks Conference - ICANN 2002 LNCS 2415*, Springer Verlag, Berlin, pp. 3–8.

Deco, G. & Rolls, E. (2003). Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex, *Eur J Neurosci* **18**(8): 2374–90.

Deco, G. & Rolls, E. (2004). A neurodynamical cortical model of visual attention and invariant object recognition, *Vision Res* **44**: 621–44.

Deco, G., Rolls, E. & Horwitz, B. (2004). ”What” and ”Where” in visual working memory: a computational neurodynamical perspective for integrating fMRI and single-neuron data, *J Cogn Neurosci* **16**: 683–701.

Del Giudice, P., Fusi, S. & Mattia, M. (2003). Modeling the formation of working memory with networks of integrate-and-fire neurons connected by plastic synapses, *J Physiol (Paris)* **97**: 659–81.

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention, *Annu Rev Neurosci* **18**: 193–222.

Destexhe, A., Mainen, Z. & Sejnowski, T. (1994). Synthesis of models for excitable membranes, synaptic transmission and neuromodulation using a common kinetic formalism, *J Comput Neurosci* **1**: 195–230.

Destexhe, A., Mainen, Z. & Sejnowski, T. (1998). Kinetic models of synaptic transmission, *in* C. Koch & I. Segev (eds), *Methods in Neural Modeling*, second edn, MIT Press, Cambridge, MA, pp. 1–25.

Duncan, J. (1996). Cooperating brain systems in selective perception and action, *in* T. Inui & J. L. McClelland (eds), *Attention and Performance XVI*, MIT Press, Cambridge, MA, pp. 433–58.

Duncan, J. & Humphreys, G. (1989). Visual search and stimulus similarity, *Psychol Rev* **96**(3): 433–58.

Eggert, J. & van Hemmen, J. (2000). Unifying framework for neuronal assembly dynamics, *Phys Rev E* **61**(2): 1855–74.

Everling, S., Tinsley, C., Gaffan, D. & Duncan, J. (2002). Filtering of neural signals by focused attention in the monkey prefrontal cortex, *Nat Neurosci* **5**(7): 671–76.

Fine, I. & Jacobs, R. (2002). Comparing perceptual learning tasks: a review, *J Vis* **2**: 190–203.

Fishman, G. (1995). *Monte Carlo: Concepts, Algorithms, and Applications*, New York: Springer Verlag.

Fourcaud, N. & Brunel, N. (2002). Dynamics of the firing probability of noisy integrate-and-fire neurons, *Neural Comput* **14**: 2057–110.

Frank, M., Loughry, B. & O'Reilly, R. (2001). Interactions between the frontal cortex and basal ganglia in working memory: A computational model, *Cogn Affect Behav Neurosci* **1**: 137–60.

Freedman, D., Riesenhuber, M., Poggio, T. & Miller, E. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization, *J Neurosci* **23**: 5235–46.

Fusi, S. (2002). Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates, *Biol Cybern* **87**(5–6): 459–70.

Fusi, S., Annunziato, M., Badoni, D., Salamon, A. & Amit, D. (2000). Spike-driven synaptic plasticity: theory, simulation, VLSI implementation, *Neural Comput* **12**: 2227–58.

Fusi, S., Asaad, W., Miller, E. & Wang, X. (2007). A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales, *Neuron* **54**(2): 319–33.

Fusi, S. & Mattia, M. (1999). Collective behavior of networks with linear (VLSI) integrate and fire neurons, *Neural Comput* **11**: 633–52.

Gerstner, W. (1995). Time structure of the activity in neural network models, *Phys Rev E* **51**: 738–58.

Gerstner, W. (2000). Population dynamics of spiking neurons: fast transients, asynchronous states and locking, *Neural Comput* **12**(1): 43–89.

Gerstner, W. & Kistler, W. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*, Cambridge University Press.

Goldstone, R. (1998). Perceptual learning, *Annu Rev Psychol* **49**: 585–612.

Hebb, D. (1949). *The organization of behavior - A neurophysiological theory*, John Wiley, New York.

Hestrin, S., Sah, P. & Nicoll, R. (1990). Mechanisms generating the time course of dual component excitatory synaptic recorded in hippocampal slices, *Neuron* **5**: 247–53.

Hille, B. (2001). *Ionic channels of excitable membranes*, 3rd edn, Sinauer Associates.

Hodgkin, A. & Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve, *J Neurophysiol* **117**: 500–44.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities, *Proc Natl Acad Sci USA* **75**: 2554–58.

Hubel, D. & Wiesel, T. (1959). Receptive fields of single neurons in the cat's striate cortex, *J Physiol* **148**: 574–91.

Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *J Physiol (Lond)* **160**: 106–54.

Jack, J., Noble, D. & Tsien, R. (1983). *Electric current flow in excitable cells*, Oxford University Press.

Jahr, C. & Stevens, C. (1990). Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics, *J Neurosci* **10**: 3178–82.

Kandel, E., Schwartz, J. & Jessel, T. (eds) (2000). *Principles of neural science*, 4th edn, McGraw Hill, New York.

Knight, B. (1972). Dynamics of encoding in a population of neurons, *J Gen Physiol* **59**: 734–66.

Koch, C. (1999). *Biophysics of Computation - Information processing in single neurons*, Computational Neuroscience, Oxford University Press, New York.

Koch, K. & Fuster, J. (1989). Unit activity in monkey parietal cortex related to haptic perception and temporary memory, *Exp Brain Res* **76**: 292–306.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd edn, Springer, Berlin.

Lapicque, L. (1907). Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarization, *J Physiol (Paris)* **9**: 620–35.

Law-Tho, D., Hirsch, J. & Crepel, F. (1994). Dopamine modulation of synaptic transmission in rat prefrontal cortex, *Neurosci Res* **21**: 151–60.

McCormick, D. (1989). GABA as an inhibitory neurotransmitter in human cerebral cortex, *J Neurophysiol* **62**: 1018–27.

McCornick, D., Connors, B., Lighthall, J. & Prince, D. (1985). Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex, *J Neurophysiol* **54**: 782–806.

McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity, *Bull Math Biophys* **5**: 115–33.

Metropolis, N. & Ulam, S. (1949). The monte carlo method, *J Amer Stat Assoc* **44**(247): 335âĂŞ41.

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

Miller, E., Gochin, P. & Gross, C. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus, *Brain Res* **616**: 25–29.

Miller, K. (1994). A model for the development of simple cell receptive fields and orientation columns through activity-dependent competition between on- and off-center inputs, *J Neurosci* **14**: 409–41.

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

Moran, J. & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex, *Science* **229**: 782–84.

Motter, B. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli, *J Neurophysiol* **70**: 909–19.

Nykamp, D. & Tranchina, D. (2000). Fast neural network simulations with population density methods, *Neurocomputing* **32**: 487–92.

Omurtag, A., Knight, B. & Sirovich, L. (2000). On the simulation of large populations of neurons, *J Comput Neurosci* **8**(1): 51–63.

O'Reilly, R. & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*, MIT press, Cambridge, MA.

O'Reilly, R., Noelle, D., Braver, T. & Cohen, J. (2002). Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control, *Cereb Cortex* **12**: 246–57.

Petersen, C., Malenka, R., Nicoll, R. & Hopfield, J. (1998). All-or-none potentiation at CA3-CA1 synapses, *Proc Natl Acad Sci USA* **95**(8): 4732–37.

Reynolds, J., Chelazzi, L. & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas v2 and v4, *J Neurosci* **19**: 1736–53.

Reynolds, J. & Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem, *Neuron* **24**: 19–29.

Ricciardi, L. (1977). *Diffusion Processes and Related Topics in Biology*, Springer-Verlag, Berlin.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. (1997). *Spikes: Exploring the Neural Code*, MIT Press, Cambridge, MA.

Risken, H. (1984). *The Fokker-Plank Equation*, Springer, New York.

Roelfsema, P. & Van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification, *Neural Comput* **17**(10): 2176–214.

Rolls, E., Aggelopoulos, N., Franco, L. & Treves, A. (2004). Information encoding in the inferior temporal visual cortex: contributions of the firing rates and the correlations between the firing of neurons, *Biol Cybern* **90**(1): 19–32.

Rolls, E. & Deco, G. (2002). *Computational Neuroscience of Vision*, Oxford University Press.

Rougier, N. & O'Reilly, R. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching, *Cogn Sci* **26**(4): 503–20.

Salin, P. & Prince, D. (1996). Spontaneous GABA-A receptor mediated inhibitory currents in adult rat somatosensory cortex, *J Neurophysiol* **75**(4): 1573–88.

Sheperd, G. (ed.) (1998). *The synaptic organization of the brain*, 4th edn, Oxford University Press.

Sigala, N. & Logothetis, N. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex, *Nature* **415**: 318–20.

Sima, J. & Orponen, P. (2003). General-purpose computation with neural networks: A survey of complexity theoretic results, *Neural Comput* **15**(12): 2727–78.

Simons, D. (2000). Attentional capture and inattentional blindness, *Trends Cogn Sci* **4**(4): 147–55.

Softky, W. & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs, *J Neurosci* **13**: 334–50.

Spitzer, H., Desimone, R. & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance, *Science* **240**: 338–40.

Spratling, M. & Johnson, M. (2004a). A feedback model of visual attention, *J Cogn Neurosci* **16**(2): 219–37.

Spratling, M. & Johnson, M. (2004b). Neural coding strategies and mechanisms of competition, *Cognitive Systems Research* **5**(2): 93–117.

Spratling, M. & Johnson, M. (2006). A feedback model of perceptual learning and categorization, *Vis Cogn* **13**(2): 129–65.

Spruston, N., Jonas, P. & Sakmann, B. (1995). Dendritic glutamate receptor channels in rat hippocampal CA3 and CA1 pyramidal neurons, *J Physiol* **482**: 325–52.

Stetter, M. (2002). *Exploration of Cortical Function*, Kluwer Scientific Publishers, Dordrecht.

Stetter, M. (2006). Dynamic functional tuning of nonlinear cortical networks, *Phys Rev E* **73**(3): 031903.

Stetter, M., Lang, E. & Obermayer, K. (1998). Unspecific long-term potentiation can evoke functional segregation in a model of area 17, *Neuroreport* **9**: 2697–702.

Stetter, M., Müller, M. & Lang, E. (1994). Neural network model for the coordinated formation of orientation preference and orientation selectivity maps, *Phys Rev E* **50**: 4167–81.

Sutton, R. (1988). Learning to predict by the methods of temporal differences, *Machine Learning* **3**: 9–44. erratum p.377.

Sutton, R. & Barto, A. (1998). *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA.

Szabo, M., Almeida, R., Deco, G. & Stetter, M. (2004). Cooperation and biased competition model can explain attentional filtering in the prefrontal cortex, *Eur J Neurosci* **19**(7): 1969–77.

Szabo, M., Almeida, R., Deco, G. & Stetter, M. (2005). A neuronal model for the shaping of feature selectivity in IT by visual categorization, *Neurocomputing* **65–66**: 195–201.

Szabo, M., Deco, G., Fusi, S., Del Giudice, P., Mattia, M. & Stetter, M. (2006). Learning to attend: modeling the shaping of selectivity in infero-temporal cortex in a categorization task, *Biol Cybern* **94**(5): 351–65.

Thorpe, S., Fize, D. & Marlot, C. (1996). Speed of processing in the human visual system, *Nature* **381**(6582): 520–22.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I. & Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval, *Nature* **401**: 699–703.

Treves, A. (1993). Mean-field analysis of neuronal spike dynamics, *Network* **4**: 259–84.

Treves, A., Panzeri, S., Rolls, E., Booth, M. & Wakeman, E. (1999). Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli, *Neural Comput* **11**(3): 601–31.

Troyer, T. & Miller, K. (1997a). Integrate-and-fire neurons matched to physiological F-I curves yield high input sensitivity and wide dynamic range, *in* J. Bower (ed.), *Computational Neuroscience: Trends in Research 1997*, Plenum Press, pp. 197–201.

Troyer, T. & Miller, K. (1997b). Physiological gain leads to high ISI variability in a simple model of a cortical regular spiking cell, *Neural Comput* **9**: 971–83.

Tuckwell, H. (1988). *Introduction to Theoretical Neurobiology*, Cambridge University Press, Cambridge.

Uhlenbeck, G. & Ornstein, L. (1930). On the theory of brownian motion, *Phys Rev* **36**: 823–41. Reprinted in 'Selected Papers on Noise and Stochastic Processes', Ed. N. Wax, Dover (New York), 1954, pp.93–111.

Wang, X. (2002). Probabilistic decision making by slow reverberation in cortical circuits, *Neuron* **36**: 955–68.

Wilson, F., Scalaidhe, S. & Goldman-Rakic, P. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex, *Science* **260**: 1955–58.

Wilson, F., Scalaidhe, S. & Goldman-Rakic, P. (1994). Functional synergism between putative gamma-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex, *Proc Natl Acad Sci USA* **91**(9): 4009–13.

Wilson, H. & Cowan, J. (1972). Excitatory and inhibitory interactions in localized populations of model neurons, *Biophys J* **12**(1): 1–24.

Wilson, H. & Cowan, J. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue, *Kybernetik* **13**(2): 55–80.

Xiang, Z., Huguenard, J. & Prince, D. (1998). GABA-A receptor mediated currents in interneurons and pyramidal cells of rat visual cortex, *J Physiol* **506**: 715–30.

Zheng, P., Zhang, X., Bunney, B. & Shi, W. (1999). Opposite modulation of cortical N-methyl-D-aspartate, receptor-mediated responses by low and high concentrations of dopamine, *Neuroscience* **91**: 527–35.