Structure and Function of Neural Networks
Controlling Evolved Autonomous Agents

Thesis submitted to the Senate of the Hebrew University
in partial fulfillment
of graduate requirements for the degree
“Doctor of Philosophy”

by

Ranit Aharonov

July 2003
Prepared under the supervision of

Prof. Hanoch Gutfreund
and
Prof. Eytan Ruppin
Contents

List of Figures iv

Abstract v

1 Introduction 1
   1.1 Thesis Outline 1
   1.2 Evolved Autonomous Agents (EAAs) 2
   1.3 EAAs and the Development of New Tools for Neuroscience 5
   1.4 Localization of Function 6
   1.5 Temporally Asymmetric Hebbian Plasticity 7

2 Evolved Autonomous Agents (EAAs): A Neuroscience Perspective 12
   2.1 Introduction 12
   2.2 The Model 13
   2.3 Agents’ Performance 16
   2.4 Exploration vs. Grazing: The Emergence of Command Neurons 17
   2.5 Location-driven Command Neurons: The SP model 18
   2.6 Memory-driven Command Neurons: The S model 20
       2.6.1 Activity Maps of the Command Neuron 20
       2.6.2 Command Neuron Activity Profiles 21
       2.6.3 The Underlying Stochastic Memory 23
   2.7 Neural Network Structure of An S-type Agent 25
   2.8 Discussion 26

3 Localization of Function in EAAs: Lessons Learned for the Lesion Approach in Neuroscience 35
   3.1 Introduction 35
   3.2 The Functional Contribution Analysis 36
       3.2.1 Definitions 36
       3.2.2 The FCA Algorithm 37
       3.2.3 Experimental Protocol 38
       3.2.4 Fast FCA: Adaptive Lesion Selection 38
   3.3 FCA of Evolutionary Neurocontrollers 39
       3.3.1 Selecting the Lesioning Method 41
       3.3.2 FCA of Agents’ Performance 43
4 High-Dimensional Localization of Function: Re-thinking the Concept of Contribution

4.1 Introduction ................................................. 57
4.2 High-Dimensional FCA ...................................... 58
  4.2.1 Motivation and Definition ............................ 58
  4.2.2 Selecting Compound Elements ....................... 59
4.3 High-Dimensional Neurocontroller Analysis .......... 62
  4.3.1 Compound Elements and Interactions .............. 62
  4.3.2 Generalization ........................................ 64
  4.3.3 Corrected Contributions .............................. 66
4.4 Discussion .................................................. 67

5 Non-Linear Temporally Asymmetric Hebbian Plasticity 68
5.1 Introduction ................................................ 68
5.2 Materials and methods ..................................... 70
  5.2.1 Temporally Asymmetric Hebbian plasticity ....... 70
  5.2.2 Mean Synaptic Dynamics ............................. 71
  5.2.3 Integrate-and-Fire Neuron ......................... 73
  5.2.4 Linear Poisson Neuron ............................... 74
  5.2.5 Mean Synaptic Dynamics for the Linear Poisson Neuron .... 75
  5.2.6 Generating Correlated Inputs ..................... 76
  5.2.7 Measuring the Performance of Learning Rules .... 76
5.3 Results ...................................................... 77
  5.3.1 The Emergence of Synaptic Patterns by Symmetry Breaking in TAH Learning ......................... 78
  5.3.2 Synaptic Representation of Input Correlations .... 96
5.4 Discussion .................................................. 103

6 Summary and Discussion ..................................... 114

Bibliography .................................................... 119
List of Figures

1.1 The paradigm of Evolutionary Autonomous Agents .......................... 4
1.2 Temporally asymmetric learning window ................................. 9

2.1 The model environment .................................................. 14
2.2 A typical evolutionary run ............................................... 15
2.3 Performance benchmarks ............................................... 16
2.4 Selectivity of command neuron activity ................................ 18
2.5 Location-dependent activity maps of the command neurons .......... 20
2.6 Activity of a stochastic memory command neuron, I ................. 21
2.7 Activity of a stochastic memory command neuron, II ............... 22
2.8 Models of command neuron activity .................................... 24
2.9 Performance decrease due to selective lesioning ...................... 25
2.10 An S-type neuro-controller ............................................ 26

3.1 Adaptive Lesioning vs. random configuration selection in FCA ...... 40
3.2 Visualization of the Adaptive Lesioning algorithm .................... 40
3.3 Comparison of biological and stochastic lesioning .................... 42
3.4 Contribution values of the agents’ neurons ........................... 44
3.5 Comparison between the single lesion approach and the FCA I .... 46
3.6 Comparison between the single lesion approach and the FCA II .... 47
3.7 The Contribution Matrix .................................................. 48
3.8 Contribution values for Grazing and Exploration ..................... 49
3.9 Comparison of synaptic contributions computed by the FCA and the synaptic weights ..................................................... 52
3.10 Agent performance as a function of synaptic lesioning degree ....... 54
3.11 Network backbone ...................................................... 55

4.1 Comparison of compound element selection algorithms in FCA ...... 61
4.2 2D and 3D FCA ....................................................... 63
4.3 Major interactions and synapses ....................................... 64
4.4 Generalization in 2D-FCA ............................................... 65
4.5 corrected contributions .................................................. 66

5.1 Non-linear TAH ......................................................... 71
5.2 Firing rate responses of a neuron driven by uncorrelated Poisson input processes ........................................................... 82
5.3 Symmetry breaking in the linear neuron driven by uncorrelated Poisson input processes ..................................................... 84
5.4 Symmetry breaking in the Integrate-and-Fire neuron driven by uncorrelated Poisson input processes ........................................ 87
5.5 Symmetry breaking in the linear neuron driven by two equally sized correlated groups .................................................. 90
5.6 Symmetry breaking in the Integrate-and-Fire neuron driven by two equally sized correlated groups ........................................ 92
5.7 Symmetry breaking in non-linear and additive TAH learning ........... 94
5.8 Effect of the updating parameter on the learned synaptic distribution ... 97
5.9 Sensitivity to the within-group correlation ...................................... 98
5.10 Sensitivity evaluated at uniformly spaced input delays .................. 101
5.11 Mapping between synaptic delays and equilibrium weights ............. 102
Abstract

This thesis deals with the relation between the structure and function of neural networks. The complex behaviors manifested by neural systems are an emergent phenomenon stemming from the interactions between many simple units – the neurons. The exact nature of behavior depends on the structure of the network, i.e. on the way in which the neurons are connected, and on the rules governing their interactions. The high complexity of biological neural networks, together with the experimental constraints on measuring both the structural and functional variables of the system, make it extremely difficult to infer their functional organization from experimental data. We therefore chose to conduct this study within the relatively tractable framework of artificial neural networks evolved to successfully control behaving autonomous agents, and harness the power of this approach to develop new quantitative methodologies for analyzing experiments performed on biological networks.

We first demonstrate (chapter 2) that the paradigm of Evolved Autonomous Agents (EAAs) is interesting and relevant for neuroscientific investigations, thus substantiating its later use as a vehicle for developing experimental analysis methodologies. We show that EAAs evolved to perform simple life-like tasks of foraging and navigation, match and even outperform memory-less algorithms designed especially for the same tasks. Analyzing the evolved neurocontrollers revealed the emergence of non-trivial structure-to-function relations, specifically, a command neuron switching the dynamics of the net-
work between radically different behavioral modes. The activity of this neuron reflects a map of the environment, acting as a location- and orientation-dependent cell. When no information about the position is available to the agent, the command neuron’s activity is based on a spontaneously evolving short-term memory mechanism, which underlies its apparent place-sensitive activity. The neural network ”brains” of the agents in these experiments have no predefined network structure. They emerge from a simulated evolutionary process, which allows the researcher to specify a desired functional behavior, rather than the structure of the network. Thus, they are on one hand unconstrained, and complex enough to be interesting, while on the other hand fully accessible to the researcher, allowing “experiments” that are impossible to perform on a biological system. The consistency of the structures emerging in our experiments and their similarity to known biological mechanisms, testify to the relevance of the evolutionary paradigm as a research tool for modeling and studying structure to function relationship in nervous systems.

In Chapters 3 and 4 we employ the evolved neurocontrollers to develop a novel quantitative approach (the FCA) for clearly defining and measuring the rather vague yet fundamental notion of the contribution of the neuronal units (structural level) to the tasks performed by the network (functional level). The proposed definition is accompanied by a method to compute the contribution based on data obtained from lesioning experiments. Quantitatively defining the contribution, allows one to naturally define a measure for the degree of localization of function in a neural network, a fundamental and debated issue in neuroscience. Applying FCA to the analysis of evolved neurocontrollers demonstrates that FCA portrays a stable set of neuronal contributions and accurate predictions on the performance levels of lesioned networks. Furthermore, we demonstrate that applying FCA for a detailed synaptic analysis of
the neurocontroller connectivity network delineates its main functional backbone – an important step in bridging between the structural and functional levels. The results presented in these chapters seriously question the adequacy of the classical single lesion analysis traditionally used in neuroscience, and show that using lesioning experiments to decipher even simple neuronal systems definitely requires a more rigorous multi-lesion analysis. The High-Dimensional FCA described in Chapter 4, generalizes the approach to identify sets of elements which interact in complex non-linear ways, and have a functional significance as a unit. Keeping in mind the goal of analyzing biological neural systems, Chapters 3 and 4 describe important conclusions and insights gained from these FCA studies about the lesion approach for studying neural structures in general. In particular, we discuss the effect of the form and degree of lesions inflicted upon the network, on the results of the analysis. Further, we demonstrate that the classical notion of contribution is unclear in many cases, and that to clearly discuss localization of function, it is crucial to find the correct functional decomposition of the network.

Beyond the innate structure of neural systems evolved to achieve the complex behaviors required for survival, neural systems are plastic. They employ learning, constantly reshaping their structure to better cope with the changing environment. To understand the roles of synaptic plasticity and learning in forming and shaping evolved neural structures, one must be able to parameterize the learning process in the genome, and allow the evolutionary process to tune learning concomitantly with tuning the initial state of the neural network. In Chapter 5 we investigate the performance of synaptic learning in representing information about the environment within the neural system. Our approach combines a synaptic plasticity rule based on recent empirical measurements with a novel model of synaptic updating functions. We
demonstrate that by correctly tuning the weight-dependence of the synaptic updating functions, the learning dynamics is stable enough to ensure faithful learning of the input activity structure, yet introduces enough synaptic competition to allow for sufficient sensitivity to changes in the input activity. Finally, we introduce a novel sensitivity measure for the performance of synaptic learning rules. Applying this measure we investigate the effect of the balance between stabilizing, competitive, and cooperative synaptic drifts on the fidelity of the synaptic learning process. This novel Hebbian learning rule is particularly fit for evolutionary studies as the crucial weight-dependence of learning is incorporated through a single parameter.
Chapter 1

Introduction

1.1 Thesis Outline

One of the fundamental questions in Neuroscience is the relation between the structure of the neural substrate underlining behavior and its function, i.e., the observed behavior. This relation has been studied from many different angles and at various levels of the system, including single and multiple cell recordings [Abeles, 1991], functional MRI [Cohen and Bookheimer, 1994], lesioning studies [Grobstein, 1990], and behavioral psychophysics [Gescheider, 1997]. However, the question is extremely complex and there seems to be a lack of a unifying terminology and methodology. Moreover, our ability to make inferences from experimental data to the functional organization of the system is at best partial.

One of the difficulties is that the systems studied are very complex, and our ability to study them is hindered by experimental constraints. For this reason, I chose to study the relation between structure and function in artificial neural networks which are relatively simple, and fully transparent in the sense that any “experimental” information can be extracted from them. However, it is imperative that the artificial networks are unconstrained by our biases and previous notions, and thus must not be designed by us. To obtain these requirements, I have focused on Evolved Autonomous Agents (EAAs). As detailed below (1.2), these are behaving agents, controlled by an artificial neural network that is evolved using genetic algorithms to solve life-like tasks. As demonstrated in Chapter 2, EAAs are not only successful in solving relatively complex tasks, but also possess interesting internal structures, making them excellent candidates to study the relation between the structure and the function of the network.

Relying on these artificially evolved networks, I then turned my attention to developing quantitative methodologies for the analysis of neural networks. The full information available about the controlling network in EAAs, combined with the relative simplicity of these networks, enabled me to quantitatively attack vague notions, such as the significance
of neural structures to functional tasks, and the distribution of function within the neural substrate (1.3). The Functional Contribution Analysis (FCA) presented in Chapter 3 clearly defines the notion of the contribution of a neural element to a task, and offers a way to compute it based on data from lesioning experiments. From this definition of contributions, indices for the localization of function and specialization of neural elements naturally emerge. The High-Dimensional FCA described in Chapter 4, generalizes the approach to identify sets of elements which interact in complex non-linear ways, and have a functional significance as a unit. Chapters 3 and 4 describe important conclusions and insights gained from these FCA studies about the lesion approach for studying neural structures.

A major future challenge in the study of EAAs is understanding the roles of synaptic plasticity and learning in forming and shaping evolved neural structures. A prerequisite for studying this question is finding a suitable learning rule which can be parametrized in the genome, allowing the evolutionary process to tune learning. A basic challenge with unsupervised learning is the balance between stability and sensitivity of the learning process. Learning must be stable enough to faithfully represent information, and yet be sensitive enough to react to changes in the input activity. In Chapter 5 we propose a parameterization of Temporally Asymmetric Hebbian learning (1.5) based on recent empirical measurements combined with a novel model of synaptic updating functions. As shown there, by correctly tuning the weight-dependence of synaptic updating, learning can be optimized to balance between stability and sensitivity and achieve faithful representation of the input activity structure. In the presented learning rule, the weight-dependence is tuned by a single parameter, making it an excellent candidate rule for implementing unsupervised learning in EAAs.

1.2 Evolved Autonomous Agents (EAAs)

Recent years have witnessed a growing interest in the study of neurally-driven Evolved Autonomous Agents (EAAs). These studies, part of the field of Evolutionary Computation, Evolutionary Robotics and Artificial Life (see [Mitchell, 1996, Langton, 1995, Fogel, 1995, Adami, 1998, Nolfi and Floreano, 2000] for general introductory textbooks), involve agents that live in a simulated environment and autonomously perform typical animat tasks like gathering food, navigating, evading predators, and seeking prey and mating partners. Each agent is controlled by an Artificial Neural Network (ANN) "brain". This network receives and processes sensory inputs from the surrounding environment and governs the agent’s behavior via the activation of the motors controlling its actions. The agents can be either software programs living in a simulated virtual environment, as is the case in our studies, or hardware robotic devices. Their controlling networks are developed
via Genetic Algorithms (GAs) [Mitchell, 1996] that apply some of the essential ingredients of inheritance and selection to a population of agents that undergo evolution.

An important aspect of EAA research is the embodiment of the evolved neural networks, i.e. developing and studying the controlling networks in a system containing a “body”. As shown in various EAA studies, this embodiment is of paramount importance for providing constraints that reduce the degeneracies involved in the neural-to-behavioral mappings [Beer et al., 1999b]. Moreover, such embodiment may have an important functional role, as complex sensory tasks can be solved not only by complex sensory processing, but also by simpler motor manipulation. For example, turning around an object to inspect it from different directions can facilitate recognition of an otherwise ambiguous pattern [Scheier et al., 1998, Beer, 2000]. This embodiment is a strong motivation to study biological systems within the framework of EAs, as biologically relevant neural network models should be studied in a comprehensive system containing not only the networks themselves but also the “bodies” in which they reside.

In a typical EAA experiment, genetic algorithms are applied to a population of agents that are evolved over many generations to best survive in a given environment (see Figure 1.1). In general, the sensors and motors can themselves be evolved, but we focus on agents with fixed sensorimotor capabilities, in which the genome solely encodes the controlling neural network. This genome is usually a vector of integer or real numbers, which can be transcribed by a predefined developmental plan into a full description of an ANN. The encoding of the network within the genome can be a ‘direct’ one, where every gene specifies a synaptic connection. This simple encoding scheme is used throughout this work. ‘Indirect’ encodings include a program for network construction, which can be quite elaborate. For example, in ‘ontogenic’ encodings, the genome expresses a program for cell division and axonal migration [Cangelosi et al., 1994, Nolfi and Parisi, 1995], whereas hybrid genotype-to-phenotype mappings encode both the synaptic weights and their learning rules [Floreano and Urzelai, 2000, Floreano and Urzelai, 2001].

The initial population consists of agents with randomly drawn genomes. The genomes are transcribed to form agents that are placed in the environment for a given amount of time, after which each agent receives a fitness score that designates how well it performed the evolutionary task. A new population of agents is generated by selecting the fittest agents of the previous generation and letting them mate, i.e., form new genomes via genetic recombination and mutations that introduce additional variation in the population (Figure 1.1). Typically, this evolutionary “search” is repeated for many generations until the agents’ fitness reaches a plateau and further evolutionary adaptation dramatically slows down. The result is a final population of best-fitted agents, whose emergent behavior and underlying neural dynamics can now be thoroughly studied in “ideal conditions”: One has full control on manipulating the environment and other experimental conditions.
More important, one has complete knowledge of the agent’s behavior on one hand, and the controlling network’s architecture and dynamics, on the other.

![Diagram](attachment:image.png)

Figure 1.1: *The paradigm of Evolutionary Autonomous Agents.***

Current EAA studies have been able to successfully evolve artificial networks controlling agents performing non-trivial behavioral tasks (see [Meyer and Guillot, 1994, Kodjabachian and Meyer, 1998, Yao, 1999, Guillot and Meyer, 2000, Guillot and Meyer, 2001] for reviews). These networks are *less biased* than conventional neural networks used in neuroscience modeling as their architecture is in general not pre-designed. They are the *emergent result* of a simplified and idealized process that models the evolution of intelligent, neurally-driven life forms. This fundamental property naturally raises the possibility of using these agents as a vehicle for studying basic questions concerning neural processing. This potential is further substantiated by two observations: Feasibility: the small size of the evolved networks, the simplicity of their environments and behavioral tasks, coupled with the full information available regarding the model dynamics, form conditions that help make the analysis of the network’s dynamics an amenable task. Relevancy: since the networks are evolved in biologically motivated animat environments, their analysis may potentially reveal interesting insights into the workings of biological systems.

Chapter 2 addresses the relevancy of EAAs to neuroscience research. The evolved agents described there manage to perform a non-trivial navigation task at levels which match or surpass the best reactive (i.e. memory-less) algorithms written for solving the problem. They also outperform solutions obtained via Reinforcement Learning techniques. From a neuroscience perspective, several noticeable results were obtained in that work. Although different successful agents had evolved very different network architectures to
cope with the same task, the behavioral strategies they used were very similar. This was true for agents with no ‘genetic similarity’ between them. Closer inspection revealed that despite the difference in the exact wiring of their neurocontrollers, these networks had similar functional components. In particular, a common structure of a command neuron was identified. This command neuron controlled switching between two distinct types of behaviors – grazing and exploration, resembling command neuron circuitry identified in several invertebrates. We found that the cues that formed the basis for the activation of the command neuron differed depending on the sensory information available to the agent. In scenarios characterized by poor sensory information the agents evolved a compensatory stochastic memory mechanism. These results, which are detailed in Chapter 2, testify to the relevance and adequacy of EAA models to neuroscientific research.

1.3 EAAs and the Development of New Tools for Neuroscientific Research

The use of EAA models as a neuroscience research tool is a very complex and challenging scientific endeavor. Even the relatively simple task of understanding the neural processing of small, fully transparent EAAs is a very difficult feat. The dual goal of analyzing evolved neurocontrollers is to uncover principles of neural processing in animat and biological nervous systems, and to develop new methods for their analysis.

Several studies have attempted to analyze neurocontrollers of evolutionary autonomous agents. In [Chiel et al., 1999b, Beer et al., 1999b], a rigorous, quantitative analysis of the dynamics of central pattern generator (CPG) networks evolved for locomotion, has been developed. The networks evolved are of very small size, composed of 3,4 or 5 neurons. A high-level description of the dynamics of these CPG networks was developed, based on the concept of a dynamical module: a set of neurons that have a common temporal behavior, making a transition from one quasi-stable state of firing to another together. Dynamical modules give new insights to CPG operation, describing them in terms of a finite state machine, and enabling a rigorous analysis of their robustness to parameter variations. In [Floreano and Mondada, 1996], the activity of internal neurocontroller neurons as a function of a robot’s location and orientation was charted by a simple form of receptive-field measurement. Neuronal functioning was generally highly distributed, but a specific interneuron that had an important role in path planning was also identified. Other researchers have studied the effects that clamping of neuronal activity has on the robot’s behavior. For example, inducing rotation, straight-line motion, or more complex behaviors such as smooth tracking of moving targets [Harvey et al., 1994]. The command neurons described in Chapter 2 were discovered by studying the agent’s behavior following single-neuron lesions and by receptive field analysis. Finally, a more “procedural” kind
of ablation, in which different processes (and not just units or links) are systematically
canceled out was recently employed in [Stanley and Miikkulainen, 2001]. Overall, these
studies have provided only glimpses of the processing in these networks.

1.4 Localization of Function

A fundamental challenge in understanding a nervous system is to characterize how its
various functions are localized or distributed among the system elements. These elements
may be single neurons, neuronal assemblies or even cortical areas, depending on the scale
on which one chooses to analyze the system. Even simple nervous systems are capable
of performing multiple and unrelated tasks, often in parallel. Each task recruits some
of the elements of the system, and often the same element participates in several tasks.
This poses a difficult challenge when one attempts to identify the roles of the network
elements, and to assess their contributions to the different tasks.

The localization of specific tasks in a nervous system is conventionally done in neuro-
science either by recording the activity of the system elements during behavior, or by
measuring the deficit in performance after lesioning specific elements. Inferring signifi-
cance from recording unit activity during behavior is problematic because correlation of
neuronal activity with performance does not necessarily identify causality. For example, it
is possible that some areas raise their activity while a given task is performed not because
they significantly contribute to the processing of that task, but rather because they are
activated by other regions that do play an important role in this task processing. To try
and uncover the units and regions whose activation really underlies a specific function or
behavior, neuroscientists have traditionally adopted a system analysis approach, where
the normal operating modes of a system are studied by lesioning and perturbing its pro-
cessing in many possible ways. These abnormal manipulations are aimed at providing
deeper insights into the system’s normal, regular, operation.

Albeit, the vast majority of lesion studies in neuroscience have been single lesion stud-
ies, in which only one network component is abolished at a time (e.g. [Squire, 1992, Farah,
1996]). Such single lesions are very limited in their ability to reveal the significance of
elements which interact in complex ways. A straightforward conceptual example demon-
strating this is when two elements have a high degree of redundancy with respect to the
processing of a function to which they equally contribute. Then, lesioning either element
alone will not reveal its true significance, since no reduction in the performance of that
function will occur. The problematic and limited value of single lesion analysis has already
been widely noted in the neuroscience literature [Sprague, 1966, Farah, 1990, Sitton et
al., 2000, Young et al., 2000]. One classical example discussed is that of the paradoxi-
cal lesioning effect [Sprague, 1966]. In this paradigmatic case, lesioning area A alone is
harmful but lesioning area A given that area B is lesioned is beneficial, hence the apparent “paradox”. Importantly, it demonstrates that looking at a single lesion alone may be misleading, as the beneficial influence of an area depends on the general state of the system. Last but not least, current lesioning studies in neuroscience have yielded mostly qualitative measures, lacking the ability to precisely quantify the contribution of a unit to the performance of the system and, perhaps even more importantly, to predict the effect of new, multiple-site lesions.

Addressing these challenges, we have developed a Functional Contribution Analysis (FCA), described in Chapters 3 and 4. The FCA framework gives a rigorous, operative definition for the neurons’ (or, cortical regions’) contributions to the performance of the system in various tasks, and an algorithm to measure these contributions via multi-lesion analysis. Acknowledging that single lesions are insufficient for localizing functions in neural systems, the FCA framework assumes an existing data set of multiple lesions to a neural system and their corresponding system performance scores in a given set of tasks. The FCA harnesses the full power of the lesion approach both to learn about the localization of functions in the intact, unperturbed network, and to predict its response to damage. The FCA operative definition permits an accurate prediction of the performance of the network after multi-lesion damage, and yields a precise quantification of the distribution of processing in the network, a fundamental open question in neuroscience [Wu et al., 1994, Thorpe, 1995]. Moreover, the generalization of the FCA to High-Dimensional FCA described in chapter 4, uses the existing data to find the relevant functional units of the network, i.e. the sets of elements that have a functional significance as a unit. This generalization allows a natural description of systems exhibiting non-linear effects of lesioning such as “paradoxical lesions”.

Although still infrequent, multiple lesion studies are now being performed in an increasing frequency in neuroscience studies of reversible inactivation [Lomber and Payne, 2001]. However, these data are still of fairly limited magnitude. Hence, the development and study of the FCA method was primarily done in the theoretical modeling framework of EAAs. EAAs enabled us to seriously study the implications of the offered definition of contribution, to test its predictive ability, and to gain insights into the very notion of contribution and distribution of function.

1.5 Temporally Asymmetric Hebbian Plasticity

The EAAs considered in Chapters 2-4, are fully determined by their genetic structure. In particular, the efficacies of the synaptic connections are genetically predetermined, and no adaptation takes place during the agent’s lifetime. Although evolution is a major and important form of adaptation, learning is crucial for adaptation to fast environmental
changes. To increase the complexity of tasks and environments for EAAs, learning will have to be incorporated in their neurocontrollers.

Learning can be divided into supervised and unsupervised learning. In supervised learning, a teacher exists, which during training provides pairs consisting of an input and a corresponding correct (target) output. In most natural cases, a teacher is not present, and the learning is driven solely by the activity within the neuronal network (activity-dependent synaptic plasticity). In such unsupervised learning it is difficult to define the goal of learning, as no target for learning is given explicitly. We discuss two rather general goals in Chapter 5, namely the activity-driven formation of maps, and the representation of input correlations. EAAs can provide a useful framework to study unsupervised learning, because there the goal of learning is defined implicitly by the fitness function, without a need to define target outputs explicitly.

The basic paradigm for unsupervised learning is Hebbian synaptic plasticity, which is considered a key concept connecting the neuronal activity in a network with the functions of memory and learning as well as the development and refinement of neuronal network architectures [Eggermont, 1990]. Hebb [Hebb, 1948] postulated that correlations in the activity of neurons trigger synaptic changes: "When an axon of cell A is near enough to excite cell B or repeatedly or consistently takes part in firing it, some growth or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased." It is intriguing that in contrast to Hebb’s explicit formulation of a causal relation between the firing of pre- and post-synaptic cells, the bulk of theoretical investigations on Hebbian synaptic learning implemented correlation-based plasticity rules that were not sensitive to the temporal ordering of pre- and postsynaptic activity. By virtue of the popular simplification “fire together, wire together”, synaptic connections were modeled to grow stronger, when the pre- and post-synaptic spike rates were positively correlated, regardless of the relative timing of individual spikes [Roberts and Bell, 2002]. This is particularly curious because a number of earlier experimental reports have clearly indicated the importance of the temporal arrangement of pre- and post synaptic activity in determining whether a connection was potentiated or depressed [Levy and Steward, 1983, Gustafsson et al., 1987, Debanne et al., 1994]. It is only recently, that measurements of temporally asymmetric spike-timing dependencies in synaptic plasticity have triggered wide theoretical interest.

Based on dual whole-cell voltage recordings of synaptically connected pairs of neurons, the dependence of synaptic plasticity on the precise timing of pre- and post-synaptic action potentials has been measured in several slice and culture preparations [Magee and Johnston, 1997, Markram et al., 1997, Bi and Poo, 1998, Debanne et al., 1998, Zhang et al., 1998, Feldman, 2000, Sjöström et al., 2001]. In these studies, the efficacy of the synaptic connection between the two cells is measured by the amplitude of EPSPs that
Figure 1.2: Temporally asymmetric learning window. The relative change of synaptic strength induced by pre- and post-synaptic spike pairs with time difference $\Delta t = t_{post} - t_{pre}$. While for negative (acausal) time differences the synapse is depressed, it is potentiated for causal spike pairs.

are evoked by single pre-synaptic test stimuli. Using the possibility to control the exact timing of individual spikes in the pre- and post-synaptic cells, it was shown that pairs of pre- and post-synaptic action potentials that occur in a causal order, i.e. the pre-synaptic spike preceding the post-synaptic action potential, within a short time interval on the order of tens of milliseconds, induce an increase in the efficacy of the synaptic connection. In contrast, the synaptic connection is depressed if the temporal order of the spike pair is reversed. Importantly, synaptic modifications increase as the magnitude of the time difference between the pre- and post-synaptic spikes becomes small, resulting in a sharp discontinuity in synaptic changes that extends approximately 5 ms around coincident pre- and postsynaptic firing. Following the findings of [Bi and Poo, 1998], Figure 1.2 shows a simplified scheme of the described spike-timing dependence of synaptic changes.

The remarkable sensitivity of the measured plasticity rules to the precise timing of individual action potentials has immediately been associated with a variety of neural coding schemes in which information is assumed to be carried by precise temporal arrangements of spikes [Bi and Poo, 2001]. Indeed, Temporally Asymmetric Hebbian (TAH) plasticity has been successfully used to model the activity-driven refinement of auditory delay-line mechanisms [Gerstner et al., 1996, Leibold et al., 2001] or neural networks that implement temporal-difference learning algorithms [Rao and Sejnowski, 2001]. Importantly, the mixture of potentiation and depression introduces competition between the afferent synapses for controlling the firing time of the target cell. As a result, TAH learning rules result in synaptic competition [Song et al., 2000], dispensing with the former necessity to invoke non-Hebbian ad hoc mechanisms such as sliding plasticity thresholds [Bienenstock...
et al., 1982] or global synaptic scaling mechanisms [Miller and MacKay, 1994, Abbott and Nelson, 2000]. As shown in [Song and Abbott, 2001], this competition in the time domain suffices to induce the emergence of ocular dominance columns in models of visual cortex networks and, additionally, gives rise to self-normalization of the output firing rate of the post-synaptic cell [Song et al., 2000]. These properties make TAH a form of learning which can potentially be implemented in EAAs without requiring complex non-Hebbian plasticity. However, in spite of the computationally interesting possibilities of this novel synapse-specific mechanism of spike-timing modulated synaptic potentiation and depression, it is important to note that its applications in models have relied on the fine tuning of the learning rule parameters. The two long standing challenges within Hebbian learning, namely stability of the synaptic learning process and the emergence of robust synaptic competition, have remained crucial issues.

On the one hand, Hebbian learning rules are plagued by a stability problem of the synaptic weights. The correlation-based plasticity mechanism implements a positive-feedback process that most efficiently strengthens those synapses that had already become effective in driving the post-synaptic neuron. Hence, to be operational, Hebbian learning paradigms generally need to be augmented by some additional mechanism that stabilizes the synaptic weight distribution such that the runaway of synaptic weights is counteracted and the learning is stabilized with respect to the impact of random noise. On the other hand, in most cases, successful learning relies on the ability of the synaptic dynamics to pick up meaningful features in the neuronal input activity. Such sensitivity, however, commonly relies on the presence of competitive processes amongst the synapses that result in some synapses growing strong on the expense of others becoming weak. Since such competitive processes are generally associated with unstable dynamics, successful Hebbian learning requires the appropriate balance of stability and sensitivity. Previous studies have shown that two different types of synaptic updating functions result in two extreme types of synaptic learning behaviors with respect to this balance. In the additive model of TAH plasticity, synaptic changes are assumed to be independent of synaptic weights and the weights are constrained to the allowed range by a clipping mechanism. In this model, the learning dynamics are unstable due to the dominance of positive feedback but can induce strong competition among synapses if the model parameters are carefully tuned. In contrast, in the multiplicative model, the scales of potentiating and depressing synaptic changes are linearly attenuated as a synapse approaches the upper or lower boundary, respectively. The weight-dependence of these updating functions stabilizes the learning, but greatly diminishes all competitive effects within the synaptic population. Based on a generalized family of updating functions, the balance between stability and competition is one of the central concerns of the treatment presented in Chapter 5. This balance is crucial for a system to be able to learn while maintaining a stable operating
In addition to understanding the interplay between stabilizing and destabilizing synaptic drifts, it is important to quantify the performance of different learning rules. Such theoretically derived optimizations of plasticity rules can give important guidance for experimental designs and measurements. In Chapter 5, we construct a quantitative measure for the performance of learning based on the learned synaptic distributions (structural level). A different measure for learning may be naturally achieved by incorporating the learning rule into behaving agents, and measuring the performance on the behavioral level, rather than on the structural level.

Previous EAA studies have been utilized to study the evolution of learning itself, and the interplay between evolution and learning (see [Nolfi and Floreano, 1999] for a review). For example, near-optimal neuronal learning rules have been evolved in a simple EAA model of reinforcement learning in bees [Niv et al., 2003]. In another study, a population of EAs was subject to both evolutionary and learning processes [Nolfi, 1997]. In a typical exploration task, the combination of learning and evolution in the agents enabled them to obtain significantly higher performance than agents with a similar genetically encoded sensorimotor subnetwork but without learning. The key to the success of the learning-able agents is their ability to develop a genetically inherited predisposition for learning. This predisposition stems from the selection of initial weights at birth that guides behavior to select the right set of inputs, thus “channeling” learning in a successful direction. This power of evolution to select specific emergent learning predispositions, points to the potential pitfalls of studying learning in isolation, as is done in conventional neural networks and connectionist models. Thus, beyond the important theoretical ground laid in Chapter 5, embedding TAH learning in EAs may reveal new aspects of the learning process, and the interplay between learning and evolution.
Chapter 2

Evolved Autonomous Agents
(EAAs): A Neuroscience Perspective

The material described in this chapter has appeared in:
R. Aharonov-Barki*, T. Beker*, and E. Ruppin. Emergence of memory-driven command
* These authors contributed in equal parts.

2.1 Introduction

As discussed in the Introduction of the thesis, the relevance of EAAs to our understand-
ing of biological neural systems has not yet gained wide recognition. There are two
fundamental questions concerning EAAs we address in this chapter: 1. Can we identify
and analyze neurons and network structures emerging in EAAs which play an important
information processing role in controlling autonomous agents? And if so, then 2. Do
structures resembling findings from biology indeed occur in EAAs? Using evolutionary
simulations we developed autonomous agents controlled by ANNs, and conducted a thor-
ough analysis of the evolved neurocontrollers. Previous works on EAAs (e.g. [Cangelosi
* et al., 1994, Beer, 1997]) analyzed the emerging neural structures and showed interesting
structure to function relationship. These works typically used very small networks
or networks with pre-determined feed-forward structure. We extend this line of work by
studying networks with tens of neurons and over 100 synapses and unconstrained recurrent
architecture. Our results strongly suggest that EAAs are relatively tractable models
to analyze, manifesting biological-like characteristics.

The EAAs we develop use unconstrained network connectivity patterns to perform simple
life-like behavioral tasks. Under these conditions, we show that networks maintaining
steady activation levels can evolve, and moreover, serve to control agents that perform at a
remarkably high level compared with algorithmic benchmarks. Non-trivial network struc-
tures evolve in these agents. We analyze these structures and demonstrate the existence
of neurons whose functional repertoire strongly resembles that of “command neurons” known from biological networks [Combes et al., 1999, Teyke et al., 1990, Nagahama et al., 1994]. The command neuron’s activity is driven either by positional information or by a short-term memory mechanism, depending on the specific sensory information available to the agent.

The emergence of location-dependent cells in EAs has previously been demonstrated by [Floreano and Mondada, 1996] who studied homing navigation of a real robot. They found a neuron in the neurocontroller that exhibits location- and orientation-dependent activity. [Jakobi, 1998] has described a simple memory-based behavior in a small, bilaterally symmetrical EAA with 10 neurons and about 20 synapses. In this chapter we revisit both of these issues, showing how they emerge concomitantly in agents with various sensory capabilities performing a task requiring simple navigation and foraging skills. We conduct a thorough analysis of the evolved networks and the computations they perform. We study the emergence of location-dependent cells also under conditions in which the agents are deprived of any positional cues, and find that the apparent place-dependent activity is actually the result of an emerging memory mechanism culminating in a single neuron. This demonstrates that different computations can result in the same phenomenon of a “place cell”. We show that the emerging place cell takes the role of a command neuron which modulates a complex switch in network dynamics between two distinct operational modes, that in turn manifest themselves in two different behavioral modes of the agent.

The chapter is organized as follows: Section 2.2 gives an overview of the model. A more elaborate description can be found in Appendix 1. Section 2.3 describes the performance levels attained by evolved agents. In section 2.4 we demonstrate the emergence of command neurons modulating the behavior, and define the two basic modes of behavior. Sections 2.5 and 2.6 investigate the properties of the command neurons under two different sensory scenarios. We show that when sensory information is scarce a memory mechanism is evolved and helps to control the behavior of the agent. Section 2.7 investigates in detail the network structure in a particular successful agent. Finally, section 2.8 discusses the results. Appendix 1 gives a self-contained, detailed description of the simulation model. Appendix 2 describes the various behavioral measures we used to quantify the agents’ two basic behavioral modes and appendix 3 gives the algorithm we wrote as a benchmark for one of the behavioral tasks.

2.2 The Model

The basic environment consists of a grid arena surrounded by walls. In this arena two kinds of resources are scattered. “Poison” is randomly scattered all over the arena. Con-
assuming this resource decreases the fitness of an agent. “Food”, the consumption of which increases fitness, is randomly scattered in a restricted “food zone” in the south-western corner of the arena. The agents’ behavioral task resembles other models found in the literature (e.g. [Nolfi et al., 1994, Miglino et al., 1996] and is seemingly very simple - to eat as much of the food while avoiding the poison. The complexity of the task in this work stems from the limited and low-level sensory information the agents have about their environment. The agents are equipped with a set of sensors, motors, and a fully-recurrent ANN controller. It is this neurocontroller that is coded in the genome and evolved; the sensors and motors are given and constant. The models we explore differ in the sensors the agents are equipped with, the size of the network (15-50 neurons), and whether or not the food zone border is marked.

![Diagram](image)

Figure 2.1: An outline of the grid arena (southwest corner) and the agent’s controlling network. The agent is marked by a small arrow on the grid, whose direction indicates its orientation. The curved lines indicate where in the arena each of the sensory inputs comes from. Output neurons and interneurons are all fully connected to each other.

The agents in all the models explored were equipped with a basic sensor we termed somatosensor, consisting of five probes. Four probes sense the grid cell the agent is located in and the three grid cells ahead of it (see Figure 2.1). These probes can sense the difference between an empty cell, a cell containing a resource (either poison or food – with no distinction between those two cases), an arena boundary and food zone boundary. The fifth probe can be thought of as a smell probe, which can discriminate between food and poison just under the agent, but which gives a random identification if the agent
is not standing on a resource. This requires sensory integration in order to identify the presence of food or poison. In addition, in some models the agents are equipped with a position sensor, effectively giving the absolute coordinates of the agent in the arena. Note that the somatosensor alone gives only purely local information to the agent, making navigation in the environment a challenging task. Moreover, even in models where the agents are equipped with a position sensor, they lack any information about orientation, so navigation remains a difficult task. The motor system allows the agent to go forward, turn 90 degrees in each direction, and attempt to eat, a costly procedure since it requires a step with no movement.

![Fitness Chart](image)

Figure 2.2: A typical evolutionary run: maximum, mean and standard deviation of fitness in a population plotted over 30000 generations of an evolutionary run. Values are plotted every 100 generations. Note that the fitness here is evaluated over a single epoch for each agent and the mean is the average of the fitness in the population (100 agents). Due to the big variety of possible environments the measured fitness is quite noisy. Throughout this work, we assess the accurate fitness of an agent by averaging over 5000 epochs.

In each generation a population of 100 agents is evaluated. The life cycle of an agent (an epoch) lasts 150 time steps, each step consisting of one sensory reading, network updating, and one motor action. At the end of its life-cycle each agent receives a fitness score calculated as the total amount of food it has consumed minus the total amount of poison it has eaten and normalized by the number of food items available to give a maximal value of 1. Simulations last a pre-defined number of generations, ranging between 10000 and 30000. Figure 2.2 shows a typical evolutionary run. The initial population consists of agents equipped with random neurocontrollers. In a typical simulation run, the average
fitness of agents in the initial population is around -0.05. As evolution proceeds better controllers emerge and both the best and average fitness in the population increase until a plateau is reached. In the next section we assess the performance of the evolved agents at these final generations by comparing them to benchmarks. For more details of the model and evolutionary dynamics see Appendix 1.

## 2.3 Agents’ Performance

In the framework described above ANN-controlled agents were evolved. We studied two types of agents: An SP-type agent possessing both a somatosensor and a position sensor, and an S-type agent possessing a somatosensor only (i.e. no positional cues available). These two types of agents were evolved in one of two possible environments. One in which the food zone borders were marked, and one in which they were not. These conditions define the four different models studied (see Figure 2.3).

In order to evaluate the performance of the evolved agents, we used two benchmarks. First, we compared their performance to the best memoryless algorithm designed by us to perform the same task. The basic idea behind all the designed algorithms we used was the same, and can be sketched as follows: “When reaching a resource – eat it if and only if it’s a food item. Unless you ‘believe’ you are inside the food zone, try to navigate into it, and ignore all resources on your sides. If you ‘believe’ you are inside the food zone..."
zone, switch to an efficient grazing mode, and try not to leave the food zone...”. This basic idea translates into different instructions, according to the available sensory input. Appendix 3 describes one of these algorithms in more detail. As a second performance benchmark we solved the same tasks using Reinforcement Learning (RL) techniques, using an $\epsilon$-greedy SARSA algorithm [Sutton, 1996, Sutton and Barto, 1998] and training for up to ten million iterations.

Figure 2.3 summarizes the performance levels achieved by the above benchmarks, and those attained by the best evolved agents, for each of the four models studied. The fitness was averaged over 5000 epochs to achieve an accurate measure. (All fitness measures throughout this work are averages of 5000 epochs, resulting in a vanishing standard deviation (less than 0.003)). As is evident, the evolved agents do well compared with the benchmarks in all possible scenarios. Since both benchmarks use no memory, a much higher score achieved by an agent indicated that it is employing some kind of memory. Indeed, as shall be seen below, the superior performance of evolved agents in models devoid of a position sensor is due to the development of memory under these scenarios.

2.4 Exploration vs. Grazing: The Emergence of Command Neurons

As a first step in the analysis, we investigated the emergent behavior. We found that for both types of agents the most successful strategy relied upon a switch between two behavioral modes – exploration and grazing. The exploration mode consists of moving in straight lines, ignoring resources in the sensory field which are not directly under or in front of the agent, and turning at walls. When food zone borders are marked agents operating in exploration mode cross them. The grazing mode, on the other hand, consists of turning to resources to the right or left in order to examine them, turning both at walls and at food zone border markings, and maintaining the agent’s location on the grid in a relatively small, restricted region. In both modes when stepping on a food item eating occurs. Exploration mode is mostly observed when the agent is out of the food zone, allowing it to explore the environment and find the food zone. Inside the food zone, however, the agents almost always display grazing behavior, which results in efficient consumption of food.

To characterize the agent’s mode in a quantitative manner, we defined behavioral measures (see Appendix 2). Using these measures and systematically clamping the activity of neurons in the network we identified a common structure in successful agents whereby the mode switch was always mediated by a central “command neuron”. The activity mode of this command neuron determines the agent’s behavioral mode. Clamping the activity of the command neuron, constantly maintaining an active or a quiescent state, reliably
Figure 2.4: Location- and Orientation-selectivity of command neuron activity in an agent with somato+position sensor. The center sub-figure shows average command neuron activity over 5000 epochs. The peripheral sub-figures show average activity when the agent is facing a given orientation. Darker means higher average activity. White grid cells near walls correspond to locations the agent never visited with the given orientation. The thick line marks the border of the food zone (not seen by the agent in this scenario). The “smearing” of grazing activity towards north and east when facing these directions helps the agent to return to the food zone after accidentally leaving it.

produces exploration or grazing modes respectively, regardless of the actual location of the agent in the environment.

As will be seen, although the functional role of the command neuron is nearly identical in all models studied, the computational basis underlying its activity is quite different. The dynamical properties of the command neuron depend on the type of sensors mounted on the agent and not on whether food zone markings exist. Hence, from this point on, we shall concentrate on the S vs. SP-type agents, without making the distinction between the marked (M) and unmarked (U) cases in each.

2.5 Location-driven Command Neurons: The SP model

Examining the networks of successful agents equipped with a position sensor reveals an important common feature: Certain interneurons have a position-sensitive response. One
of these neurons typically fires outside the food zone and remains quiescent inside the food zone and in its close vicinity. Figure 2.4 depicts the mean activity of such a neuron as a function of the agent’s location. The center sub-figure corresponds to the mean activity of the position-sensitive neuron in each grid cell, averaged over 5000 epochs. As is evident, its activity corresponds to the location of the food zone – active outside and inactive inside. Clamping this “position-sensitive cell” and measuring the behavioral mode of the agent, reveals that it acts as a command neuron. When this command neuron is active the agent assumes an exploration behavior, regardless of where it is in the environment, whereas clamping it to a silent state results in grazing behavior. Thus, the place-dependent activity map of the command neuron serves as an accurate predictor of the agent’s behavior; where the map values are low the agent will graze, whereas where they are high it will explore.

Such place-driven command neurons were found consistently in agents evolved with somato and position sensors. The computational basis for these neurons’ activity is the absolute position of the agent in the environment, and is not influenced by the existence or location of resources in the arena. Shifting the food zone to a new location, produces no change in the activity map of the command neuron, and thus no change in the behavior of the agent relative to its location (see Figure 2.5 [1b] vs. [1a]). This obviously leads to a near-zero performance level since the commanded behavioral mode is no longer adequate. Given the direct sensory information about the location in the environment, the existence of a “position-sensitive cell” in these agents is not very surprising. It is the emerging “command neuron” function of these neurons which is interesting.

Further investigation of the activity of place-sensitive command neurons revealed that they are also orientation selective. The surrounding sub-figures in Figure 2.4 depict the mean activities when the agent was facing a given orientation. As is evident, when facing east or north, the area where the command neuron is inactive (corresponding to grazing mode) is ’ smeared’ to the east or north respectively, compared to the opposite orientations. The grazing mode is thus maintained further out of the food zone when the agent is facing outward, increasing the chance to turn back and return to the food zone after accidentally leaving it, and in turn increasing these agents’ fitness scores compared with agents controlled by a non-orientation-selective programmed algorithm (see the pertaining performance comparison in Figure 2.3).
Figure 2.5: *Location-dependent activity maps of the command neurons.* Darker means higher average activity. Average is taken over 2000 epochs. **Somato+Pos Sensor (SP model):** [1a] Baseline condition, [1b] Shifted food zone [1c] Two food zones. **Somatosen- sor (S model):** [2a] Baseline condition, [2b] Shifted food zone [2c] Two food zones. In all cases agent populations were *evolved* in environments with one food zone at the south-western corner (baseline condition). Thick lines depict the food zone borders (the agents presented here were evolved in worlds with no food zone border markings, which were added here for clarity of exposition only).

2.6 Memory-driven Command Neurons: The S model

2.6.1 Activity Maps of the Command Neuron

Due to the purely local sensory information and the lack of any positional cue in the S model, it is difficult to adopt a strategy that grazes efficiently inside the food zone yet explores the environment quickly enough to reach the food zone within a reasonable time. The two atomic strategies – always moving in straight lines or always examining every resource in the sensory field – both yield near-zero performance.

Similar to the SP case, we identified in every successful neurocontroller at least one neuron whose activity was place-dependent. Figure 2.5 [2a] depicts the activity map of such a neuron, demonstrating that it is quiescent when the agent is inside the food zone and increases its activity the further the agent is from the food zone. Again, this neuron takes the role of a command neuron, i.e. when it is active the agent manifests exploratory
behavior, whereas when it is quiescent the agent switches to grazing mode. In contrast to the SP-agents, shifting the food zone from its original location to a new location (Figure 2.5[2b]) causes no malfunction, as the selective activity pattern shifts accordingly.

Since S-type agents receive no explicit sensory cues regarding their location, the intriguing question is what constitutes the computational basis for the place-dependent activity of the command neuron. The observed behavior led us to hypothesize that successful agents utilize short-term memory, maintaining grazing mode for several time-steps after eating. This hypothesis was supported by the performance level of these agents, which actually surpassed both kinds of memory-less benchmarks for the same task (Figure 2.3).

### 2.6.2 Command Neuron Activity Profiles

We now turn to study the dynamics of the S-type memory neuron in detail. To this end, we traced the average activity of the memory command neuron as a function of the elapsed time since the last eating episode (Figure 2.7, solid line, with 250 poison items). As is evident, the activity of the command neuron undergoes sharp inhibition immediately after eating, its probability of firing gradually increasing thereafter. Thus, the command neuron “remembers” the number of steps elapsed since the agent has last consumed a food item; its activity reflects a stochastic short-term memory mechanism.

![Activity profile](image.png)

Figure 2.6: *Activity of a stochastic memory command neuron:* A raster plot of 3 different epochs. Vertical bars mark steps where the neuron fired. The small triangles above them mark feeding events.

Figure 2.6 presents a raster plot of the activity of a stochastic memory command neuron, with the small triangles indicating the times of feeding events. The baseline firing
Figure 2.7: Activity profiles of a stochastic memory command neuron: The average activity level is portrayed as a function of the time elapsed since last eating episode. The agent was evolved with 250 poison items in the environment (solid line), but its activity is measured under various poison concentrations over 1500 distinct epochs.

state of the neuron is fully active (i.e., commanding exploration mode)\(^1\), and it is inhibited following feeding. The post-eating quiescence periods vary in duration. Thus, the emerging dynamics of the neuron is sustained activity, interrupted by periods of inactivity triggered by the feeding events. This behaviorally translates into sustained exploration interrupted by grazing periods of varying durations which are triggered by feeding. From the agent’s perspective this is likely to constitute an optimal strategy: As long as the agent frequently encounters food it “knows” it is in a food zone and remains in grazing mode. However, if food is not encountered for a prolonged period, the agent switches to exploration mode to search back for the food zone. The robustness of this strategy can be demonstrated by distributing the food in two designated zones in the arena. Figure 2.5\(\text{[2c]}\) shows the resulting activity map, with the two zones clearly evident. This map correctly predicts the adequate behavior of the agent. Indeed, S-type agents that evolved a memory mechanism in environments with one food zone but were evaluated in arenas with two zones, behaved adequately: Exploring the environment they reached one food zone, grazed there and after a while switched to exploration (since the food density decreased), reached the other food zone, and again switched to grazing mode. This is in sharp contrast to the case of the SP-type agents which rely on absolute location, whose behavior

\(^1\)At the beginning of each epoch the network is reset to an inactive state, resulting in the few steps of quiescence initiating each epoch.
was completely inadequate, grazing only in the south-west corner (Figure 2.5[1c]).

The stochastic nature of such a memory mechanism which emerges in a neural network with deterministic dynamics can be accounted for by the random distribution of resources in the arena, as well as the random component of input from the somatosensor. Indeed, we found that the activation of the memory neuron strongly depends on the poison distribution in the arena (Figure 2.7). For poison distributions deviating from the one the agents were evolved with, the activation as a function of elapsed time is perturbed. For higher poison concentration this merely changes the memory-span, whereas for lower concentrations it actually distorts the shape of the plot, eventually making it non-monotonous, and thus a poor correlate of the elapsed time. In the absent poison case this results in a 10% decrease in the agent’s performance. This is an interesting example of the way in which evolution harnesses harmful features of the environment: A high concentration of poison in the environment would seem like a burdening feature, the elimination of which should increase performance. However, the stochastic memory mechanism takes advantage of exactly this feature. As will be shown below, it finely tunes itself to near-optimal performance for the given “ecological niche” in which the agents evolved.

2.6.3 The Underlying Stochastic Memory

Examining the input field (post synaptic potential) of the stochastic memory neuron excluding the input from itself revealed that its firing threshold is around one standard deviation above the field’s mean. Given that the input field is noisy, a transition of the memory neuron from a quiescent state to firing will occur spontaneously. The synapse from the memory neuron to itself is strong enough to keep it above threshold once active, whereas an eating event induces a strong inhibition which shunts its activity. This corresponds well to the observed activity profiles described above.

To investigate the spontaneous return to an active state, we plotted the distribution of the post-eating quiescent period duration (Figure 2.8[a]). After a refractory period lasting one time-step, the length of the quiescence period (behaviorally commanding a grazing mode) can be roughly approximated to a first order by a continuous exponential distribution function. Thus, a first approximation for the underlying dynamics of the memory command neuron is a one-parameter geometric distribution model specifying the probability to resume firing at any quiescent step. This approximation assumes that the probability to resume firing is constant and independent of the sensory input in the step. However, further inspection reveals that the probability to resume firing does depend on the sensory input. Specifically, it is much higher near walls than in other places in the arena. Therefore, to better assess the adaptation of the emerging memory mechanism to the behavioral task, we used a two-parameter stochastic model. According to this model, the normal state of the memory neuron is active, and it is shunted right after eating with
Figure 2.8: Stochastic Behavior of the Command Neuron. [a] The distribution of the command neuron’s quiescent period durations. Solid line corresponds to an exponential distribution with \( \lambda = 0.26 \) (shifted by one time-step for the refractory period), corresponding to a mean memory maintenance of 4.85 time-steps compared with 4.74 observed experimentally. [b] A two-parameter stochastic model of the activity of the stochastic memory command neuron in the basic S model. Surface depicts the fitness attained for different values of \( p \) and \( q \). The best fitness obtained using the model was 0.39 (\( p = 0.8 \) and \( q = 0.1 \) (lower circle)). The best evolved agent scored 0.42, with \( p = 0.51 \) and \( q = 0.08 \) (higher circle, the line pointing to the fitness obtained using the same parameters in the “mounted” model). The diagonal \( p = q \) corresponds to the first-order approximation, stating that the probability to resume firing is constant.

probability one (whereas the probability to switch from activity to inactivity otherwise is zero). The probability to switch back from inactive to active state depends on two parameters: it is \( p \) if the agent senses a wall, and \( q \) otherwise. This two parameter model was then used to determine the firing state of the command neuron at every step. Note that this is not an algorithmic benchmark. The model is “mounted” on an evolved S-type agent, whose command neuron’s activity is determined by the values predicted by the model, while the rest of the network’s activity is updated naturally.

The fitness values obtained by agents whose command neuron’s activity was driven by this two-parameter stochastic model with \( p, q \) values ranging between 0 and 1 are shown in Figure 2.8[b]. The actual parameter values obtained by the evolutionary process lie on the

\[ \text{Fitness} = \begin{cases} 
0.39 & \text{if } p = 0.8, q = 0.1 \\
0.42 & \text{if evolved agent, } p = 0.51, q = 0.08
\end{cases} \]

It is straightforward to envisage a neural mechanism realizing such a two-parameter model, utilizing two different field distributions.
high ridge of the fitness surface. The best evolved agents however still obtained a higher fitness than that of the two-parameter mounted model agents (see caption). This can be explained by the fact that the two-parameter model captures the essential dynamics of the memory-driven command neuron, yet neglects certain variations in the probability to switch between firing states, which enhance the performance.

![Fitness decrease after selective clamping of interneurons](image)

Figure 2.9: Fitness decrease after selective clamping of interneurons: Each interneuron was clamped in turn to its average activity value, and the fitness of the agent measured. The decrease in the fitness is plotted for each interneuron. Only one interneuron, the command neuron (CN), significantly effects the fitness when clamped alone. Fitness was averaged over 5000 epochs, standard deviation is less than 0.5% of the fitness.

2.7 Neural Network Structure of An S-type Agent

Although successful neurocontrollers of different agents share the command neuron mechanism, they differ in their structure and in some aspects of their function. To demonstrate the evolving network’s structures we focus here on one of the evolved S-type agents which has a remarkably simple yet very successful network. Clamping the command neuron to its average activity has a marked effect on the agent’s behavior (Figure 2.9), while clamping all other interneurons to their average activity values results in no significant decrease in the agent’s fitness. The interneurons in this agent merely set the bias of other neurons, and the network can thus be reduced to the one depicted in Figure 2.10 (top).

The firing state of the command neuron, whose dynamics has been explained earlier, triggers a switch between two distinct input-output networks controlling exploration vs. grazing (Figure 2.10 (bottom)). These two basic sub-networks reside in the same network. They are modulated by the memory-based command neuron, which when active adds its set of weights to the network’s connectivity matrix. It should be noted that the
Figure 2.10: An S-type neurocontroller: $W_{cn}$ is the weight vector from the command neuron (CN) to the output neurons. $W_{in}$ is the weight matrix from the input to the output neurons. $W_{out}$ are the recurrent connections in the output layer. The rest of the neurons can be reduced to a single neuron serving as a bias input to the CN and the output neurons. In Grazing mode the activity of the CN is 0, and the resulting network is depicted on the left. In Exploration mode its activity is 1, and the equivalent network is the one on the right.

ability to discriminate between food and poison remains intact with any manipulation of interneurons’ activities. This basic ability relies on direct connections between the input and output layers, and indeed evolves at the first stages of the evolutionary process.

2.8 Discussion

This chapter presents a novel attempt to perform an in-depth analysis of structure-to-function relations in evolved neurocontrollers for autonomous agents. To this end, we analyzed the control mechanisms evolving in autonomous agents performing a simple foraging task and governed by a recurrent ANN, without any predefined network architecture. To succeed in their behavioral task, the agents developed a mechanism for switching between two distinct types of behaviors – grazing and exploration. In all four experimental scenarios examined, the evolved agents managed to closely match the best memoryless algorithms for the task, and in cases characterized by limited sensory input surpassed them by far. This was achieved with completely unconstrained network architectures.
We discussed in detail two types of evolved agents, differing in the sensory input available to them. In both cases, a similar mechanism has evolved, whereby a "command neuron" modulates the dynamics of the whole network and switches between grazing and exploration behaviors. In the case where the agents had no sensory position information, a memory mechanism emerged, which then became the basis for the place-sensitivity of the command neuron. Using a two-parameter stochastic model for the memory mechanism we demonstrated that evolution fine-tuned this mechanism towards near-optimal parameters, using the inherent environmental noise and taking advantage of features of the environment that are a-priori harmful, such as the distribution of poison in the arena. An analysis of the neurocontroller of a simple S-type agent demonstrates that the command neuron switches the dynamics between two basic input-output networks residing within the same network. Other networks controlling successful agents may be far more complex. It is the subject of further studies to fully understand all aspects of their structure and function, and will demand the utilization of more advanced analytical methods.

Several studies have previously dealt with the analysis of evolved neurocontrollers. [Cangelosi et al., 1994] describe agents coping with a foraging task that demands navigation into areas with food and water respectively according to an externally-governed motivational state. They show the emergence of two distinct neural pathways for dealing with the two states. The architecture of the neurocontrollers they develop is constrained to be feed-forward, and their primary analysis method is receptive-field analysis. In a series of works, Beer et al. use dynamical systems approach to analyze continuous-time recurrent neurocontrollers. [Beer, 1997, Chiel et al., 1999a, Beer et al., 1999a] deal with the evolution of small Pattern Generators (PGs) consisting of 3-5 neurons with recurrent connectivity. Similar to our analysis, the sensory input available during evolution is shown to affect the type of the evolving control mechanism. [Gallagher and Beer, 1999] evolves a 22-neuron symmetric feed-forward architecture for coping with a visually-guided walking task, and used selective lesioning to identify the role of single neurons in the resulting controllers. The work presented in this chapter follows similar lines, extending previous work by using unconstrained architectures in the context of a non-trivial task requiring more than a purely reactive behavior. The resulting structures are relevant to the understanding of an important biological neural archetype - that of a Command Neuron.

The idea of Command Neurons, i.e. single neurons whose activity commands a high level behavioral pattern, was first suggested some fifty years ago. Since then their existence was verified in a number of animal models, including crayfish [Edwards et al., 1999], Aplysia [Xin et al., 1996b, Xin et al., 1996a, Nagahama et al., 1994, Teyke et al., 1990], Clione [Panchin et al., 1996], crabs [Norris et al., 1994, DiCaprio, 1990] and lobsters [Combes et al., 1999]. Command neuron activity has been shown to control a variety of motor repertoires. In some cases the behavior is modulated on the basis of sensory stimuli
[Xin et al., 1996a], and in particular by food arousal [Nagahama et al., 1994, Teyke et al., 1990]. Moreover, command neurons induce different activity patterns in the same neural structures by modulating the activity of other neurons in a pattern-generating network [Combes et al., 1999, DiCaprio, 1990]. One of the mechanisms by which command neurons act on other neurons is by modulating their excitability [DiCaprio, 1990], similar to the mechanism emerging in our simulations. The resemblance between these findings and the emerging properties of networks in the simulations we described is noteworthy. Nevertheless, biological reality is much richer. E.g., chemical neuro-modulation plays an important role in the command neuron activity [Brisson and Simmers, 1998, Panchin et al., 1996], while totally absent from our model. Another example is the finding of Command Systems comprising several command neurons that act in coordination to control a varied behavioral repertoire [Combes et al., 1999, Edwards et al., 1999, Gankrelidze et al., 1995].

The current model, although simple, gives a computational insight into command neuron mechanisms. It shows a concrete model in which a single command neuron switches the dynamics of a neural network between two markedly different behavioral modes. This is achieved by dynamically setting the biases of the other neurons, thus effectively multiplexing two networks within the same set of neurons. We demonstrate that this structure does not have to be hand-crafted - it evolves spontaneously in a variety of scenarios, proving to be a robust computational mechanism. The characterization of the set of tasks that can be solved by such multiplexed networks still awaits further study. However, it is interesting to know that such non-trivial tasks indeed exist, that multiplexed networks solving them can be evolved and moreover that they can be successfully identified.

The results presented here were obtained using the crudest form of genetic encoding – direct specification of all the synaptic weights in the genome. This bears a limiting effect on the scalability and speed of the evolutionary process. With the application of more sophisticated genetic encoding schemes, such as grammatical or ontogenic encodings [Kitano, 1990, Cangelosi et al., 1994], and of more efficient selection procedures such as incremental evolution [Gomez and Miikkulainen, 1997], one may expect the evolution of larger recurrent EAs, processing more complex sensory input to achieve more intelligent behaviors. Adding learning as an additional adaptive force can enhance the efficiency of the evolutionary process (see [Hinton and Nowlan, 1987, Ackley and Littman, 1991, Nolfi et al., 1994]) and may be needed for the structural fine-tuning of large evolved networks.

The similarity to known neural mechanisms that was achieved even under the current basic and almost toy-like techniques leads us to believe that once better scalability is achieved, EAs may provide an excellent vehicle to study the fundamental problem of structure and function relation in nervous systems. The accessibility of such EAA models to thorough analysis should make them important means of investigation in the tool-chest of computational neuroscientists.
Appendix 1: Detailed Model Description

The Simulation System

A flexible simulation system was used to build a variety of evolutionary models incorporating autonomous agents acting in a life-like environment. The core system defines the basic notions of an agent, a simulation world in which several agents can coexist, and the simulation universe (in which several simulation worlds can coexist). It is implemented in C++ and runs on the Linux operating system. The specialized implementation in C++ allows for intensive optimization, resulting in running times of about 4 hours for 10,000 generations of 100 agents performing the task described above. Around the core system particular models are built. Each model defines the specific properties of simulation worlds and agents. At the level of the simulation world, this includes the geometry, appearance, and availability of different resources. At the level of the agent, it includes its sensory capabilities, its motor capabilities and the nature of the control mechanism mediating between the sensory input and motor output. Currently, the evolutionary process only affects the neural network control mechanism. A particular model also defines the specific genetic methods used in the simulation - whether it uses sexual or a-sexual breeding, the variational operators and the selection methods.

The Environment and the Behavioral Task

The agents all operate in a grid arena of size 30x30 with two kinds of resources. One resource, defined as “poison” (i.e. causing a negative reward), is randomly scattered all over the arena. A second, “food” resource bringing positive reward is randomly scattered in a restricted food zone location in the environment. In most experiments, the food zone was of size 10x10 and was located at the south-west corner of the arena. In some of the models studied the boundaries of the food zone were marked, enabling the agents to sense them, while in other models this marker was absent. A life cycle of an agent (an epoch) lasts 150 time steps, in which one motor action takes place. At the beginning of an epoch the agent is introduced to the environment at a random location and orientation.

In each generation a population of 100 agents is evaluated. Each agent is evaluated in it’s own environment, which is earlier initialized with 250 poison items and 30 food items. At the end of its life-cycle each agent receives a fitness score calculated as the total amount of food it has consumed minus the total amount of poison it has eaten divided by 30, the number of distributed food items. Thus the fitness ranges between -2.5 (eating the maximal number of poison items possible within the 150 steps of an epoch) and 1 (consuming all the distributed food items).
The Controlling Network: Structure and Dynamics

Each agent is controlled by a neural network consisting of 15 to 50 neurons (the number was fixed within a given simulation run). Out of these, $K_{in}$ neurons (5 or 7) are dedicated sensory neurons, whose values are clamped to the sensory input. Four neurons are designated as output neurons commanding the motor system. The network is composed of binary McCulloch-Pitts neurons which are fully connected, with the exception of the non-binary sensory neurons which have no input from other neurons. Network updating is synchronous. In every step a sensory reading occurs, network activity is then updated, and a motor action is taken according to the resulting activity in the designated output neurons.

The Sensory System

Each agent is equipped with a basic sensor we termed somatosensor, consisting of five probes, to each of which a sensory neuron is associated. Four probes sense the grid cell the agent is located in and the three grid cells ahead of it (see Figure 2.1). These probes can sense the difference between an empty cell, a cell containing a resource (either poison or food – with no distinction between those two cases), an arena boundary and food zone boundary. The fifth probe can be thought of as a smell probe, returning -1 or +1 if the agent is currently in a grid cell where there is poison or food respectively, and -1 or +1 randomly otherwise. Thus, the agent has to integrate the input from two sensors in order to identify the presence of food or poison. In addition, in some models the agents were equipped with a position sensor, consisting of two sensors, giving the agent’s absolute coordinates in the arena, where the origin is taken as the south-western corner. The agent has no information about its orientation.

The Motor System

The motor system of an agent consists of four motors, receiving binary commands from the four output neurons. The first motor induces forward movement when activated. Two other motors control right and left turns, inducing a 90 degrees turn in the respective direction when only one of them is activated, and maintaining the current orientation otherwise. The fourth motor controls eating, consuming whatever resource is available in the current location when activated. For eating to actually take place, however, there has to be no other movement (forward step and/or turn) in the same time step. This both enforces simple motor integration, and makes any attempt to eat a costly procedure.

---

3 The term somatosensor is inaccurate due to the smell sensor, but we shall use it for brevity of notation.
Evolutionary Dynamics

Each agent is equipped with a chromosome defining the structure of its \( N \)-neurons controlling EAA, consisting of \( N(N-K_m) \) real numbers specifying the synaptic weights. At the end of a generation a phase of sexual reproduction takes place, in which chromosomes are crossed over and then mutated to obtain the agents of the next generation. There are 50 reproduction events each generation. In each of them two agents from the parents population are randomly selected with probability proportional to their fitness.

We used uniform point-crossover with probability 0.35, after which point mutations were randomly applied to two percent of the locations in the genome. These mutations changed the pertaining synaptic weights by a random value between -0.6 and +0.6. Simulations lasted a pre-defined number of generations, ranging between 10000 and 30000. In the last generation, every agent was evaluated during 5000 epochs, on a variety of initial conditions, to accurately measure its fitness.
Appendix 2: Behavioral Indices of Exploration and Grazing

To quantify the behavior of the agents we used the following indices:

1. *Time it takes the agent to reach the food zone* for the first time in an epoch. This index should be short for exploration mode and long for grazing mode.

2. *Time it takes the agent to return to the food zone* once it leaves it. This should be short for grazing mode and long for the exploration mode.

3. *Percentage of turns to resources:* Out of all steps in which there is a resource located on either side of the sensory field, we measure the percentage of steps in which the agent turned to inspect the resource. This should be high for grazing mode, and low for exploration mode.

4. *Percentage of turning steps out of all steps in the epoch.* Exploration mode consists of moving in straight lines, and thus this measure is low for exploration mode, and high for grazing mode.

5. *Extent of arena coverage.* In exploration mode, the main goal is to cover distances, but not to waste time on densely covering the explored areas. In grazing mode, on the other hand, the searched area should be small but it should be thoroughly covered. In order to measure this, we calculated two numbers. First, the percentage of arena cells visited (out of the total number of cells). Second, the area of the minimal rectangle encapsulating all the cells visited, divided by the total area of the arena. The index is the ratio of these two numbers. A high value signals grazing mode (i.e., high coverage), whereas a low one testifies to exploration mode.

Table 2.1 depicts the values of these indices in two successful agents, one of the S-type and one of the SP-type. Note the effects of clamping the command neuron: When it is clamped to an active state, in both agents the indices of exploration are high, while clamping to a silent state induces grazing behavior. Also note that in the control state (no clamping) the behavioral mode is usually adequate, e.g. finding the food zone under control conditions takes the same time as under a clamped-to-exploration mode, etc.
<table>
<thead>
<tr>
<th></th>
<th>S-type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time to reach food zone</td>
</tr>
<tr>
<td>Control</td>
<td>55.7(±0.85)</td>
</tr>
<tr>
<td>Command neuron</td>
<td></td>
</tr>
<tr>
<td>clamped to active</td>
<td></td>
</tr>
<tr>
<td>state (exploration)</td>
<td>53.5(±0.84)</td>
</tr>
<tr>
<td>Command neuron</td>
<td></td>
</tr>
<tr>
<td>clamped to silent</td>
<td></td>
</tr>
<tr>
<td>state (grazing)</td>
<td>113.8(±1.32)</td>
</tr>
<tr>
<td></td>
<td>SP-type</td>
</tr>
<tr>
<td></td>
<td>Time to reach food zone</td>
</tr>
<tr>
<td>Control</td>
<td>45.4(±0.68)</td>
</tr>
<tr>
<td>Command neuron</td>
<td></td>
</tr>
<tr>
<td>clamped to active</td>
<td></td>
</tr>
<tr>
<td>state (exploration)</td>
<td>44(±0.66)</td>
</tr>
<tr>
<td>Command neuron</td>
<td></td>
</tr>
<tr>
<td>clamped to silent</td>
<td></td>
</tr>
<tr>
<td>state (grazing)</td>
<td>105.2(±1.37)</td>
</tr>
</tbody>
</table>

Table 2.1: *Behavioral mode indices:* For each of the agents, five indices were measured under three conditions: control (i.e. normal conditions with no clamping), the command neuron clamped to a constant active state, and to a constant silent state. Results are averages over 2000 epochs. The standard deviation of these averages is given in parenthesis.
Appendix 3: Benchmark Algorithm for the SP-model

The following is the algorithm in pseudo-code.

if(stepping on food) eat
else
    if(in front of wall)
        if(x>y and ~infoodzone) turn right
        else turn left
    else if(in front of foodzone marking)
        if(infoodzone) turn left
        else move fwd
    else if(sense resource in front-left grid)
        if(infoodzone) move fwd and turn left
        else move fwd
    else if(sense resource in front-right grid)
        if(infoodzone) move fwd and turn right
        else move fwd
else
    move fwd

Using the input from the position sensor the algorithm determines if the agent is in the food zone and takes the appropriate motor action. Loosening the “infoodzone” condition, i.e. acting in a grazing mode when the agent is within a certain range around the food zone, does not improve the performance of the algorithm.
Chapter 3

Localization of Function in EAAs:
Lessons Learned for the Lesion
Approach in Neuroscience

Most of the material described in this chapter has appeared in:
R. Aharonov, L. Segev, I. Meilijson and E. Ruppin. Localization of Function Via Lesion

3.1 Introduction

As motivated in the Introduction of the thesis, this chapter presents a new tool for localizing functions in biological and artificial neural networks, and introduces a quantitative
approach for studying the long debated issue of local versus distributed computation in
the brain. We describe here the Functional Contribution Analysis (FCA), which assigns
contribution values to the elements of a network such that the ability to predict the net-
work’s performance in response to multi-lesions is maximized. The development of the
FCA capitalizes on the availability of full information that makes EAA models a good
test-bed for studying neural processing. Our results seriously question the adequacy of
the classical single lesion analysis traditionally used in neuroscience, and show that using
lesioning experiments to decipher even simple neuronal systems definitely requires a more
rigorous multi-lesion analysis.

The chapter is organized as follows: section 3.2 describes the functional contribution
analysis, and provides the necessary definitions. In section 3.3 we apply the FCA to
evolved neurocontrollers, demonstrating its capabilities to unravel the underlying neuronal
contributions. We examine the implications of the type of lesioning method employed,
and show the FCA’s superiority over the classical neuroscience single lesion approach.
Section 3.4 studies the localization of different tasks with the FCA, and section 3.5 shows
that the FCA can be utilized for a detailed synaptic analysis of the underlying network
3.2 The Functional Contribution Analysis

The FCA aims to compute the localization of a task from a series of multi-lesion experiments. In each such experiment, a lesioning configuration specifies which of the agent’s neurocontroller units (be they neurons, synapses, or any higher-order module) is lesioned. Given this lesion, the agent is run in the environment in which it was evolved, and its performance (i.e., its fitness or any other well defined performance measure) is measured. The FCA uses this data to compute a contribution vector \( \mathbf{c} = (c_1, \ldots, c_N) \), where \( c_i \) is the contribution of element \( i \) to the task in question, and \( N \) is the number of elements in the network. The goal of the FCA is to assign the contribution values which provide the best prediction of the agent’s performance in terms of Mean Squared Error (MSE) under all possible multi-site lesions.

3.2.1 Definitions

Suppose that a multi-lesion experiment is performed where a set of elements in the network is lesioned and the neurocontroller network then performs a certain task. The result of this experiment is described by the pair \( \{\mathbf{m}, p_m\} \) where the lesion configuration vector \( \mathbf{m} \) has \( m_i = 0 \) if element \( i \) was lesioned and \( 1 \) if it was left intact, and \( p_m \) is the corresponding performance of the lesioned network, divided by the baseline performance of the fully intact network. For a system of \( N \) elements there are \( 2^N \) such possible multi-lesion experiments, i.e., \( 2^N \) possible lesioning configurations \( \mathbf{m} \).

The underlying idea of our definition is that the contributions of the units are those values which allow the most accurate prediction of performance following lesions of any degree to the system. Within the FCA framework, given a configuration \( \mathbf{m} \) of lesioned and intact units, the predicted performance \( \tilde{p}_m \) of the network is the sum of the contribution values of the intact units \( (\mathbf{m} \cdot \mathbf{c}) \), as evaluated by a non-decreasing function \( f \),

\[
\tilde{p}_m = f(\mathbf{m} \cdot \mathbf{c}).
\] (3.1)

The function \( f \) is a non-decreasing piecewise polynomial. It is non-decreasing to reflect the notion that beneficial elements (those whose lesioning results in performance deterioration) should have positive contribution values, and that negative values indicate elements that hinder performance.

Given these definitions, we seek to find the pair \( \{\mathbf{c}, f\} \), which minimizes the mean squared prediction error

\[
MSE = \frac{1}{2^N} \sum_{\{\mathbf{m}\}} (\tilde{p}_m - p_m)^2.
\] (3.2)
The vector \( \mathbf{c} \) which minimizes this error is the \textit{contribution vector} for the task tested, and the corresponding \( f \) is its adjoint \textit{performance prediction function}. Since we can arbitrarily choose the scale of \( c \) and maintain the same prediction by modifying \( f \), we normalize \( \mathbf{c} \) such that \( \sum_{i=1}^{N} |c_i| = 1 \). Thus, the optimal contribution vector \( \mathbf{c} \) and its accompanying performance prediction function \( f \) minimize the MSE of predicted versus actual performance, over all possible lesioning configurations.

The FCA is related to two known modeling and prediction methods, that of Generalized Linear Modeling and that of Projection Pursuit Regression. The Generalized Linear Model (GLM, see [McCullagh and Nelder, 1989]) approach can be succinctly defined by the relation \( g(\hat{p}_m) = \mathbf{m} \cdot \mathbf{c} \), where \( g \) is the transfer function akin to the performance prediction function \( f \) in the FCA formulation (if \( g = f^{-1} \) the above equation reduces to Eq. 3.1). In GLM, the contribution vector \( \mathbf{c} \) can be computed via a closed formula, and hence results in a much faster and deterministic computation than that performed within the FCA (see section 3.2.2). However, GLM cannot be used to predict performance since in general \( g \) is not reversible; many lesioning configurations can lead to the same performance (e.g., complete failure of the agent). Projection Pursuit Regression (PPR, see [Huber, 1985]) has a formulation very similar to the FCA. PPR approximates the response surface \( g \) by a sum of ridge functions \( g(\mathbf{x}) \sim \sum_{j=1}^{k} f_j(\mathbf{c}_j \cdot \mathbf{x}) \), which, together with the projection directions \( \mathbf{c}_j \) are found in an iterative, greedy manner, minimizing the distance from a target random variable. The FCA is hence a special case of PPR in which \( k = 1 \) and \( f \) is non-decreasing. Using only one ridge function and a non-decreasing \( f \) enables one to preserve the intuitive meaning of \( \mathbf{c} \) as the contribution values vector.

### 3.2.2 The FCA Algorithm

The contribution vector \( \mathbf{c} \) and the prediction function \( f \) for a task are computed using a training set of lesioning configurations and their corresponding performance levels in the task at hand. The FCA is a gradient descent search algorithm which iteratively updates \( f \) and \( \mathbf{c} \) until it reaches a local minimum of the training error \( \text{MSE}' \)

\[
\text{MSE}' = \frac{1}{n} \sum_{\mathbf{m} \in \mathcal{M}} [f(\mathbf{m} \cdot \mathbf{c}) - p_\mathbf{m}]^2, \tag{3.3}
\]

where \( \mathcal{M} \) is the training set and \( n \) is its size. The steps of the FCA are:

1. \textbf{Choose} a random initial normalized contribution vector \( \mathbf{c} \) for the task, and compute \( f \) as in step 4.

2. \textbf{Compute} \( \mathbf{c} \). Using the current \( f \) compute new values of \( \mathbf{c} \) by performing a gradient descent on the error \( \text{MSE}' \) (Eq. 3.3), searching for \( \mathbf{c} \) values that minimize \( E' \) using the current \( f \).
3. **Re-normalize** \( c \), such that \( \sum_{i=1}^{N} |c_i| = 1 \).

4. **Compute** \( f \). Given the current \( c \), perform isotonic regression on the pairs \( \{m \cdot c, p_m\} \) in the training set. Use a smoothing spline on the result of the regression to obtain a new \( f \) [Knott, 2000].

Steps 2 through 4 are repeated a fixed number of times.

### 3.2.3 Experimental Protocol

The FCA algorithm is stochastic, and as such is dependent on both the initial conditions and the random choices made in the gradient descent step. To reduce stochasticity, each of the MSE values reported here is a mean of 10 FCA runs. A single run consists of 10 trials, each of which is initialized with a different random contribution vector. The result of the trial with the lowest MSE on the training set is chosen for that run. We note, however, that the sensitivity to initial conditions is rather small in most cases. For each initial condition, the FCA executes a fixed number \( I \) of iterations of updating \( f \) and \( c \) (see section 3.2.2)\(^1\). Throughout this work, \( I = 150 \), except for the analysis of agent S22 where \( I = 300 \). All FCA results presented are mean and standard deviation of 10 FCA runs (i.e. a total of 100 initial condition choices). The MSE values are normalized by dividing them by the variance of the agent’s performances \( p_m \) in the test set. Thus, a normalized MSE of \( q \) means that the FCA explains a fraction \( R^2 = 1 - q \) of the variance. That is, if one would predict for any configuration a performance level equal to the mean performance of the agent, then the normalized MSE would equal 1, corresponding to \( R^2 = 0 \).

Throughout the chapter, we focus on three neurocontrollers evolved as described in the previous chapter. We study one SP-type agent, whose network consists of 10 neurons (Sp10), and two S-type agents, one whose network consists of 22 neurons (S22), and another with a 10 neuron network (S10).

### 3.2.4 Fast FCA: Adaptive Lesion Selection

For the FCA to be useful it is crucial that it generalizes well when using only a small subset of the full configuration set for training. As Figure 3.1 demonstrates, using small training sets yields good prediction capabilities for unseen configurations even when the set is randomly selected. Yet, it is important to optimize the choice of the training set, i.e. judiciously select a small subset of lesions that will maximize the FCA accuracy for a given training size. For this purpose, we have developed the *Adaptive Lesioning* (AL)

\(^1\)The number of iterations chosen is such that the MSE converges. Note that a further drift of the contribution values may exist if one continues to iterate the algorithm beyond this point. However, an appropriate re-normalization exists which reconstructs the contribution values obtained at this point.
algorithm, which iteratively selects the next lesioning configuration to be evaluated, based on the configuration set used so far:

1. Create a random initial core set of $N$ configurations\(^2\). Compute $c$ and $f$ using the FCA.

2. From all possible configurations that are not yet in the current set, find the configuration whose estimated performance (using Eq. (3.1)) is farthest from the known performance values of the configurations currently in the training set. That is, given the current configuration set $T$, choose the next lesioning configuration $m$ such that

$$ m = \arg \max_{m' \notin T} \left\{ \min_{m'' \in T} |p_{m''} - f(m' \cdot c)| \right\} \quad (3.4) $$

3. Add $m$ to $T$, recompute $c$ and $f$ using the FCA, and return to step 2.

Steps 2 and 3 are repeated until either a predefined number of training examples is reached, or the change in the test prediction error falls below a threshold criterion. As can be seen in Figure 3.1, for any training set size, the AL algorithm results on average in a lower MSE on the test set than randomly choosing the training set. Also, a very small training set (about 40 configurations) selected with the AL algorithm suffices to reach the test error achieved when training on the full configuration set (1024 configurations). Moreover, the AL algorithm is much more consistent in finding good training sets (see error bars in Figure 3.1). Random sets are prone to ineffective sampling of the full configuration space, and thus more often result in a large MSE on the test set.

The Adaptive Lesioning algorithm is closely related to the field of adaptive learning [Cohn et al., 1995, Engelbrecht and Cloete, 1999], which seeks to improve both the generalization and speed of machine learning algorithms. Attempting to uniformly sample the performance values $p$, the AL algorithm effectively samples the values of $m \cdot c$ with density proportional to the slope of $f$. Thus, the resulting FCA is accurate because it samples more data in the more crucial regions, those in which small perturbations in $m \cdot c$ result in large variations in the performance $f(m \cdot c)$ (see Figure 3.2).

### 3.3 FCA of Evolutionary Neurocontrollers

As described in the previous chapter, using evolutionary simulations, we have developed autonomous agents controlled by fully recurrent artificial neural networks. High performance levels were attained by agents performing simple life-like tasks of foraging and

\(^2\)Since $c$ has dimension $N$, the training set must consist of at least $N$ configurations. The set always includes the all-intact and the all-lesioned configurations.
Figure 3.1: *Adaptive Lesioning vs. random configuration selection.* Mean and standard deviation of the normalized test MSE vs. number of training configurations agent S10. Test set is the full $2^{10}$ configuration set. The “All lesions” dashed line denotes the test error when training on the full configuration set. The inset focuses on a subset of the same data, portraying the number of training configurations leading to significant, low test MSE values.

Figure 3.2: *Visualization of the Adaptive Lesioning algorithm.* The slope of $f$ in domain $A$ is much larger than its slope in domain $B$. As a result, a small perturbation in $m \cdot c$ where $m \cdot c \in A$ changes the performance estimation $\hat{p} = f(m \cdot c)$ more than if $m \cdot c$ were in $B$. The AL algorithm attempts to sample the $p$-axis uniformly, inducing the density of sampling along the $m \cdot c$-axis to be proportional to the derivative of $f$. 

40
navigation. Classical neuroscience methods were used to analyze the evolved neurocontrollers. Here the FCA is developed and studied in this environment and is applied to the analysis of EAA neurocontrollers, demonstrating several aspects of the approach.

The FCA requires a well defined quantitative measure for performance on each task analyzed. We consider three tasks throughout the work. First, the general survival task of maximizing fitness as defined in the previous chapter. Further, we consider two subtasks which correspond to the two emergent behaviors of successful agents, exploration and grazing, which were described in the previous chapter. Exploration performance, $p^e$, is measured by randomly placing the agent in the arena and measuring the number of steps $t$ elapsed before the agent reaches the food zone for the first time. If the agent fails to reach the food zone in less than 1000 steps, $t$ is set to 1000. The exploration performance is computed by

$$p^e = \frac{T - t}{T} \frac{\pi_b}{S} \frac{t}{T},$$

(3.5)

where $\pi_b$ is the number of poison items consumed before reaching the food zone, $S = 30$ is the total number of food items in the arena, and $T = 150$ is the original number of steps in an epoch. The first term is concerned with the speed of reaching the food zone, and becomes negative for $t > T$. The second term penalizes poison consumption which should be avoided in any behavioral mode, and is normalized to be consistent with the general performance normalization. Grazing performance, $p^g$, is measured by placing the agent in the arena for $T$ time steps. Denote by $s$ the total number of food sources eaten, and by $\pi_a$ the number of poison items consumed after reaching the food zone for the first time, then,

$$p^g = \frac{(s - \pi_a)}{S} \frac{T - t}{(T - t)/T},$$

(3.6)

where $t$ is measured as before. The numerator is consistent with the general performance normalization, and the denominator normalizes by the time spent inside the food zone.

3.3.1 Selecting the Lesioning Method

An important determinant of a lesioning analysis is the manner in which the lesioning itself is performed. We have studied two main lesioning methods, biological lesioning and stochastic lesioning. Biological lesioning refers to the classical method employed in most neuroscience lesioning experiments, where the lesioned component is ablated or cooled, i.e., completely silenced. In our model, accordingly, in biological lesioning a lesioned neuron is deactivated completely by cutting all its output connections to the rest of the network. In contrast, Stochastic lesioning is performed by making the firing pattern of the
lesioned component random rather than by completely silencing its output. That is, at every time step a lesioned neuron fires with probability equal to its overall mean firing rate in normal behavior, independent of its input field. This ensures that the lesioning does not affect the mean field of other neurons, affecting only the information content received from the lesioned neuron\(^3\). In both methods, when lesioning motor neurons, we do not alter the activity transmitted to the motors themselves. This enables us to isolate the role of the motor units in the computation of the recurrent controller networks (i.e. their contribution to other neurons), without completely immobilizing the agent.

![Graph A](image)

**Figure 3.3:** Comparison of biological and stochastic lesioning. A. MSE obtained by the biological versus stochastic lesioning for agent S10 in the general, grazing and exploration tasks. B,C. Contribution values obtained using biological lesioning (B) and stochastic lesioning (C) for agent S10 in the general fitness task.

Before advancing to study the FCA in depth as described further on in this and the next chapter, we ran a few experiments to gauge the accuracy of FCA analysis obtained with both lesioning methods. Figure 3.3A compares the test MSE obtained by the FCA on the controller of agent S10 in the general, grazing and exploration tasks, once employing

\(^3\)Similarly, in the synaptic analysis, a lesioned synapse transmits a stochastic version of the pre-synaptic activity.
biological lesioning (black bars) and once using stochastic lesioning (light bars). As is evident, the MSE obtained from biological lesioning is significantly poorer than that obtained when using stochastic lesioning. Indeed, as panels B and C demonstrate, the contribution values of the neurons computed by both lesioning methods differ, showing that the poorer prediction capability of biological lesioning stems from a real difference in the significance assigned to the system elements.

These results testify to the crucial importance of selecting the lesioning method. Specifically, they demonstrate that the classical ablation (biological) lesioning conventionally used in animal lesioning experiments might be questionable. A possible explanation for this phenomenon is that when a neuron is silenced it transfers other unlesioned neurons into a regime in which they cannot “fire” as their mean field is far below threshold. Thus, the effective part of the network which is lesioned is much larger than what was intended, and considered “lesioned” by the FCA. This results in a situation where lesions percolate to other network elements, disabling and isolating specific elements, and creating dispersed effective lesions. Stochastic lesioning also spreads the damage from lesioned neurons to neighboring neurons, but to a much lesser extent. As a result, biological lesioning is much more difficult to analyze, because the difference between the direct lesioning configuration used by the FCA and the effective one is much larger, causing FCA-like algorithms to be erratic and less accurate.

Biological lesioning is hence not fit for accurate lesion analysis (at least, in the FCA framework, which is fairly general and employs multi-unit lesions) even in very simple neurocontrollers. In view of these results, we have chosen to employ stochastic lesioning in the rest of this work. This method is definitely superior for the analysis of animat agents. We shall return to discuss these two lesioning methods from a neuroscience perspective in section 3.6.

3.3.2 FCA of Agents’ Performance

We consider first the performance obtained in the general task, i.e., maximizing fitness by consuming food and avoiding poison. We apply the FCA to the neurocontrollers of the two S-type agents (S10 and S22) and the SP-type agent (SP10) by performing lesioning experiments and testing the performance levels of the agents. For each agent, a small training set was chosen using the adaptive algorithm (section 3.2.4) or a random selection, and the FCA was applied 10 times (section 3.2.3). The result of the FCA is the contribution vector \( \mathbf{c} \) and the performance prediction function \( f \). The prediction capability of the FCA is measured by computing the MSE on a test set consisting of all \( 2^{10} \) configurations in S10 and SP10, and of 30,000 random configurations in S224. The

\[ \text{4} \] \Since obtaining the full \( 2^{22} \) configuration set is impossible, we chose a large enough set such that almost any randomly chosen set of that size results in practically the same error.
normalized MSE on these test sets were all below 0.07, demonstrating that the FCA succeeded in finding a pair \( \{ c, f \} \), which brings the error in Eq. 3.2 to a very low value even when trained on a very limited subset of the \( 2^N \) configuration space.

![Contribution values of the agents' neurons](image)

**Figure 3.4:** *Contribution values of the agents’ neurons.* The mean and standard deviation of the contribution values computed by the FCA are plotted in each panel (vertical axis). A. Agent S10. FCA trained on 45 lesioning examples obtained by the Adaptive Lesioning algorithm. B. The corresponding ten performance prediction functions. The x-axis is \( m \cdot c \), y-axis is the predicted performance \( f(m \cdot c) \) (Eq. 3.1). C. Agent SP10. FCA trained on 45 lesioning examples obtained by the Adaptive Lesioning algorithm. D. Agent S22. FCA trained on 99 random lesioning examples. The inset in panels A, C and D depicts one performance prediction function. Neurons 1-4 (on the horizontal axis) are the motor neurons (see Fig. 2.1).

Figures 3.4A, 3.4C and 3.4D depict the contribution vectors computed for the three agents. Each sub-figure depicts the mean and standard deviation of the contributions of the different neurons of the agent, over the 10 runs of the FCA. Importantly, the FCA converges to similar minima of the error in all runs, yielding consistent measures of the contributions. As expected from previous receptive field analysis of the agents (see Chapter 2), we find that the command neurons (neuron 5 in S10, neuron 6 in SP10, and neuron 8 in S22) receive significant and positive contributions in all three agents.
Interestingly, in the two S-type agents, some of the motor neurons (neurons 1 through 4) receive high contribution values, testifying to the importance of the recurrency from the output units\(^5\). This is consistent with the notion that the feedback from the motor units is required for signaling eating events, enabling the agents to switch to and remain in grazing mode. In \textit{Sp10}, where location information is available to the agent, this recurrent connection from the motor neurons is not significant. The inset in each panel plots the performance prediction function obtained by the FCA. All ten performance prediction functions look similar (as can be seen for the case of \textit{Sp10}, in Figure 3.4B), so one was chosen for clarity of exposition. The similar form of the prediction functions again testifies to the convergence of the FCA to similar minima of the prediction error. Focusing on agent \textit{Sp10} we note that when all neurons of are lesioned ($\mathbf{m} \cdot \mathbf{c} = 0$), the agents’ performance is about 0.3 of the intact network ($\mathbf{m} \cdot \mathbf{c} = 1$) baseline performance. The performance is not 0 because the input neurons are left intact, as well as the activity transmitted from the output neurons to the motors. Thus, the fully lesioned network corresponds to the minimal network of input and output neurons with no internal processing. Lesioning any of the five identified significant neurons results in a quite sharp decrease in performance as is evident from the large slope of $f$ near 1. The function allows one to fairly accurately predict the performance of any lesioned state using the contribution values depicted in panel A (Eq. (3.1)). Indeed, the normalized MSE (Eq. (3.2)) in this case is 0.015 (explaining 98.5% of the variance).

### 3.3.3 Comparison to the Single Lesion Approach

An important aspect of the FCA is the integration of information obtained from multiple lesions, going beyond the limited information available in single lesions alone. Here, we compare the FCA approach to the single lesion approach using the evolved agents. To study the significance of using multiple lesions, we compute the null hypothesis, i.e. we calculate the contribution values relying only on the performance levels obtained when single units are lesioned. The contribution of neuron $i$ is taken to be the decrease in the performance due to lesioning that neuron alone, normalized as before\(^6\). The black bars in Figure 3.5(A) depict the contributions computed from single lesions for the general performance of agent \textit{Sp10}. The light bars describe the contributions computed by the FCA, using all possible lesioning configurations. As evident, the contributions assigned by the FCA differ significantly from those obtained with knowledge of single lesion results.

Do the contribution values computed by the FCA using multiple lesions describe the system more accurately than those inferred from single lesions? The obvious criterion is

\(^{5}\)Recall that under lesioning, the motor actions governed by an output neuron are kept intact, and only the information transfer to the rest of the network is disturbed.

\(^{6}\) $c_i = (1 - p_i) / (\sum_{i=1}^{N} |1 - p_i|)$, where $p_i$ denotes the performance when neuron $i$ is lesioned alone.
Figure 3.5: Comparison between the single lesion approach and the FCA on agent S10. A. Contribution values obtained by the two methods. B. Predicted versus actual performance on the test set using single lesions information. C. Predicted versus actual performance on the test set using the FCA.

the prediction ability. In principle, the single lesion approach implies a linear prediction function, which would lead to very inaccurate performance predictions. Thus, to obtain a fair comparison, we take the contribution vector found by the single lesion analysis, and optimally fit its corresponding prediction function to the test set. This ensures that any difference in prediction capabilities stems from a difference in the contributions themselves. Panels B and C of Figure 3.5 compare the predicted versus actual performance, using single lesions (B) and the FCA (C). Clearly, single lesions do not reveal the true contributions of the neurons, testifying to the existence of some form of interaction between the effects of the different units on the network function.

Figure 3.6 compares the test MSE obtained for several agents and tasks using single lesion information (black bars) and using the FCA derived contributions (light bars). These results demonstrate again that in some agents and tasks complex interactions exist, yielding suboptimal contributions when considering only single lesions. The information from multiple lesions in such networks is indeed essential for revealing the true contributions
of the network elements. The more complex, possibly non-linear interactions are modeled well by the FCA, because although being a linear model, it is generalized by the non-linear prediction function\(^7\). However, obviously many forms of interactions cannot be modeled with the basic FCA. A classical example of such an interaction was given by Sprague in the neuroscience literature [Sprague, 1966], showing that deficits in orienting towards a stimulus, resulting from large cortical visual lesions, can be reversed through additional removal of contralateral areas. This has been known as the Sprague effect, or paradoxical lesioning [Hilgetag et al., 2000, Lomber and Payne, 2001]. To address these challenges we have generalized the FCA to include high-dimensional context-dependent interactions. This important generalization is described in the next chapter of the thesis. Note that for agent SP10, single lesion information suffices to achieve very accurate contributions, testifying to the simpler functional organization of the network\(^8\).

![Graph showing comparison of MSE between single lesion approach and FCA.](image)

**Figure 3.6:** *Comparison of MSE obtained by the single lesion approach and the FCA.* Train and test sets are the full \(2^{10}\) configuration set for S10 and SP10. For S22, the training set is a 5,000 configuration set, and the test set consists of 20,000 configurations. All predictions used an optimally fitted prediction function to the test set (see text). In most tasks the standard deviation of the MSE is too small to be seen.

\(^7\)A simple example of such non-linear interactions is a two unit system which functions perfectly unless both units are lesioned. The basic FCA models such a case by assigning both units a contribution value of 0.5, with \(f(0) = 0\) and \(f(0.5) = f(1) = 1\).

\(^8\)However, it still may be that localization of a specific task in this agent requires an FCA analysis.

47
3.4 Localization of Several Tasks

In the previous section we have discussed the results of applying the FCA to the overall performance of the agent, its fitness. We showed how the FCA can be used to find the contribution of each neuron to the task of maximizing the fitness. However, as discussed in chapter 2, the EAAs perform two additional subtasks – grazing and exploration. In this section, we show how the FCA can be used to localize these tasks in parallel.

Consider an agent with a neurocontroller network of interconnected neurons that performs a set of $K$ different functional tasks. Addressing the question of which elements contribute to which tasks, it is natural to think in terms of a contribution matrix, where $C_{ik}$ is the contribution of element $i$ to task $k$, as shown in Figure 3.7. The elements studied may be neurons, synapses or any higher-order module.

<table>
<thead>
<tr>
<th>Element</th>
<th>Task 1</th>
<th>Task 2</th>
<th>$\cdots$</th>
<th>Task K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$C_{11}$</td>
<td>$C_{12}$</td>
<td>$\cdots$</td>
<td>$C_{1K}$</td>
</tr>
<tr>
<td>2</td>
<td>$C_{21}$</td>
<td>$C_{22}$</td>
<td>$\cdots$</td>
<td>$C_{2K}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\cdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$N$</td>
<td>$C_{N1}$</td>
<td>$C_{N2}$</td>
<td>$\cdots$</td>
<td>$C_{NK}$</td>
</tr>
</tbody>
</table>

$\rightarrow S_2$

$L_1$

Figure 3.7: The Contribution Matrix. The network consists of $N$ elements and performs $K$ different functional tasks. The entry $C_{ik}$ is the contribution of element $i$ to task $k$.

As before, the data analyzed for computing the contribution matrix is gathered by inflicting a series of multiple lesions onto the agent’s neurocontroller network. Under each lesion, the resulting performance of the agent in different tasks is measured. Given this data, the FCA finds the contribution values $C_{ik}$ that provide the best performance prediction on average for all possible multi-site lesions (the $k$th column of the matrix is the contribution vector of task $k$ as described in section 3.2). Following the spirit of [Lashley, 1929], the localization $L_k$ of task $k$ can now be defined as a deviation from equipotentiality along column $k$ (e.g., $L_1$ in Figure 3.7), and similarly, $S_i$, the specialization of neuron $i$ is the deviation from equipotentiality along row $i$ of the matrix (e.g., $S_2$ in Figure 3.7).

More formally, $L_k$ is the standard deviation of column $k$ of the contribution matrix
Figure 3.8: Contribution values for Grazing and Exploration. A: Grazing task. B: Exploration task.

Divided by the maximal possible standard deviation,

\[ L_k = \frac{\text{std}(C_k)}{\sqrt{(N-1)/N^2}}. \]  

(3.7)

Note that \( L_k \) is in the range \([0,1]\) where \( L_k = 0 \) indicates full distribution and \( L_k = 1 \) indicates localization of the task to one neuron alone. Similarly, if neuron \( i \) is highly specialized for a certain task, \( C_i \) will deviate strongly from a uniform distribution, and thus we define \( S_i \), the specialization of neuron \( i \), as

\[ S_i = \begin{cases} 
2 \cdot \text{std}(|C_i|) & \text{if the number of tasks, } K, \text{ is even} \\
\frac{2\cdot\text{std}(|C_i|)}{\sqrt{(K^2-1)/K^2}} & \text{otherwise.} 
\end{cases} \]  

(3.8)

Note that \( S_i \) uses the absolute value of the contributions to reflect the intuition that the specialization of the unit is determined more by the magnitude of its contribution than by its sign. Again, \( S_i = 1 \) indicates maximal specialization. We note, however, that in principle other measures can be defined. The important point is that the contribution matrix enables one to define such quantitative measures to capture qualitative notions.

We concentrate on one agent (§10), and consider the two tasks which correspond to the two emergent behaviors of successful agents, exploration and grazing as described above (see Eqs. 3.5 and 3.6). The FCA is performed separately for each task yielding a contribution vector and a prediction function for the task. Figure 3.8 depicts the contribution values computed by the FCA for the two tasks. As evident, the localization of the two tasks within the network differs quite considerably. Notably, neuron 10 which is an interneuron, contributes positively to grazing behavior, but hinders exploration.
Given these data, we can now construct the contribution matrix of the agent, depicted in Table 3.1, along with the measures of localization and specialization computed from it. Since the size of the evolved neurocontrollers was set arbitrarily, a number of neurons do not participate in any task the agent performs, and hence the effective size of the network, considering only neurons with non-vanishing contributions, is smaller. This should be taken into account when considering task localization in the network. Thus, the effective localization $\tilde{L}_k$ of the different tasks performed by the agent (Table 3.1), is computed only on the non-vanishing neurons (neurons 1,2,3,5 and 10 in this agent), after normalizing the sum of their absolute values of contributions to 1. In this agent, exploration is more localized than grazing. Note that even though neuron 10 has contribution values of similar absolute magnitude as neuron 5, its specialization is double that of the latter, as it contributes positively to grazing behavior while actually hindering exploration.

<table>
<thead>
<tr>
<th>N</th>
<th>Grazing</th>
<th>Exploration</th>
<th>$S_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1696</td>
<td>0.5034</td>
<td>0.3338</td>
</tr>
<tr>
<td>2</td>
<td>0.2916</td>
<td>0.2344</td>
<td>0.0571</td>
</tr>
<tr>
<td>3</td>
<td>0.0878</td>
<td>0.1324</td>
<td>0.0445</td>
</tr>
<tr>
<td>4</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>5</td>
<td>0.2228</td>
<td>0.0740</td>
<td>0.1488</td>
</tr>
<tr>
<td>6</td>
<td>0.0015</td>
<td>0.0006</td>
<td>0.0009</td>
</tr>
<tr>
<td>7</td>
<td>-0.0001</td>
<td>0.0001</td>
<td>0.0002</td>
</tr>
<tr>
<td>8</td>
<td>0.0007</td>
<td>0.0001</td>
<td>0.0006</td>
</tr>
<tr>
<td>9</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>10</td>
<td>0.2257</td>
<td>-0.0547</td>
<td>0.2804</td>
</tr>
</tbody>
</table>

Table 3.1: Localization and specialization in agent S10. Each entry to the table is the contribution of the corresponding neuron to each of the two subtasks – grazing and exploration. The internal part of the table is thus the contribution matrix (see Fig. 3.7) of agent S10. The localization, $L_k$, effective localization, $\tilde{L}_k$, and specialization, $S_i$, are displayed for each task and neuron.

<table>
<thead>
<tr>
<th>$L_k$</th>
<th>0.3684</th>
<th>0.5323</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{L}_k$</td>
<td>0.1698</td>
<td>0.4691</td>
</tr>
</tbody>
</table>

50
3.5 Synaptic Analysis

The FCA is not limited to a specific level of analysis, and thus the same network can be studied concomitantly on the neuronal and synaptic level. In this section we apply the FCA to agent S10 analyzed above, to determine the contribution values of its synapses, and hence, the underlying structure of its synaptic network. The neuronal analysis presented above revealed that five out of the ten neurons are significant. Thus, as a first step we consider a reduced network consisting of only the significant neurons and their synapses. Recall, however, that insignificance of a motor neuron testifies only to the insignificance of its outgoing synapses, and not to the insignificance of its activity insofar as it controls the motor. Thus, the reduced network of agent S10 contains the six indispensable neurons (the four motor neurons, neuron 5 and neuron 10), and all the synapses connecting them (excluding the outgoing synapses from neuron 4, which has a vanishing contribution value). The reduced network has a total of 30 synapses\(^9\). The agent driven by this reduced network still attains a very high level of performance, 0.36, which is 90% of the performance obtained by the original agent, driven by the fully connected intact neurocontroller. The next step is to use the FCA to find the backbone of the network, i.e., the significant synapses within the reduced neurocontroller.

Applying the FCA for this synaptic analysis, with a training set consisting of 2000 randomly chosen configurations, the normalized MSE on a randomly chosen test set of 20000 configurations is 0.05. The training set is extremely small relative to the full configuration space (\(2^{20}\) configurations). Nevertheless, the FCA managed to reach a fairly low test error, testifying to the potential scalability of the method. Figure 3.9A depicts the synaptic contribution values obtained by the FCA. The right part of the figure (marked “Neuron”) depicts the contribution values of the neurons of the agent S10, as computed by the FCA on the neuronal level. Figure 3.9B depicts the weight values of the agent’s neurocontroller, normalized such that the sum of the absolute values of the synaptic weights is 1, also accompanied by the neuronal contributions computed by the FCA. Evidently, the network structure emerging from considering the synaptic contributions (panel A) differs quite significantly from that inferred directly from the synaptic weights (panel B) (correlation of 0.31).

Is the synaptic significance obtained by the FCA more accurate than that derived from the raw synaptic weights themselves? To answer this question we lesion the synapses incrementally and measure the deterioration of the agent’s performance. Figure 3.10 depicts the performance of the agent as a function of the number of lesioned synapses for

\(^9\)We consider here only the internal network and not the synapses coming from input neurons, as lesioning those synapses causes a degree of damage which basically causes a total malfunctioning of the agent.
Figure 3.9: Comparison of synaptic contributions computed by the FCA and the synaptic weights. A. The FCA computed contributions of the synapses and neurons (right row) of the reduced network. B. The weights of the reduced network synapses, normalized such that the sum of the absolute values of these weights is 1. Neurons 1, 2, 3, 5 and 10 are recurrently connected. Neuron 4 is the motor neuron (controlling the mouth) whose outgoing synapses are insignificant.
three methods of incremental lesioning starting from the intact reduced network of the 30 internal synapses. In the first lesioning experiment we use the synaptic contribution values computed by the FCA and lesion by ascending order of significance. The graph depicts the mean and standard deviation using contribution values from ten different FCA runs. In the second lesioning experiment we use the absolute weight as a measure of importance and lesion by ascending order of the absolute value of the weights. In the third experiment we lesion incrementally by choosing the next synapse to be lesioned randomly. The graph depicts the mean and standard deviation of 100 such choices.

As evident, the FCA is much better at identifying the significant synapses of the network. For example, leaving only 15 synapse in the entire internal network, i.e. 15% of the total internal network and 50% of the reduced network, salvages 80% of the original performance level. In contrast, using the weights as a significance measure to choose the 15 important synapses leaves only 50% of the original performance\(^{10}\). Not surprising, choosing 15 synapses randomly completely destroys the agent’s performance. In summary, the FCA proves very reliable in identifying the significant network backbone and can be used to select synapses for pruning a neural network.

Using the FCA computed synaptic contributions, a minimal effective connectivity can be inferred, to simplify the analysis of the network. This is especially important for fully recurrent networks where understanding the structure becomes a daunting task. Looking at Fig 3.9A, one observes that there are nine synapses with relatively large contribution values, compared to the rest. Figure 3.11A depicts the network formed by these nine largest synaptic contributions, and correspondingly Fig 3.11B depicts the nine synapses with largest weight values (in absolute value). Although part of the resulting backbone is similar, there are two important differences. First, although the synaptic weights imply that the connections into neuron 4 are important, the FCA correctly reveals that this is not the case. Second, the FCA analysis clearly shows the central role of the command neuron (neuron 5), a feature of the network much less apparent when considering the synaptic strengths. The functional role of the command neuron in these networks has been studied in detail in [Aharonov-Barki et al., 2001]. Examining the network backbone of Figure 3.11A one can postulate that the synapse from neuron 1 (forward step) to neuron 2 (left turn) is responsible for the move and turn behavior observed in grazing mode. Moreover, the synapses from the turn motor neurons (2 and 3) to the command neuron (neuron 5) is clearly important, as expected from previous analysis.

As discussed above the network functional connectivity emerging from the FCA analysis is more accurate in terms of being able to predict the network performance. Note, that although only nine synapses receive a high contribution value (Figure 3.9), the per-

\(^{10}\)Using the actual weight value and not the absolute value achieves even lower performance on all levels of synaptic lesioning.
Figure 3.10: *Agent performance as a function of synaptic lesioning degree*. Performance of agent S10 as a function of the number of synapses lesioned, for three different methods of incremental synaptic pruning. The performance is given relative to the performance of the original fully intact network. All methods start from the reduced network consisting of 30 synapses, i.e. 70 internal synapses lesioned. *By contribution*: For each of ten FCA runs, synapses are incrementally lesioned by ascending order of their contribution value. Line depicts mean and standard deviation of measured performance levels over the ten runs. *By absolute weight*: Performance levels following lesioning by ascending order of absolute synaptic weights. *Random*: Mean and standard deviation of performance levels over 100 random orders of synaptic lesions.

Performance of the network with only these synapses intact is still rather low (Figure 3.10). This is because other synapses are important in certain combinations, improving the performance if they are intact\(^\text{11}\). However, the nine high contribution synapses are strongly beneficial in many lesioning configurations, and are hence picked up by the basic FCA.

### 3.6 Discussion

We have presented here a functional contribution analysis that addresses the fundamental challenge of localization of function in neural networks. The FCA is based on the

\(^{11}\text{High-Dimensional FCA is necessary to reveal these combinations.} \)
Figure 3.11: *Network backbone*. A. The nine synapses of the internal network with the largest contribution values. B. Similarly for the largest synaptic weights (by absolute value). Dashed lines denote negative values. The thickness of the lines scale with their absolute value, and are normalized to have the same sum in each panel.

assignment of contribution values to the basic elements of the network, such that the ability to predict the network’s performance in response to multi-lesions is maximized. The algorithm is thoroughly examined in an EAA environment. The results shown testify that:

- In the recurrent yet simple EAA neurocontrollers studied here, the basic version of the FCA suffices to portray a stable set of contributions and yield fairly accurate multi-lesion predictions. These are significantly better than those obtained with a single lesion analysis method. The FCA may be efficiently trained on a relatively small subset of all possible multi-lesion configurations. It can be used to localize several subtasks that the agent performs, and compute the corresponding localization and specialization indices.

- FCA analysis on the synaptic level testifies to the potential scalability of the approach for larger systems. It also presents an important method for synaptic pruning and visualizing the effective “backbone” of the network architecture.

However, an important limitation of the current FCA is that some forms of interactions cannot be modeled with the basic one-dimensional approach presented here. Rather, the accurate analysis of complex networks requires high-dimensional descriptions involving compound elements that are composed of several basic units (neurons or synapses) that interact together. Such high-dimensional analysis is the topic of the next chapter, and as shown there is essential, for example, for an accurate account of subnetworks displaying the “paradoxical lesioning” effect.
The FCA has potential significance beyond the scope of EAAs. Multi-lesion analysis algorithms like the FCA will become an essential tool in neuroscience for the analysis of reversible inactivation experiments, combining reversible neural cooling deactivation with behavioral testing of animals. These methods alleviate many of the problematic aspects of the classical single lesioning technique (ablation), enabling the acquisition of reliable data from multiple lesions of different configurations (for a review see [Lomber, 1999]). If the underlying network of brain regions studied is simple enough, then perhaps the “biological” ablation method of lesioning currently employed in these experiments will suffice. However, our study suggests that if these networks are more complex, ablation lesions may result in excessive perturbations of the system and fail to reveal its normal operation (the problematic nature of analyzing lesioning data due to this excessive lesioning has also been discussed in [Young et al., 2000]). If that will turn out to be the case, then other lesioning methods that cause smaller perturbations (much alike stochastic lesioning) should be employed.

FCA-like algorithms could also be used for the analysis of transcranial magnetic stimulation studies which aim to induce multiple transient lesions and study their cognitive effects (see [Walsh and Cowey, 2000] for a review). They could also possibly be used for studying the functional efficacy of synaptic projections on different components of the neuronal dendritic tree, in parallel to recent studies examining this question with other methods [London et al., 2002]. In both cases, stochastic lesioning may prove to be the preferred lesioning method employed both experimentally and in the FCA analysis. Another potentially interesting application is the analysis of functional imaging data. One direction is to assess the contributions of each element to others, i.e., extending previous network’s effective connectivity studies employing linear models (e.g., [Friston et al., 1993]) with High-Dimensional FCA. The second direction involves the application of the FCA to the meta-analysis of functional imaging data. This will require development and study of a continuous version of the current binary FCA representations, where results of various activation studies of similar tasks are viewed as different “multiple-lesion” probes of the task in hand.

Finally, the FCA studied in this chapter has focused on studying each task separately, and then assembling this information and presenting it in a contribution matrix. The latter, however, could serve as an object of further elaborate analysis beyond the computation of localization/specialization indices. This analysis could involve a singular value decomposition of the contribution matrix, reducing it to a lower dimension and revealing possible hidden similarities in the localization of different tasks, in a manner analogous to the Latent Semantic Analysis used in Information Retrieval applications [Deerwester et al., 1990].
Chapter 4

High-Dimensional Localization of Function: Re-thinking the Concept of Contribution

The material described in this chapter has appeared in:
* These authors contributed in equal parts.

4.1 Introduction

The previous chapter introduced the basic 1-dimensional FCA, which addresses the challenges posed by the lesioning approach for the localization of function in neural systems. By utilizing multiple multi-lesion experiments, FCA assigns specific contribution values to the basic elements of the neurocontroller. These contribution values quantify the importance of each element to the task(s) the agent performs. They enable the accurate prediction of the agent’s performance in any new, unseen lesioned state. This chapter examines in more depth how the FCA may be utilized to study function localization in EAA neurocontrollers. To this end, the basic FCA is generalized to high-dimensional analysis, using high-order compound elements. Such elements are composed of conjunctions of simple elements, and enable the explicit expression of sets of neurons or synapses whose contributions are interdependent, i.e., the contribution of each of the simple elements depends on the state of the other elements in the set. This high-dimensional description is important, indeed essential, for an accurate analysis of EAA neurocontrollers, in which the interactions between elements may be high-dimensional and complex. The introduction of compound elements requires a re-thinking of the concept of the contributions of simple elements defined originally in the basic FCA. While the contributions of the simple elements can be reconstructed regardless of the dimension of the FCA, one needs to think
in terms of the contributions of compound elements to capture the interactions forming functional groups in the network.

The FCA can be applied to several different tasks performed by the agent. For each function (task), the FCA computes a separate contribution vector describing the contribution of the system elements to that specific function. Thus, it is a method for function localization, detailing which elements participate in which functions. Using the contribution vectors for the different tasks, we have defined quantitative measures of function localization and element specialization in the network, as discussed in the previous chapter. In this chapter we present results of the High-D FCA in the general task of maximizing fitness. However, since the contributions of the simple elements can be reconstructed, it also supplies the basis for measuring function localization.

High-dimensional FCA is carried out via an efficient algorithm that selects the most important compound elements out of a possibly large set of candidates. A detailed high-dimensional analysis of an agent’s neurocontroller is performed, and the structure of the emerging higher-order compound elements is explored. Capitalizing on our ability to perform multi-lesion experiments in EAAs in a computationally tractable manner, high-dimensional FCA is shown to be a new method enabling a systematic and rigorous analysis of function localization in EAA neurocontrollers.

Throughout this chapter, we focus on the analysis of agent S10. In the previous chapter, we have presented a one-dimensional analysis of this agent that localizes the two behavioral sub-tasks (grazing and exploration) and the agent’s overall survival task in its neurocontroller. Here we focus on a detailed high-dimensional analysis of the former, i.e., of the localization of its overall survival task.

The chapter is organized as follows: Section 4.2 extends the basic FCA to higher dimensions. Section 4.3 describes an in-depth high-dimensional analysis of an EAA neurocontroller. Our results and their implications to the analysis of EAAs are discussed in section 4.4.

4.2 High-Dimensional FCA

4.2.1 Motivation and Definition

The results shown in the previous chapter demonstrate that the contributions of the units of simple EAA neurocontrollers can be computed with the basic 1D-FCA. However, to deepen one’s understanding of the agents’ neurocontrollers, it is imperative to look further and examine the nature of the interactions between these units. The units of the controller may interact in such a way that the contribution value of a unit is not a single constant value but rather depends on the state (lesioned or intact) of the other units. Indeed, certain forms of interaction cannot in principle be described by the basic FCA. A
classical example of such an interaction was given by Sprague in the neuroscience literature [Sprague, 1966], showing that deficits in orienting towards a stimulus, resulting from large cortical visual lesions, can be reversed through additional removal of contralateral areas. This has been known as the Sprague effect, or paradoxical lesioning [Hilgetag et al., 2000, Lomber and Payne, 2001]. Such effects have posed an intriguing conceptual challenge: if an area is beneficial for a behavior (lesioning it hinders performance), how can its lesioning under a different condition, i.e., when another structure is already lesioned, improve the behavior? The solution is that the utility of such a unit is context-dependent – it depends on the state of the rest of the network. To address these challenges the FCA has now been generalized to include high-dimensional context-dependent interactions. This generalization is based on the usage of compound elements. Extending upon the simple, single-unit elements used in the FCA, a compound element \( \pi = \{\pi_1, \pi_2, \ldots, \pi_k\} \) denotes a specific combination of a few simple elements. The order \( k \) of a compound element is the number of simple elements composing it \( (k = |\pi|) \), and this determines the dimensionality of the FCA analysis. In the basic FCA only simple elements of order 1 are used, and hence it is denoted a 1D-FCA.

In high-dimensional FCA, we concatenate the lesioning state of the simple and compound elements to form the lesioning configuration vector \( \mathbf{m} \), such that its length is now \( N + N_c \), where \( N_c \) is the number of compound elements. A compound element is considered intact only if all of the simple elements composing it are intact. Analogously, \( \mathbf{c} \) is of length \( N + N_c \) as well, describing the contributions of both simple and compound elements. Thus, the performance prediction given by Eq. (3.1) still holds in high-dimensional FCA.

Section 3.2.4 demonstrated that the 1D-FCA generalizes well when a sample set is used for training instead of the full configuration set. Throughout this chapter, the training set used for the High-D analysis is the full configuration set, to maximize the accuracy of the results. However, in Section 4.3.2 we demonstrate that the High-D FCA also generalizes well and obtains a fairly accurate MSE with a relatively small sample of the full configuration set.

### 4.2.2 Selecting Compound Elements

Adding compound elements to the description of the system introduces the problem of selecting the most informative ones. The simplest iterative method is greedy: at each iteration one compound element is added to the set of elements until a stopping criterion is reached. To select the element to be added, the FCA is run on a training set including this candidate element and the current set of elements. The compound element which leads to the smallest MSE (on the training set) is selected and added to the element set. The cycle then repeats with the new, extended element set. The greedy algorithm is computationally very expensive, and hence impractical when the potential compound
element set to be searched is large. A faster approach, suitable only for 2D-FCA, is to compute the FCA using all order-2 elements, select those which have the highest contribution values, and then re-run the FCA using the reduced set of elements (the Highest Contributions algorithm). However, this approach requires $O(N^2)$ compound elements, and hence becomes intractable when $N$ is large. We devised a more efficient approximation algorithm, called CERE (Compound Element Error Reduction Estimation) which uses estimates of the reduction in the prediction error that each conjunction yields.

CERE estimates the prediction error when a compound element $\pi$ is added to the current set of elements by,

$$\Delta MSE_\pi = \left(\frac{1}{2^N n_\pi}\right) \sum_{m \in T_\pi} \left( [f(m \cdot c) - \bar{p}_m]^2 - [f(m \cdot c + \bar{c}_\pi) - \bar{p}_m]^2 \right),$$  

(4.1)

where $T_\pi$ is the set of lesioning configurations from the training set that match $\pi$. A lesioning configuration $m$ is said to match an element $\pi$ if all of the elements of $\pi$ are intact in $m$. The term $\left(\frac{1}{2^N n_\pi}\right)$ normalizes the error on the training set to the expected error over the complete set of $2^N$ configurations. $N_\pi$ is the number of possible configurations that match $\pi$ (equal to $2^{N-k}$, where $k = |\pi|$ is the order of $\pi$), and $n_\pi$ is the number of such configurations in the training set. $\bar{c}_\pi$ is the contribution value assigned to $\pi$, computed to maximize $\Delta MSE_\pi$ by a simple one-dimensional gradient descent. The CERE algorithm is comprised of the following steps:

1. Initialize the set of compound elements to the empty set. Compute $\{c, f\}$ using the FCA on simple elements.

2. For each candidate compound element $\pi$,
   (a) Initialize $\bar{c}_\pi$ to 0.
   (b) Update $\bar{c}_\pi$ via gradient descent to maximize $\Delta MSE_\pi$ (Eq. (4.1)):

$$\bar{c}_\pi \leftarrow \bar{c}_\pi + \epsilon \frac{\partial}{\partial \bar{c}_\pi} \Delta MSE_\pi.$$  

(4.2)

   (c) Iterate step 2b a given number of times.

3. Select the compound element $\pi$ that maximizes the reduction in the prediction error $\Delta MSE_\pi$, and add it to the set of compound elements.

4. Recompute $\{c, f\}$ using the FCA, using all $N$ simple elements and all compound elements selected so far.

Steps 2-4 repeat until a stopping criterion is met: either a predefined number of compound elements is used, or the prediction error falls below a threshold.
Figure 4.1: *Comparison of compound element selection algorithms.* Normalized test MSE vs. the number of order-1 and order-2 compound elements, for each of the algorithms (see text). “Random” denotes random selection of order-2 compound elements. The training and test set is the complete $2^{10}$ configuration set. 1D-FCA and full 2D-FCA (using all order-2 compound elements) are depicted as horizontal lines across the figure. All results are average of 10 runs.

A simpler version of CERE, called Linear CERE, is faster to compute and, as shown below, on the evolved agent analyzed here, outperforms CERE. It is essentially the same as CERE, except that $f$ is ignored, or rather, treated as if it were the identity function. Eq. (4.1) is replaced by,

$$
\Delta MSE_\pi = \left( \frac{1}{2N} \frac{N}{n_\pi} \right) \sum_{m \in T_\pi} \left( [m \cdot c - p_m]^2 - [m \cdot c + \hat{c}_\pi - p_m]^2 \right) .
$$

(4.3)

Differentiating Equation 4.3, one can obtain closed form solutions for $\hat{c}_\pi$ and $\Delta MSE_\pi$.

$$
\hat{c}_\pi = \frac{1}{n_\pi} \sum_{m \in T_\pi} (p_m - m \cdot c)
$$

(4.4)

and, by replacing $\hat{c}_\pi$ in Equation 4.3,

$$
\Delta MSE_\pi = \frac{N}{2N} \hat{c}_\pi^2 .
$$

(4.5)

The accuracy of all four algorithms in the order-2 analysis of the agent is compared in Figure 4.1. Evidently, in this example all 2D incremental algorithms significantly outperform the basic 1-dimensional method, and come close to the performance of the full 2D algorithm at fairly moderate numbers of compound elements.
4.3 High-Dimensional Neurocontroller Analysis

4.3.1 Compound Elements and Interactions

The previous chapter described agent analysis using 1D-FCA. Here we perform a high order 2-dimensional analysis of agent S10 and describe what can be learned about the system from the emerging structure of compound elements. Figure 4.2A depicts the contribution values of the simple and compound elements when eight order-2 compound elements are used to describe the agent’s performance\(^1\). The figure demonstrates that the identity of the compound elements chosen by the Linear CERE and their corresponding contribution values are stable. The addition of the eight compound elements (out of the possible 45 pairs) considerably decreases the prediction error, from the 0.0153 reached by the 1-dimensional analysis to 0.0047. Even though more than eight compound elements were chosen over all 10 runs, only eight such elements were given non-negligible contribution values. These pairs are all combinations of significant simple elements, i.e., those with non-vanishing contribution values in the original 1-dimensional analysis (Figure 3.4A). Analogously, Figure 4.2B depicts the contribution values of the simple and compound elements when eight compound elements of maximum order 3 are used, resulting in an MSE of 0.0037. Again, the selection of compound elements is very stable (there are 165 possible pairs and triplets), as are the values given to the chosen elements. The accompanying performance prediction functions are depicted in panels C and D.

What can be learned from these results? An important observation is that the contributions of the simple elements (Figure 3.4A) change significantly when higher order elements are introduced. This is because the overall significance of an element is no more measured by its single contribution. Rather, compound elements that include this unit must be considered to reveal its overall contribution. As shown in section 4.3.3, using the contribution values of the compound elements to recompute the simple elements’ contributions restores the results of the 1D-FCA. Thus, the higher order analysis singles out the interactions and modulations in the network faithfully, without losing the information about the original, basic units. In the 2D analysis, the command neuron (neuron 5) forms a significant compound element pair with every other significant neuron, two of which remain as 2D elements in the selected compound element set when 3D analysis is employed. When 3D analysis is employed, all the 2D elements selected in the 2D analysis still appear, either as pairs, or as part of a triplet in which the effect of the pair is modulated by a third element.

Figure 4.3A depicts the network whose nodes are the five significant neurons and whose edges are the eight major 2D interactions. For comparison, the network whose nodes

\(^1\)After which the reduction in the prediction error levels off.
Figure 4.2: 2D and 3D FCA. A, B. Mean and standard deviations of the contributions selected by 2D (A) and 3D (B) FCA. Results are from 10 runs of the linear CERE algorithm (limited to 8 compound elements). Both training and test sets are the complete $2^{10}$ possible configurations. Some elements were selected only in some of the runs. Hence, even though eight compound elements were selected, over all runs more than eight elements are chosen and depicted here (for calculating mean and standard deviation, the contribution values were taken as 0 when not chosen). C, D. Performance prediction function of the 2D-FCA (C) and 3D-FCA (D). All 10 runs result in similar functions. For clarity, the figure depicts one representative function.
Figure 4.3: *Major interactions and synapses.* A. The eight largest 2D contributions between the five significant neurons (by absolute value) are depicted as connections. B. The eight synapses with highest contributions (by absolute value) between the five significant neurons. C. Similarly for the largest synaptic weights (by absolute value). Dashed lines denote negative values. The thickness of the lines scale with their absolute value, and are normalized to have the same sum in each panel.

are the same but whose edges are the eight (out of the 20 possible) synapses with the highest contributions (computed using a 1D-FCA on the synapses [Aharonov et al., 2003]) is depicted in panel B. For additional comparison, the network with the eight largest synaptic weights (in absolute values) connecting these five neurons is depicted in panel C. The similarity between the two FCA-derived networks (A and B) is apparent. Both FCA analyses reveal a strong underlying edges’ backbone composed of the neuron pairs \{1, 2\}, \{2, 5\} and \{5, 3\}, with several weaker edges (two of which appear in both). In both, the importance of the command neuron (neuron 5) is apparent. In contradistinction, the structure revealed by the synaptic weights’ network (C), although similar in some ways to the FCA-computed synaptic contributions, differs in a significant manner. The importance of the command neuron is less apparent, and the strong synapse between neurons 2 and 3 does not appear important from the FCA analyses. Moreover, although the synaptic weight from neuron 1 to 10 is large (C), testifying to the fact that the activity of neuron 1 strongly influences that of neuron 10, this synapse actually receives a small contribution value and no significant interaction between the two neurons is identified (A and B). Thus, neither the significance of synapses, nor the interactions between neurons can be inferred directly by looking at the strength of the synapses in the evolved network.

### 4.3.2 Generalization

As demonstrated in section 3.2.4, the 1D-FCA generalizes well when trained on a small sample of configurations. In this section we demonstrate the generalization capabilities of the High-D FCA. We train the High-D FCA under the same conditions as presented in
Figure 4.4: *Generalization in 2D-FCA*. Comparison of mean and standard deviations of the contributions selected by 2D-FCA. The training set is a randomly chosen set of 200 configurations (black bars) or the full $2^{10}$ configuration set (gray bars). Mean is over 10 runs of the linear CERE algorithm (limited to 8 compound elements). Some elements were selected only in some of the runs. Hence, even though eight compound elements were selected, over all runs more than eight elements are chosen and depicted here (for calculating the mean and standard deviation, the contribution values were taken as 0 when not chosen).

Figure 4.2A, except that the training set is a randomly chosen set of 200 configurations. Like the 1D-FCA, the High-D FCA also generalizes well to the test set consisting of the full configuration set. This is evident from the low test MSE, 0.0074, which is 1.5 times the MSE achieved when *trained* on the full configuration set (this should be compared with the ratio of MSEs in the 1D-FCA, which is 1.3). Figure 4.4 depicts the resulting compound elements that are selected and their corresponding contribution values (as before, when computing the mean and standard deviation, if a compound element was not chosen in a specific run, its contribution is taken to be 0). The resulting contributions are compared with those obtained by using the full set, as appear in Figure 4.2A. As can be seen, the pairs selected as important are consistent, as well as the contribution values assigned to them.

The above results were obtained by randomly selecting the training set. They can be further improved upon by an adaptive algorithm similar to the one presented in section 3.2.4, which smartly selects lesion configurations to more accurately generalize from a small training sample.
Figure 4.5: Comparison of corrected contributions for different maximum order of compound elements. For maximum order 1 and 2, all compound elements were used. For maximum order 3 and 4, 100 high-order compound elements were selected by the Linear CERE algorithm.

4.3.3 Corrected Contributions

The introduction of high-order elements to the analysis blurs the meaning of the contribution values of the simple elements, as discussed above. To infer the contribution values of the simple elements from high-dimensional FCA, the contribution values of the high-dimensional elements should be “credited” back to the simple elements. The corrected contribution value $c'_i$ of the simple element $i$ is defined as

$$c'_i = \sum_{\pi, i \in \pi} c_\pi \cdot 2^{1-k},$$

where $i \in \pi$ denotes that the simple element $i$ is included in the element $\pi$, and $k$ denotes the order (number of elements) of $\pi$. The weighting by $2^{1-k}$ reflects the observation that a compound element of order $k$ is matched by $2^{1-k}$ of the configurations that match its simple elements. Clearly, $c'_i = c_i$ when no compound elements are used. Note that in general $\sum_{i=1}^N |c'_i| \neq 1$, so $c'$ is re-normalized.

Figure 4.5 compares the corrected contribution values of different high-order analyses. The figure compares 1D-FCA, 2D-FCA using all order-2 compound elements, and sets of 100 high-order compound elements selected by the Linear CERE algorithm with maximum orders 3 and 4. The similarity of all four contribution vectors is evident, testifying to the stability of the FCA: The contribution values of the simple elements can be inferred from higher-order analysis, with very similar results. Thus, as desired, high order FCA retains the contribution values of the simple elements, while singling out important interactions between those elements.
4.4 Discussion

This chapter extends the basic 1D-FCA approach to study High-Dimensional FCA in EAAs. As shown, High-Dimensional FCA becomes necessary if one strives to obtain a full and accurate description of function localization in evolutionary neurocontrollers. Efficient algorithms for the selection of the relevant subset of compound elements are presented. High-Dimensional FCA describes the system in terms of relevant functional groups – the compound elements. Each compound element is a set of elements which modulate each other’s contribution value. Moreover, when the set of selected compound elements achieves low MSE, the newly defined elements are such that now each element has a constant contribution value which is independent of the context, i.e., of the state of the rest of the system. Hence, the High-D FCA finds a new functional description of the system, which is not simply the given single elements. In this functional description the contributions of the elements are constant.

The application of High-Dimensional FCA to the analysis of evolutionary neurocontrollers leads to several novel results: 1. A rigorous, quantitative description of localization of function in the neurocontroller, in terms of the contributions of simple and compound elements composing it. 2. Accurate prediction of the effects of possible lesions on the agent’s functioning (i.e., performance). 3. Insights concerning the main subsets of simple elements in the network that interact to modulate each other’s contribution to the system.

The neuroanalysis performed here was primarily focused on a 2D description. However, it is not by any means limited to two dimensions. High-Dimensional FCA also has the potential to estimate the inherent dimension of function localization in the neurocontroller. The dimensionality can be measured by the dimension needed to accurately describe the system, i.e., by identifying the lowest dimension after which the prediction MSE is no longer significantly decreased when the dimension of the analysis is further increased. This touches upon the delicate issue of the functioning of the network as a “whole”, and upon the “complexity” of its processing. Essentially, a lesion-based approach (like that used traditionally in neuroscience, or in FCA) implicitly assumes that the network’s operation can be decomposed, i.e., viewed and understood, on its parts/units level. But if the network’s operation is irreducible, i.e., it cannot be decomposed on the level of its building blocks units, can such a lesioning approach still make sense? High-Dimensional FCA addresses this question by deriving the specific conjunction sets of elementary units that are needed to localize the network functioning, and by providing an upper bound characterizing its dimensionality. From this perspective, High-Dimensional FCA searches for an intermediate level decomposition of the network which best reflects the localization of function within it.
Chapter 5

Non-Linear Temporally Asymmetric Hebbian Plasticity

The material described in this chapter has appeared in:
* These authors contributed in equal parts.

5.1 Introduction

Correlation-based plasticity has long been proposed as a mechanism for unsupervised experience-based development of neuronal circuitry, particularly in the cortex. However, the specifics of a biologically plausible model of plasticity that can also account for the observed synaptic patterns has remained elusive. Two major issues are stability and competition [Miller and MacKay, 1994, Miller, 1996, Abbott and Nelson, 2000, Song et al., 2000, van Rossum et al., 2000, Rao and Sejnowski, 2001, van Ooyen, 2001]. If maps, such as ocular dominance maps, emerge from initially random (but statistically homogeneous) synaptic configurations by a Hebbian mechanism (but see [Crowley and Katz, 2000]), this would imply that there is an inherent instability in the dynamics of synaptic learning which destabilizes an initially homogeneous synaptic pattern. This, however, raises the question what mechanism prevents synapses from growing to unrealistic values when driven by unstable dynamics. The emergence of inhomogeneous synaptic patterns also requires a competition mechanism which makes some synapses decrease their efficacies as other synapses grow in strength. Such competition is absent in the most naive Hebb
rule, which contains only a mechanism for synaptic enhancement. Recent experiments have led to an important refinement of correlation-based or Hebbian learning, by showing that activity-induced synaptic changes can be temporally asymmetric with respect to the timing of pre- and post-synaptic action potentials with a precision of down to tens of milliseconds. Causal temporal ordering of pre- and post-synaptic spikes induces synaptic potentiation, whereas the reverse ordering induces synaptic depression [Levy and Steward, 1983, Debanne et al., 1994, Magee and Johnston, 1997, Markram et al., 1997, Bi and Poo, 1998, Debanne et al., 1998, Zhang et al., 1998, Feldman, 2000, Bi and Poo, 2001, Sjöström et al., 2001].

In this work we address the question of whether temporally asymmetric Hebbian (TAH) plasticity rules provide an adequate mechanism for unsupervised learning of input correlations. Two models of TAH plasticity have recently been studied which differ in the way they implement the weight dependence of the synaptic changes and the boundaries of the allowed range of synaptic efficacies. The additive model [Abbott and Blum, 1996, Gerstner et al., 1996, Eurich et al., 1999, Kempter et al., 1999, Roberts, 1999, Song et al., 2000, Kempter et al., 2001, Levy et al., 2001, Cateau et al., 2002] assumes that changes in synaptic efficacies do not scale with synaptic strength, and the boundaries are imposed as hard constraints. This model retains inherently unstable dynamics while exhibiting strong competition between afferent synapses. Since this model yields binary synaptic distributions, its ability to generate graded representations of input features is restricted. Moreover, due to the strong competition, patterns in the synaptic distribution can emerge that do not reflect patterns of correlated activity in the input. On the other hand, the multiplicative model ([Kistler and van Hemmen, 2000]; [Rubin et al., 2001]; see also [van Rossum et al., 2000]) assumes linear attenuation of potentiating and depressing synaptic changes as the corresponding upper or lower boundary is approached. This model results in stable synaptic dynamics. However, because of reduced competition, all synapses are driven to a similar equilibrium value, even at moderately strong input correlations. Thus, neither the additive nor the multiplicative model provide a satisfactory scenario for a robust learning rule that implements a synaptic storage mechanism of temporal structures in the inputs. Here, we introduce a non-linear temporally asymmetric Hebbian (NLTAH) model – a novel generalized updating rule which allows for continuous interpolation between the additive and multiplicative models. We demonstrate that by appropriately scaling the weight-dependence of the updating, it is possible to learn synaptic representations of input correlations, while maintaining the system in a stable regime.
5.2 Materials and methods

5.2.1 Temporally Asymmetric Hebbian plasticity

We describe TAH plasticity as a change in the synaptic efficacy $w$ between a pair of cells, where the range of $w$ is normalized to $[0, 1]$. A single pair of pre- and post-synaptic action potentials with time difference $\Delta t \equiv t_{\text{post}} - t_{\text{pre}}$ induces a change in synaptic efficacy $\Delta w$ given by

$$
\Delta w = \begin{cases} 
-\lambda f_-(w) \cdot K(\Delta t) & \text{if } \Delta t \leq 0 \\
\lambda f_+(w) \cdot K(\Delta t) & \text{if } \Delta t > 0
\end{cases},
$$

(5.1)

The temporal filter $K(\Delta t) = \exp(-|\Delta t|/\tau)$ [Song et al., 2000, van Rossum et al., 2000, Rubin et al., 2001] implements the spike-timing dependence of the learning. The time constant $\tau$ of the exponential decay determines the temporal extent of the learning window. Following experimental measurements (e.g. [Bi and Poo, 1998]), we let $\tau = 20$ ms throughout the work. The learning rate $\lambda, 0 < \lambda \ll 1$, scales the magnitude of individual weight changes. The temporal asymmetry of the learning is represented by the opposite signs of the weight changes for positive and negative time differences. The updating functions $f_+(w), f_-(w) \geq 0$, which are in general weight-dependent, scale the synaptic changes and implement synaptic potentiation for causal time differences ($\Delta t > 0$), and depression otherwise. Here, we introduce a family of non-linear updating functions in which the weight-dependence has the form of a power-law with a non-negative exponent $\mu$,

$$
f_+(w) = (1 - w)^\mu \quad \text{and} \quad f_-(w) = \alpha w^\mu,
$$

(5.2)

with $\alpha > 0$ denoting a possible asymmetry between the scales of potentiation and depression. Figure 5.1A shows the updating curves (Eqs. 5.2) for several values of $\mu$. For $\mu = 0$, the updating functions are independent of the current synaptic efficacy, and the rule recovers the additive TAH learning model. This model requires that weights, which would have left the allowed range after an updating step, are clipped to the appropriate boundary (0 or 1). The case $\mu = 1$ corresponds to the multiplicative model, in which the updating functions linearly attenuate positive and negative synaptic changes as a synapse approaches the upper or lower boundary of the allowed range. Intermediate values of the updating parameter $\mu$ determine the range of the boundary effects on the changes in $w$. Note that any non-zero $\mu$, given a sufficiently small learning rate, automatically prevents the synaptic efficacies from leaving the allowed range $[0, 1]$, thereby preventing the runaway problem of synaptic efficacies and removing the necessity of artificially clipping synaptic weights. Figure 5.1B provides an illustrative example of the effects of the parameter $\mu$ on a sequence of synaptic weight changes (see caption for details).
Following previous work (e.g. [Kempter et al., 1999]; [Song et al., 2000]; [Rubin et al., 2001], but see [van Rossum et al., 2000]), the plasticity effects of individual spike pairs are assumed to sum independently: given a post-synaptic spike, each synapse is potentiated according to Eqs. 5.1 and 5.2 by pairing the output spike with all preceding synaptic events. Conversely, a synapse is depressed when a pre-synaptic event occurs, using all pairs the synaptic event forms with preceding output spikes.

### 5.2.2 Mean Synaptic Dynamics

Since in general the spike times of the pre- and post-synaptic neurons are stochastic, the dynamics of synaptic changes is also a stochastic process. However, if the learning rate $\lambda$ is small, the noise accumulated over an appreciable amount of time is small relative to the mean change in the synaptic efficacies called the synaptic drift. This drift, denoted as $\dot{w}$, is the mean rate of change of the synaptic efficacy. Using Fokker-Planck mean field theory, the synaptic drifts are described in terms of the correlations between the pre- and post-synaptic activity [Kempter et al., 1999, Kistler and van Hemmen, 2000, Kempter et al., 2001, Rubin et al., 2001]. We consider a pair of stationary pre- and post-synaptic

---

**Figure 5.1:** Next page. *Non-linear TAH.* A Effect of the parameter $\mu$ on the updating functions $f_+(w)$ (upper half) and $f_-(w)$ (lower half) for $\mu = 0, 0.02, 0.15, 0.5, 1$ ($\alpha = 1$). As $\mu$ increases, the curves change from the constant additive updating curves ($\mu = 0$, horizontal lines at 1 and $-1$) to the multiplicative updating functions with linear weight-dependence ($\mu = 1$, straight lines with slope $-1$). B Illustration of the effect of $\mu$ on synaptic changes. At each step, one pair of pre- and post-synaptic spikes induces a change in the efficacy $w$ according to Eqs. 5.1 and 5.2. To elucidate the weight-dependency of the updating, the magnitude of the time difference between the spikes $\Delta t$ is the same in all steps, namely $-\tau$ at step 4, and $\tau$ at all other steps. For illustrative purposes, we assume that the spike pairs are sufficiently distant so that the weight change is effected only by the current spike pair. Furthermore, the effect of a single pair is unrealistically magnified ($\lambda = 0.4$). For additive updating ($\mu = 0$), all potentiating changes are of equal magnitude because $\Delta t$ is constant here and the updating is independent of the synaptic efficacy. Note, however, that the last additive efficacy change would have resulted in an efficacy greater than 1, and hence the efficacy is clipped to 1. The depression in step 4 is larger exactly by a factor of $\alpha = 1.2$. In contrast, when $\mu > 0$, although the contribution to the synaptic change from the time dependency is constant, the actual change in synaptic efficacy is not. Since potentiation is scaled by $(1 - w)^\mu$, as the synapse becomes stronger the same time difference induces a smaller change. This scaling effect is more pronounced for larger values of $\mu$ (compare the cases of $\mu = 0.5$ and $\mu = 1$).
A

\[ f_+ (w) \]

\[ -f_- (w) \]

\[ w \]

B

\[ w \]

\[ \mu = 0 \]

\[ \mu = 0.5 \]

\[ \mu = 1 \]
processes described by the pulse trains \( \rho_{\text{pre}}(t) = \sum_k \delta(t-t_{k}^{\text{pre}}) \) and \( \rho_{\text{post}}(t) = \sum_k \delta(t-t_{k}^{\text{post}}) \) with mean rates \( r_{\text{pre}} = \langle \rho_{\text{pre}} \rangle \) and \( r_{\text{post}} = \langle \rho_{\text{post}} \rangle \) and the raw cross-correlation function

\[
\Gamma_{\text{pre,post}}(\Delta t) = \langle \rho_{\text{pre}}(t) \rho_{\text{post}}(t + \Delta t) \rangle_t.
\] (5.3)

The angular brackets denote averaging over time \( t \) while keeping the time lag \( \Delta t \) between the two spike trains fixed. This cross-correlation is the probability density for the occurrence of pairs of pre- and post-synaptic spikes with time difference \( \Delta t \). Using this probability density, the synaptic drift \( \dot{w} \) is given by integrating the synaptic changes \( \Delta w \) (Eq. 5.1) over the time differences \( \Delta t \) weighted by the probabilities \( \Gamma_{\text{pre,post}}(\Delta t) \)

\[
\dot{w} = -\lambda f_-(w) \int_{-\infty}^{0} d\Delta t \ K(\Delta t) \ \Gamma_{\text{pre,post}}(\Delta t) \\
+ \lambda f_+(w) \int_{0}^{\infty} d\Delta t \ K(\Delta t) \ \Gamma_{\text{pre,post}}(\Delta t).
\] (5.4)

The integral in the first term represents the synaptic depression that stems from all input-output correlations with negative time lag, i.e. acausal correlations. These correlations are filtered by the temporal window \( K(\Delta t) \) of the learning. This contribution from the spike-timing dependence of the learning is multiplied by the weight-dependent scale of depressing synaptic changes \( f_-(w) \). Conversely, the second term represents the potentiating drift originating from causal input-output correlations which is scaled by \( f_+(w) \). Note that the weight-dependent scales \( f_+(w) \) are evaluated outside the time integrals, because, when \( \lambda \) is small, \( w \) does not change appreciably during the time scale of the temporal filter of learning. In summary, the dynamic evolution of the synaptic weights depends on the properties of the correlation between the pre- and post-synaptic activity \( \Gamma_{\text{pre,post}}(\Delta t) \), which in turn depend on the details of the spike generation mechanism of the post-synaptic cell as well as on the statistics of the afferent inputs (e.g. [Kuhn et al., 2003]).

### 5.2.3 Integrate-and-Fire Neuron

To study the implications of the above NLTAH plasticity model in a biologically motivated spiking neuron, we simulate a leaky Integrate-and-Fire neuron, with parameters similar to [Song et al., 2000]. The membrane potential of the neuron is described by

\[
C_m \frac{dV}{dt} = \frac{V_{\text{rest}} - V}{R_m} + g_{\text{exc}}(t)(E_{\text{exc}} - V) + g_{\text{inh}}(t)(E_{\text{inh}} - V),
\]

with membrane capacitance \( C_m = 200\ \text{pF} \), membrane resistance \( R_m = 100\ \text{M}\Omega \), resting potential \( V_{\text{rest}} = -70\ \text{mV} \), and excitatory and inhibitory synaptic reversal potentials \( E_{\text{exc}} = 0\ \text{mV} \) and \( E_{\text{inh}} = -70\ \text{mV} \), respectively. Whenever the membrane potential exceeds
a threshold of $-54 \text{mV}$, an action potential is generated and the neuron is reset to the resting potential with no refractory period. Modeling synaptic conductance dynamics by alpha-shaped response functions, excitatory and inhibitory conductances are given by

$$g_A(t) = \bar{g}_A \sum_{j=1}^{N_A} w_j(t) \sum_{t_j < t} (t - t_j) \exp(-(t - t_j)/\tau_A), \quad A = \text{exc, inh}$$

respectively, where the $t_j$ are the spike times of synapse $j$, and $\tau_{\text{exc}} = \tau_{\text{inh}} = 5 \text{ms}$. The values $\bar{g}_{\text{exc}} = 30 \text{nS}$ and $\bar{g}_{\text{inh}} = 50 \text{nS}$ were chosen such that the total charge injected per spike (at the threshold potential) is $Q_{\text{exc}} = 0.04 \text{pC}$ and $Q_{\text{inh}} = 0.02 \text{pC}$, respectively. While the efficacies $w$ of the $N = N_{\text{exc}} = 1000$ excitatory synapses are plastic and governed by the TAH learning rule, all $N_{\text{inh}} = 200$ inhibitory efficacies are held fixed at $-1$. In the numerical simulations, the Integrate-and-Fire neuron is driven by Bernoulli, i.e. zero-one, processes defined over discrete time bins of duration $\Delta T = 0.1 \text{ms}$, approximating Poisson spike trains with a stationary rate $r$. For the inhibitory inputs $r = 10 \text{Hz}$. All equilibrium synaptic distributions obtained with this model neuron result from an initially uniform synaptic state with all efficacies set to 0.5. In each case, the learning process (learning rate $\lambda = 0.001$) is simulated until the shape of the synaptic distribution ceases to change.

### 5.2.4 Linear Poisson Neuron

In order to investigate analytically the properties of the TAH learning rule, we consider in addition a linear Poisson neuron (see also [Kempter et al., 2001]). The spiking activity of this neuron $\rho^{\text{post}}(t)$ is a realization of a Poisson process with the underlying instantaneous rate function

$$R^{\text{post}}(t) = \frac{1}{N} \sum_{j=1}^{N} w_j(t) \rho_j^{\text{pre}}(t - \varepsilon), \quad (5.5)$$

where, as before, the $N$ pre-synaptic input spike trains and the output spikes are characterized by a series of delta-pulses, i.e. $\rho_j^{\text{pre}}(t) = \sum_k \delta(t - t_k^{\text{pre},j})$ and $\rho^{\text{post}}(t) = \sum_k \delta(t - t_k^{\text{post}})$. The parameter $0 < \varepsilon \ll \tau$ denotes a small constant delay in the output. Because this delay is small compared to the temporal window of learning, we approximate $\exp(-\varepsilon/\tau) \approx 1$ throughout this work. As before, $w_j(t) \in [0, 1]$ denotes the efficacy of the $j$th synapse. Except for Figures 5.5 and 5.9 where we investigate the large $N$ limit, we let $N = 100$ throughout this work.

In Figure 5.2A we numerically simulate the linear Poisson neuron receiving uncorrelated Poisson input spike trains. Generating the spike arrival times in continuous time (down to machine precision), the post-synaptic process defined in Eq. 5.5 is implemented by generating a post-synaptic spike with probability $w_i/N$, whenever a pre-synaptic spike arrives at a synapse $(i)$ of the neuron.
5.2.5 Mean Synaptic Dynamics for the Linear Poisson Neuron

For the Integrate-and-Fire neuron, there is no simple exact expression relating the correlations between the pre- and post-synaptic spike trains to the system parameters such as the rates and input correlations. On the other hand, because of the linear summation of inputs in the linear Poisson neuron (Eq. 5.5), this model permits the expression of the input-output correlations $\Gamma_{\text{pre,post}}(\Delta t)$ in closed form. Considering the case that all input spike trains have a common rate $r$, we obtain from Eqs. 5.3 and 5.5, that the correlation of the activity at synapse $i$ with the output activity is

$$
\Gamma_{i,\text{post}}(\Delta t) = \langle \rho_i^{\text{pre}}(t) \rho_j^{\text{post}}(t + \Delta t) \rangle_t
$$

$$
= \frac{1}{N} \sum_{j=1}^{N} w_j \langle \rho_i^{\text{pre}}(t) \rho_j^{\text{pre}}(t + \Delta t - \varepsilon) \rangle_t.
$$

Substituting the above in Eq. 5.4, and rearranging the terms, we obtain the drift of the $i$th synapse

$$
\dot{w}_i = -\lambda f_-(w_i) \frac{1}{N} \sum_{j=1}^{N} w_j \tau r^2 \left( 1 + \int_{-\infty}^{0} d\Delta t \frac{1}{\tau} K(\Delta t) \Gamma_{ij}^0(\Delta t - \varepsilon) \right)
$$

$$
+ \lambda f_+(w_i) \frac{1}{N} \sum_{j=1}^{N} w_j \tau r^2 \left( 1 + \int_{0}^{\infty} d\Delta t \frac{1}{\tau} K(\Delta t) \Gamma_{ij}^0(\Delta t - \varepsilon) \right),
$$

where we define the normalized cross-correlations between the input spike trains by

$$
\Gamma_{ij}^0(t') = \frac{\langle \rho_i^{\text{pre}}(t) \rho_j^{\text{pre}}(t + t') \rangle_t}{r^2} - 1.
$$

We denote the integrated normalized cross-correlations appearing in the above drift equation by

$$
C_{ij}^\pm = \int_{0}^{\infty} d\Delta t \frac{1}{\tau} K(\Delta t) \Gamma_{ij}^0(\pm \Delta t - \varepsilon).
$$

These matrices are the effective between-input correlations for positive and negative time lags. If $C_{ij}^+ > 0$, the activity at synapse $j$ temporally follows that at synapse $i$, such that its contribution to the post-synaptic activity results in a potentiating drift on synapse $i$. Conversely, if $C_{ij}^- > 0$, the activity at synapse $j$ precedes that at synapse $i$ and contributes to its depression. Note that the effective correlations $C_{ij}^\pm$ are zero if the $i$th and $j$th input spike trains are uncorrelated. Finally, the synaptic drifts can be written as

$$
\dot{w}_i = \frac{\lambda \tau r^2}{N} \left[ -\Delta f(w_i) \sum_{j=1}^{N} w_j + f_+(w_i) \sum_{j=1}^{N} w_j C_{ij}^+ - f_-(w_i) \sum_{j=1}^{N} w_j C_{ij}^- \right],
$$

75
with \( \Delta f = f_\text{+} - f_\text{–} \). Note that the first term in Eq. 5.8 describes competition between the synapses when \( \Delta f > 0 \); independently from the input correlations, the amount of induced depression on a given synapse \( w_i \) is large when other synapses \( w_j \) are strong. The second term represents the cooperative increase of the synaptic weights inherent in TAH learning. In contrast, the last term denotes depressive synaptic interactions stemming from negative time correlations in the input activity.

### 5.2.6 Generating Correlated Inputs

We consider input spike trains with rate \( r \) and instantaneous correlations defined by

\[
\Gamma_{ij}^0(\Delta t) = \frac{1}{r} c_{ij} \delta(\Delta t),
\]

where \( \delta(t) \) is the Dirac-delta function and \( c_{ij} \) is non-negative. In this case,

\[
C_{ij}^– = 0 \quad \text{and} \quad C_{ij}^+ = \frac{c_{ij}}{\tau r}.
\]

The backward effective correlations \( C_{ij}^– \) vanish because the argument of \( \Gamma_{ij}^0(–\Delta t – \varepsilon) \) in Eq. 5.7 is never \( 0 \). Recall that for a Poisson process \( \rho(t) \) with rate \( r \) the raw auto-correlation is \( \langle \rho(t)\rho(t + \Delta t) \rangle = r^2 + r \delta(\Delta t) \). Hence, the normalized auto-correlation \( c_{ii} = 1 \) and the between-input correlations \( c_{ij} \leq 1 \), with equality only if the two spike trains \( i \) and \( j \) are identical. In the numerical simulations we generate populations of correlated spike trains by conditioning the binwise spike probabilities at time bin \( T \) on the activity of a common reference Bernoulli spike train \( X_0(T) \) with the binwise spike probability \( r \Delta T \). To obtain a positive pairwise correlation coefficient of \( 0 \leq c = \text{Cov}(X_i(T), X_j(T))/\sqrt{\text{Var}(X_i(T))\text{Var}(X_j(T))} \) between two spike trains \( X_i(T) \) and \( X_j(T) \), the conditional probabilities \( \vartheta = P(X_k(T) = 1|X_0(T) = 1) \) and \( \varphi = P(X_k(T) = 1|X_0(T) = 0) \) for \( k = i, j \) are determined by

\[
\vartheta = r \Delta T + \sqrt{c(1 - r \Delta T)}, \quad \varphi = r \Delta T(1 - \sqrt{c}).
\]

This choice of \( \vartheta \) and \( \varphi \) for all spike trains within the correlated group, guarantees that the spike trains have rates \( r \) and an instantaneous pairwise correlation coefficient \( c \) (see Appendix). For small bin sizes this process mimics the instantaneously correlated Poisson point processes defined above (Eq. 5.9). We will also consider the case of delayed correlations of the form \( \Gamma_{ij}^0(\Delta t) = r^{-1} c_{ij} \delta(\Delta t - D_{ij}) \). These correlations are obtained by shifting the instantaneously correlated input spike trains relative to each other by a time delay \( D_{ij} \).

### 5.2.7 Measuring the Performance of Learning Rules

A natural way to measure the performance of a learning rule is to quantify its ability to imprint the statistical features of the neuronal input onto the distribution of the learnt
synaptic weights. One measure of this ability is the mutual information between the neuronal inputs and the synaptic weights. However, direct calculation of the mutual information in cases where the number of synaptic weights is large is computationally not feasible. Instead, we use here a related quantity which measures the effect of a small change in the statistics of the input on the learnt synaptic weights. We denote the features of an ensemble of neuronal inputs by the vector $\Phi = (\Phi_1, \ldots, \Phi_R)$, where the $\Phi_i$ parameterize specific input features, e.g. mean strength of the inputs or temporal correlations between different inputs. Given these features we calculate the $N \times R$ susceptibility matrix $\chi_{ij}$, the elements of which are

$$\chi_{ij} \equiv \frac{\partial w_i}{\partial \Phi_j}. \quad (5.12)$$

The $ij$-th element measures the amount of change in the $i$th synaptic efficacy which is incurred by a small change in the $j$th input feature, $\Phi_j$. A global sensitivity measure $S$ is constructed from this matrix by calculating

$$S \equiv \frac{1}{2} \ln \left| \det (\chi^T \chi) \right|, \quad (5.13)$$

where $\det(\cdot)$ denotes the determinant. The average sensitivity $S_{\text{avg}}$ is defined as $\langle S \rangle_\Phi$, where the average $\langle \cdot \rangle$ is taken over the distribution of the feature vector $\Phi$. The rational for calculating $S$ is that it is closely related to the mutual information between the input features and the weight distribution. Specifically, if the mapping from the feature space to the synaptic weight space induced by the learning dynamics is invertible, maximizing $S_{\text{avg}}$ is equivalent to maximizing the mutual information [Bell and Sejnowski, 1995, Shriki et al., 2001], in the limit of a small learning rate $\lambda$. In this work we focus on the equilibrium properties of the TAH learning rule, i.e. the weight distributions that result after the learning dynamics has converged to a stable stationary state. Therefore $\chi_{ij}$ is evaluated at the fixed point solution $\mathbf{w}^*$ of the drift equations in the linear Poisson neuron, Eq. 5.8 (see Appendix). The possibility to use the analytic expressions for $\chi_{ij}$ in calculating the sensitivity $S$ is the main advantage of using it as a measure of performance of various learning models.

### 5.3 Results

To understand learning phenomena in biological nervous systems in terms of neural network function, it is crucial to bridge the gap between the microscopic mechanisms that implement experience-based changes in neuronal signaling pathways and the macroscopic properties of the learning system composed of these pathways. In this work we focus on two general goals of learning that can be defined at the network level, and investigate the importance of the updating parameter $\mu$ of the learning rule in these contexts.
First, we consider the question of how a network can develop a functional connectivity architecture, as e.g. in ocular dominance columns. As noted in the Introduction, this type of learning task typically requires the synaptic learning dynamics to be competitive, to allow segregation between initially homogeneous synaptic populations. Moreover, it is important that the learning process is robust in the sense that the learned synaptic patterns faithfully reflect meaningful features in the neuronal input activity, rather than being dominated by contributions from random noise. Therefore, we study here how the interplay between competition and stability in TAH plasticity affects the learned synaptic distributions. In the second part of the Results we turn to the conceptually different learning task of imprinting information about a neuron's input activity into the respective synaptic efficacies. In this context, the sensitivity of the learning dynamics to features in the neuronal input becomes crucial. Thus, using the sensitivity measure introduced in the Methods, the second part of the results concentrates on a quantitative evaluation of the performance of different TAH learning rules.

5.3.1 The Emergence of Synaptic Patterns by Symmetry Breaking in TAH Learning

One of the basic requirements for the activity-driven formation of cortical maps is the ability of the learning to generate spatially inhomogeneous synaptic patterns from a population of synapses with statistically homogeneous inputs. The emergence of such symmetry breaking is an essential property of current cortical plasticity models [Miller, 1996]. In this section we study the conditions under which the TAH learning models introduced above exhibit symmetry breaking and, hence, qualify as candidate models for the development of functional maps. Moreover, since the learning dynamics may also lead to symmetry breaking that overrides the correlation structure of the afferent activity, it is important to ask what learning rules ensure a faithful representation of the input activity within the learned synaptic connections. We address these questions in three basic types of homogeneous afferent activities, that differ with respect to the correlation structure of the input spike trains: uncorrelated inputs, uniformly correlated inputs, and uniformly correlated subpopulations without correlations between the subpopulations (‘correlated subgroups’). Before treating these specific cases, we highlight the general features of the synaptic learning dynamics in a population of synapses with statistically homogeneous input activities. These results apply to all three cases of homogeneous populations of inputs.
Dynamics of a Population of Synapses with Homogeneous Inputs

To study the symmetry breaking in the synaptic patterns we consider the learning dynamics in cases where the input statistics is spatially homogeneous. This means that each input obeys the same spike statistics and has the same pattern of correlations with the other inputs. This assumption implies that the pre-synaptic rates \( r_i \) (where \( i \) denotes the index of the different afferents) are all equal. Likewise the total sum of the correlations that each input has with the rest of the inputs is the same. In particular, the mean effective causal correlations, \( C_0 \),

\[
C_0 = \frac{1}{N} \sum_{i=1}^{N} C_{ij}^+, \tag{5.14}
\]

is the same for all input channels \( i \), and similarly for the backward correlations (see also [Kempter et al., 2001]).

To understand the implications of spatial homogeneity in the pre-synaptic inputs on the learning dynamics it is useful to concentrate on the Linear Poisson neuron model (Eqs. 5.5, 5.8). For convenience we assume that all correlations between input spikes are instantaneous (see Methods, Eq. 5.10).

The important consequence of the spatial homogeneity across the pre-synaptic inputs is that the product of the effective correlation matrix \( C_{ij}^+ \) with a homogeneous vector of synaptic efficacies \( \mathbf{w}_o = (w_o, \ldots, w_o) \) remains homogeneous since \( C^+ \mathbf{w}_o = NC_0 \mathbf{w}_o \). Hence, the homogeneous synaptic state is an eigenvector of the effective correlation matrix \( C^+ \) with eigenvalue \( NC_0 \). The existence of this homogeneous eigenvector is important for the synaptic learning, because it means that in a homogeneous synaptic state \( \mathbf{w}_o \) all synapses experience identical drifts (see Eq. 5.8). Moreover, if there is a \( w_o = w^* \) such that the synaptic drifts become zero, the learning dynamics has a steady-state solution with \( w_i = w^* \) for all synapses. We call a solution where all the learned synapses are equal, a homogeneous solution. Indeed, we show in the Appendix that for all non-zero updating parameter \( \mu \) in our model (Eq. 5.2), there exists a steady-state homogeneous solution of the learning (\( \dot{w}_i = 0 \) in Eq. 5.8) with \( w^* \) being the solution of the equation

\[
\frac{f_-(w^*)}{f_+(w^*)} = \alpha \left( \frac{w^*}{1-w^*} \right)^\mu = 1 + C_0. \tag{5.15}
\]

This equation expresses the weight-dependent balance between depression and potentiation which is controlled by the mean effective correlation \( C_0 \).

Although the homogeneous synaptic steady state always exists, it may be unstable with respect to small perturbations of the synaptic efficacies, driving them into inhomogeneous states. Because of the important functional consequences of this emergence of inhomogeneous synaptic patterns at the network level, it is important to understand the features of
the learning dynamics that give rise to this phenomenon of symmetry breaking. Therefore, we analyze the effects of small deviations of the synaptic efficacies from the homogeneous synaptic steady state \( w^* \). For each synapse \( w_i \), we denote a corresponding small deviation from the homogeneous solution by \( \delta w_i = w_i - w^* \), and express its temporal evolution as a function of all deviations \( \delta w_j \). As we show in the Appendix, \( \delta w_i \) is determined by three separate contributions

\[
\delta w_i = \lambda \tau r^2 \left[ -g_0 \delta w_i - \frac{\Delta f(w^*)}{N} \sum_{j=1}^{N} \delta w_j + \frac{f_+(w^*)}{N} \sum_{j=1}^{N} C_{ij} \delta w_j \right],
\]

where \( g_0 = w^* f_+(w^*) \frac{\partial}{\partial w} [f_-(w)/f_+(w)]_{w=w^*} = \alpha \mu w^*/(1 - w^*) > 0 \) (Appendix Eqs. 5.20, 5.21). The first term is a local stabilizing term. It counteracts individual deviations from the homogeneous solution, maintaining the synaptic efficacies at the same value \( w^* \). To understand the origin of this stabilizing term in the learning dynamics we consider the effect of a single synaptic deviation \( \delta w_i \) on the balance between depression and potentiation. If a synapse is strengthened by a deviation \( \delta w_i > 0 \), the resulting scale of potentiation \( f_+(w^* + \delta w_i) \) decreases, whereas the scale of depression \( f_-(w^* + \delta w_i) \) increases (Eq. 5.2). Conversely, a weakening deviation \( \delta w_i < 0 \) shifts the balance between potentiation and depression in favor of potentiation. Since this stabilizing drift stems from the weight dependence of the ratio \( f_-(w^*)/f_+(w^*) \) (cf. Appendix), it is not present in the additive model \( (\mu = 0) \), where the \( f_\pm \) themselves are constant. The second term is proportional to the net drift \( -\Delta f(w^*) = f_+(w^*) - f_-(w^*) \). This drift is negative because at the homogeneous solution, \( f_-(w^*) > f_+(w^*) \) when depression balances the potentiating correlations (see, Eq. 5.15 and recall that \( C_0 > 0 \)). The negative drift is multiplied by the total perturbation \( \sum_j \delta w_j \), which denotes the change in the output rate due to the changes in the synaptic efficacies. Thus, this term represents the competition between the synapses. This competition results from the fact that strengthening the efficacy of any synapse increases the output rate, thereby increasing the frequency of occurrence of net negative drift in all the synapses. It is important to note, that this competition is acting between all synapses, unrelated to the correlation structure in the afferent input.

Finally, the last term is a cooperative term. Synapses which are positively correlated cooperate to elevate their weights. This cooperation is driven by the potentiating component of the TAH learning and depends on the pattern of correlations among the input channels. We emphasize that the cooperativity in the synaptic learning in general does not originate from a possible advantage of correlated synapses to drive a potentially non-linear spike generator of the post-synaptic cell, but already occurs due to an inherently increased probability of correlated synapses to precede post-synaptic spikes, even when non-linear cooperative effects in the spike generator are absent.

The stability of the homogeneous synaptic steady state results from the interplay be-
tween the stabilizing, the competitive, and the cooperative drifts in the learning dynamics. As we derive in the Appendix, perturbations of the steady state that slightly change all weights by the same amount $\delta w$ (homogeneous perturbations) decay to zero with time and, hence, do not destabilize the learning of a homogeneous synaptic distribution. In contrast, inhomogeneous perturbations, i.e. perturbations in which the deviations of the synaptic efficacies from $w^*$ are not identical, can grow exponentially through the learning dynamics and drive the system into inhomogeneous synaptic states. In the Appendix, we specifically show that the homogeneous synaptic state becomes unstable if the largest real part of all inhomogeneous eigenvalues (eigenvalues corresponding to inhomogeneous eigenvectors) of the effective correlation matrix $C^+$ is sufficiently large. Denoting this eigenvalue by $NC_1$, we find that when

$$\mathcal{K} = C_1f_+(w^*) - g_o > 0,$$  \hspace{1cm} (5.16)

the homogeneous state is unstable. This inequality means that symmetry breaking occurs whenever the cooperation between synapses $C_1$ is strong enough to outweigh the stabilizing term, $g_o$. Note that the competition coefficient $\Delta f$ does not enter directly into the stability criterion. This is because the competition term is proportional to the total weight value, and hence is not sensitive to inhomogeneous perturbations that do not change this value. Nevertheless this term has a crucial role for the stability of the learning, because it suppresses the homogeneous growth of all synapses. As is shown in the Appendix, Eq. 5.16 implies that the homogeneous solution is always stable in the multiplicative model ($\mu = 1$).

Although this analysis was performed using the plasticity equations of the linear Poisson neuron, it is qualitatively valid as well for other neuron models, as we show for specific cases. Below we study how the emergence of symmetry breaking, i.e. transitions from homogeneous to inhomogeneous synaptic distributions depends on the non-linearity of the TAH dynamics, namely the parameter $\mu$, as well as on the asymmetry between depression and potentiation $\alpha$, and on the size of the synaptic population $N$.

**Uncorrelated Inputs – Linear Neuron**

In this section we investigate the synaptic distributions that result from the TAH learning process when the post-synaptic neuron is driven by independent Poisson spike trains of equal rate $r$. For this input regime, it has been found in an Integrate-and-Fire neuron that additive learning ($\mu=0$) breaks the symmetry of the statistically identical pre-synaptic inputs and leads to a bimodal weight distribution [Song et al., 2000, Rubin et al., 2001]. On the other hand, it was shown by [Rubin et al., 2001] that multiplicative learning ($\mu=1$) leads to a unimodal distribution of synapses. As shown in the preceding section, these qualitatively different learning behaviors originate in the stabilizing effect of the weight
dependence of the synaptic changes on the homogeneous synaptic state. Here, we study the generalized non-linear TAH rule with arbitrary \( \mu \in [0, 1] \).

In the uncorrelated case, \( a_{ij} = \delta_{ij} \), and hence \( C_i^+ = \delta_{ij} / \tau r \). Both its homogeneous and inhomogeneous eigenvalues normalized by \( N \) are

\[
C_0 = C_1 = \frac{1}{\tau r N},
\]

Thus, the only cooperation in the learning dynamics stems from the positive feedback induced by the correlation of each synapse with its own contribution to the post-synaptic activity. The effects of this self-correlation on the learning dynamics decrease inversely to the effective size of the pre-synaptic population \( \tau r N \), i.e. the expected number of spikes that arrive within the learning time window.

By inserting the above expression for \( C_0 \) into Eq. 5.15, we obtain the steady state efficacy \( w^* \) of the synaptic population when the learned synaptic state is homogeneous (see Appendix, Eq. 5.19). In this case, the output rate of the linear neuron is given by this steady state efficacy times the rate of the pre-synaptic inputs \( r \) (cf. Eq. 5.5). Figure 5.2A depicts the output rate of the post-synaptic neuron as a function of the pre-synaptic input rate \( r \), for \( \alpha = 1.05 \). We focus on this value of \( \alpha \) here, because we want to compare the nonlinear rules with the additive rule. In the latter case, \( \alpha \) must be close to one; otherwise, practically all synapses will become zero (see Appendix). For \( \mu = 1 \) (multiplicative TAH), the efficacy \( w^* \) is fairly independent of \( r \) and, hence, the output rate grows linearly with the input rate. However, if \( \mu \) is sufficiently small, \( w^* \) decreases inversely with the input rate, resulting in the output rate being nearly constant.

---

Figure 5.2: **Next page.** *Firing rate responses of a neuron driven by uncorrelated Poisson input processes.* A Output firing rate of the linear Poisson neuron as a function of the input rate for selected values of \( \mu = 0, 0.024, 0.04, 1 \), with \( N = 100 \) and \( \alpha = 1.05 \). Solid lines show analytical results derived from the homogeneous synaptic state (Eq. 5.19) (which is stable for all values \( \mu > 0 \) and \( r \) shown here) or, for additive TAH (\( \mu = 0 \)), from the ratio of strong synapses given by Eq. 5.25 (see Results and Appendix). In both cases, the output rate is given by the input rate multiplied with the corresponding mean synaptic efficacy. Plot symbols depict the output rates of a numerically simulated spiking linear Poisson neuron (Methods) with parameters as in the analytical calculation and a learning rate \( \lambda = 0.001 \) (except for \( \mu = 0 \) where we used \( \lambda = 0.003 \)). B,C Results from the Integrate-and-Fire neuron with \( N = 1000 \). All results refer to the neuron after convergence of the learning process. B The output firing rate of the neuron as a function of the input rate for \( \alpha = 1.05 \). C The output firing rate as a function of \( \alpha \) for an input rate of 10 Hz.
To study the regime where the synaptic learning dynamics break the symmetry of the uncorrelated input population, we substitute Eq. 5.17 into Eqs. 5.15 and 5.16, computing the homogeneous solution $w^*$ (Eq. 5.19) and the regime of its stability. Figure 5.3A depicts the critical contour lines according to the stability condition (Eq. 5.16). Each line traces the critical combination of the parameters $\mu$ and $\tau r N$ for a fixed value of $\alpha$, such that $\mathcal{K} = 0$. Outside the corresponding contour ($\mathcal{K} < 0$) the homogeneous synaptic state is stable, and thus learning generally results in all synapses having the same efficacy. In contrast, inside the contour line, the learning dynamics induce symmetry breaking.

Figure 5.3A shows how the outcome of TAH learning depends on the effective size of the pre-synaptic population. For sufficiently small $\tau r N$, the relative contribution of each input channel to the post-synaptic activity is large and, hence, the resulting strong positive feedback drives all synapses to a stable homogeneous state near the upper boundary (Figure 5.3B, squares). In contrast, as $\tau r N$ is increased, the effect of a single synapse on the post-synaptic activity decreases. Therefore, for sufficiently large $\tau r N$, the stabilizing force induced by the weight-dependence of the synaptic changes dominates the learning dynamics for any non-zero $\mu$, resulting in a stable homogeneous synaptic state (Figure 5.3B, triangles). In between the two extremes of small and large $\tau r N$, there is a regime of intermediate effective population sizes where symmetry breaking may occur.

---

**Figure 5.3:** Next page. Symmetry breaking in the linear neuron driven by uncorrelated Poisson input processes. **A** The critical contour lines of the stability criterion $\mathcal{K} = 0$ (Eq. 5.16) for $\alpha = 1.05, 1.1, 1.25, 1.5, 2, 2.5, 3, 3.5$ (from right to left, thick line corresponds to $\alpha = 1.05$). The homogeneous solution is stable outside the contour and unstable in its interior. Inset: Critical contour for large $\tau r N$ and $\alpha = 1.05$. **B** Equilibrium synaptic weights as a function of $\mu$ for $\tau r N = 10$ (squares), 20 (circles), and 200 (triangles), with $\alpha = 1.05$. Plot symbols mark the efficacy values obtained from simulating the mean field learning dynamics of the linear neuron with $N = 100$ and the input rate $r = 5, 10, 100$ Hz, adjusted to obtain the desired value of $\tau r N$. As $\mu$ is decreased, each simulation is initialized with the synaptic efficacies of the equilibrium state obtained for the previous value of $\mu$, plus a perturbation vector ranging from -0.001 to 0.001. For small and large values of $\tau r N$ (squares and triangles) the homogeneous solution is stable for all $\mu$ (except the vanishing regime of very small $\mu < \mu_{\text{crit}}$ in the large $\tau r N$ case). For $\tau r N = 20$ (circles) the synaptic population splits into two groups (54 strong and 46 weak synapses) when $\mu$ crosses the corresponding critical contour in panel A. The solid lines depict the analytically obtained values of the equilibrium weights. For $\tau r N = 20$ (circles), the line was obtained by numerically solving the analytical expression for a bimodal synaptic state with the same split ratio as obtained in the simulations.
with the synaptic population segregating into a strong and a weak group. Such a case is shown in Figure 5.3B (circles).

Importantly, Figure 5.3A demonstrates that as the number of afferents $N$ increases, the regime of values of $\mu$ for which the homogeneous solution is unstable shrinks to zero. The inset of Figure 5.3A shows the value of $\mu$ at the border between stability and instability of the homogeneous solution, as a function of the effective population size. It is apparent that this $\mu$ decreases linearly with $1/(\tau r N)$ when $\tau r N$ is large (see also Appendix). Hence, for any sizable degree of weight-dependence and large synaptic populations, symmetry breaking does not occur.

In the purely additive TAH model, synaptic changes do not scale at all with the efficacy of a synapse and the weights have to be constrained by an additional clipping to prevent unrealistic synaptic growth. As a result, the additive learning dynamics does not posses stationary synaptic states in the above sense that the individual synaptic drifts become zero. Instead, synapses with positive drifts are held at the upper boundary, whereas synapses with negative drifts saturate at the minimum allowed efficacy. Our treatment of the additive model in the Appendix shows that the numbers of synapses gathering at the upper and lower boundaries critically depend on the ratio of depression and potentiation $\alpha$, as well as on the effective population size $\tau r N$. As in the non-linear TAH learning model, small effective synaptic populations ($\tau r N < 1/(2(\alpha - 1))$) will lead to all synapses saturating at the upper boundary due to the strong positive feedback. However, as $\tau r N$ increases beyond a critical value, the synaptic population breaks into two groups, one of which remains saturated at the upper boundary while the other, losing the competition, saturates at the lower boundary. The ratio of synapses saturating at the top boundary is $n_{up} = 1/2\tau r N(\alpha - 1)$ (Appendix). Because this ratio is inversely proportional to the input rate $r$, the output rate of the post-synaptic neuron becomes independent of the input rate, as shown in Figure 5.2A.

**Uncorrelated Inputs – Integrate-and-Fire Neuron**

We now turn to the behavior of TAH learning in the Integrate-and-Fire neuron driven by uncorrelated inputs. Figure 5.2B shows the output rate of this neuron model versus the input rate for different values of $\mu$. As the figure demonstrates, the output-rate normalization quickly deteriorates as $\mu$ departs from the additive model and synaptic changes become dependent on the efficacy of the synapse. Figure 5.2C demonstrates that the sensitivity of the output rate to the parameter $\alpha$ rapidly diminishes as $\mu$ increases. Comparing panels A and B shows the qualitative similarity between the output rate responses of the linear Poisson and the Integrate-and-Fire model neurons. Note that we have not attempted to match the overall scale of the output rates in the two models.
The output rate of the linear neuron can be arbitrarily changed by a gain factor without affecting any other results.

Figure 5.4 displays the histograms of the equilibrium distributions of learned synaptic efficacies as a function of the updating parameter \( \mu \). Recovering the behavior of additive [Song et al., 2000] and multiplicative [Rubin et al., 2001] updating models for \( \mu = 0 \) and \( \mu = 1 \), respectively, the plot reveals the transition between these models for intermediate values of \( \mu \). Specifically, it shows the emergence of symmetry breaking as \( \mu \) approaches zero.

As expected from the analysis of the linear neuron, we find that also in the Integrate-and-Fire neuron the critical value of \( \mu \) where the synaptic distribution becomes bimodal decreases as the effective population size \( \tau r N \) increases. Increasing the rate of the input processes from 10 Hz (Figure 5.4A) to 40 Hz (Figure 5.4B) lowers the first occurrence of a bimodal weight distribution from \( \mu_{\text{crit}} = 0.023 \) to \( \mu_{\text{crit}} = 0.017 \). The inset in each panel depicts the equilibrium weight distribution for the intermediate value of \( \mu = 0.019 \), showing a clearly bimodal distribution for the 10 Hz input (A) and a clearly unimodal distribution for the 40 Hz input (B). Moreover, as expected from the equations describing the homogeneous steady state in the linear neuron (Eqs. 5.15 and 5.17), the synaptic efficacy of the homogeneous state at a given \( \mu \) decreases when the input rate increases.

It is interesting to note the close similarity in the \( \mu \)-dependence of the learned synaptic distributions in the linear and the Integrate-and-Fire neurons. For example, in both cases, the critical \( \mu \) for symmetry breaking is close to 0.023 for input rates of 10 Hz (compare Figure 4A with Figure 3B (circles)). This is despite the fact that the two models have very different spike generators and different sizes of synaptic populations. The reason for this similarity is that the input-output correlations in the Integrate-and-Fire neuron with 1000 synapses turn out to match in magnitude the corresponding correlations of the linear neuron with 100 synapses (not shown).

In summary, for uncorrelated inputs and biologically realistic sizes of the pre-synaptic

---

Figure 5.4: Next page. Symmetry breaking in the Integrate-and-Fire neuron driven by uncorrelated Poisson input processes. A, B Histograms of the equilibrium synaptic distributions in logarithmic gray-scale as a function of the updating parameter \( \mu \) for \( N = 1000 \) and \( \alpha = 1.05 \). The input rate is 10 Hz in (A) and 40 Hz in (B). The arrow marks the critical \( \mu \) where the first bimodal distribution occurs. Insets: The synaptic distributions at \( \mu = 0.019 \). Note that whereas the histograms shown in the main figures are single realizations of the equilibrium synaptic distributions for each \( \mu \), the histograms shown in the insets were obtained by averaging over 30 different readouts of the converged synaptic weights taken at consecutive intervals of 500 s.
population, $N$, on the order of thousands, and for rates on the order of 10 Hz and above, the regime in $\mu$ and $\alpha$ where symmetry breaking between uncorrelated inputs as well as output rate normalization occur, is extremely narrow. Thus, the learning behavior changes qualitatively as soon as synaptic plasticity becomes weight-dependent.

**Uniformly Correlated Inputs**

We briefly discuss here the case where the pre-synaptic inputs have positive uniform instantaneous correlations, namely that for all $i \neq j$, $c_{ij}$ (Eq. 5.9) are equal. This situation may, for instance, occur when the entire pre-synaptic pool of a neuron is driven by a common source. Treating the behavior of the linear Poisson neuron, we show in the Appendix, that positive uniform correlation increases the value of the synaptic efficacy in the homogeneous synaptic steady state. Moreover, the uniform correlation does not alter the $1/(\tau r N)$ dependence of the destabilizing drifts. As a result, in non additive learning, when the effective synaptic population is sufficiently large, the homogeneous steady state remains stable for any positive uniform correlation strength. In fact, these correlations increase the stability of the homogeneous state (see Appendix) and, hence, oppose the emergence of spontaneous symmetry breaking.

**Correlated Subgroups**

We now consider afferent input activity to a neuron that is composed of $M$ equally sized groups. These groups are defined by a uniform within-group correlation coefficient $c_{ij} = c > 0$ (cf. Eq. 5.9) which is equal within all groups. For pairs of inputs belonging to different groups, the cross-correlation is zero. In this scenario, the $M$ different pre-synaptic groups compete for control over firing of the post-synaptic neuron. We first treat the linear neuron and, for simplicity, focus on the case where the overall number of pre-synaptic input channels $N$ is large. In this limit, the homogeneous and largest inhomogeneous eigenvalues of $C^+$ normalized by $N$ are

$$C_0 = C_1 = \frac{c}{\tau r M}.$$ 

Comparing these expressions with their respective values in the case of $N$ uncorrelated inputs (Eq. 5.17), we note that the learning behavior in both input scenarios is equivalent when $N$ is identified with $M/c$, the number of correlated subgroups divided by the strength of the within-group correlation. The stability of the homogeneous synaptic steady state in a large network comprising $M$ correlated synaptic subgroups each with within-group correlation $c$, behaves like in an uncorrelated network of finite size $M/c$. Thus, in the limit of a large pre-synaptic population, the correlation strength $c$ scales the effective number of pre-synaptic inputs from $M$ for $c = 1$ to infinity for $c = 0$. Importantly, the largest
inhomogeneous eigenvector is such that when the homogeneous solution loses stability, the symmetry is broken between the correlated subgroups and not within each subgroup.

The nature of the synaptic pattern that emerges once the homogeneous synaptic state loses stability depends on the number of afferent subgroups. Here we focus on the example of two equally sized subgroups, i.e. $M = 2$. A similar scenario, which is motivated by the problem of the activity-driven development of ocular dominance maps, has recently been studied by [Miller and MacKay, 1994] and [Song and Abbott, 2001]. The regimes of symmetry breaking where the learned synaptic efficacies segregate according to the two correlated input groups are depicted in Figure 5.5A (this figure is equivalent to Figure 5.3A with $c$ replacing $2/N$). Thus, symmetry breaking between two correlated subgroups can occur in non-additive TAH learning models even when the number of pre-synaptic inputs $N$ is large. This is demonstrated in Figure 5.5B (solid black line), which plots the learned synaptic efficacies as $\mu$ is varied with $c$ held fixed at 0.11. As is evident from the Figure, for this level of correlation, symmetry breaking occurs below a fairly high value of $\mu \approx 0.15$. Note that in contrast to our treatment of the uncorrelated inputs, here we

Figure 5.5: Next page. Symmetry breaking in the linear neuron driven by two equally sized correlated groups in the large $N$ limit. A The critical contour lines of the stability criterion $\mathcal{K} = 0$ (Eq. 5.16) for $\alpha = 1.25$, 1.5, 2, 2.5, 3, 3.5 (from left to right) plotted as a function of the within-group correlation $c$. The homogeneous solution is stable outside a contour line and unstable in its interior. The lines depicted in panels B, C, and D correspond to the equilibrium synaptic states obtained by numerically solving the analytical expressions for the two-group solutions with $r = 10$ Hz. Circles depict synaptic efficacies obtained by simulating the mean field learning dynamics of $N = 1000$ synapses. We set $N = 1000$ here to obtain a better approximation of the analytical results derived for the large $N$ limit. Note that small discrepancies between the analytical curves and the simulated mean field equilibrium, stemming from the finiteness of $N$ in the simulation, can be observed in panel B. B Equilibrium synaptic weights (black line) as a function of $\mu$ for $c = 0.11$ (horizontal line in panel A) with $\alpha = 1.5$. Gray lines show the equilibrium weights for $c = 0.05$, 0.06, ..., 0.2. C, D Equilibrium synaptic weights (black lines) as a function of $c$ for $\mu = 0.15$ (vertical line in panel A) with $\alpha = 1.5$ in C and $\alpha = 3$ in D. The gray lines in panel C show how the region where the homogeneous solution is unstable vanishes as $\mu$ is increased from $\mu = 0.15$ to $\mu = 0.1725$ in steps of 0.0025. The gray lines in panel D show the change in the equilibrium synaptic weights when $\alpha$ is increased from 2 (largest group separation) to 4 (smallest group separation) in steps of 0.1. The dashed lines in both panels depict the equilibrium synaptic weights for the multiplicative case ($\mu = 1$).
do not use $\alpha$ close to one, but rather set it to a generic value of $\alpha = 1.5$.

Figures 5.5C and 5.5D describe the behavior of the system as the within-group correlation is gradually turned on. As expected from the analysis of the uncorrelated input scenario, the substantial weight-dependence of the synaptic changes induced when $\mu = 0.15$ (solid black lines), yields a stable homogeneous synaptic state if the within-group correlation is sufficiently weak. However, when the correlation reaches a critical value, the homogeneous state becomes unstable and the synaptic efficacies segregate into the two input groups, with the one winning the competition suppressing the other. As the correlation increases still further, another transition may occur at a higher value of $c$, above which the homogeneous synaptic state becomes stable again. The presence of this second transition (which is discontinuous) depends on the values of $\tau r$, the expected number of input spikes per synapse arriving within its learning time window, and the ratio of depression and potentiation $\alpha$ (compare Figures 5.5C and 5.5D). Importantly, for large values of $\mu$, in particular in the multiplicative model ($\mu = 1$), the stabilizing force is so strong that the homogeneous synaptic state remains stable for all positive correlation strengths (dashed black lines in Figures 5.5C and 5.5D, see also Appendix) and no segregation is possible.

The behavior described above for the linear neuron is reproduced qualitatively in simulations of the Integrate-and-Fire neuron, as shown in Figure 5.6A. To address the question whether symmetry breaking in the Integrate-and-Fire neuron can occur also at higher values of $\mu$, we follow the linear Poisson neuron analysis shown in Figure 5.5A which suggests that increasing the value of $\alpha$, extends the $\mu$-range of bimodal synaptic distributions. Figure 5.6B displays the learned synaptic distributions as a function of $\mu$ for a within-group correlation $c = 0.05$ with $\alpha = 1.5$ and $r = 10$ Hz. Similar to the linear neuron findings shown in Figure 5.5B, symmetry breaking occurs here in a wide regime of $\mu$.

---

**Figure 5.6:** Next page. *Symmetry breaking in the Integrate-and-Fire neuron driven by two equally sized correlated groups.* Each group consists of 500 synapses receiving Poisson inputs. **A** The learned synaptic weights of the two input groups as a function of the within-group correlation $c$, for an input rate of 40 Hz, $\mu = 0.019$ and $\alpha = 1.05$. For each correlation, the group means (circles) are computed for 30 different realizations of the synaptic weight distribution, taken after convergence at successive intervals of 500 s. Error bars indicate the corresponding standard deviations. The inset shows how the regime of symmetry breaking vanishes as $\mu$ is increased through $0.019, 0.021, 0.022, 0.023, 0.024$. **B** The learned distributions of synaptic weights as a function of $\mu$ for an input rate of 10 Hz, $c = 0.05$ and $\alpha = 1.5$. In all bimodal weight distributions depicted here, the splitting corresponds to the two input groups.
To emphasize important differences between symmetry breaking in non-linear versus additive TAH learning, Figure 5.7 shows corresponding learned synaptic efficacies for selected cases of low, intermediate, and high within-group correlations. Figures 5.7A-D depict learned weight distributions from Figure 5.6A where $\mu = 0.019$. For each correlation, synaptic efficacies resulting from additive learning are depicted on the right (Figures 5.7E-H). Except in panels A and B where $c = 0$, i.e. no input subgroups are defined, the synaptic distribution of the subgroup with higher mean efficacy is depicted in light gray, while that of the subgroup with lower mean efficacy is displayed in dark gray.

Inspection of panels A and B versus E and F shows that in the regime of low correlations the learning behavior induced by the two types of plasticity is qualitatively different. While in non-linear TAH learning (panels A,B) the homogeneous synaptic state is stable and all synapses distribute around the same mean efficacy, unstable additive learning induces symmetry breaking (panels E,F). Importantly, this symmetry breaking in general does not reflect the correlation structure in the afferent input. As shown in Figure 5.7F, when the within-group correlation is 0.03, the 500 synapses of the group winning the competition (light gray), split into two fractions of 325 vs. 175 synapses, of which the larger fraction tends to zero efficacy and mixes with the efficacies of the loosing input group. In contrast, in the non-linear TAH model, unfaithful splitting of the weights occurs only for extremely small values of $\mu$, of the order of $1/\tau r N$ (compare panels A,B to E,F). This is because symmetry breaking within a uniformly correlated group does not occur for $\mu > 1/\tau r N$, and hence the weights of each subgroup remain the same.

For intermediate strengths of the within-group correlation, both learning rules induce symmetry breaking that faithfully reflects the structure of the input correlation, with the

---

**Figure 5.7: Next page.** *Symmetry breaking in non-linear and additive TAH learning.* The learned synaptic efficacies of the Integrate-and-Fire neuron driven by two equally sized correlated groups are depicted for selected within-group correlations $c = 0$, 0.03, 0.1, 0.3. Each input group consists of 500 synapses receiving Poisson inputs with rate 40 Hz (cf. Figure 6A). The asymmetry parameter $\alpha = 1.05$. For $c = 0$ (panels A and E), the histogram of the total synaptic population is depicted. For $c > 0$, the light bars describe the distribution of the pre-synaptic input group with the higher mean, and the dark bars (stacked on top of the light ones) depict the distribution of the weaker group. All histograms were obtained by averaging over 30 different readouts of the converged synaptic weights taken at consecutive intervals of 500s. **A,B,C,D** Results from non-linear TAH learning with $\mu = 0.019$. **E,F,G,H** Results from additive TAH learning, $\mu = 0$. In panel E, a total of 112 synapses are in the upper mode. In panel F, the stronger group (light bars) has a bimodal distribution with 175 synapses (35%) in the upper mode.
synaptic distributions of the two input groups well separated. This is shown in panels C and G for a correlation of $c = 0.1$. Note, however, that whereas in additive learning the efficacies of both input groups reach the respective boundaries of the allowed range (i.e. are clipped to saturation), the weights resulting from non-linear TAH learning do not saturate. As we show in the next section, this property of NLTAH plasticity enhances the sensitivity of the synaptic population to changes in the strength of the within-group correlation. Finally, when the within-group correlation is strong, in both types of learning all efficacies become large (Figures 5.7D,H).

Clearly, the detailed quantitative properties of the learned synaptic patterns, as well as the parameter values where symmetry breaking occurs, depend on the neuron model, and specifically on the spike generating mechanism. Nevertheless, the striking qualitative similarity in the findings from both neuron models investigated here suggests that the symmetry breaking induced by the within-group correlations is a general property of the non-linear TAH rule with small but non-zero $\mu$, independent of the specifics of the spike generator.

### 5.3.2 Synaptic Representation of Input Correlations

In the previous section, we studied the emergence of symmetry breaking in homogeneous synaptic populations for different types of instantaneously correlated input activity. In this section we study the more general issue of how information about the spatio-temporal structure of the afferent input is imprinted into the learned synaptic efficacies by TAH plasticity. Specifically, we investigate how the weight-dependence of the synaptic changes affects the sensitivity of the learning to features embedded in the input spikes trains.

An example of the associated phenomena is shown in Figure 5.8. Here we study the effect of weight-dependence on the steady state synaptic efficacies of the Integrate-and-Fire neuron receiving 1000 Poisson inputs that comprise a small subgroup of 50 correlated synapses ($c = 0.1$) while all other input cross-correlations are zero. In this scenario, the subgroup is statistically distinct from the rest of the synaptic population. The coherence of spikes within the subgroup increases the causal correlation of the member synapses with the spiking activity of the post-synaptic neuron. Due to the ensuing cooperation between the correlated synapses, they grow stronger than those of the uncorrelated background. Figure 5.8 shows how the strength of the stabilizing drift induced by the weight-dependence of the synaptic updating modulates the degree of separation between the two subpopulations. For decreasing values of $\mu$, learning becomes increasingly affected by the correlation structure in the input, and the separation between the subgroup and the background is more pronounced. However, below a critical $\mu$, the homogeneous state of the uncorrelated population loses stability, and splits resulting in a bimodal distribution
Figure 5.8: Effect of the updating parameter $\mu$ on the learned synaptic distribution of the Integrate-and-Fire neuron driven by a small correlated group and an uncorrelated background. The neuron is driven by 950 uncorrelated and 50 weakly correlated ($c = 0.1$) Poisson input processes with a rate of 10 Hz ($\alpha = 1.05$). Together, panels A and B cover the range of $\mu \in [0, 1]$, but note the difference in resolution.

of the background synapses. As a consequence, the representation of the afferent correlation structure in associated groups of synaptic efficacies is confounded by the mixing of the high-efficacy mode of the background with the subgroup of correlated synapses. This example raises the general problem of finding an optimal learning rule that, for a given type of input activity, compromises best between sensitivity and stability.

To address this question we need a quantitative measure for the performance of a learning rule in imprinting information about the input correlations onto the synaptic efficacies. Here we apply the sensitivity measure $S$ (Eq. 5.13, Methods), which quantifies the sensitivity of the learned synaptic state to changes in features embedded in the input correlation structure. When $S$ is high, small changes in the input features are picked up by learning and induce a large change in the learned synaptic efficacies. We emphasize that the goal of this performance measure is to quantify and compare general properties of different plasticity rules. It is therefore based only on the relation between the afferent neuronal inputs and the learned synaptic efficacies. In particular, it avoids direct reference to the neuronal output activity.
We first illustrate the application of the sensitivity measure by considering a simple example where the input feature to be represented by the learned synaptic efficacies is only one dimensional, i.e. a scalar quantity. Specifically, we apply $S$ to the scenario discussed in the previous section, of two independent input groups with within-group correlation $c$. We investigate the behavior of the linear Poisson neuron and quantify how the sensitivity of the learned synaptic distribution to the strength of the within-group correlation is affected by the weight-dependence of the synaptic changes. We consider the sensitivity of the learning as a function of $\mu$ for a fixed correlation of $c = 0.11$. As shown in Figures 5.5B and C, this correlation represents an intermediate correlation strength in the linear Poisson neuron treatment. Using the steady state synaptic efficacies from Figure 5.5B, we compute $S$ for values of $\mu$ between 0 and 0.5 (see Methods, Appendix). Figure 5.9 shows the resulting sensitivity curve. We note that each point quantifies the sensitivity of the learned synaptic weights to small changes in the correlation strength around $c = 0.11$.

![Sensitivity curve](image)

Figure 5.9: Sensitivity to the within-group correlation strength $c$ in the linear neuron receiving input from two correlated subgroups. The sensitivity is plotted as a function of $\mu$ for correlation strength $c = 0.11$ and $\alpha = 1.5$ (cf. Figure 5.5B for parameter settings, large $N$ limit).

As can be seen in Figure 5.5B, there are two qualitatively distinct regimes of synaptic distributions emerging from learning in this case. For high values of $\mu$, no symmetry breaking takes place and the correlation strength is represented by the common mean
value of the synaptic efficacies. In this regime ($\mu \geq 0.15$), $S$ decreases monotonically with increasing $\mu$ (see Figure 5.9), because the higher weight-dependence strengthens the confinement of the homogeneous synaptic state to the center range of the synaptic efficacies. For lower values of $\mu$, symmetry breaking occurs and the correlation strength is represented by the mean efficacy values of the two resulting groups. In this regime, $S$ is non-monotonic in $\mu$. For very low $\mu$, the synaptic efficacies are close to saturation at the boundaries and, hence, a change in the correlation strength cannot induce a large change in the efficacies. On the other hand, the centralizing drift induced by a large $\mu$ reduces the sensitivity. Thus, $S$ has a maximum at an intermediate $\mu$ (in the present case around $\mu = 0.02$). Finally, at the transition between the regions of homogeneous and bimodal synaptic distributions ($\mu \approx 0.15$), sensitivity is large because here a small change in $c$ may cause an abrupt and large change in the synaptic efficacies, namely a bifurcation from a homogeneous to an inhomogeneous synaptic distribution. Note, however, that this transition region in $\mu$ is narrow.

We now turn to a richer input scenario in which the afferent correlation structure is inhomogeneous and the input feature space to be represented by the learned synaptic efficacies is high-dimensional. Specifically, we consider pre-synaptic activity in which each synapse receives spike inputs with a specific relative latency with respect to the remaining synaptic population. Such latency or delay-line scenarios have been previously studied in the context of additive TAH learning [Gerstner et al., 1996, Song et al., 2000] and can, for instance, be motivated by their analogy to certain delay-line models in auditory processing [LA, 1948].

We consider the input activity to consist of $N$ time-shifted versions of one common Poisson spike train with rate $r$. Since the synaptic learning process depends on the relative timing of the input spikes, we fix one pre-synaptic input as reference, and treat the remaining $N - 1$ delays $\Delta = (\Delta_1, \ldots, \Delta_{N-1})$ as $R = N - 1$ dimensional vector of input features to be represented by the learned synaptic weights. While the delays $\Delta$ fully specify the temporal correlation structure of the neuronal input activity, $S$ measures the sensitivity of the learned synaptic efficacies to small independent changes in the individual delays. Because of the temporal sensitivity of TAH plasticity, it is intuitively clear that the learning dynamics will critically depend on the temporal scale of the relative delays. Although it is a natural choice to set this temporal scale through the standard deviation of a Gaussian distribution from which the delays are drawn [Song et al., 2000, Aharonov et al., 2001, Gütig et al., 2001], we here apply the sensitivity measure to the simpler case where we fix $\Delta$ such that the delays between the $N$ inputs are uniformly spaced at a fixed delay $\sigma/(N-1)$, i.e. $\Delta_i = i\sigma/(N-1)$ ($i = 1, 2, \ldots, N - 1$). We have checked that the qualitative behavior of $S$ in the case of a fixed delay spacing is similar to that of $S_{\text{avg}}$ (Methods) obtained from averaging over an ensemble of Gaussian delay vectors with
standard deviation $\sigma$ [Aharonov et al., 2001, Gütig et al., 2001].

We investigate here the behavior of the linear Poisson neuron. One important difference between the delay-line input scenario considered here and the input correlations treated above is that here non-zero cross-correlations between input spike trains exist also at negative time lags. Specifically, if the delays of the input activities of synapses $i$ and $j$ are given by $\Delta_i$ and $\Delta_j$, respectively, and the additional delay of the post-synaptic neuron is $\varepsilon$ (Eq. 5.5), the delay difference $\Delta_i - (\Delta_j + \varepsilon)$ determines the temporal position of the sharp peak in the otherwise zero effective correlation between the two shifted Poisson inputs (Eq. 5.7). If this delay difference is negative, the output activity contributed by the $j$th synapse lags behind the input spikes at the $i$th synapse. Hence, the $j$th synapse contributes to the potentiation of synapse $i$, and the respective effective causal correlation $C_{ij}^+$ is positive. Correspondingly, in this case the backward effective correlation $C_{ij}^-$ contributed by synapse $j$ to the depression of synapse $i$ is zero. Conversely, if the delay difference between the $i$th and $j$th input spike trains is positive, synapse $i$ is depressed by the activity of synapse $j$ because $C_{ij}^+$ becomes positive. In both cases, the magnitude of the effective correlation is scaled by the exponentially decaying time dependence of the learning rule (Eq. 5.1). The full expressions for the effective correlation matrices $C_{ij}^+$ and $C_{ij}^-$ are given in the Appendix.

To calculate $S$ for a given delay vector $\Delta$, we numerically solve the drift equation of the synaptic learning (Eq. 5.8) for the synaptic steady state. Using the resulting learned synaptic distributions, we compute the susceptibility matrix $\chi$ (Eq. 5.12, Appendix), giving $S$ (Eq. 5.13). Figure 5.10A shows the sensitivity $S$ as a function of $\mu$ for different values of the temporal delay spacing $\sigma$. The curves clearly show an optimal weight-dependence of the synaptic changes where the sensitivity peaks. For larger values of $\mu$, the performance of the learning deteriorates because the increasing confinement of the synaptic weights to the central range of efficacies restricts the sensitivity of the learning to changes in the input correlation structure. Conversely, for lower values of $\mu$, the sensitivity is impaired because the synaptic efficacies are beginning to saturate at the boundaries of the allowed range as bimodal efficacy distributions emerge. The value of $\mu$ that optimally adjusts the weight-dependence of the synaptic changes depends both on the system parameters and on the input correlations determined by the relative time delays between the inputs. Increasing $\sigma$, i.e. increasing the relative delays, weakens the effective correlations between the pre-synaptic inputs because of the exponentially decaying temporal extent of the learning rule (Eq. 5.7). Hence, a lower weight-dependence of the synaptic changes (corresponding to a lower value of $\mu$) is needed to pick up the correlations and allow sufficient sensitivity of the learning to the input delays. The effect of this change in the temporal extent of synaptic interactions on the learned efficacies is shown in Figure 5.11, which for each $\mu$ depicts all $N$ synaptic efficacies for $\sigma = \tau$ (A) and $\sigma = 4\tau$ (B). Note that because
Figure 5.10: Sensitivity in the linear neuron evaluated at uniformly spaced input delays. The delay vector $\Delta$ covers the range from 0 to $\sigma$ (see Results). The learning dynamics is simulated for the linear neuron with $N = 101$ synapses driven by the delayed Poisson inputs with a rate of 10 Hz and $\alpha = 1.5$. A The sensitivity per input feature $S/R$ as a function of $\mu$ for $\sigma/\tau = 0.25, 0.5, 0.75, \ldots, 4.75, 5$ (from bottom to top). The termination of the curves at low $\mu$ is a result of poor numerical convergence arising as the synaptic efficacies come close to the boundaries of the allowed range. B The sensitivity per feature $S/R$ as a function of the delay interval $\sigma/\tau$ for $\mu = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ (from top to bottom).

of the equidistant delays, the relation between the relative temporal position of a synapse within the pre-synaptic population and its steady state efficacy is monotonic, with the leading synapse ($\Delta = 0$) taking the largest weight. In the foreground the corresponding sensitivity curves are shown. The plots clearly demonstrate that the saturation regime, in which most synaptic weights accumulate at the boundaries of the allowed range (black
and white), begins at higher values of $\mu$ when the temporal dispersion of the inputs is small (A), i.e. the synaptic interactions are strong. The plot also reveals that in both cases for low values of $\mu$, only the leading synapse remains at a high value. Finally, it can be seen that the peaks in the sensitivity curves roughly coincide with those values of $\mu$ where the synaptic weights smoothly cover a large range of efficacies, as shown by the gradual change from dark to light values in the corresponding vertical cross sections.

![Graph](image)

Figure 5.11: Mapping between synaptic delays and equilibrium weights. The parameters are identical to those used in Figure 5.10A with $\sigma/\tau = 1$ in panel (A) and $\sigma/\tau = 4$ in panel (B). The background depicts the learned synaptic efficacies as a function of $\mu$. Each row $i$ reflects the equilibrium weight $w_i$ in gray-scale, with the upper row corresponding to the leading synapse with zero delay and the bottom row to the last synapse with delay $\sigma$. The solid lines depict the sensitivity curves from Figure 5.10A.

Finally, we ask how the learning sensitivity depends on the statistics of the input delays for a fixed value of $\mu$. To answer this question, Figure 5.10B shows $S$ as a function of the delay-line spacing $\sigma$, demonstrating that $S$ does not vary monotonically with $\sigma$, but rather has a maximum at an optimal temporal separation of the inputs. This is because tight
spacing leads to strong effective correlations between the inputs, driving the synapses towards saturation. On the other hand, loose spacing reduces the effective correlations between the pre-synaptic inputs to the extent that the learning behaves essentially as if driven by an uncorrelated pre-synaptic population.

5.4 Discussion

The understanding of activity-dependent refinement of neural networks has long been one of the central interests of synaptic learning studies. In this context, most investigations of unsupervised learning using correlation-based plasticity rules have been conducted in the framework of additive plasticity models, which do not incorporate explicit weight-dependence in the changes of synaptic efficacies. These simple models suffer from stability problems: either all synapses decay to zero or they grow without bound. An additional problem inherent to simple Hebbian models is the lack of robust competition. Indeed, it has been found that even the inclusion of synaptic depression mechanisms does not provide a robust source of synaptic competition, unless synaptic plasticity is fine-tuned to approximately balance the amount of potentiation and depression [Miller, 1996].

Recent studies of experimentally observed temporally asymmetric Hebbian learning rules have added two new ideas. One idea is that under these plasticity mechanisms, synapses compete against each other in controlling the time of firing of the target cell and, thus, engage in competition in the time domain. While TAH learning rules are indeed inherently sensitive to temporal correlations between the afferent inputs, we have shown here that this sensitivity alone is not sufficient to resolve the problems associated with either stability or competition. In the additive model of TAH plasticity, hard constraints need to be imposed on the maximal and minimal synaptic efficacies to prevent the runaway of synaptic strength. In addition, as was shown here, in this model synaptic learning is competitive, only when the ratio between depression and potentiation is fine-tuned, and even then the emergent synaptic patterns do not necessarily segregate the synaptic population according to the correlation structure in the neuronal input. The second idea is that TAH rules would exhibit novel behavior due to the role of the non-linear spike generation mechanism of the post-synaptic cell [Song et al., 2000]. In fact, we have shown in this work that the qualitative features of TAH plasticity are strikingly insensitive to the non-linear integration of inputs in the target cell (see also [Kempter et al., 2001]). For the parameter choices studied, the properties of the synaptic steady states in the Integrate-and-Fire neuron are qualitatively similar to those found in a linear input-output model for neuronal firing. Nevertheless, we note that there are substantial quantitative differences between the two models, particularly with respect to the parameters $\tau r N$ and $c$ which effectively determine the correlations between the pre- and post-synaptic spike trains.
Although a quantitative analysis of these differences is beyond the scope of our work, such a study might reveal interesting insights into the quantitative effects of the details of the post-synaptic spike generator on the learned synaptic distributions. Further, it is possible that the details of the spike-generation mechanism will affect the transient phase, i.e. the dynamics, of the synaptic learning process.

From the present work we conclude that some of the underlying difficulties in correlation-based learning are alleviated by non-linear plasticity rules such as the NLTAH rule. The non-linear weight-dependence of the synaptic changes provides a natural mechanism to prevent runaway of synaptic strength. As in additive TAH learning, synaptic competition is provided by the mixture of depression and potentiation. However, in NLTAH plasticity, the balance between depression and potentiation is maintained dynamically by adjusting the steady state value of the synaptic efficacies. Indeed, we have shown that this competition is sufficient to generate symmetry breaking between two independent groups of correlated pre-synaptic inputs. However, in order for this to occur, the stabilizing drift induced by the weight-dependence of the synaptic changes should not be too strong. In particular, the simple linear weight-dependence (Eq. 5.2 with $\mu = 1$) assumed in the original multiplicative model, is incapable of breaking the symmetry between competing input groups. In fact, we have shown that with $\mu = 1$ the homogeneous synaptic state is stable for any pattern of homogeneous input correlations, provided there are no negative correlations in the afferent activity. The present power-law plasticity rule with $0 < \mu < 1$ provides a reasonable balance between the need for a stabilizing force and a potential for spontaneous emergence of synaptic patterns. Our study of symmetry breaking between two competing groups of correlated synapses is inspired by the activity-dependent development of ocular dominance selectivity. This scenario has also been studied recently by [Song and Abbott, 2001] using the additive version of TAH plasticity. In their model, achieving a faithful splitting between the two competing input groups with weak correlations requires relatively tight tuning of the depression to potentiation ratio, $\alpha$.

One of the surprising results of our investigation is the possibility that when the correlation within input groups is made strong, the stability of the homogeneous synaptic state may be restored. We have shown that this apparently counterintuitive behavior, predicted by the analytical study of the mean synaptic dynamics of the linear Poisson neuron, is also seen in simulations of the full learning rule in the Integrate-and-Fire neuron. It would be interesting to explore possible experimental testing of this result, perhaps in the context of the development of ocular dominance. In this work we have limited ourselves to correlated subpopulations of inputs with positive within-group correlations. However, in general, negative correlations are an additional potential source of competition [Miller and MacKay, 1994]. Furthermore, we have not addressed the important issue of competition between synapses that target different cells. Lateral inhibitory connections between
target neurons may provide a source of such competition.

The last part of the present work addresses situations with inhomogeneous input statistics. Different inputs are distinct in their temporal relationship to the rest of the input population. Here the issue is not whether a spatially modulated pattern of synaptic efficacies will form through TAH learning, but rather whether this pattern will efficiently imprint the information embedded in the input statistics. To quantify the imprinting efficiency of the learning rule, we introduced a new method for measuring learning rule sensitivity. In the present context, this measure quantifies the amount of information about the temporal structure in the inputs that a TAH rule can store. Using this method to study the novel class of NITAH plasticity rules introduced here, we find that the optimal learning rule depends on the input statistics, in the present example on the characteristic time scale of the temporal correlations between the inputs. This finding suggests that biological systems may have acquired mechanisms for metaplasticity to adapt the learning rule to slow temporal changes in the input statistics. It should be pointed out that the sensitivity measure \( S \) focuses entirely on how the learned synaptic distribution changes as a result of changes in the correlation pattern among the input channels. It does not, however, address the problem of 'readout', namely how the resulting changes in the synaptic distribution affect the firing pattern of the output cell. A measure which takes the post-synaptic spike train into account, will in general depend on the details of the spike generating mechanism, rather than only capture the properties of the learning rule. In general, however, any readout mechanism will depend on the information that is available in the learned synaptic state. Hence, if the learning itself is insensitive to changes in the input features, the synaptic efficacies will fail to represent these changes and no readout mechanism will be able to extract them. The sensitivity measure \( S \) therefore provides an upper bound on the learning performance of the full neural system (including readout). In summary, while quantitative claims about the optimality of specific learning rules have to consider specific readout mechanisms, our study of the general properties of the investigated plasticity rules provide general insights into the mechanisms that enable unsupervised synaptic learning to remain sensitive to input features during learning.

Present experimental results [Bi and Poo, 1998] based on the averaging of individual efficacy changes in different synapses, suggest the possibility that indeed the ratio of depressing and potentiating synaptic changes increases in a stabilizing fashion as synapses grow stronger (cf. [van Rossum et al., 2000]). However, available data do not provide conclusive evidence regarding the details of the weight-dependence of the efficacy changes. Our work clearly demonstrates the importance of the weight-dependence of the TAH updating rule. Synaptic learning rules that implement a stabilizing weight-dependence of the type introduced in this work, have several advantageous properties for the learning in neural networks. Specifically, our results predict, that synaptic changes should neither
be additive nor multiplicative, but rather feature intermediate weight-dependencies that could, for instance, result from a gradual saturation of the potentiating and depressing mechanisms. It will be interesting to see whether future experimental results will confirm such prediction. In this context, it is also important to note that recent experiments and modeling studies reveal important non-linearities in the accumulation of synaptic changes induced by different spike pairs [Castellani et al., 2001, Senn et al., 2001, Sjöström et al., 2001], as well as evidence for complex intrinsic synaptic dynamics that challenges the simple notion of a scalar synaptic efficacy [Markram and Tsodyks, 1996]. The theoretical implications of these sources of non-linearity and intrinsic dynamics remain to be explored.
Appendix

Generating Correlated Spike Trains

We show here that two spike trains that are generated by conditioning their binwise spike probabilities on the activity of a common reference spike train $X_0(T)$ as described in Methods, have a pairwise correlation coefficient $c$. For clarity, we denote $X_i(T)$ simply by $X_i$. The pairwise correlation coefficient is defined by $\text{Cov}(X_i, X_j)/\sqrt{\text{Var}(X_i)\text{Var}(X_j)}$, where the covariance is $\text{Cov}(X_i, X_j) = E[X_iX_j] - E[X_i]E[X_j]$. Since $X_i$ is either 0 or 1, $E[X_i] = P(X_i = 1)$, and therefore

$$E[X_i] = P(X_i = 1) = P(X_0 = 1)P(X_i = 1|X_0 = 1) + P(X_0 = 0)P(X_i = 1|X_0 = 0)$$

$$= r\Delta T\theta + (1 - r\Delta T)\varphi$$

$$= r\Delta T,$$

where in the final step we use Eqs. 5.11. Note that because $E[X_i] = r\Delta T$, the spike train $X_i$ has rate $r$. Similarly,

$$E[X_iX_j] = P(X_i = 1, X_j = 1)$$

$$= P(X_0 = 1)P(X_i = 1, X_j = 1|X_0 = 1) + P(X_0 = 0)P(X_i = 1, X_j = 1|X_0 = 0)$$

$$= r\Delta T\theta^2 + (1 - r\Delta T)\varphi^2$$

$$= (r\Delta T)^2 + c(r\Delta T)(1 - r\Delta T).$$

Finally, since $\text{Var}(X_i) = (r\Delta T)(1 - r\Delta T)$, the binwise correlation coefficient becomes

$$\frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}} = \frac{(r\Delta T)^2 + c(r\Delta T)(1 - r\Delta T) - (r\Delta T)^2}{(r\Delta T)(1 - r\Delta T)} = c.$$

Homogeneous Synaptic Steady State for a Homogeneous Population of Synapses

We derive the homogeneous synaptic steady state solution by setting $\dot{w}_i = 0$ and $w_i = w^*$ in Eq. 5.8 with $C_{ij}^{-} = 0$,

$$\frac{\lambda\tau^2}{N} \left[ f_+(w^*) \sum_{j=1}^{N} C_{ij}^{+} w^* - \Delta f(w^*) \sum_{j=1}^{N} w^* \right] = 0.$$

Using Eq. 5.14, this yields

$$f_+(w^*)w^*NC_0 - \Delta f(w^*)w^*N = 0,$$

which, discarding the trivial solution where all synapses are zero, implies that the homogeneous steady state

$$f_+(w^*)C_0 - \Delta f(w^*) = f_+(w^*)(1 + C_0) - f_-(w^*) = 0.$$  \hfill (5.18)
From this, we find that the homogeneous solution is given by \( w_i = w^* \), where \( w^* \) is the solution to Eq. 5.15. Since \( C_0 \) is positive, this equation has a unique solution \( 0 < w^* < 1 \) for any \( \mu > 0 \). Note, however, that in the additive model where \( \mu = 0 \), the ratio \( f_-(w)/f_+(w) = \alpha \) and, hence, in general there is no homogeneous synaptic steady state unless \( w^* = 0 \) or all weights are clipped to the upper boundary.

For uncorrelated input activity \( C_{ij}^+ = \delta_{ij}/(\tau r) \) and hence \( C_0 = 1/(\tau r N) \). Substituting \( C_0 \) into Eq. 5.15, we obtain the homogeneous solution for this input scenario,

\[
    w^* = \frac{1}{1 + \alpha^{1/\mu} \left( 1 - \frac{1}{1 + \tau r N} \right)^{1/\mu}}.
\]

(5.19)

Stability of the Homogeneous Synaptic Steady State

We analyze the stability of the homogeneous synaptic steady state by deriving the time evolution of small perturbations \( \delta w_i = w_i - w^* \) of the synaptic efficacies \( w_i \) around the steady state value \( w^* \). If these perturbations decay to zero, the homogeneous steady state is stable. For small perturbations, the time evolution is given by

\[
    \dot{\delta w}_i = \sum_{j=1}^N \frac{\partial \delta w_i}{\partial w_j} \delta w_j.
\]

Using the expression for the synaptic drifts from Eq. 5.8, we obtain

\[
    \delta w_i = \frac{\lambda \tau r^2}{N} \left[ -N g_0 \delta w_i - \Delta f(w^*) \sum_{j=1}^N \delta w_j + f_+(w^*) \sum_{j=1}^N C_{ij}^+ \delta w_j \right]
\]

where

\[
    g_0 = w \left[ \frac{d}{dw} \Delta f(w) - C_0 \frac{d}{dw} f_+(w) \right] \bigg|_{w=w^*}
    = w \left[ \frac{d}{dw} f_-(w) - (1 + C_0) \frac{d}{dw} f_+(w) \right] \bigg|_{w=w^*}
    = \frac{w}{f_+(w)} \left[ f_+(w) \frac{d}{dw} f_-(w) - f_-(w) \frac{d}{dw} f_+(w) \right] \bigg|_{w=w^*}
    = w f_+(w) \frac{d}{dw} \left[ \frac{f_-(w)}{f_+(w)} \right] \bigg|_{w=w^*}
\]

(5.20)

In the second step, we used Eq. 5.15 to substitute for \( 1 + C_0 \). Note that the factor \( g_0 \) is proportional to the derivative of the ratio between the scales of negative and positive synaptic changes with respect to the weights. Hence, \( g_0 = 0 \) in the additive model. For \( \mu > 0 \),

\[
    g_0 = w (1 - w) \mu \frac{d}{dw} \left[ \alpha \left( \frac{w}{1 - w} \right)^\mu \right] \bigg|_{w=w^*}
    = \alpha \mu \frac{w^{*\mu}}{1 - w^*}
\]

(5.21)
is positive because $0 < w^* < 1$. In matrix notation the time evolution of a synaptic perturbation $\delta \mathbf{w}$ can be rewritten as

$$\dot{\delta \mathbf{w}} = \frac{\lambda \tau \nu^2}{N} J \delta \mathbf{w}$$

with the matrix

$$J_{ij} = -\delta_{ij} N g_o - \Delta f(w^*) + f_+(w^*) C^+_{ij}.$$

If the eigenvalues of $J$ are negative, all perturbations of the homogeneous state are attenuated by the learning dynamics, and, hence, this synaptic state is stable. In contrast, if any eigenvalue of $J$ is positive, a perturbation along the direction of the corresponding eigenvector will grow exponentially. The matrix $J$ has a homogeneous eigenvector with eigenvalue $-N g_o - N \Delta f(w^*) + N C_0 f_+(w^*)$, which using Eq. 5.18 reduces to $-N g_o$. Since this eigenvalue is always negative, the homogeneous component of any perturbation $\delta \mathbf{w}$ decays to zero with rate $\lambda \tau \nu^2 g_o$. In contrast, the temporal evolution of the strictly inhomogeneous component of $\delta \mathbf{w}$ (whose elements sum to zero) comprises a spectrum of rates which are determined by the various eigenvalues of $J$ that correspond to inhomogeneous eigenvectors. The largest inhomogeneous eigenvalue of $J$ is $f_+(w^*) N C_1 - N g_o$, where $N C_1$ denotes the largest inhomogeneous eigenvalue of $C^+$. Hence, the homogeneous synaptic state is stable, if $f_+(w^*) N C_1 - N g_o < 0$, which gives the stability criterion stated in the Results (Eq. 5.16). Inserting Eqs. 5.15 and 5.21 into the criterion 5.16, we obtain an upper bound for $\mu_{\text{crit}}$, the largest value of $\mu$ for which the homogeneous solution is unstable

$$\mu_{\text{crit}} < \frac{C_1}{1 + C_0}. \quad (5.22)$$

An important observation is that $C_1 \leq C_0$, and hence this bound is necessarily smaller than 1, implying that the homogeneous solution is always stable in the multiplicative model where $\mu = 1$. To see that $C_1 \leq C_0$, recall that since $N C_1$ is an eigenvalue of $C^+$, there is an eigenvector $\mathbf{v}$ such that $C^+ \mathbf{v} = N C_1 \mathbf{v}$. Specifically, for $v_m$, the largest component of $\mathbf{v}$, this implies that $N C_1 v_m = \sum_{j=1}^{N} C^+_{mj} v_j$, and hence, $N C_1 = \sum_{j=1}^{N} C^+_{mj} (v_j / v_m) \leq \sum_{j=1}^{N} C^+_{mj}$. But since $\sum_{j=1}^{N} C^+_{mj} = N C_0$ (Eq. 5.14), this yields $C_1 \leq C_0$.

For uncorrelated input activity the above bound for $\mu_{\text{crit}}$ becomes

$$\mu_{\text{crit}} < \frac{1}{1 + \tau \nu N^2}. \quad (5.23)$$

where we used $C_0$ and $C_1$ from Eq. 5.17. Hence, for large $\tau \nu N$, the regime of $\mu$ where symmetry breaking exists vanishes at least with $1/(\tau \nu N)$.

**Additive TAH in the Linear Neuron – Uncorrelated Inputs**

The drift of the $i$th synapse of a neuron receiving uncorrelated inputs and implementing
the additive model, is given by setting \( \mu = 0 \) in Eq. 5.8 with \( C_{ii}^+ = 1/(\tau r) \) and all other effective correlations equal 0

\[
\dot{w}_i = \lambda \tau r^2 (1 - \alpha) \left( \frac{1}{N} \sum_{j=1}^{N} w_j \right) + \frac{\lambda r w_i}{N}.
\]  

(5.24)

As explained above, this linear system has no steady state. Imposing the boundary conditions by clipping the efficacies, results in all synapses taking the value of either 0 or 1. Thus, the learned synaptic distribution is fully described by \( n_{up} \), the ratio of the synapses that are saturated at the upper boundary. For a ratio \( n_{up} \) to be consistent, the drift of a synapse with efficacy 0 must be non-positive, while the drift of a synapse with efficacy 1 must be non-negative. From imposing these conditions in Eq. 5.24 we get,

\[
\tau r^2 (1 - \alpha) n_{up} \leq 0 \quad \text{and} \quad \tau r^2 (1 - \alpha) n_{up} + \frac{r}{N} \geq 0.
\]

The first inequality implies \( \alpha \geq 1 \) if \( n_{up} > 0 \). It is important to note that the regime of \( \alpha < 1 \) would simply yield saturation of all efficacies at the upper boundary since all synapses experience a positive drift. The second condition yields

\[
n_{up} \leq \frac{1}{\tau r N (\alpha - 1)}.\]

However, it can be shown using methods similar to [Rubin et al., 2001], that

\[
n_{up} = \frac{1}{2 \tau r N (\alpha - 1)},
\]

where if this quantity is larger than 1, \( n_{up} = 1 \). Therefore, if \( \alpha \leq 1 + 1/(2 \tau r N) \), all synapses will saturate at the upper boundary, whereas if \( \alpha > 1 + 1/(2 \tau r) \), even a single synapse at the upper boundary will experience a negative drift, and hence no synapse will saturate at 1.

Moreover, the firing rate of the linear neuron is given by

\[
r^{\text{post}} = n_{up} r = \frac{1}{2 \tau r N (\alpha - 1)},
\]

which is independent of the input firing rate \( r \) (except for very low rates, where all synapses become strong, i.e. \( n_{up} = 1 \), and \( r^{\text{post}} = r \)). Thus, output rate normalization is a property of the linear neuron when the additive model is used.

**Uniformly Correlated Inputs in the Linear Neuron**

When the pre-synaptic inputs are uniformly correlated, namely \( c_{ij} = c \geq 0 \) for all \( i \neq j \) \( (c_{ii} = 1) \), the effective correlation matrix \( C_{ij}^+ = (c + \delta_{ij}(1 - c))/(\tau r) \), and hence

\[
C_0 = \frac{1 + c(N - 1)}{\tau r N} \quad \text{and} \quad C_1 = \frac{1 - c}{\tau r N}.
\]  

(5.26)

110
Because \( C_0 \) increases with \( c \), the correlation increases the value of the synaptic efficacy in the homogeneous synaptic state \( w^* \) (Eq. 5.15). To see the effect of the correlation on the stability of the homogeneous solution, note that because \( C_1 \) decreases with \( c \), the correlation decreases the value of \( K \) (Eq. 5.16), and hence, increases the stability of the homogeneous solution. Moreover, since both \( C_0 \) and \( C_1 \) decrease with \( 1/(\tau r N) \), for large \( \tau r N \) the critical \( \mu \) (Eq. 5.22) approaches zero. Thus, for any \( \mu > 0 \), the homogeneous synaptic state is stable when the effective population is sufficiently large.

\section*{Computing the Susceptibility Matrix \( \chi \) for the Linear Neuron}

Here we compute the susceptibility matrix \( \chi \) (Eq. 5.12, Methods) used in the Results to evaluate the sensitivity measure \( S \), for the learning process in the linear Poisson neuron model. This matrix is obtained by the implicit function theorem. In the synaptic steady state \( (w^*) \), the synaptic drifts are zero by definition, and hence,

\[
\frac{d\hat{w}_k}{d\Phi_j} \bigg|_{w^*} = 0 \quad (k = 1 \ldots N, \ j = 1 \ldots R),
\]

where \( R \) is the dimension of the space of input features. Using

\[
\frac{d\hat{w}_k}{d\Phi_j} \bigg|_{w^*} = \frac{\partial \hat{w}_k}{\partial \Phi_j} \bigg|_{w^*} + \sum_{i=1}^{n} \chi_{ij} \frac{\partial \hat{w}_k}{\partial w_i} \bigg|_{w^*} = 0 \quad \text{with} \quad \chi_{ij} = \frac{\partial w_i}{\partial \Phi_j} \bigg|_{w^*}
\]

and denoting

\[
M_{ij} = \frac{\partial \hat{w}_i}{\partial w_j} \bigg|_{w^*}, \quad M^0_{ij} = \frac{\partial \hat{w}_i}{\partial \Phi_j} \bigg|_{w^*}
\]

we obtain

\[
\chi = -M^{-1}M^0.
\]

The matrices \( \chi \) and \( M^0 \) are of dimensions \( N \) by \( R \), and \( M \) is an \( N \) by \( N \) matrix. Below we derive \( \chi \) for the two input scenarios studied in the Results.

\section*{Two Correlated Input Groups}

For the case of two correlated subgroups, the sensitivity to the within-group correlation is measured. Hence, the input feature is \( \Phi = c \) with \( R = 1 \). Using Eq. 5.8 with \( C_{ij} = 0 \), \( C_{ij} = c/(\tau r) \) if \( i \neq j \) are in the same subgroup \( (C_{ii} = 1/(\tau r)) \), and \( C_{ij} = 0 \) otherwise, we derive the expressions for \( M \) and \( M^0 \):

\[
M_{ii} = \frac{\tau r^2}{N} \left[ (1 - w_i)^\mu - \alpha w_i^\mu \right] + \frac{rc}{N} (1 - w_i)^\mu + \frac{r}{N} (1 - c) (1 - w_i)^\mu - \mu \frac{\tau r^2}{N} \left( \sum_{k=1}^{N} w_k \right) \left[ (1 - w_i)^{\mu-1} + \alpha w_i^{\mu-1} \right]
\]

\[
- \mu (1 - w_i)^{\mu-1} \frac{r}{N} \left[ (1 - c) w_i + \sum_{k \in G_i} w_k \right]
\]

111
\[ M_{ij} = \frac{\tau r^2}{N}[(1 - w_i)^\mu - \alpha w_i^\mu] + \frac{r c}{N}(1 - w_i)^\mu \quad j \in \mathcal{G}_i \]

\[ M_{ij} = \frac{\tau r^2}{N}[(1 - w_i)^\mu - \alpha w_i^\mu] \quad j \notin \mathcal{G}_i \]

\[ M^\circ_{ii} = (1 - w_i)^\mu \frac{r}{N} \left[ \sum_{k \in \mathcal{G}_i} w_k - w_i \right], \]

where \( j \in \mathcal{G}_i \) if synapse \( j \) and \( i \) are in the same group.

**Delayed Poisson Inputs**

For a neuron receiving time shifted versions of a common Poisson spike train \( \rho_0^{\text{pre}} \), the input processes are \( \rho_0^{\text{pre}}(t) = \rho_0^{\text{pre}}(t - \Delta_i) \) (\( i = 1, \ldots, N \)), where \( \Delta_i \) is the delay of the activity at synapse \( i \). Substituting these into Eq. 5.6, we obtain the effective correlation matrices (Eq. 5.7)

\[ C_{ij}^\pm = \Theta \left( \mp (\Delta_i - \Delta_j - \varepsilon) \right) (\tau r)^{-1} \exp \left( \mp (\Delta_i - \Delta_j - \varepsilon)/\tau \right), \quad (5.27) \]

where \( \Theta \) denotes the Heaviside step function \((\Theta(x) = 1 \text{ if } x \geq 0 \text{ and } \Theta(x) = 0 \text{ otherwise})\). In this case, the input features \( \Phi = \Delta \) and \( R = N - 1 \). Based on Eqs. 5.8 and 5.27 we derive the expressions for \( M \) and \( M^\circ \). For compactness, we define the matrix of relative delays \( D_{ij} = (\varepsilon + \Delta_j - \Delta_i) \), with \( \Delta_N = 0 \) corresponding to the delay of the reference spike train. Thus,

\[ M_{ii} = \frac{\mu}{N} \left[ -\mu(1 - w_i)^{\mu-1} \left[ \tau r^2 \sum_j w_j + r \sum_{j \neq i} w_j \Theta(D_{ij}) \exp \left( -\frac{D_{ij}}{\tau} \right) + rw_i \right] 
- \alpha \mu w_i^{\mu-1} \left[ \tau r^2 \sum_j w_j + r \sum_{j \neq i} w_j \Theta(-D_{ij}) \exp \left( \frac{D_{ij}}{\tau} \right) \right] 
+ \tau r^2 \left[ (1 - w_i)^\mu - \alpha w_i^\mu \right] + (1 - w_i)^\mu r \right] \quad i = 1 \ldots N \]

\[ M_{ij} = \frac{\mu}{N} \left[ \tau r^2 \left[ (1 - w_i)^\mu - \alpha w_i^\mu \right] - \alpha \mu w_i^\mu r \Theta(D_{ij}) \exp \left( \frac{D_{ij}}{\tau} \right) \right] 
+ (1 - w_i)^\mu r \Theta(D_{ij}) \exp \left( -\frac{D_{ij}}{\tau} \right) \quad i, j = 1 \ldots N, \ i \neq j \]

\[ M^\circ_{ii} = \frac{\mu}{N} \left[ (1 - w_i)^\mu \left[ \frac{r}{\tau} \sum_{j \neq i} w_j \Theta(D_{ij}) \exp \left( -\frac{D_{ij}}{\tau} \right) \right] 
- \alpha w_i^\mu \left[ -\frac{r}{\tau} \sum_{j \neq i} w_j \Theta(-D_{ij}) \exp \left( \frac{D_{ij}}{\tau} \right) \right] \right] \quad i = 1 \ldots N - 1 \]

\[ M^\circ_{ij} = \frac{\mu}{N} \left[ (1 - w_i)^\mu \left[ -\frac{r}{\tau} w_j \Theta(D_{ij}) \exp \left( -\frac{D_{ij}}{\tau} \right) \right] 
- \alpha w_i^\mu \left[ \frac{r}{\tau} w_j \Theta(-D_{ij}) \exp \left( \frac{D_{ij}}{\tau} \right) \right] \right] \quad i = 1 \ldots N, \ j = 1 \ldots N - 1, \ i \neq j \]

112
where $\Theta$ denotes the Heaviside step function, and all sums are taken over the $N$ weights. The expressions are evaluated at the synaptic steady state $w^*$, which is obtained by numerically simulating the learning equations for all $N$ synapses.
Chapter 6

Summary and Discussion

This thesis studies the relation between structure and function in neural systems. The notion that the function of the system emerges from its structural organization is at the basis of neuroscientific research. Studying this relation in biological nervous systems is very difficult due to the extreme complexity of the systems, and the variety of technical barriers involved. We therefore chose to investigate the same question within the more tractable framework of artificial neural networks evolved to successfully control behaving autonomous agents (EAAs).

We first demonstrate that EAAs are indeed an interesting and relevant model, which is on one hand unconstrained and complex enough to be interesting, while on the other hand is fully accessible to the researcher, allowing “experiments” that are impossible to perform on a biological system. We then developed the Functional Contribution Analysis (FCA), which enables one to measure the contribution of the units (structural level), to any specified task performed by the system (functional level), as well as the degree of function localization in the network. Although developed and studied in EAAs, this method is applicable to biological systems, and we believe will become an important tool in analyzing biological data. The last chapter is devoted to the study of Temporally Asymmetric Hebbian Learning, with the future aim of incorporating this learning mechanism into the evolved networks. EAAs provide a natural framework to study unsupervised learning, defining the goal of learning through the fitness function. In addition, the FCA may prove a powerful tool in better understanding the dynamics of learning, and how it shapes the network interactions and function distribution to achieve the behavioral goal.

The main objective of the first part of the results (Chapter 2) is positioning EAAs as a viable neuroscience research tool, thus substantiating its later use in the development of the FCA. Non-trivial structure-to-function relations were uncovered in EAAs using classical neuroscience analysis tools such as receptive field mapping. The EAAs in these experiments had no predefined network architecture. Rather, they were shaped by evo-
olution in reaction to requirements imposed by the behavioral task. In order to succeed in their behavioral task, the agents developed a mechanism for switching between two distinct types of behaviors – grazing and exploration. In all four experimental scenarios examined, the evolved agents managed to closely match the best memory-less algorithms for the task, and in cases characterized by limited sensory input surpassed them by far.

Two types of evolved agents, differing in the sensory input available to them, were analyzed in detail. In both cases, a similar mechanism has evolved, whereby a command neuron modulates the dynamics of the whole network and switches between grazing and exploration behaviors. Agents which were deprived of sensory position information evolved a memory mechanism, which became the basis for place-sensitivity of the command neuron. A two-parameter stochastic model for the memory mechanism demonstrated that evolution fine-tuned this mechanism toward near-optimal parameters, using the inherent environmental noise and taking advantage of features of the environment that are a-priory harmful, such as the distribution of poison in the arena. An analysis of the neurocontroller of such an agent demonstrated that the command neuron essentially switches the dynamics between two basic input-output networks residing within the same network.

In nature, command neuron mechanisms are known to exist in many invertebrates, including crayfish [Edwards et al., 1999], Clione [Panchin et al., 1996], crabs [Norris et al., 1994, DiCaprio, 1990], Aplysia [Xin et al., 1996b, Xin et al., 1996a, Gamkrelidze et al., 1995, Nagahama et al., 1994, Teyke et al., 1990], and lobsters [Combes et al., 1999]. Command neuron activity has been shown to control a variety of motor repertoires, and to induce different activity patterns in the same neural structures by modulating the activity of other neurons in a pattern-generating network [Combes et al., 1999, DiCaprio, 1990]. The EAA models presented here, although simple, give a computational insight into command neuron mechanisms. For the first time, a concrete model is presented, in which a single command neuron can switch the dynamics of a small neural network between two markedly different behavioral modes. This is achieved by dynamically setting the biases of other neurons in the network, thus effectively multiplexing two networks within the same set of neurons. As we have shown, this structure does not have to be hand-crafted – it evolves spontaneously in a variety of scenarios, proving to be a robust computational mechanism. The dynamics of command neuron switching in some of the models was based on a simple form of a stochastic memory mechanism.

These results show that EAAs have the potential to become an important modeling tool, as well as a useful test-bed for novel analysis methods to study neural systems. This claim is based on two key features of the methodology. First, EAAs present neural structures that are crafted by forces which resemble, in a simplified manner, the forces forming natural nervous systems. This is in contrast to most artificial neural network models, which reflect synthetic human notions about computation. Second, they provide
an excellent vehicle for analysis. On the one hand, being embodied in agents with sensory-motor capabilities, they can be subjected to standard neuroscience analysis methods. On the other hand, and unlike natural nervous systems, they are free from any restriction in terms of the information that can be gathered about network dynamics.

The second and third parts of the results (Chapters 3 and 4) present and explore a novel approach for employing lesion analysis to address the fundamental challenge of localizing functions in a neural system. The approach was developed and thoroughly examined in EAAs. We described the Functional Contribution Analysis (FCA) which assigns contribution values to the elements of the network such that the ability to predict the network’s performance in response to multi-lesions is maximized. We have demonstrated that the FCA portrays a stable set of neuronal contributions and accurate multi-lesion predictions, which are significantly better than those obtained based on the classical single lesion approach. The FCA was also utilized for a detailed synaptic analysis of the neurocontroller connectivity network, delineating its main functional backbone. Further, we have presented a generalization of the basic FCA to high-dimensional analysis, using high-order compound elements. Such elements are composed of conjunctions of simple elements. Their usage enables the explicit treatment of sets of neurons or synapses whose contributions are interdependent, a prerequisite for localizing the function of complex neuronal networks. High-Dimensional FCA was shown to significantly improve the accuracy of the basic analysis, and to provide new insights concerning the main subsets of simple elements in the network that interact in a complex non-linear manner.

An important aspect of the FCA is that it enables one to freely define the elements studied, and the way a lesion effects each element (the “whom” and the “what”). Re. the “whom”, it is possible to study the contributions of synapses, neurons, clusters of neurons, as well as specific parameters of the neuronal activities (e.g. the decay constant of a neuron or the parameters of a learning rule). Re. the “what”, the researcher is free to choose the type of the lesioning, as well as its degree, e.g. the degree of disruption of the unit’s output [Keinan et al., 2003b]. This is made possible in the FCA because the configuration vector is defined as a general mask, which simply indicates which elements are lesioned. In the current analysis, an element can be either lesioned or left intact. This can be easily extended to allow continuous configuration vectors, so that a continuum of lesioning possibilities is considered.

Identifying the importance of elements of a system to varying tasks is often used as a basis for claims concerning the degree of localization versus distribution of computation in that system (see, for example [Wu et al., 1994]). The distributed representation approach hypothesizes that computation emerges from the joint interaction between numerous simple elements [McClelland et al., 1986, Hopfield, 1982]. The local representation hypothesis suggests that activity in single neurons represents specific and well-defined con-
cepts (the grandmother cell notion) or performs specific computations (see [Barlow et al., 1972]). This fundamental question of distributed versus localized computation in nervous systems has attracted ample attention, from Lashley’s seminal paper [Lashley, 1929] to today’s ongoing controversies [Downing et al., 2001, Haxby et al., 2001, Cohen and Tong, 2001]. However, there seems to be a lack of a precise and rigorous definition for these terms [Thorpe, 1995]. The ability of the FCA to quantify the contribution of elements to tasks enables a precise definition of the distribution of processing within the network.

Importantly, the analysis attempts applied to animat systems, raise interesting new questions and give rise to new insights concerning the basic concepts which guide conventional thinking in neuroscience about animate neural processing. As we have shown explicitly in Chapter 3, employing single lesions to analyze a neural system is significantly lacking, and may be misleading. Moreover, our results testify to the crucial importance of selecting the lesioning method. Specifically, they demonstrate that the classical ablation lesioning conventionally used in neuroscience, that is, completely silencing the lesioned unit(s), might be questionable. Using lesioning experiments to decipher even simple neural systems requires a more rigorous multi-lesion analysis, such as offered by the FCA. Finally, the results and discussions presented in Chapters 3 and 4, demonstrate that the concept of contribution of a unit is a delicate issue, and may be inadequate in cases. For example, in systems manifesting “paradoxical lesions”, it is unclear what the contribution of an element is, because it depends in general on the state of the rest of the system. However, as we show in Chapter 4, it is possible to define a stable set of contributions (i.e. independent of the state of the rest of the system), if the system is decomposed into compound elements rather than the original units. Thus, to clearly discuss the contributions of the system’s elements to its functions, it is crucial to find the correct functional decomposition.

The ability to work in an EAA environment greatly enhances the development of neuroanalysis algorithms like FCA and High-Dimensional FCA. Simply, one needs access to the full information characterizing such embodied neurocontrollers in order to develop these methods (see [Ruppin, 2002] for a detailed discussion). However, the FCA has potential significance beyond the scope of EAAs. Multi-lesion analysis algorithms like the FCA can become a useful tool in neuroscience for the analysis of reversible inactivation experiments, combining reversible neural cooling deactivation with behavioral testing of animals [Lomber, 1999]. Indeed, they have now already been successfully applied to such data [Keinan et al., 2003a], revealing the underlying “paradoxical interactions”. As discussed in Chapter 3, ablation lesions may result in excessive perturbations of the system and fail to reveal its normal operation (the problematic nature of analyzing lesioning data due to this excessive lesioning has also been discussed in [Young et al., 2000]). If that will turn out to be the case, then other lesioning methods that cause smaller perturbations
(much like stochastic lesioning) should be employed. Other possible applications include the analysis of transcranial magnetic stimulation studies which aim to induce multiple transient lesions and study their cognitive effects (see [Walsh and Cowey, 2000] for a review). Such applications should prove useful in obtaining insights to the organization of natural nervous systems and settling the long-lasting debate about local versus distributed computation in animate systems.

In summary, the FCA provides a quantitative way of measuring how functions are localized and distributed among network elements, yielding useful insights into the organization of both animat and animate neural systems. Nevertheless, these are merely the first steps in our quest for understanding the operation of neurocontrollers. The way in which information is represented, encoded and manipulated on the neural, assembly and network levels remains an open question.

One of the major simplification made in the EAA models presented here is the fact that the synaptic matrix of each neurocontroller is determined solely by evolution, and fixed throughout the life of the agent. In real nervous systems, evolution shapes the basic neural templates, but the network continuously evolves throughout the life of the organism through learning and synaptic plasticity. The interplay between these two levels of neural development is one of the open questions in our understanding of how neural structures are formed. To study this question within the EAA framework, suitable learning rules should be incorporated into the scheme. A basic problem in this process is the balance between stability and sensitivity of the learning process. That question is tackled in the fourth part of the results (Chapter 5). Approaching the problem of balancing stability against sensitivity, we combine the empirically measured temporally asymmetric spike-timing dependent Hebbian learning rules with a novel generalized model of synaptic updating functions. These updating functions implement additional scaling factors for synaptic modifications, such that the resulting learning rule incorporates an explicit dependence on the current synaptic weights. Using a one-parameter model for the functional form of this weight-dependence, the learning dynamics are stabilized through the attenuation of the scales of synaptic potentiation and depression as the synaptic weights approach the upper or lower boundary, respectively. By varying the weight-dependence parameter, we interpolate between the predominantly used additive model [Abbott and Blum, 1996, Gerstner et al., 1996, Eurich et al., 1999, Kempter et al., 1999, Roberts, 1999, Song et al., 2000] and the multiplicative model [Kistler and van Hemmen, 2000, Rubin et al., 2001].

We studied the learning dynamics in two model systems, namely a linear Poisson neuron, which is accessible to a full analytical study, and a biologically inspired conductance based Integrate-and-Fire neuron. Through the analytical treatment of the linear Poisson neuron in Fokker-Planck mean field theory, the three main constituents of the learning dynamics were identified: a stabilizing drift that results from the weight-dependence of
the updating, a competition term that scales with the total synaptic weight, and a cooperation term that reflects the correlation structure of the spike input. Adjusting the degree of weight-dependence in the plasticity rule allows us to quantitatively understand the interplay between these various drift terms. The crucial importance of the weight-dependence for the learned synaptic distributions was demonstrated in three different input scenarios consisting of uncorrelated input activity, uniformly correlated input activity, and correlated synaptic subgroups. Moreover, we showed that for intermediate weight-dependencies, the learning process dynamically adjusts the balance between depression and potentiation, resulting in stable yet sensitive learning. The analytical findings were reproduced in simulations of the Integrate-and-Fire neuron, demonstrating that the results are strikingly robust with respect to the post-synaptic spike generation mechanism.

We also introduced a novel sensitivity measure for synaptic learning rules that is exclusively based on the learned distributions of synaptic efficacies. Capturing the amount of information about a given input feature that a learning rule is capable of imprinting onto the learned synaptic weights [Bell and Sejnowski, 1995], we compared the performance of temporally asymmetric Hebbian plasticity as a function of the weight-dependence and the properties of the neuronal activity that drives the learning process. Using two different input scenarios, we showed that for a given input statistics, the learning process can indeed be optimized by adjusting the degree of weight-dependence implemented by the updating functions. The proposed performance measure for learning is based on the learned synaptic weights, because defining a more global goal of unsupervised learning in neural networks is a difficult challenge. EAAs offer an opportunity for a natural environment to study the performance of different learning rules.

The novel non-linear TAH rule is particularly fit for evolutionary studies as the crucial weight-dependence of the learning is incorporated through a single parameter. As we have demonstrated, the optimal value of this parameter depends on the statistics of the inputs, i.e. on the nature of the environment and task. Thus, it is a strong candidate for evolutionary optimization, and may serve as an interesting model to study the interaction between learning and evolution (see [Nolfi and Floreano, 1999]). In addition, since the functional organization of networks endowed with synaptic plasticity is dynamic, the FCA may prove a powerful tool in understanding how learning effects the network interactions to achieve the behavioral goal. For example, one can follow the contributions of the units, as well as the identity of compound elements, throughout the learning process. Such studies may reveal the importance of the learning process in shaping the functional organization of the network, a crucial step in understanding the relation between the structure and function in neural systems. It remains for future studies to incorporate unsupervised learning into EAAs, and employ the FCA for their analysis.
Bibliography


[Gütig et al., 2001] Robert Gütig, Ranit Aharonov, Stefan Rotter, Ad Aertsen, and HaimSompolinsky. Generalized synaptic updating in temporally asymmetric hebbian learn-


Y. Xin, K. R. Weiss, and I. Kupfermann. A pair of identified interneurons in Aplysia that are involved in multiple behaviors are necessary and sufficient for the arterial-shortening component of a local withdrawal reflex. *Journal of Neuroscience*, 16(14):4518–4528, 1996.

