A THEORY OF INFORMATION PROCESSING FOR ADAPTIVE SYSTEMS: INSPIRATION FROM BIOLOGY, FORMAL ANALYSIS AND APPLICATION TO ARTIFICIAL SYSTEMS

A DISSERTATION SUBMITTED
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DOCTORAL DEGREE

Departamento de Ingeniería Informática
Escuela Politécnica Superior

UAM

UNIVERSIDAD AUTÓNOMA DE MADRID

Manuel Antonio Sánchez-Montañés Isla
June 2003

Advisers: Fernando J. Corbacho Abelaira and Juan A. Sigüenza Pizarro

Acknowledgements

This thesis is the result of tons of enthusiasm and effort. I hope you enjoy reading it.
I would like to thank Javier De Felipe, Federico Morán, Ricardo Martínez Murillo, Pablo Varona, Enrique Muro and Juan Alberto Sigüenza whose advice and help has been decisive for the beginning of this journey.
I am very grateful to Fernando Corbacho whose help and dedication has been crucial for this thesis. Many of the ideas presented here have arisen from our endless discussions about the brain.
Thanks to Paul Verschure and Peter König from the Institute of Neuroinformatics. I learned a lot from them during my visits to INI which I really enjoyed (in spite of Paul's strange music CDs).
It is also a pleasure to thank Tim Pearce for his advice, support and help from which this thesis has greatly benefited.
The help and contributions of many people were also decisive for this thesis. I would like to mention specially Francisco Rodríguez, Ramón Huerta, Luis F. Lago-Fernández and Pablo Varona.
Thanks to my family and my very good friends Nazareth Castellanos and Raúl Medina for all their love, support and help.
I would like to acknowledge my lab colleagues Philip, Luis, Marta, David, Naza, Víctor, Luis and the rest of colleagues at the department for all the good moments I have spent here. Thanks to Juana for all the patience she has had with me.
This thesis was supported by grants from MEC and MCyT.
Finally, I would like to acknowledge the music of Kraftwerk and Jarre for showing me that doing science can be very exciting and funny !

Resumen

En esta tesis se presenta una medida formal del procesamiento de información que realiza un sistema adaptativo. Dicha medida posibilita el desarrollo de un nuevo marco teórico que permite tanto el análisis de sistemas complejos adaptativos existentes, como el diseño de nuevos sistemas artificiales complejos. La medida introducida no depende de la implementación ya que depende de las estadísticas globales de la interacción del sistema con su entorno.

El marco teórico desarrollado puede ser usado para analizar cómo los sistemas perceptuales biológicos construyen representaciones internas óptimas de su entorno. Adicionalmente el marco teórico propuesto permite realizar predicciones nuevas y puede ayudar a comprender cómo los diferentes sistemas perceptivos realizan el análisis de su entorno.

Desde un punto de vista más analítico el marco teórico propuesto permite la construcción de un mapa con diferentes algoritmos existentes en computación neuronal y en sistemas artificiales de aprendizaje. Por ejemplo PCA, análisis discriminante de Fisher, C4.5, etc. De esta forma el marco teórico puede ayudarnos a clarificar las analogías y funcionalidades diferentes de los distintos algoritmos.

Finalmente el marco teórico expuesto da pie a la creación de potentes algoritmos para el diseño de sistemas adaptativos complejos. Por ejemplo presentamos un nuevo método de construcción de árboles de decisión obtenido de la teoría desarrollada que combina varias características de métodos ya existentes. También presentamos un nuevo algoritmo para la extracción de características no lineales en problemas de clasificación.

Abstract

This thesis presents a formal framework based on a new information processing measure that allows both the analysis and the design of complex adaptive systems. The introduced processing measure depends on both the level of complexity of the internal representation of the system and on the task it must perform. This measure is implementation independent since it depends on the global statistics of the interaction between the adaptive system and its environment.

The framework can be used to analyze how different biological perceptual systems construct optimal internal representations of the environment. Additionally the proposed framework allows new predictions and may shed a new light on how the different perceptual systems perform the analysis of their environment. From a more analytical point of view the proposed framework allows the construction of a map of several of the existing algorithms in neural computation and machine learning, for instance, PCA, Fisher discrimination analysis, C4.5, etc. and can help to elucidate their common analogies and different functionalities. Finally the proposed framework gives rise to the construction of new powerful algorithms for the design of adaptive complex systems. For instance we present a new learning method for decision tree construction derived from our framework which contains several features of existing methods. We also present a new algorithm for constructing optimal nonlinear representations (non linear feature extraction) in classification problems.

Contents

Part I

Motivation and Introduction to the problem

Chapter 1

Introduction

1.1 The biological systems as efficient adaptive systems

The thesis starts with a theoretical study of biological systems since they are very efficient in their interaction with their environment. In order to perform optimally they must construct adequate internal representations of the complex sensory information they receive [Barlow, 1961, Atick, 1992]. The construction of an efficient information representation has several advantages. First, *lower computational and energetic cost*: natural stimuli come in a highly inefficient form since they tend to possess statistical regularities. For example in natural images near pixels are very correlated in space, time and color [Ruderman, 1994] so the representation formed by the global activity of the photoreceptors is highly inefficient. As a clear example of this inefficiency just consider the high compression rates of audio-visual data achieved by MPEG, usually around a 30:1 ratio [Furht, 1998]. Thus a recoding strategy of these signals into less redundant codes makes the subsequent processing of these signals simpler and less energy consuming [Attneave, 1954, Barlow, 1961, Atick, 1992, Baddeley, 1996]. Second, the internal representation might have dramatic implications to an animal's *ability to learn the relations between the elements* in the environment [Barlow, 1989]. For instance all the aspects of a natural image are the result of the objects present

in the scene. The ability of the animal to learn functional relations between the objects depends crucially on its ability to represent the objects as independent entities. Third, the internal representation is critical for the *generalization abilities* of the system and must depend on the task. Some variations (e.g., distance to an object and thus its retinal size) are not important for one aspect (e.g., recognizing its identity) and thus represent noise but can be all decisive for other aspects (e.g., grasping it). Thus, the system must be able to represent the information appropriately depending on the task at hand. Therefore a study of principles of organization in those systems will provide us with useful insights about what the principles of organization of an optimal adaptive system should be.

Over the last decades the amount of experimental data about the nervous system has rapidly increased. This has allowed the emergence of theoretical neuroscience, a branch which studies the functioning of neural systems using computer and theoretical models. There are many aspects of neural systems which have not been experimentally addressed yet. Nevertheless, theoretical models can be useful and lead to testable predictions if the level of description of the model fits the experiments it tries to explain. For a review of the different types of models of neural systems see [Koch & Segev, 1998] and [Arbib, 1998].

We start the thesis analyzing the representation of information at the olfactory epithelium. This structure is composed by millions of olfactory receptor neurons, which represent the first stage of processing in the olfactory system [Kandel *et al.*, 2000]. The representation of the information in this structure is thus critical for an optimal performance of the system. This is specially important here, since over 10000 odorous compounds are known to exist in nature, and these can occur in multitude of combinations [Pearce *et al.*, 2002]. However, most olfactory receptors have unspecific receptive fields, responding to a large variety of chemical compounds [Sicard & Holley, 1984]. The question of whether this unspecificity results from some physical constraint placed upon chemical transduction, or on the contrary it is beneficial to system performance, is unclear.

We show that the neural configuration which maximizes the information transfer in a simple model of the olfactory epithelium has very similar properties to the real

system: the receptive fields show bipolarity, unspecificity, and homogeneous distribution. Thus we can say that the unspecificity of the receptive fields is beneficial for the performance of the system, optimizing the information transfer to higher steps. However the receptive field of each receptor is determined by the genetic processes that occurs when it is formed [Reed, 2000]. Therefore the representation at the olfactory epithelium is genetically encoded and does not result as a consequence of learning or experience. Then we proceed to study the representation of the information in auditory and visual cortices where plasticity and learning occurs and the internal representation is more ellaborated.

Over the past years neuroscientists have gained insight in the neural mechanisms responsible for the ability of learning and adaptation in biological systems (for a review see for example [Alkon *et al.*, 1991, Buonomano & Merzenich, 1998]). A number of models of learning have been proposed with different desirable properties [Sejnowski, 1977, Stent, 1973, Bienenstock *et al.*, 1982, Brown & Chattarji, 1998, Fregnac, 1998]. However recent physiological results on neurons in cortex give a richer picture where temporal relationships at the millisecond scale between the signals a neuron receives are critical for the plasticity dynamics [Markram *et al.*, 1997, Zhang *et al.*, 1998, Bi & Poo, 1998]. Since these mechanisms can be critical for understanding the emergence of internal representations in the cortex, we introduce these mechanisms in realistic cortical models.

As it was shown, the internal representations formed in the computer models are very similar to the biological systems, and emerge as a consequence of the cooperation and competition at several levels. The receptive fields formed by these mechanisms give rise to neurons with more specific receptive fields than those seen in the olfactory model. Near neurons in space tend to code the same feature while far neurons tend to code different aspects of the information. Thus the information is represented by *functional groups of neurons* (for related concepts see for example [Hebb, 1949, Abeles, 1991, Tononi & Edelman, 1998]). Due to this specificity, the internal representations in the auditory and visual cortex are *sparse codings*, that is, the stimulus is represented by only a few active cells out of a potentially much higher number [Olshausen & Field, 1996, Baddeley, 1996] as in contrast with the olfactory

system. This strategy is efficient in the sense that it minimizes the complexity and energetic cost of the code while maximizing the representational accuracy for natural images [Olshausen & Field, 1996, Baddeley, 1996].

1.2 Derivation of general principles from the biological system analysis

The above considerations suggest that if a general framework of information processing exists then it should be valid for describing any complex system with an arbitrary internal code and arbitrary implementation details. This suggests that the general theory should be expressed in terms of the internal states of the system and not on physical parameters related to the specific implementation of the system. Therefore our framework should be *implementation independent.* On the other hand the notion of *low complexity of the representation* is a key ingredient in the cortical representations of the visual and auditory cortices [Barlow, 1989, Olshausen & Field, 1996, Sánchez-Montanés *et al.*, 2002]. Thus the biological systems reduce the intrinsic complexity of the sensory input by constructing higher level representations of the information (e.g. "edge detectors" in the visual cortex versus "pixel detectors" in the array of photoreceptors at the retina [Bear *et al.*, 1996]). This is intuitive since the ultimate goal of the animal is to solve the tasks the environment imposes and thus an efficient representation of the information which captures the regularities of the environment is critical [Barlow, 1989]. Thus the concept of *complexity reduction* seems to be another basic ingredient in our desired theory. Finally, the experimental observations in the auditory cortex and our theoretical analysis show that the representation of information in that structure is modulated by experience as the behavioral importance of the stimuli change in time. Thus the internal representation is biased to behaviorally important stimuli. These experiments reveal the important principle that internal representations, even at primary levels of processing, are influenced by the tasks the environment imposes on the animal. This seems very reasonable since the resources that a biological system has

are limited. The animal should focus on that part of the information which is really relevant for the task due to the high complexity of the environment and the large amount of different stimuli it receives.

Apart from these considerations about the limitation of resources, it is critical for a proper generalization to focus on the relevant part of the information neglecting its spurious and noisy aspects. These considerations conducts us to the principle that the concept of task should be a crucial element in a general theory of information processing, that is, the new information processing measure should be *task-dependent.*

1.3 The general framework

The aim of this thesis is to provide a new formal framework that allows both the analysis of existing adaptive complex systems as well as the design of artificial complex systems. In this context information theory quantifies the performance of such systems as a function of their global statistical properties but not their implementation details [Cover & Thomas, 1991]. Thus maximization of mutual information and related concepts such as Fisher Information seem very appropriate to describe the global properties of sensory systems, becoming very successful in the description of some specific neural systems [Atick, 1992, Deco & Obradovic, 1996, Borst & Theunissen, 1999, Dayan & Abbott, 2001]. However, we will show that this approach is not valid for describing the properties of higher level representations in some systems such as the auditory cortex. This and some other general conclusions obtained from our study of the biological systems lead us to propose a general framework of information processing in adaptive systems.

Given these conclusions and the ones obtained in section 1.2 we derive a general mathematical framework which contains the essence of these principles. In order to make it as general as possible we use the notions of *agent* and *environment* valid for any system that interacts with its environment. We introduce the concept of the *amount of effective information processing* (ΔP) performed by a part $\mathcal{W}$ of the agent which has to solve the task imposed by the environment. This notion does not depend on the specific implementation but on the global statistical relations between

that part of the agent and the environment. Then the crucial notion of distance to the task emerges, which depends on both the level of uncertainty and complexity that the information processed by $\mathcal{W}$ has with respect to the task. Since we express this theory using the general notions of agent and environment and the mathematical tools the theory uses are implementation-independent, the framework can be used to analyze the interaction of any complex adaptive system with its environment (whether biological or artificial) as well as to obtain new optimal learning strategies for artificial systems.

1.4 Validation of our theory with the biological models

Since biological systems are very efficient interacting with their environment we would expect them to build internal representations which maximize the amount of effective information processing (ΔP). As we show the theoretical configuration which maximizes ΔP in the auditory cortex has very similar properties to the biological system. On one hand, it predicts that neurons in auditory cortex should respond to specific frequency bands. On the other hand, it predicts that stimuli are internally represented by an amount of resources proportional to their behavioral importance but not to their probability of occurrence. These properties derived from a theoretical analysis are very similar to experimental observations [Weinberger, 1993, Kilgard & Merzenich, 1998].

The theoretical configuration which maximizes ΔP in the olfactory epithelium is composed by a repertoire of neurons showing maximum diversity in their pattern of responses, bipolar sensitivities and unspecific receptive fields. All these properties have been reported by experimenters [Sicard & Holley, 1984, Schild & Restrepo, 1998, Sanhueza *et al.*, 2000].

Finally, in preliminary work we have obtained that maximizing ΔP in a simplified model of the retina [Atick & Redlich, 1990] we obtain similar properties to ganglion cells. Therefore we conclude that ΔP seems to be maximized in biological systems,

which constitutes a validation of our theory and demonstrates the potentiality of our framework to study and understand biological systems.

1.5 Known machine learning algorithms as particular solutions in the framework

There is an enormous variety of machine learning techniques useful for different problems (for a review see [Mitchell, 1997]). For example, a usual classification of machine learning algorithms divides them in supervised, unsupervised and reinforcement learning techniques [Mitchell, 1997, Sutton & Barto, 1998]. This thesis shows how within the same framework both supervised and unsupervised learning algorithms emerge. Additionally, optimal internal representations for reinforcement learning techniques can also be derived within this framework.

Specifically we demonstrate how our framework can be used to obtain the optimal learning algorithm for an autonomous artificial system under different conditions. For instance if a noisy linear system processes a gaussian signal in order to transmit as much information as possible about it, then principal component analysis emerges as one of the solutions which maximize ΔP. On the other hand, if the task is to classify the signal in different classes then Fisher discriminant analysis [Duda & Hart, 1973] arises as the optimal solution when there is high overlapping between classes and the statistics are well represented by gaussians. Classical learning algorithms for tree construction are also obtained by maximization of ΔP in classification problems. For example the basic algorithm of C4.5 [Mitchell, 1997] is obtained as a special case when the complexity of the internal representation is not taken into account.

Thus our theory can serve as a unifying framework which allows to create a map of different machine learning techniques which can help us to elucidate their fundamental analogies and differences.

1.6 Utility of the framework for obtaining new learning schemes for artificial systems

Finally we demonstrate the utility of the framework for developing new optimal learning schemes. For example we show that the principle of ΔP applied to decision trees construction induces a new learning algorithm which combines the good features of known methods such as *information gain* and *gain ratio* [Mitchell, 1997]. It also shows a natural ability of *early stopping* and as a particular case contains the [LópezdeMántaras, 1991] distance for attribute selection. On the other hand when the principle of ΔP is used in a nonlinear layer of neurons for a classification task, a nonlinear feature extraction algorithm emerges which shows very interesting features. The amount of resources used by the system is automatically adjusted to the required precision and the complexity of the problem, providing a useful strategy for avoiding overfitting. Thus the algorithm selects the more efficient representation for a given accuracy. The algorithm also shows a natural tendency to maximize the margins of the decision frontiers which provides a natural link with support vector machines [Vapnik, 1998].

1.7 General sketch of the thesis and methodology

The thesis starts with the study of the internal representation of information in biological sensory systems. On one hand, we use abstract models which try to capture the global properties of the system. These models are mathematically simple but allow us to make concrete testable predictions about the system. We analyze them using information theoretical measures which describe the performance of the system based on its global statistics and not on implementation specific details. On the other hand, we use biologically realistic models including specific details about the dynamics of the neurons. These models will allow us to understand more specific questions such as how the internal representations are developed.

From the analysis of both kinds of models we obtain a set of desirable properties that a general framework to analyze and design complex systems should have.

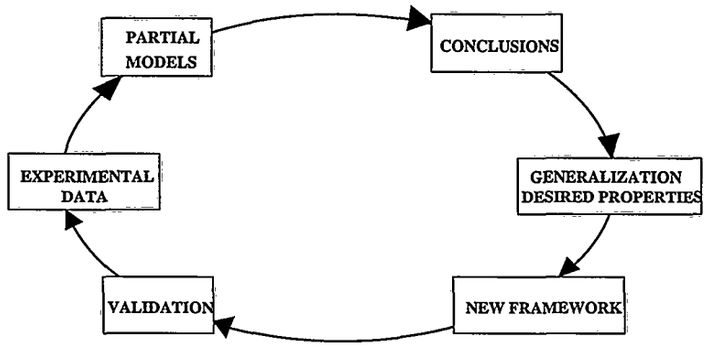

Figure 1.1: General scheme of the thesis

Then we derive a theoretical framework which satisfies the required properties and is expressed using mathematical tools which are implementation independent. This allows the generality of the theory and its application in other contexts such as machine learning problems.

Finally, the framework is validated in two different directions. First, we show that the application of the theory to artificial problems leads to the emergence of very well known algorithms of machine learning, thus demonstrating the validity of the theory in these problems. Moreover, it allows to obtain new optimal learning algorithms from which two different examples are shown in the thesis. Second, we show that, when applied to biological systems, our theory predicts an optimal configuration with very similar properties to the biological system. This provides another validation of our theory. Finally we discuss how this framework can be used in the design and analysis of new biological experiments.

Part II

Study of neural biological systems

Chapter 2

Global properties of optimal internal representations: The olfactory system

2.1 Context

In this chapter we will analyze the representation of information at the olfactory epithelium. This structure is composed by the olfactory receptor neurons (ORNs), which represent the first stage of processing in the olfactory system [Kandel *et al.*, 1991]. The representation of the information in this structure is thus critical for an optimal performance of the system. This is specially important here, since over 10000 odorous compounds are known to exist in nature, and these can occur in multitude of combinations [Pearce *et al.*, 2002].

We will use tools from information theory which will allow us to analyze global properties of the system without the need of detailed models. Importantly, the concrete tool we use, Fisher Information, can be linked directly with the psychophysical discriminability of the animal between individual stimulus components [Seung & Sompolinsky, 1993, Dayan & Abbott, 2001]. This allows the interpretation of the properties of the internal representations from the behavioral point of view.

Part of the results of this chapter have been presented elsewhere:

[Sánchez-Montanés & Pearce, 2001, Sánchez-Montanés & Pearce, 2002, Pearce & Sánchez-Montanés, 2002].

2.2 Introduction

The activity of the array of olfactory receptors at the olfactory epithelium provides a population coded representation of the odor stimulus. The extent of the broad tuning underlying olfactory perception of general odors in mammals was vividly demonstrated by Sicard and Holley after taking 74 olfactory neurons at random from the olfactory epithelium of the frog and exposing them *in vitro* to a series of single chemical compounds [Sicard & Holley, 1984]. Figure 2.1 shows the degree of unspecific tunings observed across the neuron population, the spot size relating to the spiking frequency produced by the cell in response to a single chemical compound. The results showed conclusively how odor perception in mammals is supported by neurons that have overlapping unspecific sensitivities of varying degrees to groups of compounds, each here with distinct tunings.

Sicard and Holley later concluded from an analysis of the 60 olfactory neuron responses shown, that no neuron pairings within this randomly selected sub-population displayed identical tunings to the odors presented. This seemingly implied a bewildering diversity of olfactory neuron tunings suggesting a lack of order in the encoding of odor information. However the level of receptor protein diversity was later quantified by Buck and co-workers who cloned 18 different members of an extremely large multigene G-protein coupled receptor family, thought to be responsible for the transduction of olfactory stimuli [Buck & Axel, 1991]. They provided an estimate of the number of receptor classes based on the sequence homology of this small sub-population to be between 300–1,000 in mice. It is now clear that the pattern of activation across a number of such olfactory neurons enables the brain to interpret complex molecular stimuli.

Given such enormous diversity of olfactory neuron responses and receptor protein types, combined with their broadly unspecific tunings, a key question that needs to be addressed is how does the detection performance of a sensory system responding

to many input stimuli depend upon the degree of specificity and on the distribution of receptive fields of the underlying receptor population? We see that for the general odor discrimination task nature deploys largely unspecifically tuned olfactory neurons – yet what, if any, performance advantage does this provide? In the case of the olfactory system these are complex issues since the stimulus virtually always comprises large numbers of chemical components.

In this chapter we apply the Fisher information concept to this multi-component chemical stimulus case in order to quantify any performance advantage. Unlike previous studies examining the role of the receptive field widths on detection performance [Dayan & Abbott, 2001, Seung & Sompolinsky, 1993, Pouget *et al.*, 1999, Zhang & Sejnowski, 1999, Wilke & Eurich, 2002] we neither impose a particular shape on the receptive fields nor a homogeneous distribution.

2.3 Basic model of the olfactory epithelium

The olfactory epithelium is composed by a population of noisy sensory neurons which codes the olfactory stimulus as an activity pattern. Each neuron in the model has different parameters which determine its pattern of response to the stimuli (figure 2.2).

By varying these parameters we can study which configuration is optimal. This population sends the information to further processing steps. In order to perform adequately, these higher levels need implicitly or explicitly to make a correct estimation of the odor components of the stimulus. Therefore, a good representation of the information at the sensory neuron population is critical for the goodness of the estimation. As we will show in the following section the mathematical tool we will use, Fisher Information, lets us study the optimality of the sensory neuron population without the need of modeling the further processing steps.

2.3.1 Neuron model

In order to experiment with different receptive field distributions and their effect on the global reconstruction error, we consider a sensory system consisting of an arbitrary population of 100 neurons. The input to the system is a combination of many single odor components (individual chemical compounds). Hence, the dimension of the input space is usually high. We model the input as a vector $\vec{s}$ of which component j is the level of concentration of the single chemical compound j. As such, we consider the stimulus to be multidimensional as opposed to previous work where the stimulus was considered to be scalar [Dayan & Abbott, 2001, Seung & Sompolinsky, 1993, Pouget *et al.*, 1999].

We now model the response of the ith sensory neuron to $\vec{s}$. For simplicity, we approximate this as linear

$$r_i = \vec{a}_i^T \vec{s} + b_i + \eta_i, \tag{2.1}$$

where r_i is the firing rate of the neuron i, $\vec{a}_i$ is the vector of sensitivities of this neuron to the different single chemical compounds (which we call "receptive field"), b_i is its spontaneous firing rate, and η_i is its zero-mean noise. Note that the linear simplification is equivalent to requiring that the firing rate of the neuron is scaled with the stimulus intensity, which is a reasonable assumption for moderate stimulus concentrations. Also note that here we are not imposing anything about the shape of the receptive fields (RFs) themselves, which are determined uniquely by the sensitivity vectors $\vec{a}_i$. Finally, we make the approximation that the noise within each neuron is Gaussian and independent. This is a reasonable assumption since the receptor neurons have no lateral connections at the epithelium level and the noise processes are likely to be local to the neuron.

2.3.2 Fisher information

When a multi-component stimulus $\vec{s}$ is exposed to the system, each neuron k responds with firing rate r_k following some probability distribution $p(r_k|\vec{s})$ defined by its tuning to the stimulus and its intrinsic noise. An optimal estimator that uses the population response $\vec{r}$ for reconstructing the stimulus $\vec{s}$ (figure 2.2) should give the

correct stimulus values on the average over a large number of repeated presentations. That is, the mean estimate for repeated presentations of the same stimulus $\vec{s}$ should be equal to $\vec{s}$. We call to this type of estimator an "unbiased estimator". Moreover, the estimate should be as close as possible to the applied stimulus when the presented stimulus is fixed (minimum variance) [Deneve *et al.*, 1999].

The entries of the Fisher information matrix (FIM), $J_{ij}(\vec{s})$, are defined as [Cover & Thomas, 1991]

$$J_{ij}(\vec{s}) = \int d\vec{r}\, p(\vec{r}|\vec{s}) \left(\frac{\partial}{\partial s_i} \ln p(\vec{r}|\vec{s})\right) \left(\frac{\partial}{\partial s_j} \ln p(\vec{r}|\vec{s})\right). \tag{2.2}$$

Then for every unbiased estimator that uses the population response $\vec{r}$ for reconstructing the stimulus $\vec{s}$

$$\mathrm{var}(\hat{s}_i|\vec{s}) \geq (J^{-1}(\vec{s}))_{ii}, \tag{2.3}$$

where "var" means variance, and $\hat{s}_i$ is the estimation of the component i of $\vec{s}$, $i = 1, \ldots N$ (note that the variance of an unbiased estimator is just its squared error). This well-known result is the "Cramér-Rao bound" [Cover & Thomas, 1991] and limits the performance of any unbiased estimator. An estimator is said to be optimal if its variance is equal to this lower bound. Using eq. 2.3, we can calculate the minimum estimator variance across all of the stimulus components

$$\mathrm{var}(\hat{\vec{s}}|\vec{s}) = \sum_{i=1}^{N} \mathrm{var}(\hat{s}_i|\vec{s}) \geq \sum_{i=1}^{N} (J^{-1}(\vec{s}))_{ii} = tr(J^{-1}), \tag{2.4}$$

so the performance of the best unbiased estimator that can be built is defined by the entries of the FIM, J_{ij}. Importantly, the psychophysical discriminability across a range of individual stimulus components in the animal can be linked directly to the optimal reconstruction error defined by this equation [Seung & Sompolinsky, 1993, Dayan & Abbott, 2001].

Establishing a theoretical limit on the accuracy of a neural code is interesting, but it can be irrelevant if there are no biophysically reasonable schemes for implementing an optimal, or near optimal, decoding method using real neural circuitry

[Abbott & Sejnowski, 1999]. Pouget et al. have shown how maximum likelihood estimation, which achieves the maximum possible accuracy for any unbiased estimator, can be performed using real neural circuits [Pouget *et al.*, 1998], and thus establishes that the theoretical limit corresponding to the Fisher information is achievable [Abbott & Sejnowski, 1999]. Our interest is then in finding the tunings of the population of receptors which minimize the right hand side of eq. 2.4.

2.3.3 Optimization Methods

The FIM of the system can be calculated as [Sánchez-Montanés & Pearce, 2002]:

$$J = \sum_{i=1}^{R} \frac{1}{\sigma^2} \vec{a}_i \vec{a}_i^T, \tag{2.5}$$

where R is the number of receptors in the population. We have assumed that the noise in different neurons have the same variance σ^2. This equation gives the sum of the independent contributions to the Fisher information from each sensor. Our goal is to find the set of receptive fields $\{\vec{a}_1, ..., \vec{a}_R\}$ that minimizes the trace of J^{-1}, which bounds the optimal reconstruction error of the system (eq. 2.4). The free parameters to be optimized are then the individual sensitivities, a_{ij} $(i = 1...R;\ j = 1...N)$ of the population. Therefore, we do not impose any particular distribution on the RFs, as opposed to previous works where it is usually assumed a priori an homogeneous distribution within the population [Seung & Sompolinsky, 1993, Zhang & Sejnowski, 1999, Abbott & Dayan, 1999, Pouget *et al.*, 1999]. Hence we can study which distribution of RFs is optimal in terms of detection performance.

If there are no additional constraints on the system, this function has no global minimum since $\mathrm{tr}(J^{-1}) \to 0$ as $|\vec{a}_i| \to \infty$ and $|J| \neq 0$. Therefore, we should bound our search space in order to find the optimal configuration. This is defined by the physical constraints placed on the system, which we take as $-c \leq a_{ij} \leq c$. This can be interpreted as a system where each neuron can have olfactory receptor proteins interacting with intracellular currents of any type (all excitatory, all inhibitory, or mixed). The biological plausibility of these constraints are discussed in the conclusions section. Because the RFs are linear, the value of c has no effect on the optimal

configuration, so we take $c = 1$. Note that the numerical value of σ^2 is irrelevant for the optimization since it represents a constant factor multiplying the global function (eq. 2.5).

The optimal error is then calculated as a function of the number of single odor components ("input dimension"). The global optimization is done using a standard genetic algorithm [Levine, 1998]. The default parameters are used with the algorithm (see the user manual at [Levine, 1998]).

2.3.4 Optimal theoretical configuration and the real system

Bipolarity and unspecificity of the receptive fields

The results show that the neurons of the optimal system configuration can be influenced by any individual compound (Figure 2.3 A) – no neurons assume zero sensitivity to any of the input dimensions.

Note that some stimuli can excite the neuron (positive sensitivities) while others can inhibit it (negative sensitivities). In any case, the sensitivities have maximum gain, which is 1 within the constraints imposed. The distribution of RFs across the population shows an exact Poisson distribution (Figure 2.3 B). Therefore, the probability for an arbitrary sensitivity to be 1 or -1 is independent on the values of the other sensitivities for each neuron, and is 0.5. This leads to the system having a mixture of all kinds of receptive fields (Figure 2.3 B). However, through chance, RFs with similar numbers of positive and negative sensitivities are more common than RFs with a dominating sign in the sensitivities (Figure 2.3 B). Interestingly, this configuration can be easely constructed with a local stochastic process which selects each sensitivity to be 1 or -1 with the same probability.

The question of how much better is this unspecific configuration compared to a specific configuration is addressed in figure 2.4. In the specific configuration each neuron is assumed to respond to only one single odor component with maximum gain; that is, all the sensitivities are null but one. In this case each single chemical component is assumed to be detected by an equal number of specific neurons. This can be shown to be the configuration which minimizes the optimal estimation error (eq.

2.4) given that the receptive fields are specific. Figure 2.4a shows that the unspecific configuration is much better than the specific one, and this difference increases linearly with the input dimension (Figure 2.4b).

Distribution of the receptive fields

The number of potentially different receptive fields, assuming that each sensitivity can be arbitrarily either 1 or -1, is shown in Figure 2.5 A (dashed line) as a function of the input dimension. The optimal system configuration reaches this limit with low input dimensions (Figure 2.5 A). For higher input dimensions, the number of potentially distinct RF configurations is greater than the number of neurons. In this case the number of RFs saturates the maximum allowable vertices of the search space reachable by the system (Figure 2.5 A). Therefore, there is maximum diversity in the RFs configuration for every input dimension. Moreover, these different RFs are homogeneously distributed across the population (each different configuration is used by the same number of neurons as the others), see Figure 2.5 B.

2.4 Expanded model of the olfactory epithelium

In section 2.3 we have shown that the optimal theoretical configuration is a mixture of receptive fields of greatest possible diversity, and their distribution across the population follows the maximum homogeneity principle. We showed that this configuration can be easily constructed by a hypothetical local process to each neuron which selects randomly binary sensitivities for each single chemical compound.

However ORNs do not have the degrees of freedom necessary to independently select their tunings to each stimulus dimension in this way. When an ORN is formed it selects at most one g-protein coupled receptor gene for expression from a superfamily of approx 320 (in humans). Each one of these genes determines a characteristic receptive field [Reed, 2000]. Therefore, the ORNs must select between fixed sets of sensitivities to the complex stimulus by selecting one gene from the superfamily, which fixes the tuning to the universe of possible chemical stimuli. Interestingly, early

experimental studies using PCR methods clearly demonstrate an homogeneous distribution of ORNs expressing a given receptor gene within each of the four zones of the olfactory epithelium [Chess *et al.*, 1994]. These and other more recent experiments have lead to the conclusion that the selection of ORNs is controlled locally and is stochastic within distinct subsets of the receptor superfamily ([Serizawa *et al.*, 2000]; see [Kratz *et al.*, 2002] and [Mombaerts, 2001] for discussion).

In this section we will investigate how a restriction on the number of different tuning curves affect the results of our model. Concretely, we will analyze how this restriction affects the detection performance of the system and whether the general principles listed above (unspecificity in the RFs, homogeneity in their distribution) do still ensure an optimal information representation.

2.4.1 Mathematical model

The number of different receptor genes (the *gene pool*) that may be expressed in the population is constrained to be equal to a parameter M which we will vary in the simulations. Thus each one of these genes has an associated receptor tuning $\vec{u}_i$. Now if we consider the set of receptor gene tunings to the stimulus within across the gene pool, $\{\vec{u}_1, ..., \vec{u}_M\}$, we can express the Fisher information matrix of the system as

$$J = R \sum_{k=1}^{M} p(k) \frac{1}{\sigma^2} \vec{u}_k \vec{u}_k^T, \tag{2.6}$$

in analogy with equation 2.5. R is the total number of ORNs in the population; $p(k)$ is the fraction of ORNs expressing receptor gene k, hence having a receptive field $\vec{u}_k$, and σ^2 is the noise variance in each receptor neuron.

As in the previous section we want to determine the configuration of the receptive fields that minimizes the optimal estimation error (that is, minimize $tr(J^{-1})$). In order to understand the implications of a finite gene pool and the homogeneous expression of the receptor genes in the biological system we will compare three different situations:

1. There is no constraint on the size of the gene pool nor imposed homogeneity in the RF distribution (*unconstrained pool size* situation). The parameters to

optimize are then the individual sensitivities of the neurons, which corresponds to the basic model studied in section 2.3.

2. There is a constraint in the gene pool size and the percentage of neurons that express a given gene is the same for all the genes (*homogeneous gene expression* situation). This represents the biological situation, where we assume that the evolution has selected an optimal gene pool. The free parameters to optimize in our model are then the individual sensitivities of the pool u_{kj} ($k = 1 \ldots M$, $j = 1 \ldots N$).

3. There is a constraint in the gene pool size but the percentages of neurons that express a given gene can be different (*unconstrained gene expression* situation). The free parameters to optimize in our model are then the individual sensitivities of the pool u_{kj} and the fractions $p(k)$ subject to $\sum_k p(k) = 1$ and $p(k) \geq 0$.

In all the cases the individual sensitivities are constrained to be in the $[-1, 1]$ interval as in the previous section, representing excitatory or inhibitory ORN responses. The global minimization of $tr(J^{-1})$ is done using a standard genetic algorithm [Levine, 1998].

2.4.2 Results with the extended model of the olfactory epithelium

We will first comment the properties of the optimal configuration for the homogeneous gene expression (which corresponds to the biological situation). Analogously to the results of section 2.3 the receptive fields of the optimal theoretical configurations have binary sensitivities (1 or -1) with maximum gain within the imposed constraints (figure 2.6 A). As before we will define the width of a receptive field as the fraction of stimuli to which the neuron responds positively. Then the set of optimal receptive fields associated with the gene pool is a mix of broad and narrowly tuned receptive fields (figure 2.6 B). Since the gene expression is constrained to be homogeneous we have that the receptive fields of the neural population is also a mix of broad and narrowly tuned receptive fields in accordance with the results of section 2.3.4. On

the other hand, the number of different receptive fields in the pool is equal to its maximum value (M) for any pool size (figure 2.6 C). This means that the principle of maximum diversity derived with the basic model is also valid when a restriction in the size of the gene pool exists, which corresponds to the biological situation. How is the performance of this system compared to a system with unrestricted gene expression or with unrestricted pool size ?

The detection performance of the optimal configurations in both homogeneous and unconstrained gene expression cases approach the performance of the solution with no constraint in the pool size (figure 2.7 A). The detection performance of the optimal configuration with homogeneous gene expression is very similar to the solution with unconstrained gene expression, being practically equal for a pool size greater than 200. This is shown in detail in figure 2.7 right, where we plot the relative error in the homogeneous case respect to the error in the unconstrained case $\frac{\epsilon^2_{hom}-\epsilon^2_{unc}}{\epsilon^2_{unc}}$. We see that indeed the optimal error of the homogeneous case converges to that of the optimal unconstrained solution. Thus the detection performance of both solutions is practically identical for biological relevant numbers of gene pool size (*ca.* 320 in humans).

With all these considerations we conclude that the principle of maximum diversity of receptive fields and maximum homogeneity of their expression conducts to a configuration which optimizes the detection performance even if a restriction in the gene pool size exists. Moreover, the performance of the solution with restricted pool size is almost identical to the performance when there is not this restriction.

2.5 Conclusions

We have created a simplified model of the olfactory epithelium. Then we use information theory in order to study the configuration which maximizes the information contained in the receptor population about the stimulus. In concrete we calculate the Fisher information of the internal representation, which through the Cramér-Rao bound limits the performance of any unbiased system

which tries to estimate the real stimulus components from the internal representation [Cover & Thomas, 1991]. Fisher information can also be directly related to the limit of psychophysical discriminability of the animal to different stimuli [Seung & Sompolinsky, 1993, Dayan & Abbott, 2001].

- **Bipolar sensitivities and conferred biological advantage** - It might be expected that permitting bipolar sensitivities in the neurons to each of the stimuli (as opposed to purely excitatory or inhibitory responses) produces the best overall detection performance. As Schild and Restrepo (1998) state

  > "Differential stimulation or suppression of olfactory neurons by odors could be used by the olfactory system as a mechanism for contrast enhancement. In addition, the responses resulting from simultaneous stimulation and inhibition of different neurons by one odorant could be contrasted in the olfactory bulb in such a way that low odorant concentrations could be detected at signal levels that could not be resolved from noise in a purely excitatory system."

  Interestingly, such suppressive or inhibitory effects of particular odor on olfactory sensory neurons has only recently been observed in mammals, whereas it is much more common in amphibians. For this reason the role of inhibitory receptor responses across species is only now becoming clear and the prediction of our model for requiring both forms of sensitivity exist difficult to verify from the published experimental data. Even so, it is clear that excitatory ORN responses are far more common. In any case, the conclusions obtained from our model do not depend on bipolar sensitivities since the case with only excitatory sensitivities distributed following a Poisson distribution (each sensitivity has a chance of 0.5 of being 0, and 0.5 of being 1) is still better than the specific case, but worst than the bipolar optimal configuration (data not shown).

  There is direct evidence for such bipolar sensitivities in biological olfactory sensory neurons that might result from multiple second messenger signaling pathways (mediated by cAMP and IP_3 for example) within the neuron and

perhaps even multiple receptor types within a single neuron. Figure 2.8 clearly shows both inhibitory and excitatory responses to different single compounds observed in a single neuron. More recent evidence for bipolar sensitivities is given by Sanhueza et al., who observed both an excitatory cAMP-dependent current and inhibitory Ca^{2+}-dependent current in a single olfactory sensory neuron of the rat [Sanhueza *et al.*, 2000].

- **Unspecific tuning and conferred biological advantage** - Our results show that the optimal configuration is composed by a mix of neurons with broad and narrowly tuned receptive fields. This result is in accord with [Wilke & Eurich, 2002] where it was shown that a population of neurons with unimodal tuning curves obtain a more accurate representation of the stimulus if the widths are not homogeneous. In our model the optimal configuration of unspecific tunings produce better sensing performance than the imposed specific tuning case for any given dimensionality of stimulus (figure 2.4). Such an arrangement might be preferred by nature in circumstances where sensitivity to each of a large number of stimuli is important, which we describe as the general odor discrimination task. This has the added benefit that the system is able to respond to unseen or even entirely novel stimuli and so is extremely broadly tuned to its environment (for example see [Laurent, 1999] for a discussion). Hence this may explain why olfactory neurons have unspecific receptive fields.

  The mixed RFs scenario observed in our results presents the intriguing possibility of clustering of sensitivities towards groups of compounds of interest to the animal that might make arise phylogenetically. Superclusters of similar sequence homology in olfactory 7-transmembrane receptors observed from recent phylogenetic studies might reflect this aspect of our model [Zozulya *et al.*, 2001].

- **Stochastic gene selection and conferred biological advantage** - We have shown that the detection performance for an homogeneous gene expression is nearly optimal. Thus a local stochastic gene selection process giving rise to homogeneous gene expression across the receptor population can potentially lead

to a configuration with near optimal detection performance. Critically a homogeneous constraint allows the gene selection process to be a completely random local process. A key point here, since no overall control is required across the population. Interestingly, there are evidences that such a local and stochastic process exists [Serizawa *et al.*, 2000, Kratz *et al.*, 2002, Mombaerts, 2001]. We show that this arrangement is seemingly adequate to obtain near optimal detection performance – a surprising result considering the inherent randomness of the process. A local selection process also becomes fundamental when we consider that there is a continual turnover of ORNs over the lifetime of the animal and so gene selection is an ongoing process. Under these circumstances a locally defined selection process would also make sense for maintaining stable detection performance over time.

- **Model properties** - In order to make the global search of the RF space feasible, some simplifications have been made to the model. For example, the neurons have been approximated as linear elements. It is reasonable to assume a linear model for low concentrations where the firing rate is approximately dependent linearly on the number of sites filled on the receptor and there are sufficient sites relative to molecules such that no competition occurs for sites between compounds.

  It is important to remark that we have avoided describing the input to the olfactory system as a scalar, as has been done in previous work [Abbott & Dayan, 1999, Seung & Sompolinsky, 1993, Pouget *et al.*, 1999]. Using the representation we have chosen it is possible to account for both the odor's intensity as well as its chemical composition. In addition, our choice is able to account for the case when several simultaneous stimuli are present, which can not be described by just a scalar.

  The receptive fields of all the neurons are subject to optimization. Therefore, we neither impose a functional form on the RFs nor that these are homogeneously distributed, as opposed to previous papers [Abbott & Dayan, 1999, Seung & Sompolinsky, 1993] where these assumptions lead to mathematical

simplifications. Here the optimization is carried out using a numerical technique. An analytical approach is also possible, but these approaches depend critically on the particular constraints placed upon the sensitivities.

2.5.1 Discussion

Summarizing, our results show that the information representation at the olfactory epithelium is optimal when a) all the input dimensions are coded by each of the neurons in the population; b) maximum allowable gain of the sensitivities occurs in each case; c) the diversity of tunings across the population is maximum; and d) the spread of tunings across the stimulus space is homogeneous.

Since many of these features have been observed experimentally [Sicard & Holley, 1984, Chess *et al.*, 1994, Buck & Axel, 1991] we can say the olfactory epithelium seems to follow the *principle of maximum information transfer*, defined as the maximization of the information that the receptors activity carries about the stimulus. Therefore classical information theory is an appropriate tool to describe this system.

The olfactory epithelium represents the information using a *population code*, where the information about a stimulus is coded internally not by a particular neuron but by the global activity of the network. Therefore, the "symbols" of the internal "alphabet" of the system are the global internal states. This is interesting since in describing this system we can not refer to the concept of individual neurons but to the notion of global internal state. The statistics of these internal states and their correlation with the different stimuli are the key ingredients which determine the performance of the whole system.

Interestingly we have shown that a local stochastic gene selection process giving rise to homogeneous gene expression across the receptor population can potentially lead to a configuration with near optimal detection performance.

The conclusions we get from the study of the olfactory epithelium provide us useful insights about the principles of organization in the olfactory system, and the kind of tools we need in order to formalize a general framework of information processing.

However, the internal organization of the olfactory epithelium arises as a consequence of genetic processes [Reed, 2000]. Therefore it is "programmed" in the animal, that is, it is previously determined. Since we are interested in understanding how optimal internal representations can be learned, we will study in next chapter two different systems where plasticity dynamics have been described by the experimentalists, namely the visual and auditory cortices.

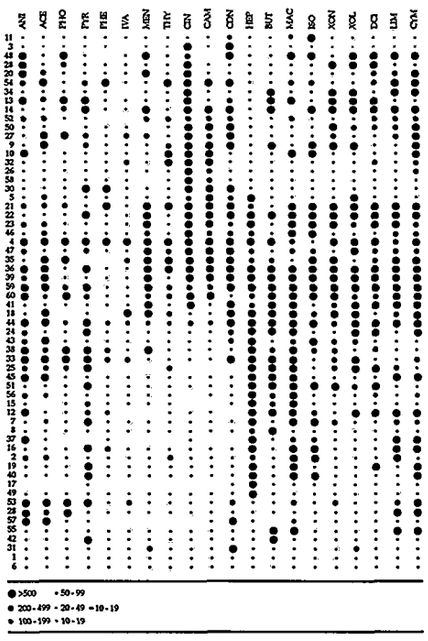

Figure 2.1: **Diagrammatic representation of olfactory sensory neuron activity following stimulation.** The spot size is roughly proportional to spike frequency (spike/min). Neurons are identified by a serial number in the left column. Note the mix of highly specific to highly unspecific neuron responses to this subset of all possible odors. Importantly, 14 receptors within this randomly selected subpopulation of olfactory sensory neurons failed to respond to any of the odors presented (not shown). ACE - acetophenone, ANI - anisole, BUT - n-butanol, CAM - DL-camphor, CDN - cyclodecanone, CIN - 1,8-cineole, CYM - p-cymene, DCI - D-citronellol, HEP - n-heptanol, ISO - isoamyl acetate, IVA - iso-valeric acid, LIM - D-limonene, MAC - methyl-amylketone, MEN - L-menthol, PHE - phenol, PHO - thiophenol, PYR - pyridine, THY - thymol, XOL - cyclohexanol, XON - cyclohexanone. Adapted from [Sicard & Holley, 1984].

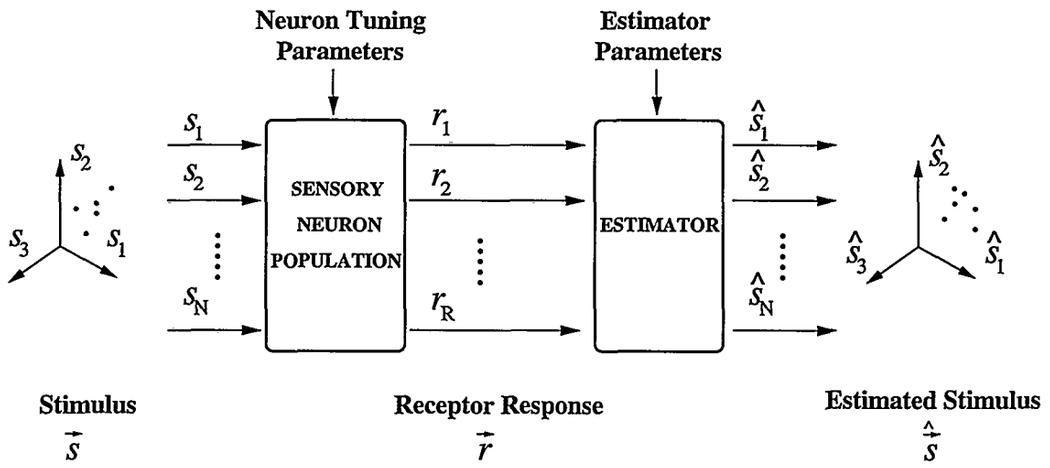

Figure 2.2: **Model of the olfactory system:** the olfactory epithelium is a population of sensory neurons which codes the olfactory stimulus as an activity pattern which is processed by further processing steps. In order to perform adequately, these higher levels need implicitly or explicitly to make a correct estimation of the odor components of the stimulus. Therefore a good representation of the information at the sensory neuron population is critical for the goodness of the estimation.

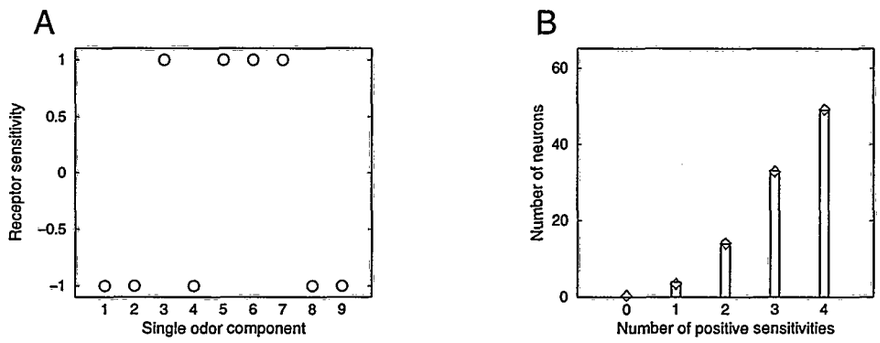

Figure 2.3: Receptive field unspecificity in the optimal OSN configuration. **A:** odor sensitivities of an arbitrary neuron, input dimension = 9. **B:** number of neurons in the population with the same number of positive sensitivities (input dimension = 9). In this plot we have taken into account that a receptive field $\vec{a}_i$ is totally equivalent to $-\vec{a}_i$ for the Fisher information Matrix (see eq. 2.5). Therefore, if $\vec{a}_i$ has more than 4 positive sensitivities, it is multiplied by -1. Diamonds: theoretic distribution considering that the probability for a sensitivity to be 1 or -1 is 0.5 (Poisson distribution)

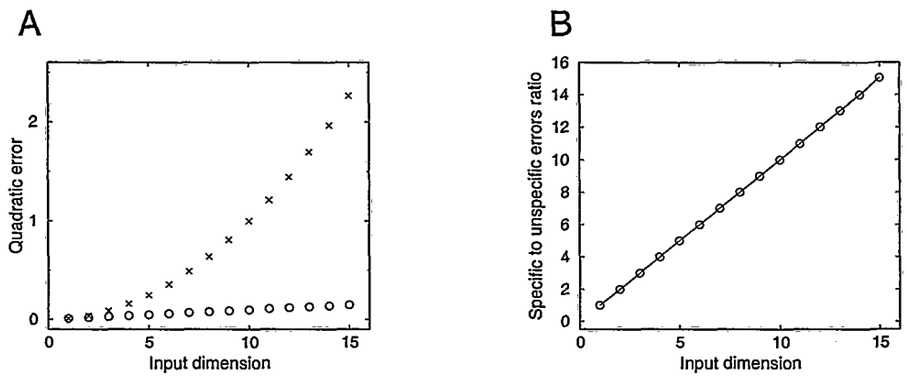

Figure 2.4: **A:** Minimum expected squared error in the reconstruction of $\vec{s}$ in units of noise variance. Crosses: specific RFs case. Circles: unspecific RFs case. **B:** Specific squared-error to unspecific squared-error ratio in the reconstruction of $\vec{s}$. The psychophysical discriminability across a range of individual stimulus components in the animal can be linked directly to the global reconstruction error defined by this equation [Dayan & Abbott, 2001].

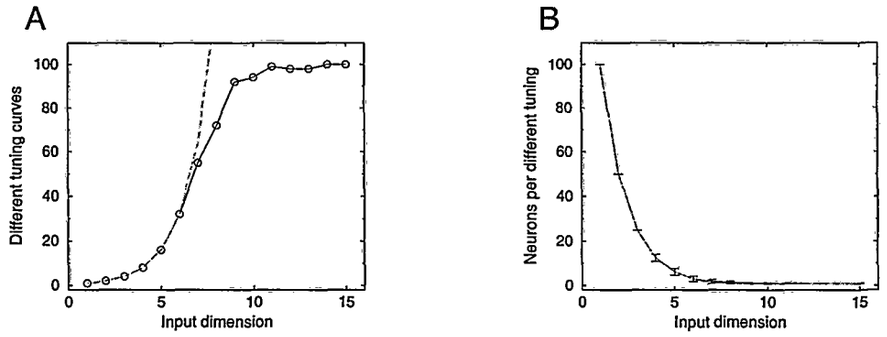

Figure 2.5: Receptive fields distribution in the optimal OSN configuration. **A:** Number of different tuning curves as a function of the input dimension (N). Dashed: number of theoretically different tuning curves assuming that each sensitivity can be arbitrarily either 1 or -1. Since $\vec{a}$ and $-\vec{a}$ are considered as the same receptive fields, this number is 2^{N-1}. **B:** Number of different tuning curves per sensor as a function of the input dimension. Vertical bars indicate the standard deviation. Dashed: theoretical line assuming an homogeneous distribution of the different receptive fields.

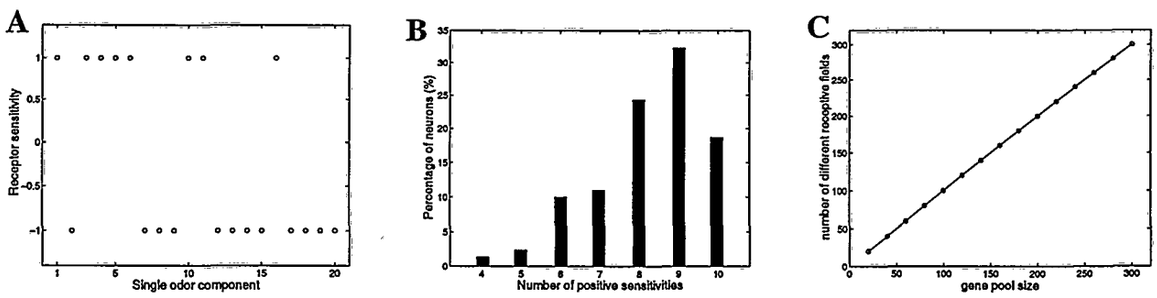

Figure 2.6: Properties of the optimal configuration for an homogeneous gene expression. The input dimension is chosen as 20. **A:** Odor sensitivities of an arbitrary neuron. **B:** Percentage of genes which code a receptive field with a given number of positive sensitivities. Since a change $\vec{u}_i \to -\vec{u}_i$ is irrelevant for the Fisher information (eq. 2.6) we normalize the global sign so that if $\vec{u}_i$ has more than 10 positive sensitivities it is multiplied by -1. **C:** Number of different receptive fields as a function of the gene pool size.

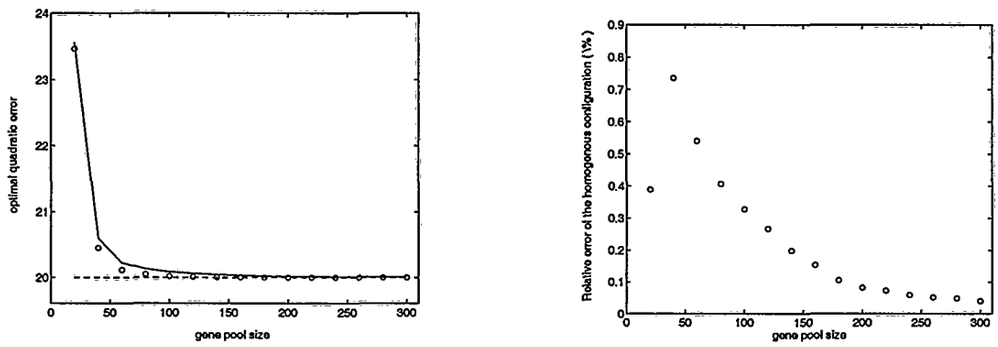

Figure 2.7: **A:** reconstruction error of the optimal configuration for an arbitrary input dimension (N) equal to 20. Solid: homogeneous gene expression. Circles: unconstrained gene expression. Dashed: unconstrained pool size. The optimal global error is shown as a function of the receptor pool size M (note that the unconstrained pool size configuration has not dependence on M). **B:** relative error between the homogeneous and the unconstrained gene expression situations.

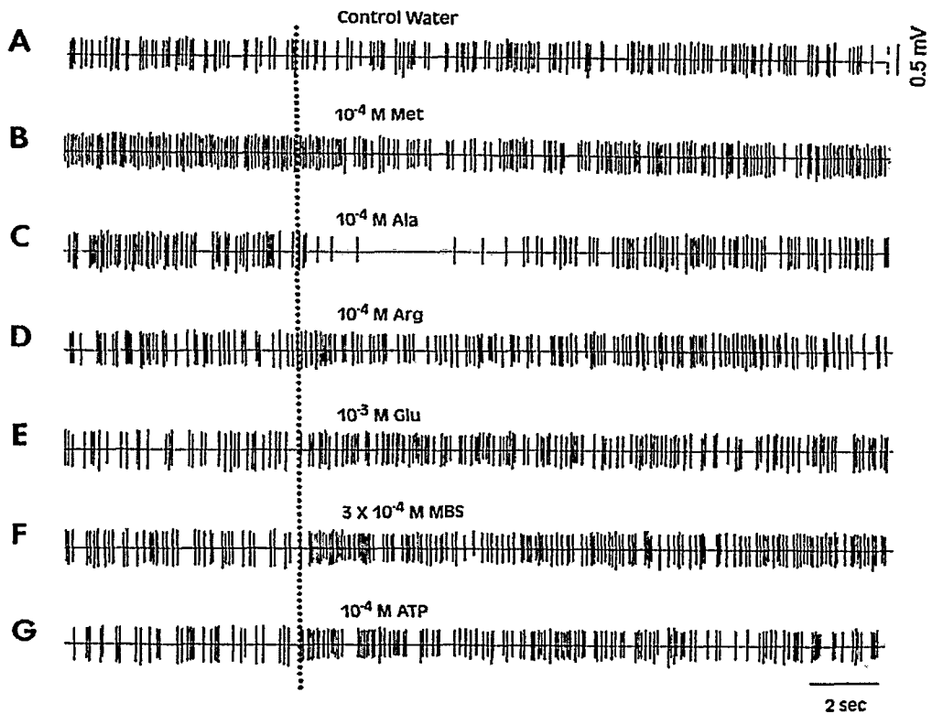

Figure 2.8: Extracellular recordings of a single olfactory sensory neuron of channel catfish, *Ictalurus punctatus*, to 6 odor stimuli and a water control. (a) no significant change from spontaneous activity to water control. (b) inhibitory response to 10^{-4}M methionine (Met). (c) inhibitory response to 10^{-4}M alanine (Ala). (d) excitatory response to 10^{-4}M arginine (Arg). (e) excitatory response to 10^{-3} glutamic acid (Glu). (f) excitatory response to 3×10^{-4}M MBS (sodium salts of cholic acid, taurocholic acid, and taurolithocholic acid each at 10−4M). (g) excitatory response to 10^{-4}M ATP. Vertical dotted line indicates beginning of neural responses as defined by onset of simultaneously recorded electrolfactogram response (local field potential). Reproduced with permission from [Schild & Restrepo, 1998].

Chapter 3

Study of biological systems with plasticity mechanisms: The visual and the auditory cortex

3.1 Context

In this chapter we will study two different systems where plasticity dynamics are crucial for the development of optimal internal representations. Our objective is to study the basic principles responsible of the generation of these optimal representations. In order to achieve this, we will use detailed models with realistic dynamics, where all the different aspects and parameters are taken from experiments described in the literature. First, we will study the mechanisms of self-organization and the emergence of receptive fields and sensory maps in a model of the visual cortex. The results are very similar to the properties observed in the biological systems. Specifically, our results are in agreement with other theoretical studies ([Miller *et al.*, 1989, Wimbauer *et al.*, 1997, Miyashita *et al.*, 1997]). Next, we will study the development of receptive fields in the auditory cortex. We will show that the introduction of neuromodulatory mechanisms provides the model with a mechanism that assigns to each stimuli an amount of neuronal resources depending on its behavioral importance, as observed in experiments with biological systems. However,

we can not interpret this using information theory and a concept of "meaning" must be introduced in order to explain these observations. Finally, we will expose our conclusions and will suggest general principles which will be properly formalized in next chapter.

Part of the results of this chapter have been present elsewhere:
[Sánchez-Montanés et al., 1999] for the visual cortex model and [Sánchez-Montanés et al., 2000, Sánchez-Montanés et al. 2001, Sánchez-Montanés et al., 2002] for the auditory cortex model.

3.2 Introduction

Over the past years neuroscientists have gained insight in the neural mechanisms responsible for the ability of learning and adaptation in biological systems (for a review see for example [Alkon *et al.*, 1991, Buonomano & Merzenich, 1998]). The substrate of learning in these systems is thought to be provided by the mechanisms which regulate the change of synaptic efficacies of the connections among neurons [Martin *et al.*, 2000, Tsien, 2000]. In his seminal work D.O. Hebb proposed that neurons which are consistently coactivated strengthen their coupling [Hebb, 1949] and form associative networks. Since then many experiments have addressed different mechanisms which regulate changes in synaptic efficacies dependent on specific properties of pre- and postsynaptic activity [Bliss & Collingridge, 1993, Buonomano & Merzenich, 1998]. Based on these experiments, a number of Hebbian learning rules have been proposed with different desirable properties [Sejnowski, 1977, Stent, 1973, Bienenstock *et al.*, 1982, Brown & Chattarji, 1998, Fregnac, 1998].

These learning rules have been considered physiologically realistic when they only rely on signals which are available to the synapse locally in time and space. However, recent physiological results on neurons in cortex give a richer picture. These studies demonstrate, firstly, that an action potential triggered at the axon hillock propagates not only anterogradely along the axon, but also retrogradely through the dendrites [Stuart & Sakmann, 1994, Buzsaki & Kandel, 1998]. Secondly, on its

way into the dendrite the action potential may be attenuated or blocked by inhibitory input from other neurons [Spruston *et al.*, 1995, Tsubokawa & Ross, 1996]. Thirdly, it has been demonstrated that these backpropagating action potentials directly affect mechanisms regulating synaptic plasticity [Markram *et al.*, 1997] which depends on post-synaptic calcium dynamics [Köster & Sakmann, 1998]. Fourthly, the temporal relationship between the backpropagated action potential and the synaptic activity can determine whether potentiation or depression occurs, potentiating only the synapses whose activity occurs previously to the postsynaptic activity [Markram *et al.*, 1997, Zhang *et al.*, 1998, Bi & Poo, 1998].

In addition, the dramatic effect of even single inhibitory inputs on the calcium dynamics in the dendritic tree, in particular in its apical compartments, suggests that regulation of synaptic plasticity can be strongly influenced by inhibitory inputs [Larkum *et al.*, 1999]. Thus, the backpropagating action potential can make information on the output of the neuron available locally at each of its afferent synapses, and inhibitory inputs onto a neuron can in turn regulate the effectiveness of this signal.

The above described mechanisms make a change in synaptic efficacy dependent on the temporal relation between pre- and post-synaptic activity. On one hand, the synaptic efficacy will be strongly affected by the temporal relation between presynaptic and postsynaptic activity, potentiating only the synapses whose activity have really contributed to the postsynaptic activity [Markram *et al.*, 1997, Zhang *et al.*, 1998, Bi & Poo, 1998]. Therefore the plasticity dynamics do not depend merely on the correlation between pre- and postsynaptic activity, but is able to distinguish between cause and effect. On the other hand, the synaptic efficacy will also depend on the relation between the inhibition and excitation a neuron receives and its own activity. Neurons which fire with the shortest latency to a stimulus will receive inhibition after they have generated backpropagating action potentials. In this case active synapses can be potentiated [Larkum *et al.*, 1999]. For instance, neurons which fire late to a stimulus would receive inhibition before they have generated a spike. Their backpropagating action potentials are modulated by this inhibition preventing potentiation of their active synapses. This dynamic seems to be reflected in the physiology of the visual system where the optimality of the tuning of a neuron seems to be directly

reflected in its response latency to a stimulus [König *et al.*, 1995]. Given the above mechanism this would imply that the optimally tuned neurons prevent further learning by other neurons in the map.

Synaptic plasticity, however, is not only dependent on the dynamics of the local network but also on modulatory signals [Abbott, 1990]. For instance, the basal forebrain is a subcortical structure which sends to the whole cortex connections of two different types: cholinergic (they use acetilcholine as the chemical transmitter) and inhibitory [Kandel *et al.*, 2000]. Since these connections are arranged in a diffuse way and their influence is not related to the specifics of a given stimulus they act as a kind of global signal. The action of these connections is a necessary ingredient for the induction of cortical representations following monocular deprivation [Singer & Rauschecker, 1982]. In addition, it may switch between storage and recall modes in the hippocampus [Hasselmo, 1993], and it gates the plasticity of receptive fields of neurons in the primary auditory cortex during classical conditioning [Weinberger, 1993, Bakin & Weinberger, 1996, Kilgard & Merzenich, 1998]. These results support the suggestion that modulatory substances can act as a "print now" signal gating synaptic plasticity [Singer *et al.*, 1979] which remarks the behaviorally important events.

The dynamics of the synaptic mechanisms which regulate the plasticity in the network will then depend on all these local factors which in turn are affected by the network dynamics. The ability of the network to learn and adapt to the environment will then emerge from the temporal and spatial interactions of all these mechanisms. In this chapter we will study how these factors determine the development of optimal internal representations in sensory systems. In order to do this we will construct realistic models which take into account these factors and will study them in realistic conditions using real stimuli in real time.

3.3 Model of the primary visual cortex

The visual information that arrives to the eyes is detected by the cones and rods [Kandel *et al.*, 1991]. The information taken by these photoreceptors is processed

in the retinal circuits before it arrives to the ganglion cells, which are the output neurons of the retina [Kandel *et al.*, 1991]. Most ganglion cells in the mammalian retina have a center-surround receptive field of one of the following types: ON (the neuron is maximally activated when a small spot of light is projected in the center of its receptive field) or OFF (analogously with a small spot of dark) [Bear *et al.*, 1996].

The ganglion cells send excitatory connections to the lateral geniculate nuclei of the thalamus, which after process this information sends it to the primary visual cortex [Kandel *et al.*, 1991]. The neurons in the geniculate nuclei have center-surround receptive fields similar to the ganglion cells. However, the neurons in the primary visual cortex have more complex response patterns since many of them respond selectively to specific orientations [Hubel & Wiesel, 1959] and / or specific directions of movement which occur at the center of their receptive field [Hubel & Wiesel, 1962]. Moreover, there is a topological organization so that neurons that are close in the network tend to respond to similar stimuli. This gives rise to orientational and directional maps with complex properties [Kandel *et al.*, 1991, Bear *et al.*, 1996]. Are these properties programmed in the animal or on the other hand are they learned by visual experience ?

Many cells in the primary visual cortex of cats are both orientionally and directionally selective already before eye-opening [Albus & Wolf, 1984, Braastadt & Heggelund, 1985, Hubel & Wiesel, 1963, Movshon & van Sluyters, 1981, Sherman & Spear, 1982]. However, the selectivity of neurons for oriented stimuli at the time of eye opening increases dramatically after the onset of visual experience [White *et al.*, 2001]. Therefore the new-born animal has a rough internal representation which is subsequently refined by visual experience. There are experimental evidences that support the hypothesis that the rough initial maps are developed before eye-opening by activity-dependent plasticity processes. For example, the major development of orientational selectivity in ferrets depends on neural activity [Chapman & Stryker, 1993]. On the other hand, directional selectivity can be abolished by strobe-rearing during the critical period [Blakemore & van Sluyters, 1975, Humphrey & Saul, 1995]. Therefore these experiments suggest the existence of a self-organization process in the visual cortex

which depends on the neural activity which gives rise to these maps. The existence of spontaneous activity waves in the developing retina [Meister *et al.*, 1991] has been proposed as the driving force that would activate this self-organization process [Miller *et al.*, 1999].

In this section we demonstrate that the spontaneous activity waves in the retinas present in the prenatal animal can be the responsible of the emergence of these mappings and representations from scratch. The other basic ingredients is a local plasticity mechanism which separates cause and effect [Markram *et al.*, 1997, Zhang *et al.*, 1998], and the presence of dynamics of cooperation and competition at different levels (neighbors, synapses, time). We will see that these different dynamics emerge naturally from a simple local plasticity mechanism and the interaction between neighboring neurons. Therefore, we conclude that the internal representation is implicitly programmed in the sense that the existence of the activity waves is programmed, but all the exact details are adjusted, self-organized. The visual waves can then be seen as a mechanism that provides the animal a "starting point" for its optimal internal representation.

3.3.1 The model

Structure of the network

The structure of the model, which is an extension of [Sánchez-Montanés *et al.*, 1999], is schematized in Fig. 3.1. It includes two retinas (left and right), two thalamic populations (left and right), and cortical excitatory and inhibitory neural populations. Each of these populations consists of a 2D arrangement of integrate and fire neurons. Our functional model of each retina consists of a layer of 70x70 cells where each cell can fire spontaneously a burst of activity. The probability of a neuron to fire a burst is defined by a Poisson distribution of frequency 0.001 Hz. Each cell is connected to its 28 nearest neighbors, allowing the spontaneous activity to spread in traveling waves similar to the activity patterns observed during development even prior to the opening of the eyes [Meister *et al.*, 1991].

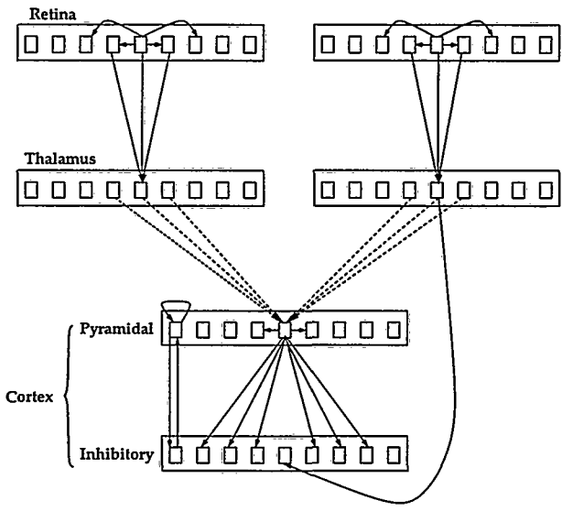

Figure 3.1: Schema of the network (we omit many connections and show a 1D projection of the network for clarity of display). The figure focuses on the connections corresponding to the central pyramidal unit omitting its intercolumnar connections only displayed for the leftmost column. Plastic connections are displayed by dashed lines.

Each retina projects retinotopically to 9 cells in the corresponding thalamic population. The waves of activity in each retina give rise to coarser waves of activity in the corresponding thalamus where active units produce short trains of spikes. The next two populations represent the pyramidal and inhibitory neurons in the striate cortex. Each one of the 40x40 pyramidal cells receives input from 29+29 (left and right) thalamic cells while each of the 40x40 inhibitory cells receives input from 1+1 thalamic units. Neighboring pyramidal cells are connected to neighboring thalamic cells, that is, we assume that a rough retinotopy is prewired before directional and orientation selectivity selforganizes. Each pyramidal cell sends excitation to its 53 nearest neighbors, receiving inhibition from its nearest inhibitory cell. Thus the long range connections between pyramidal cells and inhibitory cells of other farther microcolumns implement the competition process required for self-organization [von der Malsburg, 1973].

The model of the neurons

The dynamics of the membrane potential of neuron j in population a, $V_j^a(t)$, is defined as:

$$\dot{V}_j^a(t) = -\tau^{-1} V_j^a(t) + C^{-1} \sum_{b,i} I_{i \to j}^{b \to a}(t) - C^{-1} \Delta Q_{sp}\, \delta(t - t_{spike_j}) \tag{3.1}$$

where τ is the time constant; C is the membrane capacitance; $I_{i \to j}^{b \to a}(t)$ represents the current injected by the synapse from neuron i in population b. In case V is greater than the threshold (V_{Th}) the neuron emits a spike (t_{spike} is current time) and the membrane potential is reset to 0 by injecting the charge ΔQ_{sp} instantaneously (represented by the Dirac delta function, δ). The dynamics of the membrane potential also includes an absolute refractory period.

The dynamics of the synapses formed by neurons of population b with neurons of population a are modeled using a first order approximation:

$$\dot{I}_{i \to j}^{b \to a}(t) = -\tau^{-1} I_{i \to j}^{b \to a}(t) + \gamma^{b \to a} w_{i \to j}^{b \to a}\, \delta(t - t_{spike_i}) \tag{3.2}$$

where $\gamma^{b \to a}$ is a constant gain factor that defines the type of connection (positive for excitatory, negative for inhibitory) and its maximum gain; $w_{i \to j}^{b \to a}$ is a variable ranging from 0 to 1 that expresses the efficacy of the synapse.

The model is implemented using mathematical algorithms which maximize the speed and computational efficiency [Sánchez-Montanes, 2001].

Plasticity dynamics

We have concentrated on the plasticity of the connections projecting from the thalamus to the pyramidal cells. The plasticity mechanism is based on the physiological data reported by [Markram *et al.*, 1997, Zhang *et al.*, 1998]. The decision to increase or decrease the thalamo-pyramidal weights is determined by the temporal relation between the presynaptic and postsynaptic spikes [Markram *et al.*, 1997, Zhang *et al.*, 1998]: if the presynaptic neuron fires before the postsynaptic one, the synapse is potentiated. Otherwise, the weight is depressed (fig. 3.2). In addition, an

heterosynaptic LTD mechanism is included: all the synapses that are not active when the postsynaptic cell fires (in a symmetric temporal window of 200 ms) are depressed by a constant factor.

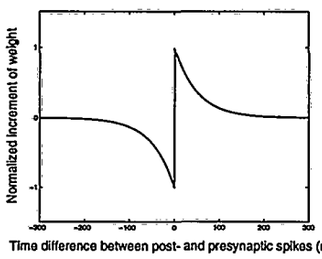

Figure 3.2: Plasticity dynamics. Both the sign (homosynaptic LTP/LTD) and amplitude of the change in synaptic strength depends on the time difference between the post- and presynaptic spike, Δt. Thus $\Delta w = \alpha\, sign(\Delta t) \exp(-\tau^{-1}|\Delta t|)$, with $\alpha > 0$ and $\tau = 50$ ms.

3.3.2 Results

The initial plastic weights for the connections arriving from the thalamus are randomly chosen in a small range. Thus, the excitation and inhibition that comes into a pyramidal cell when a stimulus is presented does neither appreciatively depend on its orientation nor its direction. As a result, microcircuits initially respond indiscriminately to all the analyzed stimulus features. When a front wave comes into the neuron receptive field, synapses from thalamic neurons that first fired and made the pyramidal neuron fire are potentiated (fig. 3.3). On the other hand, those connections from thalamic neurons firing later and those from not active neurons are depressed. This creates a small bias in the receptive field so that now this neuron fires slightly better to stimuli with characteristics similar to the wave front (fig. 3.3).

The action of this mechanism during hundreds of waves, together with the competition between microcircuits implemented by the long range connections, makes the pyramidal cells eventually differentiate into different recognition cells (i.e. with different feature specificity). After self-organization takes place, the fully connected system converges to a configuration of microcircuits where some connections have

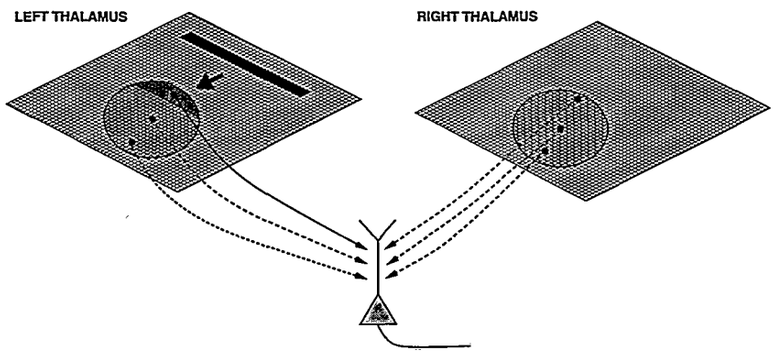

Figure 3.3: Synaptic changes induced by a front wave coming into the receptive field of a pyramidal neuron. Those synapses from neurons in the activated thalamus which fire before the pyramidal neuron are potentiated (solid line). The other synapses from both the activated and non-activated thalami are depressed due to homosynaptic and heterosynaptic LTD respectively (dashed lines).

been potentiated and some others have died away. The self-organization process creates an asymmetry between the thalamo-cortical synapses that come to a pyramidal cell that breaks the initial isotropy. Thus at the end of this process the temporal relation between feedforward excitation and intracortical inhibition arriving at the pyramidal cell depends critically on the stimulus direction. This accounts for the directional selectivity of the neuron [Douglas & Martin, 1991] (see fig. 3.4).

The competition between different neurons and the spatial correlation of the input patterns accounts for the specialization of each unit to a different orientation [von der Malsburg, 1973]. Moreover, all possible orientations and directions are represented after the self-organization (fig. 3.5). Finally, the heterosynaptic LTD allows competition between input coming from different retinas that finally gives rise to the ocular dominance organization [Miller *et al.*, 1989]. This final state is stable in the sense that no further stimulation will cause any change in the circuit. Due to the relative small number of neurons in the simulation, a correlation of the optimal features with position in the population can be observed. Work in progress addresses the reduction of this correlation by mechanisms such as stronger competition, etc.

Our model of spiking neurons with plausible synaptic dynamics can thus account for the simultaneous development of both orientational and directional selectivity,

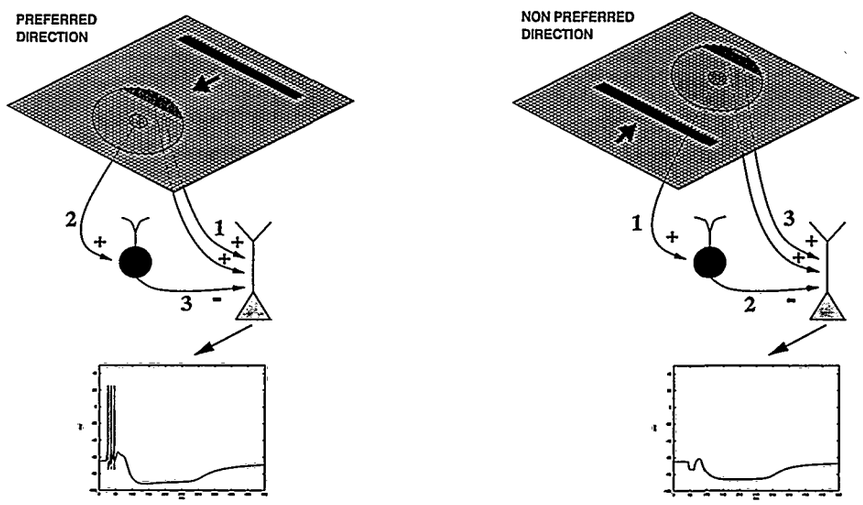

Figure 3.4: Microcircuit structure after self-organization. Each neuron finally receives input from just a particular thalamus (ocular dominance), which is plotted in the figure. The neuron specializes to a particular orientation and direction: the orientational selectivity depends on the alignment within the receptive field and the directional selectivity arises from a spatial shift between the thalamic input and the intracortical inhibition.

as well as a topographical organization, obtaining similar results to those obtained by more abstract models [Wimbauer *et al.*, 1997, Miyashita *et al.*, 1997]. In these previous works the existence of thalamic cells with different response latencies is crucial for the development of directional selectivity. Thus this feature arises from an asymmetry in the contribution of cells with different latency to the feedforward excitation of cortical cells. However, in our model the directional selectivity emerges from the temporal interaction between feedforward excitation and intracortical inhibition [Douglas & Martin, 1991]. This mechanism and the latency-based could not interfere but act synergetically. Accordingly we would expect to obtain similar results if thalamic cells with different latencies are incorporated in our model. Moreover, a plasticity mechanism similar to the one described in this paper can be implemented in the lateral intracortical connectivity, creating neural circuits selective to direction [Rao & Sejnowski, 2000]. In addition, our model develops ocular dominance columns [Miller *et al.*, 1989, K. Obermayer, 1995, Andrade & Morán, 1996]. The development of all these different features with the same plasticity mechanism suggests

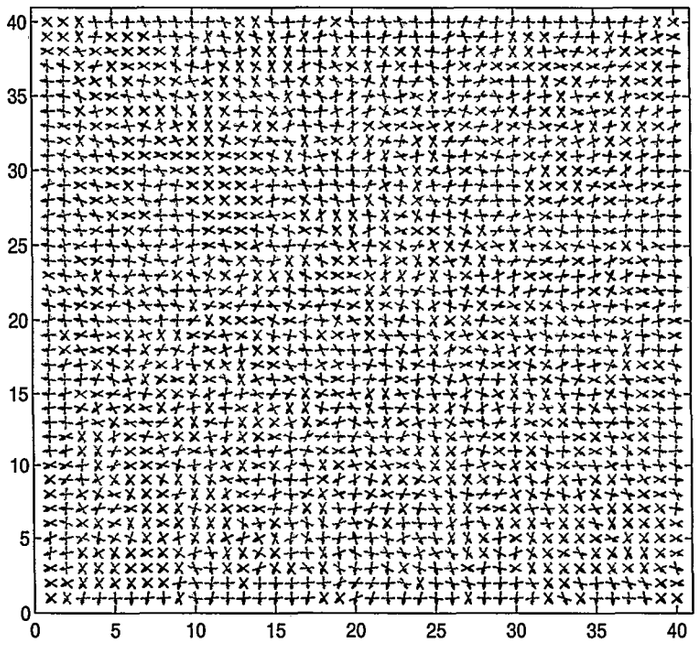

Figure 3.5: Optimal stimuli for each pyramidal neuron after self-organization. X and Y axis are the coordinates of the neuron in the pyramidal population. Bars indicate the optimal orientation while arrows indicate the optimal direction. Features corresponding to neurons that respond only to the left retina are plotted in black, while those corresponding to neurons responding to the right retina are plotted in gray.

that the general neural and synaptic mechanisms underlying developmental processes for all kinds of selectivity might be similar.

3.4 Model of the primary auditory cortex

The primary auditory cortex of mammals is composed by neurons which are selective to specific frequencies [Bear *et al.*, 1996]. This internal representation is not static but can change through experience. For example, in classical conditioning experiments where tones are paired with aversive stimuli such as a footshock the receptive fields of many neurons suffer a shift to these tones [Weinberger *et al.*, 1993]. These changes are retained indefinitely [Weinberger, 1993] due to long-term changes in the neural circuit. Subsequently it was shown that the aversive stimulus could be replaced by

direct stimulation of the basal forebrain [Kilgard & Merzenich, 1998]. In these latter experiments it was shown that more neurons in the primary auditory cortex would respond to the reinforced frequency while the representation of the others was not increased (see figure 3.6).

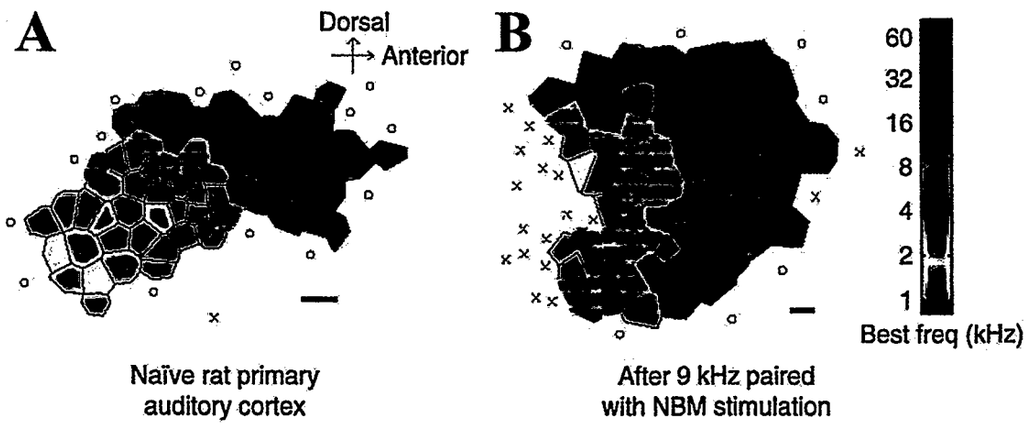

Figure 3.6: Map reorganization in the auditory cortex gated by subcortical activity. Each neuron has a preferred tone which maximizes its response. The figure shows the distribution of preferred tones in $A1$ as a function of the neuron location in this cortex. **A:** Naive rat. **B:** Rat which has been placed in a controlled environment where the tones 1 kHz, 9 kHz and 30 kHz have been presented in random sequences during weeks. The animal had an electrode implanted in its basal forebrain which was activated everytime the 9 kHz tone was presented. In experiments where this electrode is not implanted, the experimenters report maps similar to **A.** (Adapted from [Kilgard & Merzenich, 1998]).

In this section we will study the mechanisms which give rise to this process of learning from experience. We will develop a model of primary auditory cortex that includes realistic neural and synaptic dynamics, and modulatory signals coming from the basal forebrain. Since we are interested in validating the model using stimuli and conditions as close as possible to those present in the brain, we have implemented the model in a setup which allows a real-time simulation using real-world stimuli.

We demonstrate that this biologically realistic real-time neuronal system forms stable receptive fields similar to those present in the animal. We will show that in our model the representation size can be biased by global modulatory signals acting on the local learning mechanism, which is in accord with the experimental observations

[Weinberger, 1993, Bakin & Weinberger, 1996, Kilgard & Merzenich, 1998]. Finally we will interpret the results of our model from an information theoretical point of view studying the optimality of the global detection performance of the system.

3.4.1 Realistic model implementation

Hardware setup

All experiments are conducted in a standard office environment with a room size of about $30m^2$. The analog audio signals are sampled using a microphone (ME64, Sennheiser, Wedemark, Germany) at 44.1 kHz and digitized with 16 bit resolution on an interface card (Soundblaster, Creative Technology Ltd, Singapore, Singapore). On each block of 1024 sampled signals a digital FFT is computed [Frigo & Johnson, 1998]. Input to the model is provided by the absolute values of the first 128 FFT coefficients. The whole system for the control of the setup, the stimulus generation protocol, the simulation, and data acquisition is defined within the distributed neural simulation environment IQR421 [Verschure, 1997] using three Pentium III 450 MHz PCs (fig. 3.7 A).

The network

The neural network is a very rough sketch of the mammalian auditory system and includes five sets of integrate and fire neurons: an input population, a thalamic population, cortical excitatory and inhibitory neurons and an additional neuron representing the basal forebrain (fig. 1 B). All neurons are simulated in strict real time, i.e. simulated biological time matches 1:1 spent physical compute time. The dynamics of the neurons are the same as in the visual cortex model (eq. 3.1). In table 3.1 we show the concrete values of the parameters for each population.

The dynamics of the synapses formed by neurons of population b with neurons of population a are modeled as in the visual cortex model (eq. 3.2).

There are three types of connections in the model: (1), non-plastic (w is constant and equal to 1); (2), subject to short-term plasticity; and (3), subject to long-term

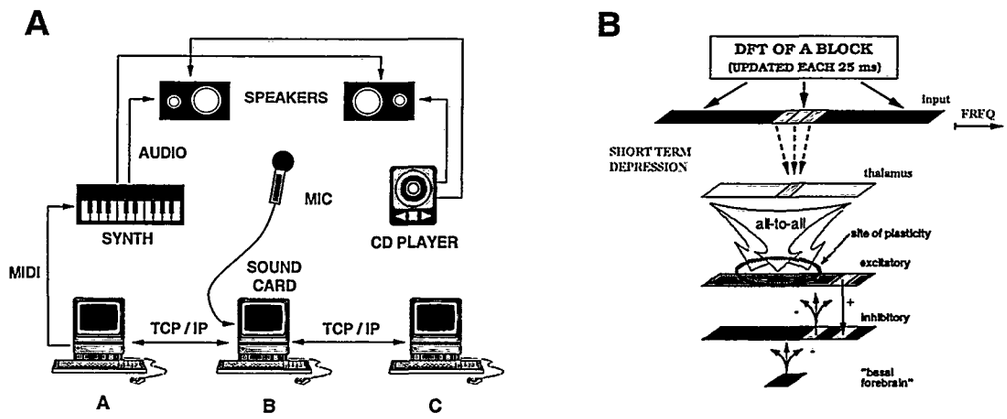

Figure 3.7: **A**: Schema of the hardware. The sounds are generated by a pair of speakers that receive audio signals either from a synthesizer which is controlled by a computer using the MIDI protocol (A) or from a CD-player. The microphone sends the audio signal to the soundcard installed in computer B; this computer calculates the FFT and sends this data to C, where the neural model is implemented. Each computer communicates to the others through TCP/IP (connections between computers A and C are omitted for clarity). The whole system is controlled by the distributed neural simulation environment IQR421. **B**: Schema of the neural model. Each one of the first 128 FFT coefficients excites its corresponding input neuron. The output of each group of 3 of these neurons converges into a thalamic neuron through short-term depressing synapses. Each cortical excitatory neuron receives excitation from the whole thalamic population. These synapses are subject to long-term synaptic plasticity. Each cortical excitatory neuron is connected to a cortical inhibitory neuron which sends back inhibition to the whole excitatory population. Finally, the unit representing the basal forebrain activity sends inhibition to the cortical inhibitory population. See text for further details.

plasticity. A transmission delay of 2 ms is taken into account in all the connections. In table 3.2 the numerical details for each type of connection are given.

Each input neuron receives an excitation proportional to the absolute value of its respective analog Fourier coefficient (frequencies up to a quarter of the Nyquist frequency, 5.5 kHz). Thus we simulate in a first approximation the spectral decomposition achieved by the cochlea and subcortical nuclei [Kelly, 1985]. The spectra of all the sounds used in the experiments are kept in this range. Because it takes 23 ms to sample a block of audio signals, the input to these neurons is updated every 24 ms and kept constant until the next update. The time spent in FFT computation is

	Input	Thalamic	Cortical excitatory	Cortical inhibitory	BF
size	128	43	36	36	1
τ (ms)	19	19	19	19	19
ARP (ms)	6	10	10	6	8

Table 3.1: Parameters of the different populations. "BF": Basal forebrain. "ARP": absolute refractory period.

Connection	Plasticity	Connections	$C^{-1}\gamma$	$\tau(ms)$
Input $\rightarrow$ Thalamus	Short-term	3 to 1	.15	19
Thalamus $\rightarrow$ C. excitatory	Long-term	all to all	.1	19
C. excitatory $\rightarrow$ C. inhibitory	No	1 to 1	1	19
C. inhibitory $\rightarrow$ C. excitatory	No	1 to all	-.0025	19
BF $\rightarrow$ C. inhibitory	No	1 to all	-.06	19

Table 3.2: Parameters of the connections between populations. "C. excitatory": cortical excitatory. "C. inhibitory": cortical inhibitory. "BF": Basal forebrain. The connection strength, $C^{-1}\gamma$, is given in units of the postsynaptic threshold V_{Th}.

very low compared to the sampling time.

Each thalamic neuron receives excitation from 3 input neurons in a tonotopic manner (fig. 3.7 B). This convergence of information allows to process a broad frequency band with a reduced number of neurons, making real-time processing possible. The details of this connectivity, however, are not critical to the performance of the model.

The synapses from input neurons are subject to short-term depression [Varela *et al.*, 1997], making the efficacy of the synapse dependent on previous presynaptic activity:

$$\dot{w}_{i\rightarrow j} = \tau_d^{-1}(1 - w_{i\rightarrow j}) - f\, w_{i\rightarrow j}\, \delta(t - t_{spike_i}) \tag{3.3}$$

τ_d defines the recuperation time of the synapse (4 s). f defines the speed of adaptation, being 0.1.

Each cortical excitatory neuron receives excitatory input from all thalamic neurons and in turn projects to one cortical inhibitory neuron. All cortical inhibitory neurons project to all cortical excitatory neurons. The synaptic strengths $w_{i\rightarrow j}$ of

the connections from thalamic neurons to cortical excitatory neurons are initially random, with values between 0.7 and 0.8 (homogeneous distribution); therefore, the receptive fields of the excitatory cortical neurons are initially diffuse. These synapses are subject to long-term synaptic plasticity (see Learning Dynamics). To model the context of a larger network, we added an independent excitatory input to each cortical neuron which is firing at 10 Hz following a Poisson distribution. Finally, the unit representing basal forebrain activity sends inhibitory connections to the cortical inhibitory neurons [Freund & Gulyas, 1991, Freund & Meskenaite, 1992].

Learning dynamics

The synaptic strength of the thalamic projections to the cortical excitatory neurons evolves according to a modification of a recently proposed model of synaptic plasticity [Körding & König, 2000, Sánchez-Montanés *et al.*, 2000]:

1. When the backpropagating action potential and the presynaptic action potential arrive within a temporal association window W (i.e. the absolute value of the time difference between the two events is smaller than $W = 20\,ms$), the efficacy of the respective synapse is increased [Gerstner *et al.*, 1993, Markram *et al.*, 1997, Magee *et al.*, 1998, Bi & Poo, 1998]:

$$\Delta w_{ij} = \alpha \frac{t_0}{t_0 + |t_i - t_j|} \tag{3.4}$$

   t_i is the time when the postsynaptic cell fires, and t_j is the time when the action potential of the presynaptic cell arrives at the synapse.

2. If the backpropagating action potential and the afferent action potential occur within the temporal association window W, but the inhibitory input attenuates the backpropagating action potential [Spruston *et al.*, 1995, Tsubokawa & Ross, 1996], the efficacy of the respective excitatory synapse is decreased:

$$\Delta w_{ij} = -\beta \frac{t_0}{t_0 + |t_i - t_j|} \tag{3.5}$$

3. In case of non-attenuated backpropagating action potentials which do not coincide with presynaptic activity, synaptic efficiency decreases with a constant amount:

$$\Delta w_{ij} = -\eta \tag{3.6}$$

Thus, in this learning rule the changes of synaptic efficacy are crucially dependent on the temporal dynamics in the neuronal network. In our model we used the values $\alpha = 0.02$, $\beta = 0.005$, $\eta = 0.01$, $t_0 = 10\,ms$. The weights are kept by saturation in the 0-1 range.

Training protocol and analysis

The network is trained with different types of acoustic stimuli. First, we use a commercial CD that is continuously played for 2.5 hours. In this experiment the synaptic weights are sampled at intervals of 20 seconds for further analysis. In the second set of experiments we use a music synthesizer (QS8, Alesis, Santa Monica, USA) for generating the stimuli. Simple sinusoids are played for a few minutes together with continuous low-band noise. The noise is obtained by passing white noise through a low-pass linear filter with a cut-off frequency of 600 Hz. Network activity, synaptic weights, and sound frequency and amplitude are continuously recorded for further analysis. All the parameters of the model are kept constant over all experiments and the learning mechanism is continuously active. Data analysis is performed using a commercial software package (MatLab, Math Works, Massachusetts, USA).

3.4.2 Results

Development of specific receptive fields presenting real stimuli

In the real world events do not occur in isolation but are combined in a variety of ways. In the first experiment we assess whether our model is able to develop specific and stable representations under these circumstances. The initial weights of the synapses from thalamic neurons to cortical excitatory neurons are randomly chosen in the range of 0.7 - 0.8 (fig. 3.8 A); this makes the initial receptive field of all the

cortical excitatory neurons diffuse and no knowledge about the stimuli is put into the network. The network is exposed for 2.5 hours to the music from the CD ('Cabo do Mundo" by Luar na Lubre, Warner Music Spain, 1999). The CD style is celtic music played with traditional instruments, vocals, drums and synthesizers. The CD is available worldwide by music stores such as Amazon.

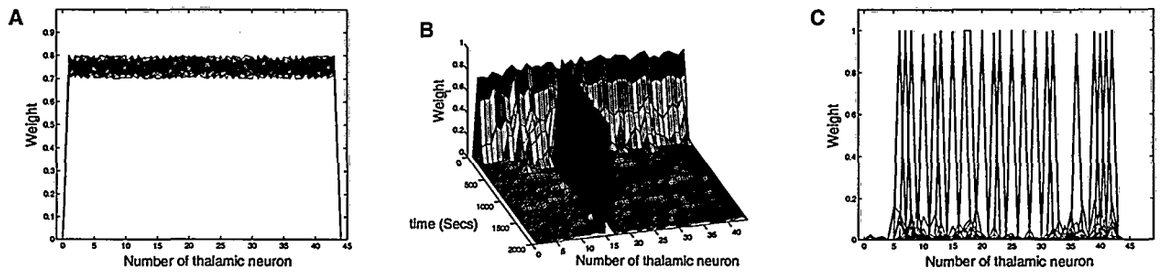

Figure 3.8: Receptive field dynamics under continuous stimulation with music. **A:** superposition of the initial receptive fields of every cortical excitatory neuron. **B:** evolution of the receptive field of one of the cortical excitatory neurons. **C:** superposition of the final receptive fields of every cortical excitatory neuron after 2.5 hours of stimulation.

In this period the learning mechanism continuously acts on the synaptic efficacies of the thalamo-cortical projections shaping the receptive fields of the cortical neurons.

Due to the short-term depression in the projection from the input neurons to the thalamic neurons, not the absolute intensity but the fast dynamics of the different frequency components is transmitted to the cortical neurons. However, due to the initial homogeneous connections from thalamic neurons to cortical excitatory neurons, most of these excitatory neurons are active, resulting in a high level of inhibition in the network. This inhibition leads to an attenuation of most backpropagating action potentials within the excitatory neurons and, thus, to a depression of thalamo-cortical synapses (fig. 3.8 B, 0-200 Sec). With the decrease of the activity level, inhibition is reduced as well, and some synapses are potentiated, leading to the formation of well defined receptive fields (fig. 3.8 B, 200-500 Sec). After 30 minutes most neurons have highly specific and stable receptive fields which practically cover the full frequency spectrum presented to the system. In addition, the different receptive fields provide a practically homogeneous coverage of the stimulus space (fig. 3.8 C).

The ability of the network to develop receptive fields which cover the full range of presented frequencies is the result of a competitive process. Neurons with a receptive field which is specific to the provided input respond with a short latency after stimulus onset. This in turn drives the inhibitory population rapidly, shunting the back propagating actions potentials in those neurons which are not effectively representing the input, preventing a change in synaptic efficacy to occur in their afferents.

These results demonstrate that this local learning mechanism allows single neurons to develop specific receptive fields within minutes, which are for realistic input conditions stable over hours. In addition, at the level of the network it allows the full range of inputs to be represented.

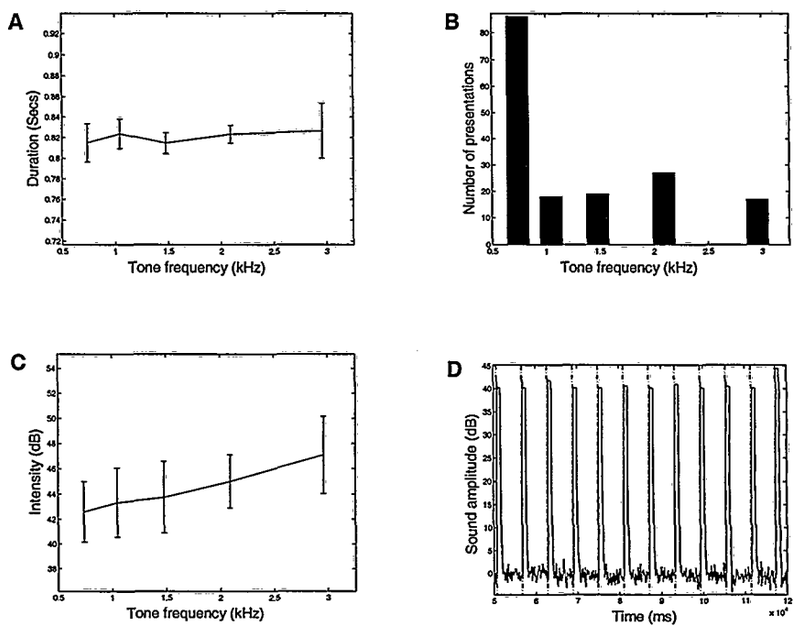

Figure 3.9: Stimulus statistics using a pseudo-random sequence of 5 different tones (0.74, 1.05, 1.48, 2.09 and 2.96 KHz). The probability of occurrence is 1/2, 1/8, 1/8, 1/8 and 1/8 respectively. **A**: mean duration of each stimulus. **B**: number of presentations of each stimulus. **C**: mean intensity of each stimulus. The 0 dB level is chosen as the averaged level of noise in the room. **D**: sound amplitude over time.

Dynamic modulation of representation size

The brain uses global signals to provide information on the behavioral relevance of events. These signals can affect local mechanisms which govern changes in synaptic plasticity. An example of such a system is the basal forebrain, mentioned in the introduction. It was recently shown that the paired activation of this structure with a particular tone induces an enlargement of the representation of this tone in the primary auditory cortex [Kilgard & Merzenich, 1998] (see figure 3.6). This change in representation size does, however, not affect the size of other representations in the cortical map and the presentation of unpaired tones does not seem to affect the organization of this cortical area.

We investigate our model using an equivalent stimulation protocol. Sinusoidal tones with frequencies of 0.74, 1.05, 1.48, 2.09 and 2.96 kHz are generated on a digital synthesizer. These frequencies are presented in a pseudo-random order with an average duration of 0.8 Sec (fig 3.9 A) and a probability of occurrence of 1/2, 1/8, 1/8, 1/8, and 1/8 respectively (fig. 3.9 B). In these experiments the signal-to-noise ratio is above 30 dB (fig. 3.9 C, D). High learning rates are used, $\alpha = 0.05$, $\beta = 0.4$, $\eta = 0.1$, in order to demonstrate the ability of the learning mechanism to learn with few stimulus presentations.

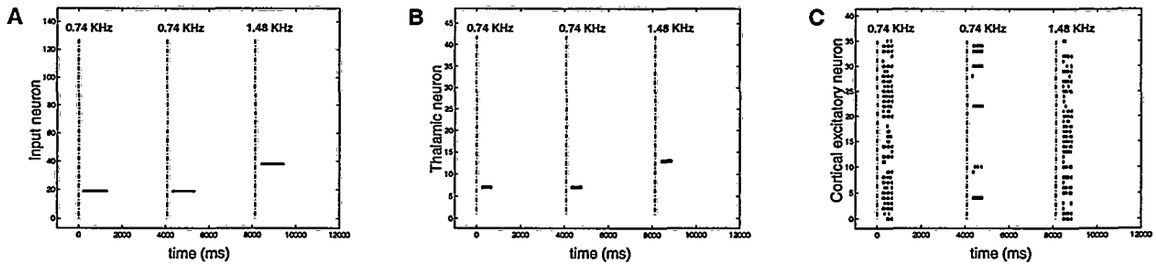

Figure 3.10: Raster of network activity responding to 3 tones in sequence (0.74 kHz, 0.74 kHz, 1.48 kHz). Time zero corresponds to the onset of the first tone. Vertical dashed lines represent the onset of each tone. **A**: input population. **B**: thalamic population. **C**: cortical excitatory population.

Figure 3.10 shows a typical example of the responses in the network after the presentation of the sequence 0.74, 0.74, 1.48 kHz. When a tone is presented, typically

1-3 neurons fire in the input population and 1-2 neurons in the thalamic population, depending on the intensity of the sound. As observed in the first experiment, nearly all cortical excitatory neurons respond initially (fig. 3.10 C, 0-1000 ms; fig. 3.12 A) to a novel stimulus. However, after a few presentations, the number of neurons which respond to this stimulus stabilizes (fig. 3.12 A).

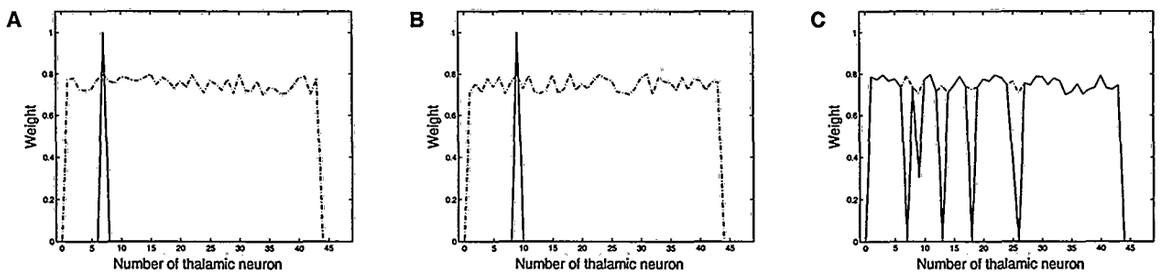

Figure 3.11: Initial (dashed line) and final (solid line) receptive fields of some neurons. **A**: neuron finally selective to the 0.74 kHz tone. **B**: neuron finally selective to the 1.05 kHz tone. **C**: neuron that finally does not respond to any of the 5 tones used in the training. The final receptive field of each neuron is either selective to one tone (A, B) or insensitive to any tone used in the training (C).

The developed receptive fields are specific: a neuron that responds to one tone does not respond to any of the others (fig. 3.11 A, B). Furthermore, the size of the representation of each tone, i.e. the number of neurons responding to it, does not depend on its probability of occurrence (fig. 3.12 A, B).

These results demonstrate that the learning rule is robust and can handle inhomogeneities in the occurrence of different stimuli. In addition, it shows the ability of the network for dynamic recruitment [Körding & König, 2000]. That is, those neurons that do not develop specific receptive fields remain "unspecific", while loosing any sensitivity to frequencies represented by other neurons in the population (fig. 3.11 C). These unspecific neurons can be activated by novel tones and develop receptive fields specific to them. Hence, the network has the ability to "reserve" neurons for representing future novel stimuli.

As a next step, comparable to recent physiological experiments done by Kilgard and Merzenich [Kilgard & Merzenich, 1998], we pair one of the rare stimuli (2.09 kHz) with the activation of the basal forebrain unit. Basal forebrain stimulation

occurs simultaneously with stimulus onset. After a paired presentation 22 neurons develop specific receptive fields to this tone (their receptive field are similar to those in fig. 3.11 A and B, data not shown). The number of neurons responding to this tone is stable since in the following paired presentations this number does not decrease (fig. 3.12 C). Therefore we see that the size of representation of this rare tone is much increased compared to the previous experiment where the basal forebrain remained inactive (fig. 3.12 D). In that experiment the representation size of this stimulus stabilizes after several presentations in 4 neurons (fig. 3.12 B).

The basal forebrain input hyperpolarizes the cortical inhibitory neurons, delaying their activity with respect to the cortical excitatory neurons by about 6 ms and, thus, enlarging the temporal window for the backpropagating action potential to induce the potentiation of synaptic efficacies. This results in an increase in the representation size of this stimulus. This effect is independent on the presentation frequency of the stimulus (data not shown) and does not affect the size of the representation of the other stimuli. With no basal forebrain activation, the final number of specific neurons responding to the tones 0.74, 1.05, 1.48, 2.09, and 2.96 kHz is 2, 3, 2, 3 and 3 neurons respectively (fig. 3.12 B). In the experiment where the basal forebrain is paired with the 2.09 kHz tone, the number of specific neurons responding to the tones is 2, 4, 3, 22 and 4 neurons respectively (fig. 3.12 D).

When pairing is discontinued after presentation 22, the size of the representation of the previously paired tone is reduced and reaches a size comparable with the representation of the other tones (2, 3, 2, 2, and 2 neurons respectively). Thus, the learning rule dynamically modifies the size of representation of the stimuli according to their behavioral importance, represented by the level of activity in the basal forebrain. This effect is independent on the probability of occurrence of the stimuli (fig. 3.12 D). In addition, the dynamic modification does not affect the representations of other stimuli (fig. 3.12 D).

Learning in the presence of acoustic noise

As an additional control we investigate the properties of the proposed learning rule using stimuli with acoustic noise of greater amplitude in a non-overlapping frequency

band. We use the same protocol as in the previous experiment (see figs. 3.13 A, B), while a continuous low-band noise is played by the synthesizer. The global signal-to-noise ratio of all the stimuli is close to 1 (fig. 3.13 C, D).

The noise continuously excites the input neurons corresponding to the lowest frequencies (fig. 3.14 A). This in turn drives the thalamic neurons tuned to low frequencies leading to a response of all cortical excitatory neurons at the first presentation due to their initially diffuse receptive fields. After a few seconds, however, the efficacy of the synapses from the input population to the thalamic population, which transduce the presented frequencies, diminishes due to their short-term depression. This prevents continuously present harmonics from further activating the thalamic and cortical populations.

However, as shown in Fig. 3.14 B not all aspects of the continuously presented stimulus are filtered out. This is due to fluctuations in harmonics which have a small contribution to the signal and are not filtered out by the short-term depressing synapses. The weak contribution of these harmonics makes the corresponding input neuron fire at a low firing rate. As a result fluctuations in the harmonics of the noise are processed by the cortical network, mixed with the information about the tones presented to the system. Furthermore, these fluctuations in the noise can activate those input neurons that are activated when the 0.74 kHz tone is presented (fig. 3.14). Therefore we see that the noise overlaps with the signal both temporally and spatially.

Hence, one would expect that the continuously presented noise would interfere with the development of receptive fields specific to the tones. However, those thalamocortical synapses that transduce information about the noise tend to get weaker. In appendix C.1 we show analytically that the learning rule decorrelates signals that are independent, in this case the fluctuations in the spectrum of the noise and the tones played by the synthesizer. Therefore, this learning mechanism decorrelates the noise from the receptive fields of the cortical excitatory cells sensitive to tones. Effectively, we see in (fig. 3.15 A) that the receptive fields of the neurons that fire to the tones are decorrelated from noise.

A few neurons develop receptive fields specific to frequencies that are part of the noise: two are finally selective to frequencies lower than 0.7 kHz and one is selective

to 0.90 kHz (fig. 3.16). These neurons, however, do not respond to any of the tones (fig. 3.15 B). Finally, the remaining neurons do not respond to either the tones or the noise, remaining "unspecific" (fig. 3.15 C).

This ensures the ability of the network to learn future tones. In conclusion, the system proved to be robust against high noise levels and the results obtained are similar to those without noise (fig. 3.16).

3.4.3 Analysis of the model using Fisher information

In this section we will analyze our model of the auditory system using tools from information theory. We will see that a concept of *behavioral meaning of the stimuli* needs to be included in order to explain the experiments. Then with this notion the predictions using information theory are in accord with the biological experiments and our model results.

Mathematical model

As in chapter 2, we will use Fisher Information to calculate the configuration of receptive fields which maximizes the detection performance of the system. We will model the spectral content of the stimulus as a vector of N components $\vec{s}$, each one representing the amount of signal in a certain frequency band (for example, $[\vec{s}]_1$ is the amount of signal in the 0-10 Hz band, $[\vec{s}]_2$ represents the 10-20 Hz band and so on). Note that with this representation we account for both the amplitude of the sound as well as its spectral content. In addition, this choice accounts for the case when several simultaneous stimulus are present (which can not be described by just a scalar).

Now we model the response of the *ith* neuron of A1 to the stimulus $\vec{s}$. For simplicity, we consider this as linear:

$$r_i = \vec{a}_i^T \vec{s} + b_i + \eta_i, \quad i = 1 \ldots R \tag{3.7}$$

where $\vec{a}_i$ is the vector of sensitivities of the neuron to the different frequency bands, b_i is the bias of the neuron, η_i is its noise, and R is the number of neurons in

the population. Note that the linear simplification is saying that the output of the neuron is scaled with the stimulus intensity. We are not imposing anything about the shape of the RF, which is determined by the tuning vector $\vec{a}_i$. Finally, we make the approximation that the noise in the neurons is gaussian and independent.

Analogously to section 2.3.2, the Fisher Information Matrix of the system can be calculated as:

$$J = \sum_{i=1}^{R} \frac{1}{\sigma^2} \vec{a}_i \vec{a}_i^T \tag{3.8}$$

where σ^2 is the noise variance in the neurons.

As we showed in section 2.3.2 the performance of any unbiased estimator is limited by the Cramér-Rao bound:

$$var(\hat{\vec{s}}|\vec{s}) \geq tr(J^{-1}) \tag{3.9}$$

Our goal is then to find the set of receptive fields which maximize the trace of J^{-1} which is directly related to the detection performance of the system (see section 2.3.2 for details).

Optimization Methods

The set of receptive fields was optimized using a genetic algorithm. We have chosen this method in order to avoid local minima since the dependency of the Fisher Information on the RFs is non-linear. With no additional constraints the optimization problem is ill-defined since the trace of J^{-1} scales down as the norm of $\vec{a}_i$ increases and thus no global maximum exists. Analogously to section 2.3.2 we need to put additional constraints in order to define the problem properly.

The neurons of the auditory thalamus are tuned to specific frequencies, responding maximally when the preferred frequency is present. Since the auditory cortex receives excitatory connections from these neurons we can assume $a_{ij} \geq 0$. On the other hand the synapses that converge into a cortical neuron are in continuous competition so that just a set of them are potentiated while the others are depressed, as we have

seen in the previous section. We can take into account this situation introducing the restriction $|\vec{a}_i| \leq c$.

We will consider a population of 100 neurons and an input dimension of 10 (number of frequency bands). This arbitrary election has been done based on computational efficiency reasons since these parameters are not critic for the results we will show.

Optimal configuration when the modulatory signals are not taken into account

At the end of the optimization process the receptive field of each neuron, given by $\vec{a}_i$, converges so that only one of its components is not null (figure 3.17). The sensitivity of this component acquires the maximum value within the imposed constraints. Therefore, each neuron specializes to a certain frequency band (figure 3.17). The number of neurons per each frequency band is constant in the optimal configuration. That is, the representation of these frequency bands is homogeneously distributed along the population (figure 3.18 A). This situation corresponds to section 3.4.2 where we observed the same distribution of receptive fields. Therefore we conclude that the plasticity mechanisms which give rise to this internal representation are maximizing the representational accuracy of the stimuli.

Expanding the Fisher Information concept to incorporate the "behavioral meaning" of the stimuli

The maximization of Fisher information in our model of auditory system leads to an homogeneous internal representation independently on the input statistics of the stimuli, in correspondence with the biological system [Kilgard & Merzenich, 1998] and the results of the realistic model. This is interesting since in preliminary work we have observed that the maximization of other theoretical measures such as mutual information leads to configurations which depend on the stimuli statistics.

However we have seen that in situations where a stimulus is remarkably important the internal representation is biased to it (section 3.4.2). This seems intuitive since

the amount of resources in a biological system are limited and therefore it should assign to each stimulus an amount corresponding to its importance. How can we account for this feature in our theoretical analysis ?

Let us make the simplification that different frequency bands code different information. This is true in the experimental situation we are modeling [Weinberger, 1993, Kilgard & Merzenich, 1998]. We introduce the cost of the error in the estimation of frequency band i as:

$$k_i(s_i - \hat{s}_i)^2 \tag{3.10}$$

where s_i is the ith component of $\vec{s}$ (which corresponds to the ith frequency band) and k_i is a positive number which models "how important" is a good estimation of that frequency band.

Then the expected cost of estimation of band i, given that stimulus $\vec{s}$ is presented, is:

$$C_i(\vec{s}) = k_i \left\langle (s_i - \hat{s}_i)^2 \mid \vec{s} \right\rangle \tag{3.11}$$

Thus the total expected cost of the system given $\vec{s}$ can be written as:

$$C(\vec{s}) = \sum_{i=1}^{N} C_i(\vec{s}) = \sum_{i=1}^{N} k_i \left\langle (s_i - \hat{s}_i)^2 \mid \vec{s} \right\rangle = \sum_{i=1}^{N} k_i \cdot var(\hat{s}_i|\vec{s}) \tag{3.12}$$

Making use of the Cramér-Rao bound for separated components (eq. 2.3), we obtain that for unbiased estimators $C(\vec{s})$ satisfies the unequality:

$$C(\vec{s}) \geq \sum_{i=1}^{N} k_i \left[J^{-1}(\vec{s})\right]_{ii} \tag{3.13}$$

Note that for the special case $k_i = c$ (all the frequency bands are equally important) we have an expression proportional to 3.9. Thus the optimization of the original equation 3.9 can be seen as the optimization of detection performance when all the frequency bands are equally important. This leads to an optimal solution where the RFs are homogeneously distributed over the frequency bands.

On the other hand, when one frequency band is more important than the others

the optimal representation is not homogeneous. In figure 3.18 we show the optimal configuration obtained by minimization of 3.13 when the 8th frequency band is more important than the others.

In general, the level of representation of a given frequency band increases as its relative importance respect to the others increases. This is interesting since in our model of spiking neurons we observe the same when the level of activity in the basal forebrain for a stimulus is higher than for the others. This parallelism between our theoretical analysis and the realistic neural model allows us to interpret the level of activity in the basal forebrain from the behavioral point of view as the "cost" that it has for the animal to make a bad estimation of the stimulus.

3.5 Conclusions

We have analyzed two different sensory systems using realistic models with similar dynamics. Both of them show the emergence of realistic sensory maps as a consequence of self-organization (visual cortex model) or experience (auditory cortex model). We have shown that one of the basic ingredients in the development of these internal representations is the interaction between mechanisms of competition at several levels: inhibition between far neighbors, competition cause-effect at the level of the synapses and competition between potential learners based on learning modulation.

The spike-timing dependent plasticity implements a competition mechanism for the synapses that converge into the same neuron. Synapses whose activity occurs immediately before the activity of the neuron are potentiated while the others are depressed. Therefore only those synapses which have contributed effectively to the activation of the neuron are potentiated. This mechanism is the responsible of the emergence of directional selectivity in the visual cortex model as it was explained. On the other hand, the heterosynaptic LTD plasticity allows the competition between input coming from different retinas that finally gives rise to the ocular dominance organization. Finally, the topological structure of the network appears as a result of the spatial correlation induced by the activity patterns, and from the competition between far neurons through inhibition. We have made more experiments where we

use a high learning rate, resulting in an abrupt topology (data not shown).

In the model of auditory cortex the mechanisms of neural competition make the neurons develop receptive fields very rapidly while remaining stable: neurons that respond faster to the stimulus show rapid acquisition, while neurons responding late will suffer strong depression of their activated sensory synapses extinguishing their response to a future presentation of the stimulus. Another consequence of this competition is that the receptive fields of the neurons tend to be non-overlapping. Using low learning rates, however, would diminish the 'average competition' allowing receptive fields to overlap. The details of this process depend on the details of the stimulus statistics.

If stimuli are not presented alone but mixed in different combinations (e.g. by using a typical music CD) the system achieves a 'sparse' representation of the environment that minimizes the redundancy while covering the complete stimulus space. Interestingly, this is the type of representation that the visual cortex seems to use [Olshausen & Field, 1996], having the advantages of minimizing the energy consumed [Baddeley, 1996] while minimizing the reconstruction error [Olshausen & Field, 1996]. In addition, it is important to obtain low-redundancy codes ('minimum entropy codes') in order to make the processing by higher stages as simple as possible [Barlow, 1989].

However, in both models of visual and auditory cortices there is not just a neuron which codes a stimulus feature but a group of neurons. In other words, the mechanisms of competition and cooperation involve the simultaneous interaction of many neurons, which results in a representation of the information by groups of neurons. Neurons corresponding to the same group code similar aspects of the information, while neurons of different groups code different aspects. We can describe this situation as a complexity reduction in the representation of the information, where each different representation symbol is the activity of a group of neurons. This conclusion is in accord with the observations in the model of the olfactory epithelium. Therefore we can not talk about redundancy minimization between individual neurons but redundancy minimization between *functional groups of neurons.*

On the other hand the model of auditory cortex creates an internal representation

of the information which depends on the task to perform: the stimuli which have a behavioral meaning (correlation with "pain") are much more represented (more neurons process them) than the neutral ones. This same property is observed in the biological system [Kilgard & Merzenich, 1998, Weinberger, 1993]. This will increase the signal-to-noise ratio for those stimuli, resulting in a better detection. On the other hand, the model predicts that if a stimulus changes from "important" to "neutral", its internal representation will change accordingly, so that the number of neurons which will process it will decrease. Moreover, if the stimulus statistics changes so that a new stimulus not present before (and therefore not represented) occurs, then the system will assign new resources to it. That is, in the primary auditory system the extraction of information and its optimal representation depends on two different aspects: 1) the intrinsic structure of the information, and 2) the task the animal has to perform with that information.

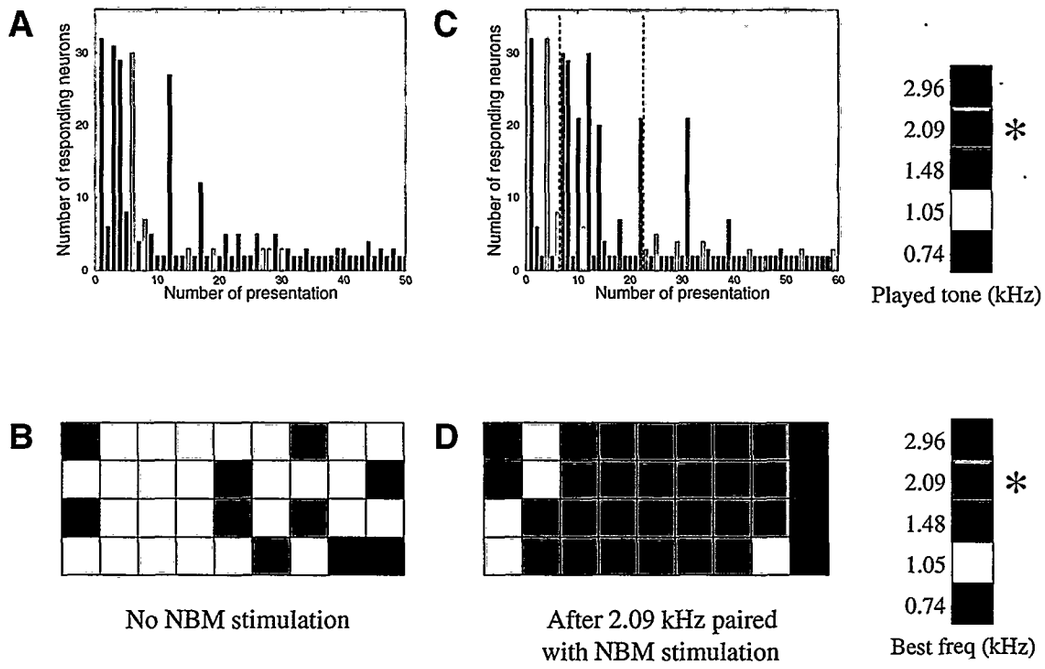

Figure 3.12: Response of the cortical excitatory neurons during training. **A**: Number of neurons responding to each tone in a pseudo-random sequence of consecutive presentations. In each presentation a tone is randomly chosen from the set (0.74, 1.05, 1.48, 2.09 and 2.96 KHz) with a probability of (1/2, 1/8, 1/8, 1/8, 1/8) respectively. The color of a bar indicates which tone is presented and its height represents the number of neurons which respond to it. **B**: distribution of the preferred frequency of the 36 cortical excitatory neurons after training, without basal forebrain activity. Each square corresponds to a neuron. Color indicates the preferred stimulus frequency. For better visibility, the neurons are arranged in 4 rows in order of increasing preferred stimulus frequency. Neurons marked in white are not selective to any of the used tones. Thus, this representation might be compared to a top view onto the primary auditory cortex as used by [Kilgard & Merzenich, 1998]. **C**: same as A, but now one of the rare stimuli (2.09 KHz) is paired with basal forebrain activity (both the start and ending of the pairing phase are indicated by vertical dashed lines). **D**: same as B, for the experiment described in panel C. The receptive fields are measured after the last paired presentation (presentation 22).

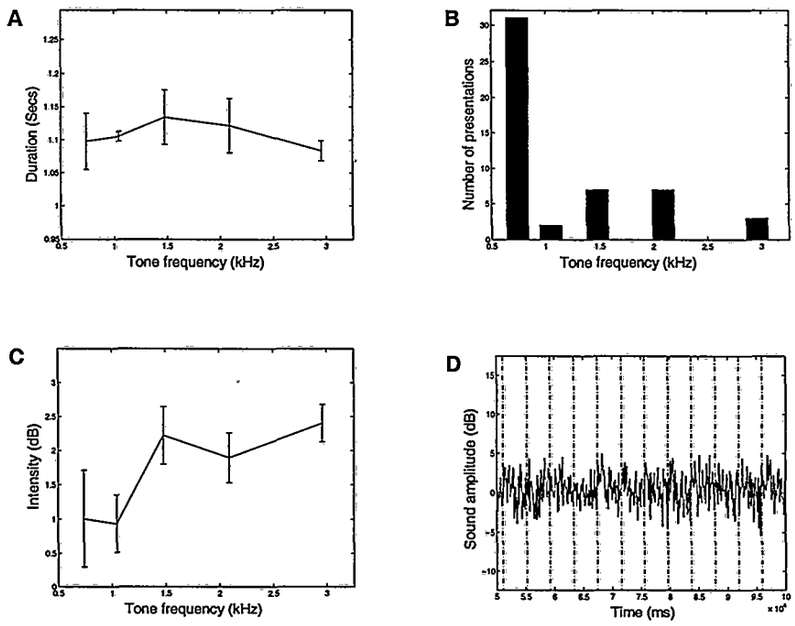

Figure 3.13: Stimulus statistics using a pseudo-random sequence of 5 different tones (0.74, 1.05, 1.48, 2.09 and 2.96 KHz) and very loud noise as background. The probability of occurrence is 1/2, 1/8, 1/8, 1/8 and 1/8 respectively. **A**: mean duration of each stimulus. **B**: number of presentations of each stimulus. **C**: mean intensity of each stimulus. The 0 dB level is chosen as the averaged level of noise in the room. **D**: sound amplitude over time.

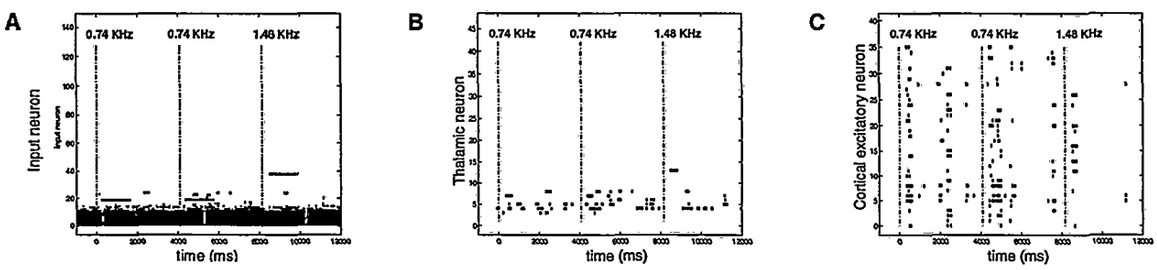

Figure 3.14: Raster of network activity following the presentation of three tones (0.74 kHz, 0.74 kHz, 1.48 kHz) with very loud noise as background. Time zero corresponds to the onset of the first tone. Vertical dashed lines represent the onset of each tone. **A**: input population. **B**: thalamic population. **C**: cortical excitatory population.

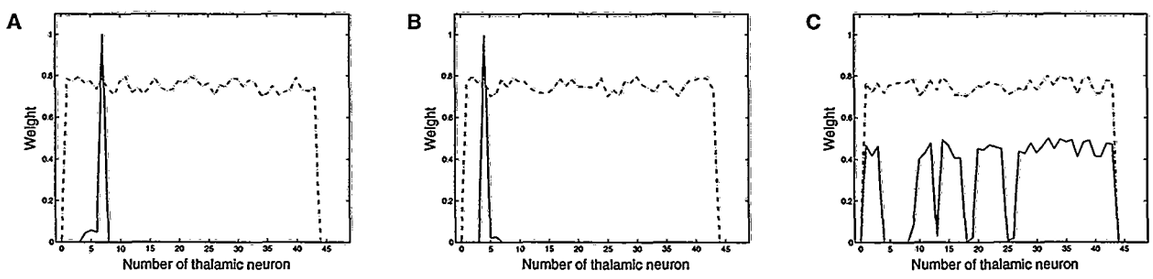

Figure 3.15: Initial (dashed line) and final (solid line) receptive fields of selected neurons in the experiment with very loud noise as background. **A**: neuron finally selective to the 0.74 kHz tone. **B**: neuron finally selective to the 0.40 kHz component of the noise. **C**: neuron that finally does not respond to any of the 5 tones used in the training.

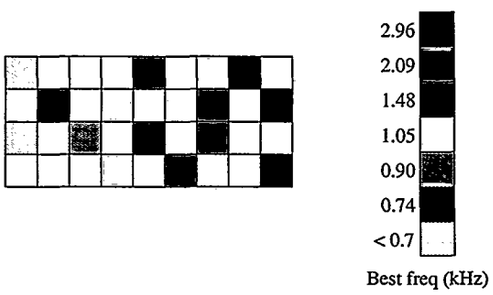

Figure 3.16: Distribution of the preferred frequency of the 36 cortical excitatory neurons after training with a sequence of tones with very loud noise as background using the same convention as in Figure 6 B, D. Each tone in the sequence is randomly chosen from the set (0.74, 1.05, 1.48, 2.09 and 2.96 KHz) with a probability of (1/2, 1/8, 1/8, 1/8, 1/8) respectively. The displayed receptive fields were stable and resulted after 50 presentations. Neurons marked in gray and gold are selective to frequencies which are part of the noise: gray indicates a preferred frequency lower than 0.7 kHz, and gold indicates a preferred frequency of 0.90 kHz.

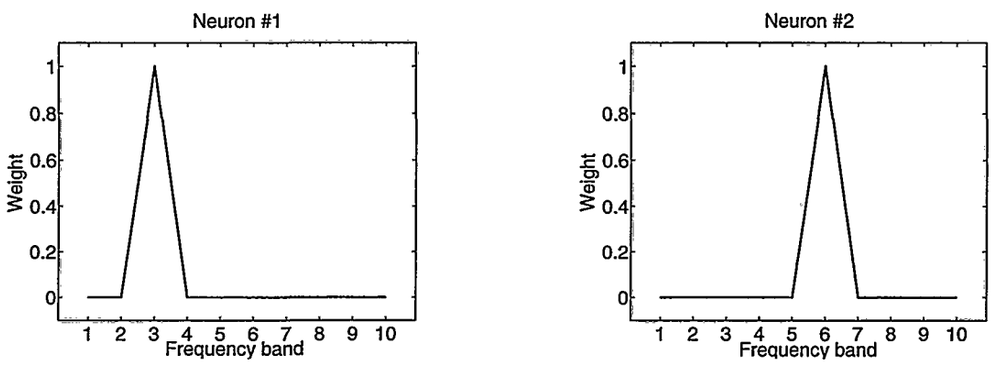

Figure 3.17: Receptive fields of two neurons after maximization of Fisher Information

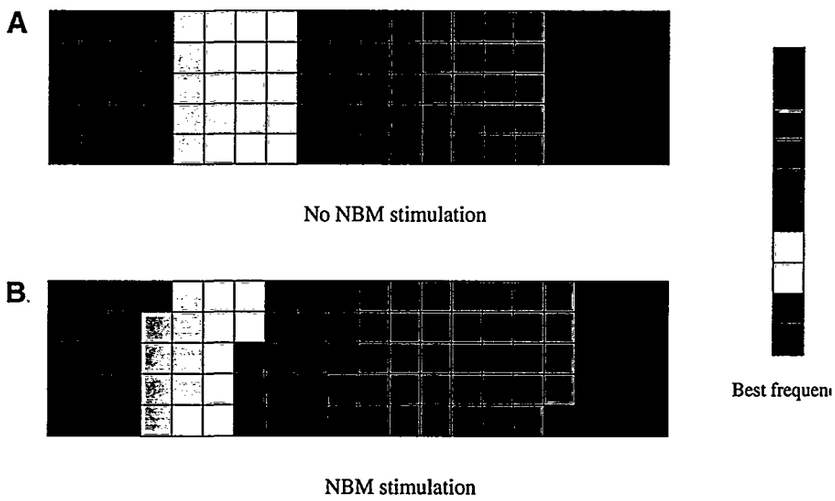

Figure 3.18: RF distribution of an array of neurons obtained through Fisher Information maximization. **A**: all the frequency bands have similar behavioral importance. **B**: the 8th frequency band is more important than the others.

Chapter 4

General conclusions obtained from the biological models

4.1 Conclusions

4.1.1 Different encoding strategies

As we have shown the olfactory epithelium is mainly composed by receptors with broad tuning curves, that is, they respond to a large variety of different stimuli. Therefore the information about an odor is represented not by a particular neuron but by the simultaneous activity of many different neurons. We demonstrated that, given the biological constraints, this scheme is optimal from an information transfer point of view. This form of representing the information contrasts with the internal representations developed by the visual and auditory cortex models.

In the visual system model the mechanisms of competition lead to a representation where each neuron focus on a particular aspect of the stimulus. The features coded by neurons which are located in far positions tend to be different and therefore their activity tend to be uncorrelated. This strategy constitutes a *sparse code* where the image is represented by only a few active cells out of a potentially much higher number [Olshausen & Field, 1996, Baddeley, 1996]. This strategy is efficient in the sense that it minimizes the complexity and energetic

cost of the code while maximizing the representational accuracy for natural images [Olshausen & Field, 1996, Baddeley, 1996]. In Barlow's terms [Barlow, 1989], it constitutes a minimum-entropy code. It is important to remark that this optimality is achieved under the assumption that the decoding of the image by further processing steps of the system is linear [Olshausen & Field, 1996].

The results we obtain in the auditory model when it is trained with real stimuli are similar to the visual system: the mechanisms of learning through competition lead to a representation which covers the whole stimulus space while maximizing the spareness. Thus the internal representations in the visual and auditory cortex are *population codes* in the sense that the stimulus is described by separated neurons which code different features in parallel different, in analogy with the olfactory epithelium. On the other hand, the representation is a sparse code in the sense that neurons that code different features tend to be uncorrelated along different stimuli [Olshausen & Field, 1996], in contrast with the olfactory epithelium.

As an extreme case of sparse coding there are neural systems where the neurons respond to very specific stimuli. This is the case of the neurons of the inferotemporal cortex which are tuned to views of complex objects such as faces, being very weakly activated by other objects [Bruce *et al.*, 1981, Rolls, 2000] (cf. *grandmother cells*).

If we want to develop a general framework applicable to systems which use different strategies for the representation of the information, the theory should be able to cope with these different situations. How could it do that ? The statistics of the internal states of the system and their correlation with the different stimuli is the key ingredient that determines the performance of the whole system, no matter if in each internal state many neurons are active or not (implementation specificities). Thus we introduce the following concept:

- *We call "structural property" a feature that depends on the statistics of the states of the system but not on the implementation details*

For instance, a population of neurons with identical response pattern are structurally equivalent to just one neuron since the number of global states is the same as in a single neuron.

Then the desired theory in order to be general should work with structural properties of the system and not on the implementation specificities.

- *the desired general framework must depend on measurable quantities that do not depend on the specific system implementation. That is, they should depend on the structural properties of the system but not on implementation details.*

Considerations such as spareness [Olshausen & Field, 1996] and factoriality of the code [Barlow, 1989] depend on the particularities of the implementation such as consumed energy and should appear as particular requirements for each different system.

Interestingly, there exist concepts in information theory such as "entropy", "Bayes error" and "Fisher information" which allow us to ask and study global aspects of the system without the need of referring to the physical implementation of the system [Cover & Thomas, 1991]. That is, they describe structural properties of the system. *These are the kind of tools we need if we want to have a theoretical framework of information processing that does not depend on the particular implementation of the system.*

4.1.2 Is maximum information transfer a general principle of organization for adaptive systems ?

Let us consider a system which transforms an input x into y. In section 2.5.1 we defined the *principle of maximum information transfer* as the maximization of the information contained in y about x. We showed that the first stage of the olfactory system seems to follow this principle since the theoretical configuration which maximizes the information about the stimuli has very similar properties to the real system. A variation of this principle which also maximizes code sparseness explains the receptive fields of simple neurons in the primary visual cortex [Olshausen & Field, 1996]. It is intuitive that these systems seem to maximize the information transfer since they are the first sensory stages of the organism and we would expect them to communicate as much information as possible to the rest of the system.
Is this principle of maximum information transfer valid for any processing stage of the biological system ?

In the auditory cortex model we showed that the neural model creates an internal representation of the information which depends on the task to perform: the stimuli which have a behavioral meaning (correlation with "pain") are much more represented (more neurons process them) than the neutral ones. This same property is observed in the biological system [Weinberger, 1993, Kilgard & Merzenich, 1998]. Therefore even at primary stages of processing the auditory system does not act as a mere communicator but as an active agent which processes the different aspects of the stimuli correspondingly with their relation to the task.

4.1.3 The concept of task and its implications

The model of the auditory cortex predicts that if an auditory stimulus changes from "important" to "neutral", its internal representation will change accordingly, so that the number of neurons which will process it will decrease. Moreover, if the stimulus statistics changes so that a new stimulus not present before (and therefore not represented) occurs, then the system will assign new resources to it. That is, in this system the internal representation depends on two different aspects: 1) the intrinsic structure of the information, and 2) the task the animal has to perform with that information.

All this guide us to the following conclusion: the internal representation of information in a biological system is not static, but can change with time, adapting to the different tasks the environment imposes to the animal. This seems intuitive since the resolution of a given problem depends crucially on an adequate representation of the information. If we choose it correctly, the resolution of the task will be easy in that representation. In general we can then say that learning of a new task = learning of the optimal information representation + learning of the solution of the problem expressed in the new coordinates. Thus we require for our new framework *to take into account the task(s) to be solved by the system.*

4.1.4 The concept of complexity reduction

We have seen that the representation of information in the visual cortex is qualitatively different than in previous processing stages (retina and geniculate nuclei). Neurons have more elaborated receptive fields, responding to higher level features such as orientation and motion direction. These features are the basic primitives from which higher concepts such as "contour" and "form" are posteriorly elaborated.

In general, higher processing steps should filter those aspects of the sensory information which are not important and thus may represent noise, while focusing on the relevant parts of the information. Hence we introduce the concept of *complexity reduction* as the filtering of the spurious part of the information, which makes it statistically simpler.

However the notion of "what is important" depends on the task at hand as we have seen before. Therefore we conclude that *an optimal processing system should reduce the complexity of the information it receives while preserving the aspects related to the task.*

4.2 Towards a general theory: necessary concepts

In order to generalize the ideas introduced previously to any complex adaptive system we will introduce the following concepts which we will develop in the next chapter:

The concepts of agent and environment

An autonomous agent is a system which inhabits a dynamic, unpredictable environment in which it tries to satisfy a set of time-dependent goals or motivations [Maes, 1994]. The agent makes actions which affect the environment and vice versa. Note that the concept of agent is thus very general and contains very different situations, for example an animal, a robot which interacts with its physical environment, an adaptive artificial algorithm which tries to minimize the classification error in a data set, etc.

The concept of global task of the agent

The global task of the agent consists in learning how to interact with the environment in order to satisfy its time-dependent goals or motivations. This can be formalized as the maximization of certain goal function which can change in time.

Optimality of a part of the agent in the context of the global task

In order to reach its goals efficiently, the subsystems which compound the agent must also perform optimally. For instance, in a pattern recognition problem it is critical the way the information is preprocessed and internally represented at the different stages of processing.

The concept of structural property

In section 4.1.1 we defined a *structural property* a feature of the system that depends on the statistics of its global states but not on the implementation details. Since we want our general theory to be implementation independent, it should be expressed in terms of structural properties.

Part III

Formal analysis of complex adaptive systems

Chapter 5

Theoretical framework

5.1 The problem of information processing in an autonomous system

5.1.1 The general model of an adaptive autonomous system

We consider an agent that interacts in an environment [Shen, 1994, Maes, 1994], (see figure 5.1). The internal state of the agent at time t is $\Phi(t)$. The agent can interact physically with the environment. We call $\vec{a}(t)$ the vector of "actions" of the agent on the environment at time t. By action we mean *any* interaction that changes the state of the environment (e.g. forces, transmission of signals). Conversely, the environment can interact with the agent affecting directly or indirectly its internal state. We call these signals $\vec{u}$.

The dynamics of the agent is thus given by the equation $\Phi(t+1) = m(\Phi(t), \vec{u}(t))$ where m is a possibly stochastic function. Finally, we call $\mathcal{T}(t)$ the state of the total system (agent + environment).

The objective of the agent is to interact with the environment so that a particular "goodness" function $\mathcal{H}$ is optimized. In general, this function depends on the global state of the system as well as on how that state has been achieved (speed for example). If this is the case, in order to describe this situation we only need to expand the state vector $\mathcal{T}$ with additional states which take into account the path followed by other

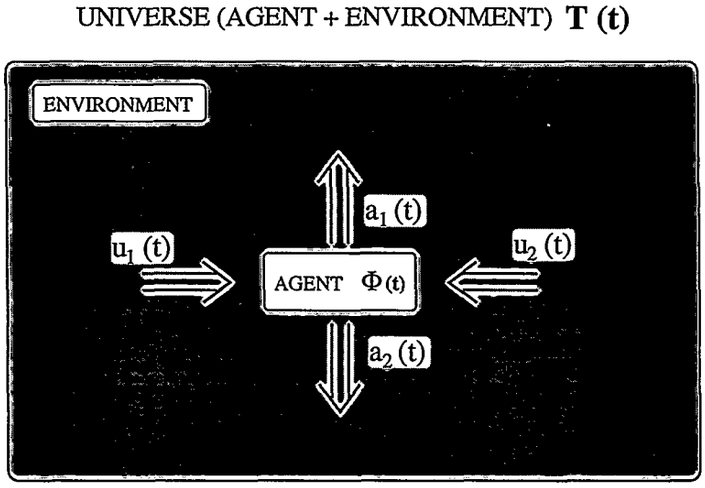

Figure 5.1: An agent interacts with its environment in order to maximize a global goal. The global state of the universe (agent + environment) at time t is represented by the the state vector $\mathcal{T}(t)$. The variables which define the internal state of the agent form the state vector $\Phi(t)$. In order to achieve its goal, the agent performs actions a on the environment. In turn, the environment can act on the agent modifying its internal state through the actions u.

states. That is, without loss of generality we can consider $\mathcal{H} = \mathcal{H}(\mathcal{T})$. In some specific situations $\mathcal{H}$ depends on just the internal state of the agent Φ. Notice that the form of the dependency $\mathcal{H}(\mathcal{T})$ may change with time (dynamic environments with changing objectives). For each particular problem, we assume there is a unique optimal action g for each different state $\mathcal{T}$ so that $\mathcal{H}$ is globally optimized.

The *policy* of the agent is the function $\mathcal{P} : \Phi \rightarrow a$ that describes which action the agent performs in each situation. This could be a stochastic function, which only depends on Φ since the agent's decision depends only on local information. The explicit task of the agent consists in performing the optimal action $g(\mathcal{T})$ corresponding to the current state $\mathcal{T}$.

5.1.2 The problem of learning in an autonomous system

The explicit task of the agent consists in performing the optimal action $g(\mathcal{T})$ corresponding to the current state $\mathcal{T}$. In order to do this, the system must detect and process adequately the information in the environment that is relevant for its task, which we have shown to be implicitly determined by the function $\mathcal{H}$ to optimize. If the agent captures the important aspects and regularities of the environment-agent dynamics related to the task, it will perform and generalize adequately. Therefore, the agent must learn two different mappings: on one hand m, which is the internal state dynamics, and the internal policy $\mathcal{P}$. Both together will determine the optimal interaction of the agent with its environment and should be learned simultaneously.

The focus of this thesis is the definition of a new information processing measure that allows to measure the optimality of m for a given task. Preliminary results have been presented elsewhere [Sánchez-Montanés & Corbacho, 2002, Sánchez-Montanés & Corbacho, 2003]. Let us consider the diagram in figure 5.2. The vectors x and y represent all the information that S_2 receives and sends respectively. Note that since the dynamics of S_2 is arbitrary, memories and internal recurrences are also considered. We want to measure the amount of information processing that the sub-system S_2 performs, given that it transforms x into y and the explicit goal of the overall system is g. We shall denote this amount of information processing by $\Delta P(x \rightarrow y|g)$.

Since we want an implementation independent processing measure, we will analyze information theory since it has been successful in the design of information measures that are implementation independent.

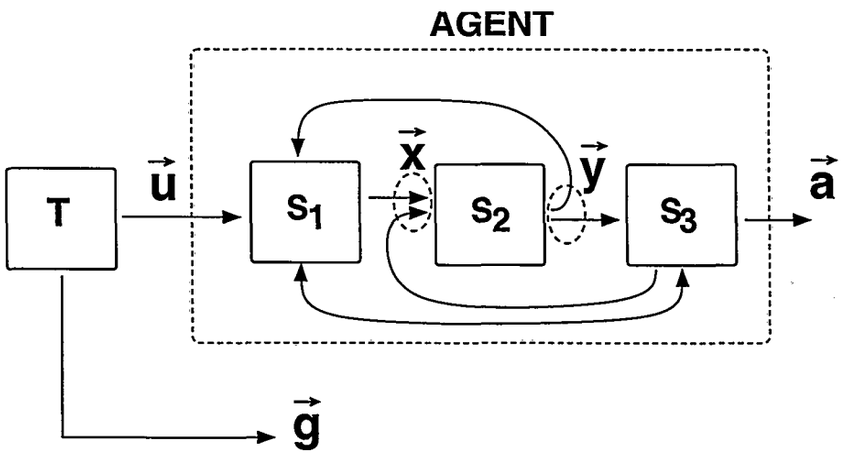

Figure 5.2: Schema of the agent. The explicit task of the agent is to perform the optimal action g for each state $\mathcal{T}$. We want to measure the amount of information processed by a given part S_2 of the agent which receives $\vec{x}$ and sends $\vec{y}$. Note that $\vec{x}$ contains all the information that S_2 receives from the rest of the system. Direct transmission of $\vec{u}$ to S_2 is modeled as a bypass of information though S_1. The vector $\vec{y}$ contains all the information that S_2 sends to the rest of the system. $\vec{x}$, $\vec{y}$ and $\vec{g}$ represent any kind of data: static/temporal, symbolic/numerical, scalar/vectorial, etc.

5.2 Communication versus Information Processing

5.2.1 The three different levels of communication in the system as defined by Weaver

Communication can be defined in a broad sense as the procedure by which a physical (part of a) system A affects another B. Accordingly, Weaver distinguishes three different levels of communication [Shannon & Weaver, 1949]:

1. Technical problems: how accurately can the symbols of communication be transmitted?

2. Semantic problems: how precisely do the transmitted symbols convey the desired meaning?

3. Effectiveness problems: how effectively does the received meaning affect the receiver's conduct in the desired way ?

From this point of view, the problem of any autonomous agent living in an active environment can be seen as a mixture of these three different levels: the sensors as well as interconnected parts of the agent must communicate precisely the information whatever the implementation is, the receptors of that information (actuators, preprocessing steps, etc.) must interpret correctly the signals they receive, and finally the actions the agent does on the environment can be seen as signals communicated in order to affect it in the desired way.

As Weaver [Shannon & Weaver, 1949] already pointed out, classical information theory deals mainly with the technical problem. In the classical approach the amount of information decreases when it undergoes any processing [Cover & Thomas, 1991], that is, processing is passive instead of being active (i.e. elaborating the data and approaching the goal hence, posing a paradox for an information processing system). In this regard a perfect communication channel has maximal mutual information yet minimal information processing.

We claim that even nowadays a shift of view is necessary to take into proper consideration within the information theoretical framework the other problem levels such as the knowledge of the receiver. This thesis provides a step towards dealing with the semantic and the effectiveness problems by making optimal coding depend on the specific task(s) to be solved by the system as well as on the specific knowledge about the environment/receiver in order to affect it in the desired way.

5.2.2 Structural uncertainty and spurious information about the task

In this section we will introduce two concepts which will be useful for the development of our theory. We will show that these two concepts are actually the two sides of the

same coin.

Let us consider the four systems depicted in Figure 5.3. Each quadrant of the

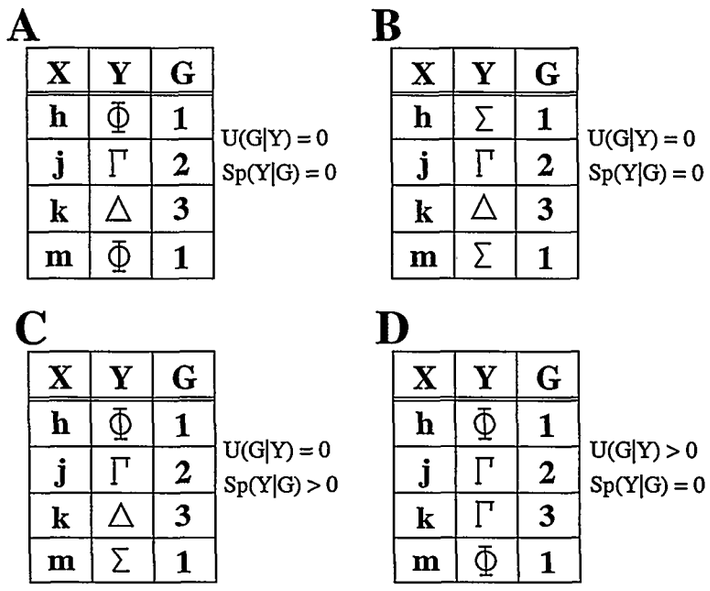

Figure 5.3: State tables for four different systems, A, B, C, D. X represents the input space, Y represents the internal space of representation of the system and G represents the goal space.

figure represents the diagram with the states of a system which transforms x in y with the goal of achieving G. It is easy to see that system A has 0 uncertainty about G since given Y the goal G is completely characterized. Moreover, the output y of the system has 0 amount of spurious information about G since each goal is associated with only one y. Formally we will write $U(G|Y) = 0$ and $Sp(Y|G) = 0$.

Similarly, system B has 0 uncertainty and 0 spurious information about g since its outputs simply consist of a relabeling of system A. On the other hand system C has larger redundancy $Sp(Y|G) > 0$ since there is a spurious state at Y for $G = 1$. Contrarily system D has 0 redundancy but it has larger uncertainty $U(G|Y)$ since two different states of Y give rise to the same state at G for $G = 1$.

Interestingly the amount of unnecessary, spurious information in Y about G, can be seen as the level of uncertainty in Y given G. For example, if we know in system

C that g is 1, we have still the uncertainty that y is Φ or Σ. Hence, if we quantify the level of uncertainty about the task by $U(G|Y)$, the natural measure of spurious information is then $U(Y|G)$. This is what we call the *uncertainty-redundancy duality principle.*

Finally we will proceed to define the measure of structural uncertainty. In order to be a proper measure of *structural uncertainty* we require it to satisfy the following properties:

1. **Independency on the specific implementation:** $U(B|A)$ must depend on the structural relationship between A and B but not on the specific implementation. Therefore $U(B|A)$ must depend on the statistical relationship between the different states of A ($\{a_1, a_2, ...\}$) and the different states of B ($\{b_1, b_2, ...\}$) no matter what they are or represent (neurons, electronic devices, numbers, symbols, vectors, scalars, etc.). Thus the set of probabilities $p(a_i, b_j)$ should determine completely $U(B|A)$.

   X and Y are called to be *structurally equivalent* if $p(x_i, y_j) = p(x_i) = p(y_j)$ for every pair i, j. That is, knowing X completely determines Y and vice versa. Hence if C is structurally equivalent to A and D is structurally equivalent to B we require that $U(B|A) = U(D|C)$.

2. **Positiveness:** For every A and B, $U(B|A) \geq 0$ must hold.

3. **Zero uncertainty:** $U(B|A) = 0$ if and only if B is completely determined by A, that is, if $p(a_i) > 0$, then there is only a possible state in B. This condition can be expressed as *if* $p(a_i) > 0$ *then* $p(b_j|a_i) = 0$ *or* 1. As a particular case, $U(A|A) = 0$.

4. **Triangular inequality:** $U(A|C) \leq U(A|B) + U(B|C)$ for every A, B and C

5. **A close processing system never reduces the structural uncertainty about the objective:** Let us consider a system where all the information it receives is A, transforming it in B. Informally, the structural information in B about another external variable G can not be greater than the information which was implicitly or explicitly present in A. Therefore $U(G|B) \geq U(G|A)$.

5.3 Desired properties for the new information processing measure

This section lists a set of desired properties that, we claim, the new information processing measure must have if it is to be considered for the design of any adaptive information processing system (whether artificial or biological).

(a) It should take into account the task(s) to be solved by the agent. The input to the agent can be statistically rich and complex, yet it may be mostly useless if it is not related to the task (non-reversibility property).

(b) It must be a measurable quantity that does not depend on the specific system implementation, that is, it should depend on the structural properties of the states of the agent and not on local properties dependent on implementation details.

(c) It should take into account how much the input data (number of transformations) has to be processed in order to extract the relevant information for the task.

(d) It should be null for a perfect communication channel (in the classical sense, i.e. exact copy of the input message) and maximal for the case of perfect transformation to the objective alphabet (active property).

(e) It should be a compromise between reduction of spurious information and extraction of the relevant part of the information.

(f) It should account for uncertainties introduced by different means, such as: lost of meaningful information, environmental noise, etc.

(g) It should be able to deal with systems composed of stochastic elements.

5.4 Specific requirements for the new information processing measure

Next we would like to impose a set of requirements that this new measure should have in order to be regarded as a candidate for an active general information processing measure. From these requirements we will derive a specific expression for ΔP:

Effective processing measure

It must be an effective processing measure, that is, ΔP must depend on x (the input), y (the output) and g (the goal) but it must not depend on the information processing path taken to go from x to y ($x \rightarrow y$), that is,

$$\Delta P(x \rightarrow y|g) = \Delta P(x \rightarrow w|g) + \Delta P(w \rightarrow y|g) \tag{5.1}$$

for all x, y, w, g, that is,

$$\Delta P(x \rightarrow y|g) = f(x,g) - f(y,g) \tag{5.2}$$

where the function $f(y,g)$ defines a sort of distance function that should depend on the structural properties of the system (as expressed in the section on desired properties) of the states of the system y and the structural properties of g, the structural properties being determined by the global statistical properties.

Maximum value for ΔP

The maximum value for $\Delta P(x \rightarrow y|g)$, when x and g are fixed, must occur when $y = \hat{g}$ and as a consequence $f(y,g) \geq f(g,g)$. So that f can be chosen, without loss of generality, such that

$$f(y,g) \geq 0; f(g,g) = 0 \tag{5.3}$$

for all y, g. And lastly the maximum value for $\Delta P(x \rightarrow y|g)$ when g is allowed to vary for all x y and g is

$$\Delta P(x \rightarrow y|g) \leq \Delta P(x \rightarrow y|y) \tag{5.4}$$

and as a consequence $f(x,g) \leq f(x,y) + f(y,g)$ which corresponds to the triangular inequality. So that taking into account the triangular inequality and the previously defined properties in expression 5.3, it can be concluded that f is a *pseudo-distance* function.

Maximum value for ΔP in a subset of different solutions

Suppose we have a set of different systems $\mathcal{A} = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, ...\}$ which transform $\vec{x}$ in $\vec{y}$, all with the same goal $\vec{g}$. Since the dimensionality and statistical structure of $\vec{y}$ may be different in these systems, their corresponding ΔP's may be also different. There could not exist no system in $\mathcal{A}$ which reaches $\vec{y} = \vec{g}$. Which is the system in $\mathcal{A}$ with greatest ΔP ? We require it to be the system $\mathcal{S}_m$ whose $\vec{y}$ is "closest" to $\vec{g}$. That is, no other system in $\mathcal{A}$ has less uncertainty about G than $\mathcal{S}_m$ has and less spurious information about G than $\mathcal{S}_m$ has. As we have seen in section 5.2.2, the level of spurious information the system has about G can be expressed as $U(Y|G)$.

Thus the requirement that the system with maximum ΔP is the one with $\vec{y}$ "closest" to $\vec{g}$ can be formulated as:

if $\mathcal{S}_m$ maximizes $\Delta P(\vec{x} \to \vec{y}|\vec{g})$ in $\mathcal{A}$ for a given $\vec{x}$ and $\vec{g}$, no other configuration in $\mathcal{A}$ satisfies neither $U(G|Y) < U(G|Y)_{\mathcal{S}_m}$ and $U(Y|G) \leq U(Y|G)_{\mathcal{S}_m}$, nor $U(G|Y) \leq U(G|Y)_{\mathcal{S}_m}$ and $U(Y|G) < U(Y|G)_{\mathcal{S}_m}$.

Noting that $\Delta P(\vec{x} \to \vec{y}|\vec{g}) = f(\vec{x}, \vec{g}) - f(\vec{y}, \vec{g})$ (equation 5.2) and since $\vec{x}$ and $\vec{g}$ are kept constant, the previous requirement can be rewritten as:

if $\mathcal{S}_m$ minimizes $f(\vec{y}, \vec{g})$ in $\mathcal{A}$ for a given $\vec{g}$, no other configuration in $\mathcal{A}$ satisfies neither $U(G|Y) < U(G|Y)_{\mathcal{S}_m}$ and $U(Y|G) \leq U(Y|G)_{\mathcal{S}_m}$, nor $U(G|Y) \leq U(G|Y)_{\mathcal{S}_m}$ and $U(Y|G) < U(Y|G)_{\mathcal{S}_m}$.

Interestingly, if $\mathcal{S}_m$ minimizes $U(Y|G) + \beta U(G|Y)$ for a given $\beta > 0$, then there does not exist any configuration in $\mathcal{A}$ such that $U(G|Y) < U(G|Y)_m$ and $U(Y|G) \leq U(Y|G)_m$, or $U(G|Y) \leq U(G|Y)_m$ and $U(Y|G) < U(Y|G)_m$ (see appendix C.3 for details).

Therefore, it is natural to define the pseudodistance f as

$$f(\vec{y}, \vec{g}) \equiv U(Y|G) + \beta U(G|Y) \tag{5.5}$$

with $\beta > 0$. Since the uncertainty measure satisfies the basic requirements $U(A|B) \geq 0$, $U(A|A) = 0$ and $U(A|C) \leq U(A|B) + U(B|C)$ then all the requirements we did about ΔP are also satisfied.

5.5 Properties of the new information processing measure

Given eqs. 5.2 and 5.5 the measure of information processing is:

$$\Delta P(\vec{x} \rightarrow \vec{y}|\vec{g}) = [U(X|G) - U(Y|G)] - \beta\,[U(G|Y) - U(G|X)] \qquad (5.6)$$

The first term is the reduction of spurious information. We call this the "complexity reduction" term since a reduction of spurious information implies a reduction of the complexity of the information (section 4.1.4). The second term in eq. 5.6 represents the loss of information about the goal (uncertainty creation). This measure has the following properties:

1. The term of loss information is never negative in closed systems. This is a direct consequence of the property 5 we required for U (section 5.2.2).

2. The term of complexity reduction can be negative (complexity grows) only in stochastic processes

   Proof:

   Let us consider a system $\mathcal{S}$ which transforms $\vec{x}$ into $\vec{y}$ with overall goal $\vec{g}$. The term of complexity reduction is $U(X|G) - U(Y|G)$. The uncertainty measure was required to satisfy $U(A|C) \leq U(A|B) + U(B|C)$ (property 4 in 5.2.2). Therefore, $U(Y|G) \leq U(Y|X) + U(X|G)$. This can be rewritten as $U(X|G) - U(Y|G) \geq U(Y|X)$. If $\mathcal{S}$ is a deterministic process, then Y is completely determined by X, and thus $U(Y|X) = 0$ (property 3 in 5.2.2). This together with the previous property determine $U(X|G) - U(Y|G) \geq 0$. As a conclusion, the term of complexity reduction is never positive in a deterministic process.

3. $\Delta P = 0$ in a perfect communication channel

   Proof:

   In a perfect communication channel that transforms X into Y we have $H(Y|X) = 0$ and $H(X|Y) = 0$ where H denotes the Shannon conditioned

entropy [Cover & Thomas, 1991]. But $H(A|B) = 0$ only if the knowledge of the state of B determines completely the state of A [Cover & Thomas, 1991]. Therefore, the knowledge of X determines completely Y and vice versa, and then we say that X and Y are *structurally equivalent* (property 1, section 5.2.2). Then our U satisfies $U(X|G) = U(Y|G)$ and $U(G|X) = U(G|Y)$ (property 1, 5.2.2) which implies $\Delta P = U(X|G) - U(Y|G) + \beta\,(U(G|X) - U(G|Y)) = 0$.

4. For a fixed goal and input statistics x, the global maximum of ΔP along all possible systems occurs when Y is a relabeling of g.

   Proof:

   Since $\Delta P = f(X,G) - f(Y,G)$ and X and G are fixed, the maximum value of ΔP occurs when $f(Y,G)$ is minimum. Suppose Y is a relabeling of G. Then $U(Y,G) = 0$ and $U(G,Y) = 0$, and therefore, $f(Y,G) = 0$, which is the minimum value of f since this quantity is ≥ 0. Therefore we conclude that a system whose Y is a relabeling of G maximizes globally ΔP.

5. For a fixed goal, input statistics x, and a subset of systems $\mathcal{A} = \{\mathcal{S}_1, \mathcal{S}_2, ...\}$ which transforms x into y, if $\mathcal{S}_m$ is the system with greater ΔP, then there is no other system in $\mathcal{A}$ which simultaneously has less uncertainty and less complexity (section 5.4).

Note that all these properties are independent on the concrete election of U as long as it satisfies our requirements (section 5.2.2).

5.6 Choice of the measure of uncertainty

Several known uncertainty measures such as Shannon's entropy and Bayes error satisfy the properties required by a structural uncertainty measure and therefore constitute proper measures of uncertainty in our theory (see appendix C.4 for a detailed proof for the Shannon's conditioned entropy).

The value for ΔP and its specific interpretation depends on the concrete choice of the uncertainty measure. In this thesis we choose the conditional Shannon's entropy

$H(A|B)$ as $U(A|B)$ because it is easy to obtain analytical expressions in the theoretical examples we show. We want to remark again that this is not the only possible choice and there is a family of different ΔP measures.

5.7 The Principle of Maximization of ΔP for Adaptive Systems

We claim that the agent should maximize ΔP in order to achieve the maximum efficiency in its interaction with the environment. Then the agent will minimize the uncertainty about the optimal action it has to do while minimizing the unnecessary complexity. The latter can be seen as minimizing the uncertainty the environment has about the internal state of the agent. Therefore, the agent will be efficient if it achieves a strong "coupling" with the task it has to solve.

How can the agent learn to perform efficiently ? There are two possible situations:

a) the optimal policy for some states is communicated by the environment to the agent, or

b) the agent has to find out which actions are more appropriated for each situation

In this thesis we will illustrate our ideas with examples of the first case. Specifically, we consider problems where there is only one optimal action for each state. We will then analyze two types of problems: classification (the optimal action is to guess correctly the real class of each pattern) and autoencoder (the optimal action is to reconstruct the input with the required precision). Thus the agent is provided with a set of examples together with their corresponding "optimal actions" which has to learn in order to generate the optimal actions on unseen examples (generalization).

For dealing with situation b) the agent must mix the optimization of ΔP (construction of optimized representations of the environment) with the problem of learn which actions in the past have been responsible of the value of $\mathcal{H}$ now. This can be done mixing known strategies of reinforcement learning with the maximization of ΔP (future work).

5.7.1 Interpretation for β

The pseudodistance f, defined by eq. 5.5, is a weighted sum of the level of uncertainty about the goal and the amount of spurious information. The value of β is determined by the accuracy with which the agent has to solve he task. If β is low, configurations with low uncertainty about the goal are more efficient even if their complexity is not small. However, if β is high, systems with low uncertainty are more efficient even if they are very complex.

Due to noise and uncertainties in the input it is impossible to perform without errors in real world problems. Suppose the autonomous agent tries to maximize its efficiency optimizing ΔP. Then a too high value for β is problematic since then the agent would be very complex having many parameters to tune, increasing the risk of overfitting and poor generalization. On the other hand, having a too small β is also undesirable since then the agent will have a high error.

Therefore β should be chosen at intermediate values, corresponding to the desired accuracy. As shown in the results section 6.3.2 the agent can learn this metaparameter from its interaction from the environment: if the agent has too many errors, then β will increase. It then makes it possible to learn the problem with a minimum structural complexity given the desired precision, optimizing the efficiency of the agent.

5.7.2 Communication theory in relation to the proposed framework

Let us consider a system whose goal is to transmit the input it receives ($g = x$). Then,

$$\Delta P = (U(x|g) - U(y|g)) + \beta\,(U(g|x) - U(g|y)) =$$

$$= (U(x|x) - U(y|x)) + \beta\,(U(x|x) - U(x|y)) = -U(y|x) - \beta U(x|y) \qquad (5.7)$$

Since the uncertainties and β are ≥ 0, for this special case we get $\Delta P \leq 0$. The first term, $-U(y|x)$ is due to complexity creation by the channel.

The other term, $U(x|y)$, is the uncertainty that the receiver has about the original signal x. This is due to the noise and loss of relevant information in the channel.

The global optimum value of ΔP occurs when y is a relabeling of x, that is, when the information is transmitted without any error nor redundancies. However, real communication channels have noise, and ΔP can not achieve its maximum value. Therefore, ΔP represents a trade-off between $U(y|x)$ and $U(x|y)$. The interpretation of this trade-off is straightforward: in order to obtain better communication (i.e. reducing $U(x|y)$) , we need to make the channel more sophisticated, increasing the complexity term $U(y|x)$.

In case we are interested in transmitting the signal with very high quality, even if the channel is very complex, then $\beta >> 1$, and therefore the problem of maximizing ΔP is equivalent to maximize $U(x|y)$. In case we choose the conditioned Shannon's entropy as our measure of uncertainty, we have $U(x|y) = H(x|y)$. Since the mutual information between x and y satisfies $I(x;y) = H(x) - H(x|y)$, and $H(x)$ is fixed in our problem, we conclude that the problem of maximizing ΔP is equivalent to maximize the mutual information $I(x;y)$, which is a fundamental concept in the design and analysis of communication channels [Cover & Thomas, 1991].

Chapter 6

Results

In the following sections we will show the natural emergence from the information processing measure of the following different techniques:

1. **Linear systems**
   (a) Linear autoencoder
   (b) Optimal linear transformations for classification
2. **Non linear systems**
   (a) Decision Trees Construction
   (b) Deterministic and stochastic layer of binary neurons
   (c) Non Linear Feature Extraction for classification

Before starting the analysis of systems with continuous dynamics, we will discuss about the correct manner of quantifying the uncertainties in such systems. Through all this chapter we will use the compact notation for gaussian probability distributions introduced in appendix C.2, and the properties listed there. Finally we will analyze the olfactory and auditory systems in the context of the new framework.

6.1 Analysis of systems with continuous dynamics

6.1.1 The problems with differential entropy

The Shannon's entropy of a discrete random variable x is defined as [Cover & Thomas, 1991]:

$$H(x) = -\sum_i p(x_i) \log p(x_i) \tag{6.1}$$

Let us consider now that the random variable is continuous, hence characterized by its probability density function $p(\vec{x})$. If we discretize this variable using a bin of Δ, we can calculate the entropy of this discretization version using eq. 6.1. However, $\lim_{\Delta \to 0} H^{\Delta}$ will diverge to ∞ [Cover & Thomas, 1991]. That is, the entropy of a continuous variable is ∞ since we need infinity information to describe the infinity digits of its state.

However, the limit $\lim_{\Delta \to 0} H^{\Delta} + \log \Delta$ exists and can be expressed as [Cover & Thomas, 1991]:

$$\lim_{\Delta \to 0} H^{\Delta} + \log \Delta = -\int p(x)\, dx \equiv h(x) \tag{6.2}$$

where $h(x)$ is called the "differential entropy of x". This can be used as a measure of uncertainty and some of its properties are analogous to those of the entropy of a discrete variable. Moreover, we can define analogously the conditioned differential entropy as [Cover & Thomas, 1991]:

$$h(y|x) \equiv -\int \int p(x,y) \log \frac{p(x,y)}{p(x)}\, dy\, dx \tag{6.3}$$

with $p(x,y)$ being the joint probability density function of the two random variables.

However, the differential entropy can be negative and has problems with singularities [Cover & Thomas, 1991]. For example, if x is a continuous random variable, $H(x|x) = -\infty$. This problem can be alleviated assuming a statistically independent noise n which is added to the variable x. Then $H(x+n) - H(n) \geq 0$ but the problem

of $H(x+n|x+n) = -\infty$ remains. Another important problem is that the differential entropy depends on the scale. That is, $h(a \cdot x) \neq h(x)$, with a being a constant [Cover & Thomas, 1991], which makes it dependent on the implementation. Thus differential entropy is not appropriated in our theoretical framework.

Here we propose a different solution which solves this problem. In order to obtain analytical results in these systems with continuous dynamics, we discretize the different variables and use the Shannon entropy for discrete variables which is well defined. In the different analytical problems shown in this thesis these continuous variables are assumed to be well described by multidimensional gaussian distributions.

6.1.2 Uniform quantization of gaussian variables

Let us consider a random continuous variable y with gaussian pdf:

$$p(y) = \mathcal{G}(\sigma^2, y) \tag{6.4}$$

For clarity reasons we have assumed its average to be 0 since all the results in this section do not depend on it. Now we fix a bin Δ and quantize y using this bin:

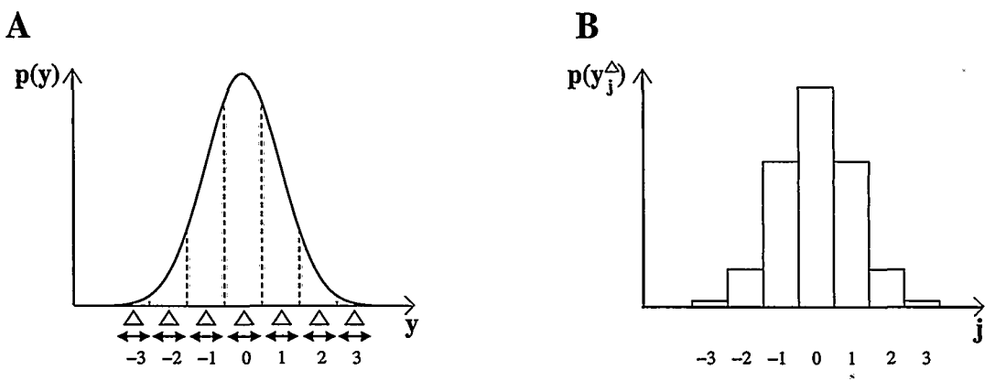

Figure 6.1: Uniform quantization of a gaussian variable. **A:** Original pdf of the gaussian variable. We discretize it with uniformly with bin Δ. The index of each interval denotes the index of the correspondingly discretized symbol: **B:** histogram of probability of the quantized symbols.

Let us call y^Δ to this quantized variable. Then the probability of the symbol y_j^Δ is given by:

$$p(y_j^\Delta) = p\left(y \in [\Delta(j-.5), \Delta(j+.5))\right) = \int_{\Delta(j-.5)}^{\Delta(j+.5)} p(y)\, dy \tag{6.5}$$

This integral can be rewritten as:

$$p(y_j^\Delta) = \frac{1}{2}\left[erf\left(\frac{\Delta}{\sqrt{2}\sigma}(j+.5)\right) - erf\left(\frac{\Delta}{\sqrt{2}\sigma}(j-.5)\right)\right] \tag{6.6}$$

where erf is the so-called *error function* which is defined as [Arfken, 1985]:

$$erf(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2}\, du \tag{6.7}$$

Entropy of a discretized variable

The entropy of the discretized variable is:

$$H(y^\Delta) = -\sum_j p(y_j^\Delta) \log p(y_j^\Delta) \tag{6.8}$$

where $p(y_j^\Delta)$ is given by eq. 6.6. Thus the entropy we want to measure is given by a complex equation which can not be simplified. It can be numerically evaluated but we need an analytical expression in order to perform our study. We will consider a different way of quantizing the variable y which will allow us to obtain much simpler equations. Then we will compare the results of our approximation with the exact entropy (eqs. 6.6 and 6.8).

Let us consider a stochastic discretization of y. Now each symbol is not defined by an interval but by an "activation function" defined by a gaussian function of dispersion σ_a. Points where the activation function is higher are points where that symbol is more likely given y. The distance between peaks of consecutive activation functions is Δ. The activation function of symbol j is thus given by $\mathcal{G}(\sigma_a^2, y - j\Delta)$. The interval where this symbol is the most likely given y is then $[(j-.5)\Delta,\ (j+.5)\Delta]$, which corresponds to a stochastic version of the deterministic uniform quantization

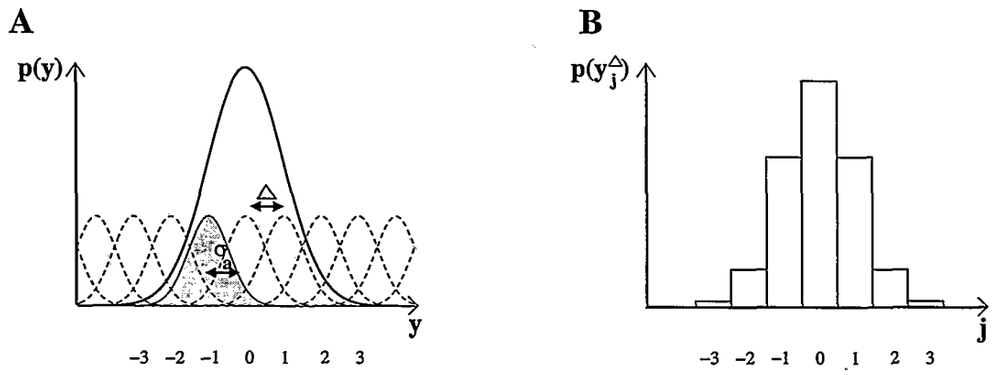

Figure 6.2: Stochastic quantization of a gaussian variable. **A**: each discrete symbol is characterized by an "activation function" which determines the probability of the symbol given y. **B**: Histogram of the probabilities of the symbols.

(eq. 6.5).

Note that the activity function can not be directly interpreted as a probability since, given y, the probabilities of all the stochastic symbols must sum 1. Thus the probability of symbol j given y is:

$$p(y_j^\Delta|y) = \frac{\mathcal{G}(\sigma_a^2, y - j\Delta)}{\sum_{i=-\infty}^{\infty} \mathcal{G}(\sigma_a^2, y - i\Delta)} \tag{6.9}$$

Now we proceed to calculate the denominator. Notice that $\mathcal{G}(\sigma_a^2, y - i\Delta) = \frac{1}{\sigma_a}\mathcal{G}(1, i\frac{\Delta}{\sigma_a} - \frac{y}{\sigma_a})$ (property C.16). Let us consider the infinite summatory $|a| \sum_{n=-\infty}^{\infty} \mathcal{G}(1, an - b)$, which in expanded notation is:

$$|a| \sum_{n=-\infty}^{\infty} \mathcal{G}(1, an - b) = \frac{|a|}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} e^{-\frac{(an-b)^2}{2}} \tag{6.10}$$

This summatory can be rewritten as:

$$\frac{|a|}{\sqrt{2\pi}} e^{\frac{-b^2}{2}} + \frac{|a|}{\sqrt{2\pi}} \sum_{n=1}^{\infty} \left[e^{-\frac{(an-b)^2}{2}} + e^{-\frac{(an+b)^2}{2}} \right] \tag{6.11}$$

Note that all the terms in the summatory are positive. If $a \neq 0$, they also decay faster than the terms of the sum $\sum_{n=1}^{\infty} e^{-a^2 n}$ which is convergent as long as $a \neq 0$. Then this series is also convergent if $a \neq 0$.

In the limit of $|a| \to 0$ we have:

$$\lim_{|a|\to 0}\left(\frac{|a|}{\sqrt{2\pi}}\sum_{n=-\infty}^{\infty} e^{-\frac{(an-b)^2}{2}}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-b)^2}{2}}\,dx = 1 \tag{6.12}$$

Therefore, in a neighborhood of $|a| = 0$ the summatory can be approximated as:

$$|a|\sum_{n=-\infty}^{\infty} \mathcal{G}(1, an-b) \simeq 1 \tag{6.13}$$

How precise is this approximation ? In figure 6.3 we show the mean absolute error of the approximation. As we can see, the error is null or insignificant for $\frac{\Delta}{\sigma_a} < 2$.

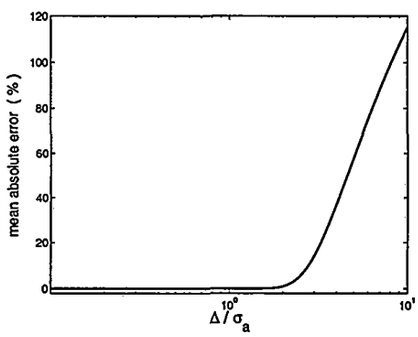

Figure 6.3: Absolute deviation of the series $\frac{\Delta}{\sigma_a}\sum_{n=-\infty}^{\infty}\mathcal{G}(1, \frac{\Delta}{\sigma_a}n-b)$ from 1. This error has been averaged along all the possible values of b. The error starts to be appreciable at $\frac{\Delta}{\sigma_a} > 2$.

Therefore if $\Delta < 2\sigma_a$ we can approximate with a very high degree of precision eq. 6.9 as:

$$p(y_j^{\Delta}|y) = \frac{\mathcal{G}(\sigma_a^2, y-j\Delta)}{\sum_{i=-\infty}^{\infty}\mathcal{G}(\sigma_a^2, y-i\Delta)} \simeq \Delta\cdot\mathcal{G}(\sigma_a^2, y-j\Delta) \tag{6.14}$$

Now we proceed to calculate the entropy of the discretized variable. First we will calculate the probability of the discretized symbol $p(y_j^{\Delta})$ as:

$$p(y_j^\Delta) = \int_{-\infty}^{\infty} p(y_j^\Delta|y)\cdot p(y)\,dy = \Delta\int_{-\infty}^{\infty} \mathcal{G}(\sigma_a^2, y-j\Delta)\cdot\mathcal{G}(\sigma^2,y)\,dy = \Delta\,\mathcal{G}(\sigma_a^2+\sigma^2, j\Delta) \tag{6.15}$$

where we have used the theorem of "ands" of gaussians (eq. C.20). We can rewrite this equation in a compact manner:

$$p(y_j^\Delta) = \Delta\cdot\mathcal{G}(\sigma_a^2+\sigma^2, j\Delta) = \frac{\Delta}{\sqrt{\sigma_a^2+\sigma^2}}\,\mathcal{G}\left(1, \frac{j\Delta}{\sqrt{\sigma_a^2+\sigma^2}}\right) = \gamma\,\mathcal{G}(1,\gamma j) \tag{6.16}$$

where we have defined $\gamma \equiv \frac{\Delta}{\sqrt{\sigma_a^2+\sigma^2}}$ and used property C.16.

Then $-\log p(y_j^\Delta) = -\log(\gamma\,\mathcal{G}(1,\gamma j)) = -\log\frac{\gamma}{\sqrt{2\pi}} + \frac{1}{2}\gamma^2 j^2 = \frac{1}{2}(\log(2\pi\gamma^{-2}) + \gamma^2 j^2)$. Thus we can write:

$$H(y^\Delta) = -\sum_{j=-\infty}^{\infty} p(y_j^\Delta)\,\log\,p(y_j^\Delta) = \frac{1}{2}\sum_{j=-\infty}^{\infty}(\log(2\pi\gamma^{-2}) + \gamma^2 j^2)\,\gamma\,\mathcal{G}(1,\gamma j) =$$

$$= \frac{1}{2}\left(\log(2\pi\gamma^{-2})\right)\,\gamma\sum_{j=-\infty}^{\infty}\mathcal{G}(1,\gamma j) + \frac{1}{2}\gamma\sum_{j=-\infty}^{\infty}\gamma^2 j^2\,\mathcal{G}(1,\gamma j) \tag{6.17}$$

Remember our approximation 6.13, rewritten as $\sum_{i=-\infty}^{\infty}\mathcal{G}(1,\gamma j) \simeq \frac{1}{\gamma}$ (notice we have taken $y = 0$ in 6.13). Taking the derivative respect to γ and rearranging we obtain:

$$\sum_{i=-\infty}^{\infty}\gamma^3 j^2\,\mathcal{G}(1,\gamma j) \simeq 1 \tag{6.18}$$

In figure 6.4 we see that this approximation is valid as long as $\gamma < 2$, that is, $\Delta < 2\sqrt{\sigma_a^2+\sigma^2}$.

Using these results in eq. 6.17 we get:

$$H(y^\Delta) \simeq \frac{1}{2}\left(\log(2\pi\gamma^{-2})\right) + \frac{1}{2} = \frac{1}{2}\log(2\pi e\gamma^{-2}) \tag{6.19}$$

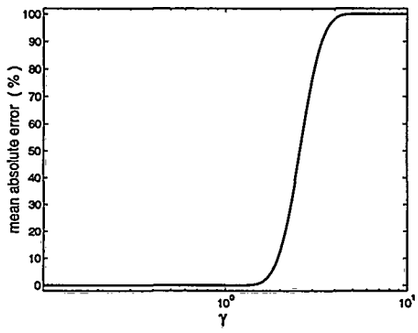

Figure 6.4: Absolute deviation of the series $\sum_{i=-\infty}^{\infty} \gamma^3 j^2 \mathcal{G}(1, \gamma j)$ from 1. The error starts to be significant for $\gamma > 2$.

Remembering that we defined γ as $\frac{\Delta}{\sqrt{\sigma_a^2+\sigma^2}}$ and rearranging we have:

$$H(y^{\Delta}) \simeq \frac{1}{2} \log 2\pi e \left(\left(\frac{\sigma_a}{\Delta}\right)^2 + \left(\frac{\sigma}{\Delta}\right)^2 \right) \tag{6.20}$$

The $\frac{\sigma_a}{\Delta}$ ratio controls the degree of overlapping between the activation functions and thus making it larger will make our stochastic quantization cleaner. However we can not make this ratio arbitrarily large since then the activation functions will not cover uniformly the space (figure 6.5).

We will determine which value of this ratio lets us obtain a good approximation of the entropy of the deterministic uniform quantization $H(y^{\Delta})_d$ (eqs. 6.8 and 6.6). Since this entropy tends to 0 when Δ tends to infinity we adjust the ratio in order to conserve this property. Thus we should choose $2\pi e \left(\frac{\sigma_a}{\Delta}\right)^2 = 1$, obtaining the approximation:

$$H(y^{\Delta})_d \simeq \frac{1}{2} \log \left(1 + 2\pi e \Delta^{-2} \sigma^2\right) \tag{6.21}$$

We will show that this approximation is also valid in the limit $\Delta \to 0$. Then the exact entropy $H(y^{\Delta})_d$ tends to $H(y) - \frac{1}{2} \log \Delta^2$ as $\Delta \to 0$ [Cover & Thomas, 1991], where $H(y)$ is the differential entropy of the continuous variable y, which in this case is equal to $\frac{1}{2} \log 2\pi e \sigma^2$ [Cover & Thomas, 1991]. When Δ is very small our equation 6.21 is:

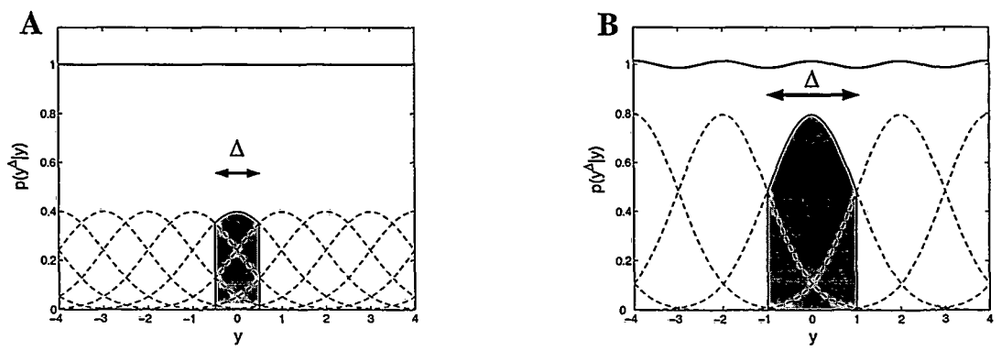

Figure 6.5: Superposition of the activation functions. **A**: $\Delta = 1$ and $\sigma_a = 1$. The superposition of the activation functions is perfectly uniform and equal to 1. The zone were symbol 0 is more likely given y is that marked in gray. **B**: $\Delta = 2$ and $\sigma_a = 1$. The superposition is approximately uniform. The gray area (zone were symbol 1 is more likely given y) is greater, the code is more deterministic (less superposition).

$$\frac{1}{2}\log\left(1+2\pi e\Delta^{-2}\sigma^2\right) \simeq \frac{1}{2}\log\left(2\pi e\Delta^{-2}\sigma^2\right) = \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log\Delta^2 \qquad (6.22)$$

which coincides with the exact entropy.

In figure 6.6 we show the comparison between our approximation and the numerical computation of the exact solution (eqs. 6.8 and 6.6).

As we can see our approximation is very accurate for $\frac{\Delta}{\sigma_a} < 3$, then it tends to 0 as the real entropy does but in a slower manner.

We want to mention that the election we made $2\pi e\left(\frac{\sigma_a}{\Delta}\right)^2 = 1$ implies $\frac{\Delta}{\sigma_a} = \sqrt{2\pi e}$, which is out of the range $\frac{\Delta}{\sigma_a} \leq 2$ for which our approximations where accurate. However, the accuracy lost in our approximation is compensated by a smaller overlap between the activity functions, allowing a good estimation of $H(y^\Delta)_d$.

Extension to multidimensional variables

Here we consider the quantization of a gaussian variable of N dimensions with pdf:

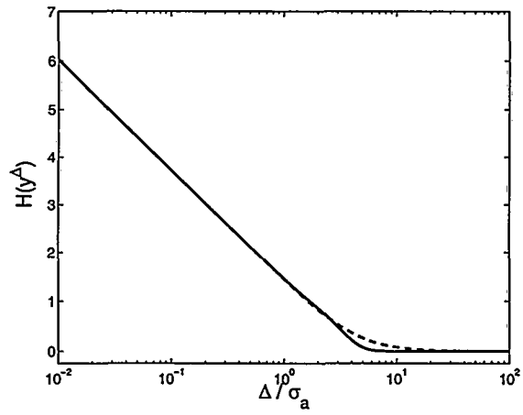

Figure 6.6: Quality of our approximation to $H(y^{\Delta})$. Solid: exact value computed numerically (eqs. 6.8 and 6.6). Dashed: approximation (eq. 6.21).

$$p(\vec{y}) = \mathcal{G}(E, \vec{y}) \tag{6.23}$$

with E being an $N \times N$ covariance matrix. Now the quantization is defined by N independent directions, each one with its own quantization bin Δ_α. The quantization is thus defined by a matrix Δ where its columns $\vec{q}_\alpha$ are the different quantization directions and the norms of these columns are the corresponding quantization bins Δ_α.

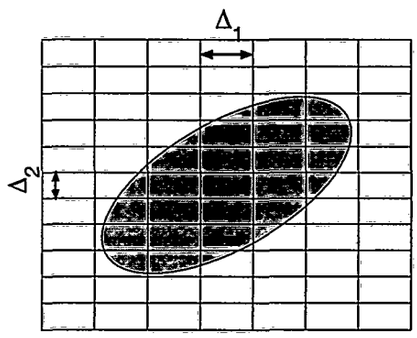

Figure 6.7: Uniform quantization of a multidimensional gaussian variable. The quantization directions are chosen as the axes. Each direction has its own quantization bin. The discrete symbols resulting from the quantization are then the different boxes.

In appendix C.6 we demonstrate how equation 6.21 can be generalized to the multidimensional case obtaining the approximation:

$$H(y^{\Delta}) \simeq \frac{1}{2}\log\det(I + 2\pi e\, Q^{-T} E\, Q^{-1}) \tag{6.24}$$

Note that equation 6.21 for the scalar variable is a special case of this general equation.

Noise - quantization duality

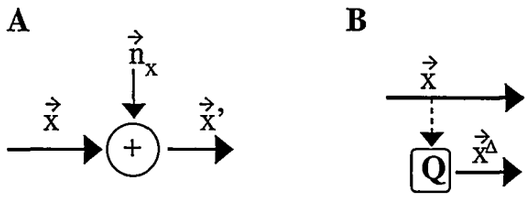

Figure 6.8: Noise-quantization duality

Consider the process in figure 6.8 A. The vector $\vec{x}$ is corrupted by noise $\vec{n}_x$ yielding the noisy variable $\vec{x}' = \vec{x} + \vec{n}_x$. The noise has covariance matrix N. Intuitively, the noise introduces a granularity in the data yielding "confidence levels", therefore quantizing it. Let us check this: the amount of information which is not corrupted by the noise is given by the mutual information $I(\vec{x}; \vec{x}+\vec{n}_x) = H(\vec{x}') - H(\vec{x}'|\vec{x})$ which after simplifications can be written as:

$$I(\vec{x}; \vec{x}') = \frac{1}{2}\log\det(I + N^{-1}C) \tag{6.25}$$

where C is the covariance matrix of $\vec{x}$.

On the other hand, consider the process in figure 6.8 B. The information content in $\vec{x}$ is now measured quantizing it previously with a quantization matrix given by Q. Then, the amount of information about $\vec{x}$ which remains in the quantization is $I(\vec{x}; \vec{x}^{\Delta}) = H(\vec{x}^{\Delta}) - H(\vec{x}^{\Delta}|\vec{x}) = H(\vec{x}^{\Delta})$ since, given $\vec{x}$, we completely know $\vec{x}^{\Delta}$. If we calculate $H(\vec{x}^{\Delta})$ using eq. 6.24 we have:

$$I(\vec{x};\vec{x}^{\Delta}) = \frac{1}{2}\log\det(I + 2\pi e Q^{-T}CQ^{-1}) = \frac{1}{2}\log\det(I + 2\pi e(Q^{T}Q)^{-1}C) \qquad (6.26)$$

where we have used property C.2.4.

Comparing this equation with 6.25 we conclude that, from an information preserving point of view, a variable corrupted by noise with variance N is equivalent to a quantization of the original signal with quantization matrix $Q = \frac{1}{\sqrt{2\pi e}}N^{1/2}$.

Therefore, we have:

$$H(\vec{x}^{\Delta}) = \frac{1}{2}\log\det(I + N^{-1}C) \qquad (6.27)$$

6.2 Analysis of linear systems

6.2.1 Linear system with linear objectives

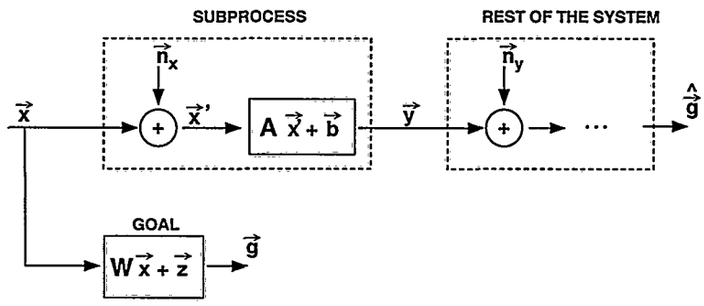

Figure 6.9: The global processing system tries to estimate a goal g which depends linearly on the input x. The subprocess we consider has sensors which introduce a noise n_x in the signal, therefore limiting the precision in the knowledge about x. The rest of the system reads the output of the process with a precision constrained by the noise n_y.

Consider the system in figure 6.9 where a layer of noisy linear neurons responds

to the stimulus $\vec{x}$ as:

$$\vec{y} = A(\vec{x} + \vec{n}_{\vec{x}}) + \vec{b} \tag{6.28}$$

where $\vec{y}$ is the vector of the responses in the layer and $n_{\vec{x}}$ is the noise in the input due to noisy receptors.

We assume for simplicity that the noise is zero-mean normal distributed with covariance matrix N_x. The input statistics are assumed to be well represented by a multidimensional gaussian of covariance matrix C. The objective of the system is to achieve the goal $W\vec{x} + \vec{b}$ with a confidence interval given by Δ_g. In order to perform our analysis we need to discretize the continuous variables in the problem obtaining $\vec{x}^{\Delta}$, $\vec{y}^{\Delta}$ and $\vec{g}^{\Delta}$. The quantization in $\vec{g}^{\Delta}$ is given by the desired precision Δ_g.

In 6.1.2 we saw that the noise in the reception of $\vec{x}$ introduces a natural quantization $\vec{x}^{\Delta}$. Additionally, the noise in the reception of $\vec{y}$ by the rest of the system introduces a natural quantization $\vec{y}^{\Delta}$. Alternatively, we can consider that this noise does not exist but a limited precision in the generation or reception of $\vec{y}$.

In order to quantify the information processing measure in our system we need to calculate:

$$\Delta P(\vec{x}^{\Delta} \to \vec{y}^{\Delta} | \vec{g}^{\Delta}) = d(\vec{x}^{\Delta}, \vec{g}^{\Delta}) - d(\vec{y}^{\Delta}, \vec{g}^{\Delta}) =$$

$$= H(\vec{x}^{\Delta} | \vec{g}^{\Delta}) + \beta H(\vec{g}^{\Delta} | \vec{x}^{\Delta}) - H(\vec{y}^{\Delta} | \vec{g}^{\Delta}) - \beta H(\vec{g}^{\Delta} | \vec{y}^{\Delta})$$

In appendix C.7.1 we develop in detail the mathematical analysis which conducts to the derivation of ΔP, obtained as:

$$\Delta P(x^{\Delta} \to y^{\Delta} | g^{\Delta}) = \frac{1}{2} \ln \frac{(\det(I + D))^{1+\beta}}{(\det(I + S))^{\beta}} + \frac{1}{2} \ln \frac{(\det(I + V\Phi V^T))^{\beta}}{\det(I + VV^T)^{\beta+1}} \tag{6.29}$$

where $D = N_x^{-1}(C^{-1} + 2\pi e W^T N_g^{-1} W)^{-1}$, $S = N_x^{-1} C$, $\Phi = (N_x + N_x D)^{-1/2}(C + N_x)(N_x + N_x D)^{-1/2}$ and $V = N_y^{-1/2} A (N_x + N_x D)^{1/2}$.

The second term in the summation 6.29 is the one which determines the maximization of Δ since the other does not depend on the responses of the neurons (matrix A).

It can be easily proved that if R is a rotation in the space of neurons, then the solution $\hat{V} = RV$ has exactly the same ΔP than V. Thus there does not exist a unique optimal configuration but a family of optimal solutions. In appendix C.7 we derive the optimal family of solutions, which can be described as:

- Take the eigenvectors (normalized) $\vec{u}_i$ of Φ with greatest eigenvalues λ_i which satisfy

$$\lambda_i > 1 + \frac{1}{\beta} \tag{6.30}$$

  In case there are none, take $A = 0$. If the number of neurons is less than the number of eigenvectors satisfying the requirement, take the eigenvectors with greatest eigenvalues

- Assign **only** one neuron to one of the selected eigenvectors $\vec{u}_i$. The receptive field of this neuron i is given by:

$$\vec{a}_i = \left(\beta - \frac{\beta+1}{\lambda_i}\right)^{1/2} N_y^{1/2} \cdot \vec{u}_i \cdot (N_x + N_x D)^{-1/2} \tag{6.31}$$

Any optimal solution is then a rotation in the space of neurons of this basic solution.

Specific results for the linear autoencoder

In this specific case, the objective of the system is to reconstruct the original signal x with precision given by Δ_g. Therefore, $W = I$, and for clarity reasons we define the symbol $\Delta_x \equiv \Delta_g$ for the required reconstruction precision. In appendix C.7.4 we show that the optimal configuration is defined by:

- Let us define $a_i \equiv \frac{\sigma_{Ci}^2}{\sigma_x^2}$ and $b \equiv \frac{\Delta_x^2}{2\pi e \sigma_x^2}$. Take the eigenvectors of C with greatest eigenvalues σ_{Ci}^2 which satisfy

$$a_i > \frac{1 + b + \sqrt{(1+b)^2 + 4\beta b}}{2\beta} \tag{6.32}$$

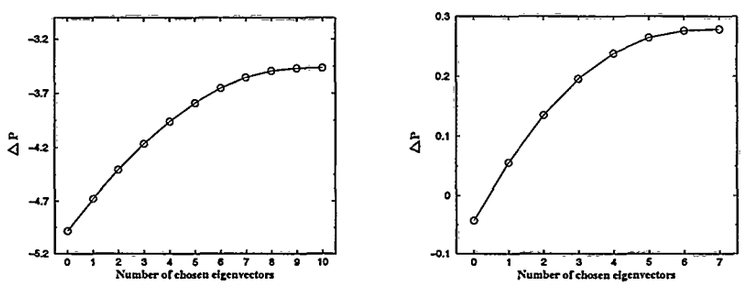

Figure 6.10: Contribution of the chosen eigenvectors in the optimal configuration of the linear autoencoder. Left: $b = \frac{1}{2}$. Right: $b = 2$.

In case there are none, take $A = 0$. If the number of neurons is less than the number of eigenvectors satisfying the requirement, take the eigenvectors with greatest eigenvalues

- Assign **only** one neuron to one of the selected eigenvectors making its receptive field proportional to the eigenvectors using gain

$$\frac{\sigma_y}{\sigma_x}\sqrt{\frac{\beta a_i^2 - (a_i + b + a_i b)}{(a_i + 1)(a_i + b + a_i b)}} \tag{6.33}$$

The family of optimal solutions are then rotations of this optimal configuration.

Similar results are obtained when maximizing the mutual information between $\vec{x}'$ and $\vec{y}$ imposing additional constraints to the system [Campa *et al.*, 1995]. Note that we do not need to impose any constraint in order to obtain these results.

In figure 6.10 we show the optimal configuration when x has 10 components with the set a_i homogeneously distributed between 2 and 8, and $\beta = 1$. If the required discretization is chosen to be smaller than the input noise σ_x ($b = \frac{\Delta_x^2}{\sigma_x^2} = \frac{1}{2}$) the system performance ΔP is negative (6.10 left). Although all the eigenvectors are chosen and contribute to make ΔP larger the system can not communicate the input with the desired precision. On the contrary, if $b = 2$ the optimal ΔP is positive, and needs the selection of only 7 eigenvectors (figure 6.10 right).

It is instructive to consider the limit of input noise negligible respect to the input statistics ($\sigma_x << \sigma_{c_i}\ \forall i$). Then, $\lambda_i = \sigma_{c_i}^2\left(2\pi e \Delta_{x_i}^{-2} + \sigma_{c_i}^{-2}\right) = 1 + 2\pi e \frac{\sigma_{c_i}^2}{\Delta_{x_i}^2}$. On the other

hand, the receptive fields gains are $\frac{\sigma_y}{\sigma_{c_i}}\sqrt{\beta\frac{\sigma_{c_i}^2}{\Delta_x^2}-1}$. Due to the output discretization, the number of effective output symbols N_o is the square root of the output variance divided by the discretization bin σ_y, that is, $N_o = \frac{\sqrt{var(y)}}{\sigma_y}$, which can be calculated as $N_o = \sqrt{\sum\left(\beta\frac{\sigma_{c_i}^2}{\Delta_x^2}-1\right)}$. As we see, the number of output symbols is directly controlled by β. This allows us a straightforward interpretation of eq. 6.30: since that equation can be rewritten as $\beta\frac{\sigma_{c_i}^2}{\Delta_x^2}-1>0$, we conclude that a given eigenvector is chosen only if it contributes to create a not null number of output symbols.

The optimal configuration is thus equivalent to PCA [Oja, 1982] where the number of eigenvectors is determined by the input and noise statistics as well as by the desired precision. Moreover, PCA can be seen as a special case of ICA [Bell & Sejnowski, 1995][Amari *et al.*, 1996] where the statistics of the input sources are gaussian. Therefore, we expect to obtain similar results to ICA when applying the new information processing measure in the non-gaussian statistics case.

6.2.2 The problem of classification with a linear system

In this case the global objective of the system is to classify the inputs in N_C different classes. If the input statistics are constant (the agent can not act neither in previous stages of processing nor in changing the input statistics) the problem of maximizing ΔP is equivalent to minimize $d(y,g) = \beta H(y) - (\beta+1)H(y|g)$. Since the goal is a discrete variable we can write:

$$H(y|g) = \sum_{i=1}^{C} p(c_i)\cdot H(y|c_i) \tag{6.34}$$

where $p(c_i)$ is the a priori probability of class i [Cover & Thomas, 1991].

As in the previous section, we have a continuous system whose internal representation must be quantized in order to use our measure. On the other hand, we assume that we can describe satisfactorily the system dynamics using second-order statistics. Then:

$$p(\vec{y}|g_i) = \mathcal{G}(S_{w_i}, \vec{\mu}_i) \tag{6.35}$$

where $\vec{\mu}_i$ is the average of $\vec{y}$ given that the patterns are of class g_i. That is:

$$\vec{\mu}_i = \langle \vec{y} \rangle_{c_i} \tag{6.36}$$

where $< \cdot >_{c_i}$ indicates "expected value given class i". On the other hand, the intraclass scatering matrices S_{w_i} are the covariance matrices for a given class, that is:

$$S_{w_i} = \left\langle (\vec{y} - \vec{\mu}_i) \cdot (\vec{y} - \vec{\mu}_i)^T \right\rangle_{c_i} \tag{6.37}$$

Similarly to the previous sections, we must discretize the variables in our system in order to proceed. Note that since g is already discrete, it does not need further discretization. Using the equations 6.24 and 6.34 we get:

$$H(y^{\Delta}|g) = \frac{1}{2} \sum_{i=1}^{N_C} p_{c_i} \log \det(I + Q_y^{-T} A S_{w_i} A^T Q_y^{-1}) \tag{6.38}$$

Similarly we get

$$H(y^{\Delta}) = \frac{1}{2} \log \det(I + Q_y^{-T} A S_c A^T Q_y^{-1}) \tag{6.39}$$

with S_c being the total scattering matrix, given by:

$$S_c = \left\langle (\vec{y} - \vec{\mu}) \cdot (\vec{y} - \vec{\mu})^T \right\rangle$$

with $\mu = \langle \vec{y} \rangle$. Then $d(y, g)$ can be expressed as

$$\frac{1}{2} \log \frac{\prod_{i=1}^{N_C} \det(I + Q_y^{-T} A S_{w_i} A^T Q_y^{-1})^{(\beta+1)p_{c_i}}}{\det(I + Q_y^{-T} A S_c A^T Q_y^{-1})^{\beta}} \tag{6.40}$$

Since this expression is well defined for all matrices A, and tends to ∞ as at least one of its components goes to ∞ (we suppose S_{w_i} of full rank) the expression reaches its global minimum in a finite point, where the gradient respect to A is null. Then, defining $V \equiv Q_y^{-T} A$ we get:

$$(\beta+1)\sum_{i=1}^{N_C} p_{c_i}\cdot(I+V\,S_{w_i}\,V^T)^{-1}\,V\,S_{w_i}-\beta\cdot(I+V\,S_c\,V^T)^{-1}\,V\,S_c=0 \qquad (6.41)$$

This equation is valid for any number of classes and neurons. Since in principle there are no preferred directions in the internal representations, we fix $Q_y = q_y \cdot I$ with q_y being a scalar.

For a problem with two classes and one processing unit this can be rewritten after simplifications as:

$$(a\,S_{w_1}+b\,S_{w_2})\vec{w}=S_b\vec{w} \qquad (6.42)$$

where the interclass scattering matrix S_b is defined as $S_b \equiv \sum_{i=1}^{2} p_{c_i}(\mu_i-\mu)\cdot(\mu_i-\mu)^T$, and we have used the property $S_c = S_w + S_b = S_{w_1} + S_{w_2} + S_b$. The other parameters are $a=\frac{q_y^{-2}\vec{w}^T S_c\vec{w}+1}{\gamma p_2\left(q_y^{-2}\vec{w}^T S_{w_1}\vec{w}+1\right)}-\frac{1}{p_1}$, $b=\frac{q_y^{-2}\vec{w}^T S_c\vec{w}+1}{\gamma p_0\left(q_y^{-2}\vec{w}^T S_{w_2}\vec{w}+1\right)}-\frac{1}{p_1}$, and $\gamma\equiv\frac{\beta}{\beta+1}$.

The parameters a and b are positive if there is enough interclass overlapping, which on the other hand corresponds to the situation where our approximation $p(y) = \mathcal{G}(S_c, \vec{\mu})$ holds. Then the optimal configuration is an eigenvector of $(aS_{w_1}+bS_{w_2})^{-1}S_b$. This corresponds to the solution of the classical Fisher discrimination analysis but with modified a priori probabilities (the normalized parameters $\frac{a}{a+b}$, $\frac{b}{a+b}$ play this role) [Duda & Hart, 1973]. In some special cases this solution is the same as the Fisher discriminator with the same a priori probabilities, for example the case when $S_{w_0} = \lambda S_{w_1}$.

6.3 Analysis of nonlinear systems

6.3.1 Construction of decision trees

In this section we apply the new information processing measure to the induction of decision trees. The output of a decision tree for an input pattern is the terminal node that classifies that pattern [Quinlan, 1986]. We would like $\Delta P(X \to Y|G)$ to be maximized for the induced tree. Since $\Delta P(X \to Y|G) = d(X,G) - d(Y,G)$, the

maximization of this quantity is equivalent to the minimization of $d(Y,G)$ since the input statistics are constant in this context. It follows that $d(Y,G)$ can be written as (see the appendix C.9 for the details):

$$d(Y,G) = d(Y_{noN},G) + p(N)\left(H_N(Y|G) - \beta I_N(Y;G)\right) \tag{6.43}$$

where $d(\vec{Y}_{noN}),G$ is the distance of the tree without the subtree N to G, and $H_N(Y)$ and $I_N(Y;G)$ are computed using the local statistics in N.

We see that contribution of the subtree N to the global distance depends on its local information. In case we are asked to expand a node we should choose that one which contributes to make $d(Y,G)$ smaller, that is, which maximizes the functional

$$\beta I_N(A;G) - H_N(A|G) \tag{6.44}$$

Therefore it is natural to define a greedy construction algorithm that starts with a root node, choosing the expansion A that maximizes 6.44, and then use it recursively in the children subtrees. Note that if this quantity is negative, it will contribute to make eq. 6.43 greater. Therefore, if we reach a node where all possible expansions make eq. 6.44 negative we stop expanding that branch. As it can be seen, the new measure provides a method for constructing decision trees as well as a natural stopping criteria to avoid overfitting. This is in contrast with many other algorithms which use a local information gain measure such as

$$Gain_N(A) \equiv H_N(G) - H_N(G|A) = I_N(A;G) \tag{6.45}$$

in order to evaluate the goodness of an expansion A in node N (such as ID3, C45, C5 [Mitchell, 1997]). Note that the maximization of equation 6.44 with very high β is equivalent to the maximization of the information gain 6.45. Since β represents how much the uncertainty about the goal is weighted in ΔP respect to the complexity term (section 5.7.1), we can interpret the information gain as a maximization of ΔP when the complexity of the resulting tree is not taken into account.

Since the information gain 6.45 is always ≥ 0, it achieves the value 0 when the

number of examples in the node to expand is one or all the examples at the node have the very same class [Mitchell, 1997]. Therefore a recursive application of the gain information criteria would make the tree expand since all the examples at a terminal node have the same class. This produces very complex trees in general that need to be post-pruned [Quinlan, 1986]. Moreover this procedure has difficulties with attributes with many possible values [Mitchell, 1997] (since it has a bias to them and they are not very informative). For this reason in the literature the gain ration is defined [Mitchell, 1997] as

$$Gain\,ratio_N(A) \equiv \frac{I_N(G;A)}{H_N(A)} \tag{6.46}$$

to overcome this problem, but still it has the problem of being always a positive quantity. Interestingly, another technique suggested to solve the problem of attributes with many values [LópezdeMántaras, 1991] is a special case of our measure with $\beta = 1$.

The greedy maximization of the new proposed measure for the induction of decision trees can be seen as a technique which combines the good features from the information gain, the gain ratio and early stopping. Figure 6.11 displays the results of applying C4.5 to the gaussians database without pruning. The figure on the middle corresponds to the zoom in of the figure on the left. The figure on the right is the result of applying the new measure to the induction of the tree. As it can be observed no pruning is needed.

It is important to mention that our measure can be used to compare different potential expansions formed by complete subtrees. This allows to perform the tree expansion mixing hill-climbing with depth search. The amount of depth-search respect to hill-climbing can be easily tuned just by adjusting the maximum depth of the "candidate subtrees" to expand a node in each iteration.

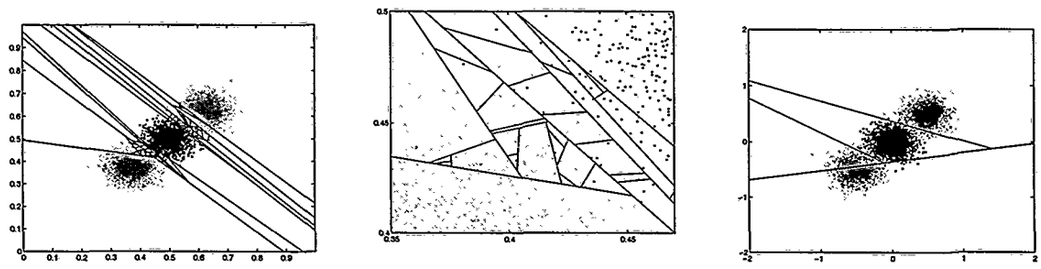

Figure 6.11: Overfitting with a classical inductive decision tree. The figure on the left displays the results of applying C4.5 to the gaussians database without pruning. The figure on the middle corresponds to the zoom in of the figure on the left. The figure on the right is the result of applying the new measure to the induction of the tree.

6.3.2 Perceptron Learning

Perceptron for a simple classification task

Here we investigate learning the optimal structure of a network of nonlinear units in a simple classification task. The dataset *gaussians* consists of three equiprobable clusters of data elements belonging to two different classes. There are three mutually exclusive processes which generate vectors $(x1, x2)$ following gaussian overlapping distributions. Two of the processes are considered of class "A" (grey) while one of them is considered of class "B" (black). The goal of the global system is thus to predict, given a new example $(x1, x2)$, to which of the 2 classes it belongs to.

We consider that in our global system the first processing step is a layer of nonlinear neurons. The output of the *ith* classifier (y_i) is 1 in case $\vec{m}_i\vec{x}+b_i > 0$, 0 otherwise, where $\vec{x}$ is the input pattern. The binary vector composed by all the classifiers outputs $\vec{Y}$ determines the achievable accuracy of the rest of the system as well as the amount of processing it has to do.

The adaptive system must find the configuration that maximizes ΔP. The classifiers configurations have been generated by searching the parameter space by means of a genetic algorithm [Levine, 1998] due to its global search properties. The β parameter is chosen as 4, the number of examples used in the optimization is 10000. All the parameters and initial random weights in the genetic algorithm are equal in all

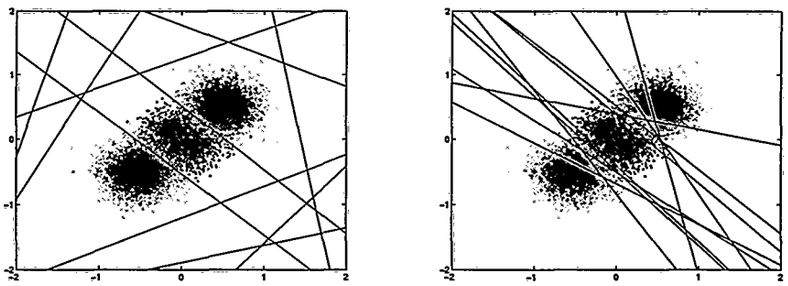

Figure 6.12: Comparison of the solutions that optimize the new measure with respect to the solutions that optimize mutual information for the perceptron case. The goal is to separate points of class "gray" from points of class "black". The points are generated by three equiprobable clusters defined by gaussian distributions. **Left.** Results when using ten processing elements as the maximum number of resources with the new information processing measure. Notice that the new measure needs to use only two out of the ten maximum number. **Right.** Similarly for mutual information. Notice that mutual information uses the ten classifiers.

the simulations performed with the gaussians data set. We have performed several computer experiments with different random seeds and parameters leading to the same results.

We have compared the solutions that optimize the new measure with respect to the solutions that optimize the mutual information between Y and G. For the case of the new measure, processing is equal to complexity reduction minus loss of important information. Yet, mutual information only takes into account uncertainty minimization ignoring the reduction of complexity. Figure 6.12 A displays the configuration selected when using a pool of ten nonlinear units. Note that the optimal configuration only uses two of them since the output of the rest is kept constant. However, if mutual information $I(Y;G)$ is chosen as the objective function to maximize we obtain a configuration where all the resources are used (figure 6.12 right). This is due to the fact that mutual information only takes into account uncertainty minimization ignoring extraction of redundancies for this simple task.

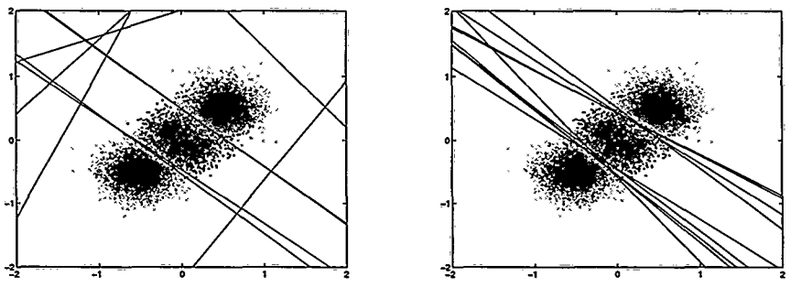

Figure 6.13: Results when using ten stochastic processing elements as the maximum number of resources with the new information processing measure. Left: Level of noise = 2 %. Right: level of noise = 20 %.

Stochastic neurons and population coding

In this section we consider the same classification problem as in the previous section but now the neurons are stochastic. The output of each neuron is computed as previously, but then each neuron switches its output with certain probability. The optimal system is again calculated using a genetic algorithm. All the parameters and initial random weights in the genetic algorithm are equal to the simulations in the previous section. Notice that for a level of noise of 2% the system uses more than 2 classifiers (figure 6.13 Left). Also notice that the new measure begins to use more resources to account for the noise in the input data. If the noise is increased up to 20% we see that the optimal system uses ten classifiers (figure 6.13 Right).

Perceptron for Proben1 tasks

In this section we consider the *cancer1* and *heart1* databases from the *proben1* archive [Prechelt, 1994]. As before we calculate the configuration of the perceptron which maximizes ΔP. When the optimization process finishes, we assign to each of the different spatial regions delimited by the classifiers the more frequent class in that zone. Each concrete value of β determines a corresponding optimal configuration which has a particular classification error in the validation set. In general high values of β determine more complex representations, and thus more likely to suffer overfitting. Thus the optimal value of this metaparameter is adjusted by cross-validation in a

separate validation set, yielding the following algorithm:

1. Initialize $\beta = \beta_0$
2. Optimize ΔP using the training set
3. Label each zone with the more frequent class there
4. Compute classification error in the validation set
5. If it is satisfactory, compute the error in the test set and stop. Otherwise select a new β and go to 2

For this problem we search the optimal β using a greedy strategy (*gold section* [Press *et al.*, 1992]). This metaparameter can be adjusted in other ways such as in an on-line manner which does not distinguishes between training and validation sets ("trial and test technique").

	Distance		Mutual Information		Proben1
Data Set	Test error (%)	C	Test error (%)	C	Test error
gaussians	5.2	2	5.4	10	-
cancer1	0.57	2	11.49	10	1.149
cancer2	4.598	1	14.94	10	5.747
heart1	22.17	5	57.4	15	20.00

Table 6.1: Test errors for the different databases. C is the number of used classifiers out of the maximum (15 for the *heart1* database, 10 for the others). Note that the maximization of the mutual information between Y and G conducts to configurations which use all the available resources.

Table 6.1 we show our results. In all cases we have compared the solutions that optimize the new measure with respect to the solutions that optimize the mutual information between Y and G. The new measure proves to be clearly superior under conditions of noise, overfitting and allocation of optimal number of resources.

6.3.3 Non Linear Feature Extraction for classification

General algorithm

In this case we have a global system whose objective is to classify the inputs in N_C different classes. In order to achieve its goal efficiently, it needs to build optimal internal representations of the data. In this section we will study the optimal internal representation for a layer of nonlinear neurons. The measure of information processing for this layer is:

$$\Delta P(\vec{x} \rightarrow \vec{y}|g) = d(\vec{x},\vec{g}) - d(\vec{y},\vec{g}) = d(\vec{x},\vec{g}) + \beta H(\vec{y}^{\Delta}) - (\beta+1)H(\vec{y}^{\Delta}|g)$$

The representation will be optimal if ΔP is maximized. Since we will concern with the optimization of the internal representation $\vec{y}$, $d(\vec{x},\vec{g})$ is constant. Therefore, the functional to maximize for practical purposes is:

$$\beta H(\vec{y}^{\Delta}) - (\beta+1)H(\vec{y}^{\Delta}|g)$$

The response of each processing unit to the input $\vec{x}$ is nonlinear:

$$y_k(\vec{x}) = f_k(\vec{x},\vec{\alpha}_k)$$

with the vector $\vec{\alpha}_k$ representing the parameters of that processing unit.
Since we have discrete goals:

$$H(y|g) = \sum_{i=1}^{C} p(c_i) \cdot H(y|c_i) \tag{6.47}$$

On the other hand we will assume the internal states dynamics to be satisfactorily described by second-order statistics. Then:

$$p(\vec{y}) \simeq \mathcal{G}(Sc, \vec{y} - \overline{\vec{y}})$$

with S_c being the scattering matrix of the internal states:

$$Sc = \left\langle (\vec{y} - \overline{\vec{y}})(\vec{y} - \overline{\vec{y}})^T \right\rangle$$

Then, using the equation 6.24 we get $H(\vec{y}^{\Delta}) = \frac{1}{2}\log\det(I + \tilde{Q}_y^{-2}Sc)$ with Q_y being a quantization matrix. Using the same strategy in $H(\vec{y}|c_i)$ we get:

$$H(\vec{y}^{\Delta}|c_i) = \frac{1}{2}\log\det(I + \tilde{Q}_y^{-2}Sw_i)$$

with Sw_i being the intraclass scattering matrices of the internal representation:

$$Sw_i = \left\langle (\vec{y} - \overline{\vec{y}})(\vec{y} - \overline{\vec{y}})^T \right\rangle_{c_i}$$

Then the functional to maximize is given by:

$$\beta\frac{1}{2}\log\det(I + \tilde{Q}_y^{-2}S_c) - (\beta + 1)\sum_{i=1}^{N_c} p_{c_i}\frac{1}{2}\log\det(I + \tilde{Q}_y^{-2}S_{w_i}) \qquad (6.48)$$

Since in principle there are no preferred directions in the internal representations, we fix $Q_y = q_y \cdot I$ with q_y being a scalar. The concrete election of q_y is not important, we have seen that $q_y = .1$ works fine.

Gradient algorithm

For clarity, we will introduce several new symbols in our notation as in [Cruz & Dorronsoro, 1998]. There are N_c classes and N_i examples in class i. The number of total examples N will be $N = \sum_{i=1}^{N_c} N_i$. We will order the examples in each class so that $\vec{x}_j^i$ is the jth training example in class i. The specific order will be irrelevant for our purposes. Correspondingly, the response of the layer to such an example will be denoted by $\vec{y}_j^i$. Finally, $\vec{\mu}$ is the total average of $\vec{x}$:

$$\vec{\mu} = \frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\vec{x}_j^i$$

and $\vec{\mu}_i$ is the average of $\vec{x}$ in class i:

$$\vec{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \vec{x}_j^i$$

The scattering matrices are then determined by:

$$S_c = \frac{1}{N} \sum_{i=1}^{N_c} \sum_{j=1}^{N_i} (\vec{y}_j^i - \vec{\mu})(\vec{y}_j^i - \vec{\mu})^T \tag{6.49}$$

and

$$S_{w_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} (\vec{y}_j^i - \vec{\mu}_i)(\vec{y}_j^i - \vec{\mu}_i)^T \tag{6.50}$$

We have implemented two different classes of algorithms in order to find the parameters $\vec{\alpha}_i$ which maximize ΔP: genetic algorithms and gradient search. In both cases the equations we have derived allow us to design optimal internal representations for a classification problem with an arbitrary number of classes.

The gradient of the functional 6.48 is derived in appendix C.8:

$$\frac{\partial \Delta P}{\partial \alpha_{kz}} = \frac{1}{N} \sum_{i=1}^{N_c} \sum_{j=1}^{N_i} \frac{\partial \left[\vec{y}_j^i\right]_k}{\partial \alpha_{kz}} \cdot \left[\beta(\tilde{Q}_y^2 + S_c)^{-1}(\vec{y}_j^i - \vec{\mu}) - (\beta + 1)(\tilde{Q}_y^2 + S_{w_i})^{-1}(\vec{y}_j^i - \vec{\mu}_i)\right]_k \tag{6.51}$$

where $[\cdot]_k$ indicates the kth component of the vector.

The computational complexity of a gradient descent of ΔP is similar to other algorithms for nonlinear feature extraction (see [Cruz & Dorronsoro, 1998] for instance). Note that equation 6.51 is general from several points of view: first, it is valid for any number of classes. Second, it is valid for any nonlinear activation function. We show examples where we use sigmoidal activation functions but we could have used splines and other nonlinearities. Third, there can be any mix of neurons with different nonlinear activation functions (e.g. sigmoidal + linear + splines).

Concrete choice of the activation functions

In the examples we show we use sigmoidal activation functions. That is, the response of neuron k to the input is given by $y_k = sigm(\vec{\alpha}_k \cdot \vec{x})$, with $sigm(x) = \frac{1}{1+e^{-x}}$ (sigmoidal function). Therefore, $\frac{\partial y_k}{\partial \alpha_{kz}} = y_k(1 - y_k)x_z$. In order to take into account the bias constant for each neuron, we expand $\vec{x}$ including a constant component equal to 1.

The problem of the three gaussians

We have used the learning algorithm with the classification problem introduced in section 6.3.1. The learning algorithm is very fast, allowing for high learning rates and converging in a relative small number of epochs.

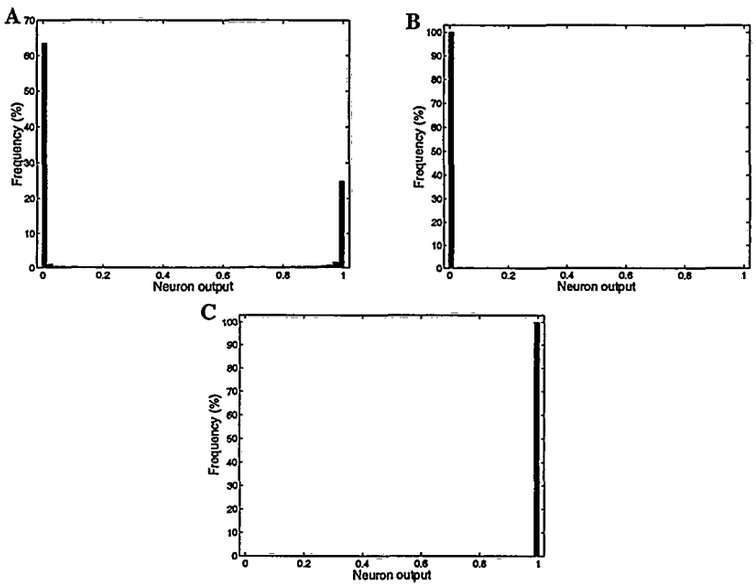

Figure 6.14: Two different types of neurons in the optimal configuration. A: the histogram of the activity is composed by two sharp peaks (in 0 and 1): the activity is binary. B and C: the activity of the neuron is constant (0 or 1).

After training, the neurons can be divided in two categories: neurons with binary activity (fig. 6.14 A), and neurons with constant activity (fig. 6.14 B and C). If we

look at the parameters the network have learned, α_k, they have very high absolute value. This means that the optimization process introduces a high gain in the sigmoidal function, so that $sigm(x) \simeq 1$ for $x > .5$, and 0 for $x < .5$. This can be interpreted so that the optimization of ΔP leads to a natural discretization of the neurons activity. In other words, from the continuum of different possible states of each neuron, the optimal solution uses only two. This will reduce dramatically the complexity of the internal states. Moreover, it will make us able to easily interpret the internal states of the system as we will see. This important property will be also present in the following examples in this section.

In figure 6.15 we see the optimal internal representation found by the gradient descent algorithm for the problem of the three gaussians.

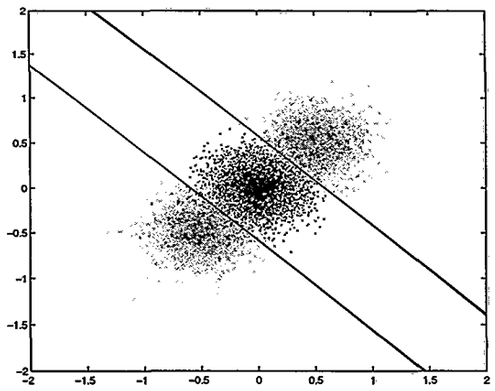

Figure 6.15: Optimal solution for a layer of 4 neurons. The lower decision frontier is in reality the overlapping of two neurons. This same effect occurs with the upper frontier which is the overlapping of the other 2 neurons.

Each line represents the "decision frontier" of each neuron. Points in one side of the frontier will make the neuron output be 0, and points in the other side will make it output 1 ($\beta = 10$).

From the 4 different neurons of the layer, the optimal solution we show uses 2 + 2 (completely overlapping between them). Structurally, it is equivalent to a solution with only 2 neurons. We also have observed optimal solutions where the "redundant neurons" are pushed away from the central area, therefore having constant activity.

Noisy xor

We have seen that the algorithm we developed founds the optimal representation for the problem of the three gaussians. Now we check if it is able to cope with strong nonlinear problems such as xor. In figure 6.16a we show the optimal configuration found by the algorithm ($\beta = 5$) for a layer of 4 neurons. As in the previous section, each one of the two decision frontiers is the superposition of two neurons.

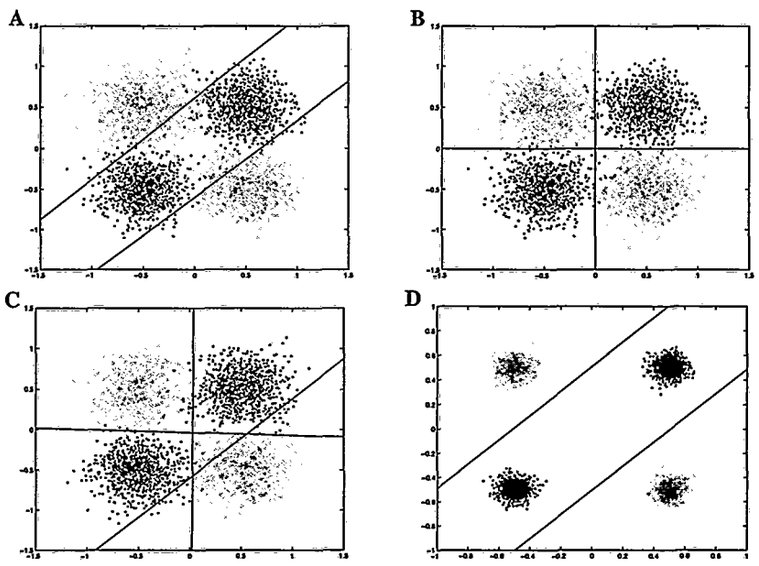

Figure 6.16: **A:** optimal solution. **B:** the configuration with vertical and horizontal rfs is not optimal. **C:** example of local maximum. **D:** problem with lower gaussian dispersion.

Thus the optimal configuration separates completely the problem by diagonals. The "factorial solution", (fig. 6.16 B) is not optimal since has greater ΔP. This is because now the number of internal states is greater (4 compared to 3), yielding similar classification quality. Therefore, the factorial solution is structurally more redundant, thus having worst ΔP. Sometimes the gradient algorithm stops at local maxima (fig. 6.16 C). This problem disappears when a global search using a genetic algorithm is performed.

In general we observe that the algorithm tends to maximize the margins in the decision frontiers. This is clearly shown in figure 6.16 D where we have reduced the dispersion of the gaussians for a better appreciation of this effect.

The ring problem

Finally, we will test our algorithm with a "hard" problem from the point of view of the activation functions we are using. That is, a problem with non trivial solutions in terms of linear decision frontiers. The task is to discriminate between a cluster of points distributed along a noisy ring ("class A"), and a cluster of points centered in it ("class B") (see figure 6.17). The problem is intrinsically nonlinear and does not admit trivial solutions in terms of linear decision frontiers.

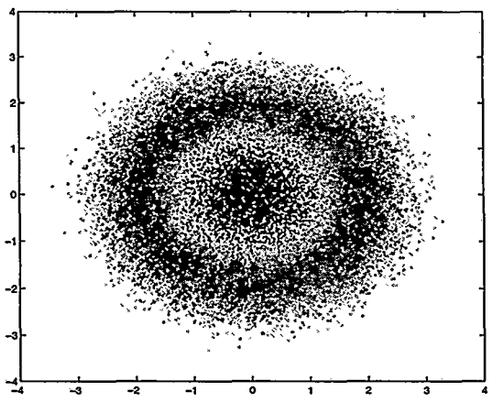

Figure 6.17: The ring classification problem. The task is to discriminate between class "grey" and class "black"

In figure 6.18 we see the optimal solution with $\beta = 20$. Each of the lines is in reality the exact superposition of two neurons. The other 4 are put far from the data so that they remain virtually constant. If now we increase β to 28, the optimal solution forms a hexagon. Finally, with $\beta = 60$ the optimal solution uses all the available resources forming a polygon of 20 sides with better accuracy in the discrimination but higher complexity.

Therefore, β acts as a "structural complexity tuner". That is, making it greater

we achieve more complex solutions. Note that this complexity is not just the number of neurons that are used but the way they are used (cf *structural complexity*).

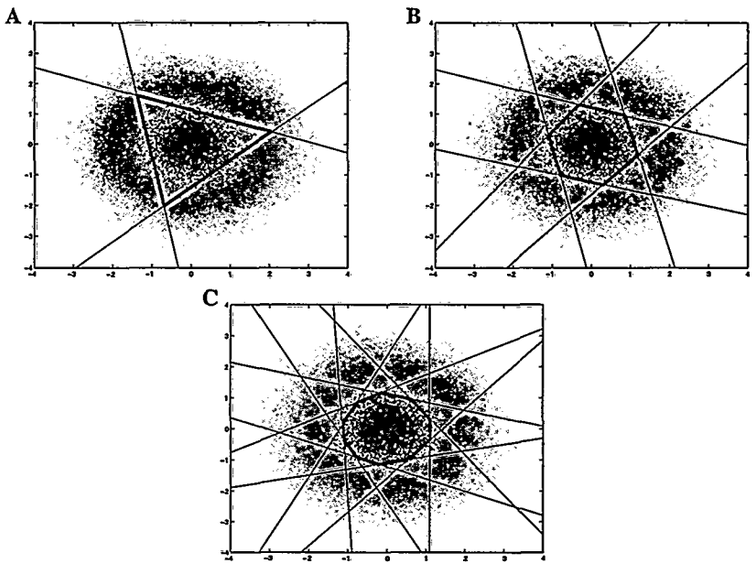

Figure 6.18: **A:** optimal solution for $\beta = 20$. **B:** $\beta = 28$. **C:** $\beta = 60$.

6.4 The olfactory and auditory systems in the context of the new framework

In chapter 4 we showed that the principle of maximum information transfer is valid only for certain sensory systems. Additionally, this principle alone is not well-defined and needs of additional constraints. Otherwise, the gains of the receptive fields in the optimal configuration are ∞ (see sections 2.3.3 and 3.4.3, and the reference [Campa *et al.*, 1995] for example). In this section we will analyze the biological systems exposed in the first part of the thesis in the context of the new framework. We will use the basic scheme shown in figure 6.19 where we consider a system which is trying to transmit as much information as possible about the aspects of the original signal $\vec{x}$ which are relevant for the task.

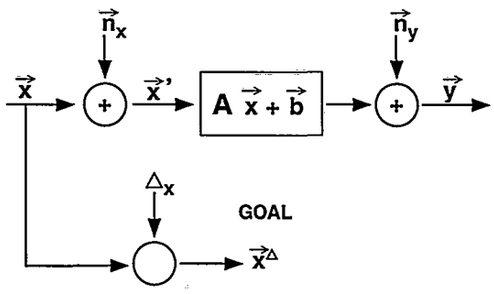

Figure 6.19: Schema used for modeling sensory biological systems. The goal of the system is to communicate the input $\vec{x}$ with a degree of precision given by the matrix Δ_x. The most important aspects of the information are required to be transmitted with higher precision.

The vector $\vec{n}_x$ represents any kind of noise (hardware or semantical) present in the ideal signal the system is trying to transmit. On the other hand the neurons in the system are intrinsically noisy, which is represented by the stochastic term $\vec{n}_y$.

In section 6.2.1 we have shown that the principle of maximization of ΔP applied to this kind of system is always well-defined and conducts to a family of optimal solutions. However these optimal solutions could not be constructed by the biological system due to physical constraints such as limited gain in the neurons or energy consumption. In this section we will show that the principle of maximization of ΔP combined by biological restrictions conducts to general properties very similar to those observed in biology.

6.4.1 The olfactory epithelium

Now we will proceed to derive the optimal theoretical configuration which maximizes ΔP in our model of the olfactory epithelium (section 2.4). The biological constraints we consider are a limited size in the gene pool and a maximum gain in the sensitivities of the neurons ($-c \leq a_{ij} \leq c$). Since we do not know the matrix of correlations of the input (C) we have performed several simulations with randomly generated matrices. In figure 6.20 we show the general properties of the optimal configuration using a genetic algorithm [Levine, 1998]. In general we observe no qualitative change for different choices of C and Δ_x as long as β is high enough.

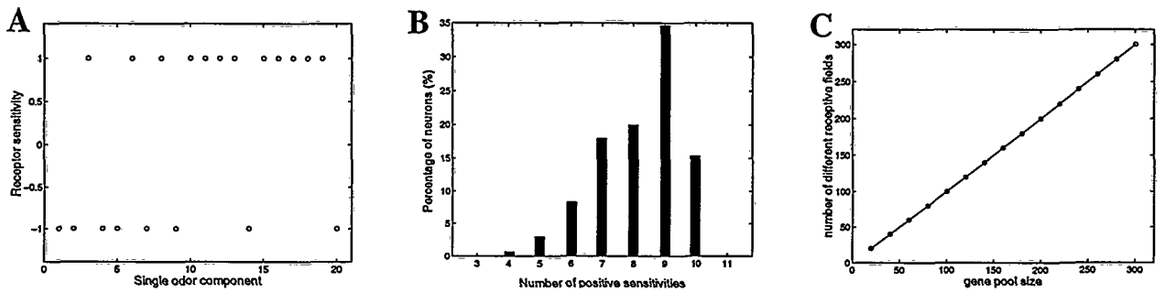

Figure 6.20: Properties of the optimal configuration for an homogeneous gene expression. The input dimension is chosen as 20. **A:** Odor sensitivities of an arbitrary neuron. **B:** Percentage of genes which code a receptive field with a given number of positive sensitivities. Since a change $\vec{u}_i \to -\vec{u}_i$ is irrelevant for the Fisher information we normalize the global sign so that if $\vec{u}_i$ has more than 10 positive sensitivities it is multiplied by -1. **C:** Number of different receptive fields as a function of the gene pool size.

The optimal configuration thus shows properties very similar to those exposed in section 2.4.

6.4.2 The auditory cortex

Now we proceed to analyze the auditory cortex using our framework. Then we will compare the results with our realistic model. Let us consider that the goal of the system is to reconstruct each frequency band i with a given level of accuracy Δ_i. The different accuracies are imposed by the environment though the tasks the animal has to solve. Let us describe the stimuli statistics using second-order statistics. Then our results about the autoencoder (section 6.2.1) are valid in this situation. For simplicity in the equations we will consider that the receptor's noise variance σ_x is very small compared with the variance of the stimuli. Then there is a family of optimal solutions, one of them being characterized by:

- Define $a_i \equiv \frac{\sigma_{C_i}^2}{\sigma_x^2}$ and $b \equiv \frac{\Delta_x^2}{\sigma_x^2}$. Take the eigenvectors (normalized) $\vec{u}_i$ of Φ with greatest eigenvalues λ_i which satisfy $\lambda_i > 1 + \frac{1}{\beta}$. In case there are none, take $A = 0$. If the number of neurons is less than the number of eigenvectors satisfying the requirement, take the eigenvectors with greatest eigenvalues

- Assign only one neuron to one of the selected eigenvectors using gain $\frac{\sigma_y}{\sigma_x}\sqrt{\beta\frac{\sigma_x}{\Delta_x}}$

Then any optimal solution is an orthogonal transformation of this basic configuration in the space of neurons.

Since the eigenvectors of Φ are the eigenvectors of the stimulus correlation matrix C (section C.7.4), we will first determine which form has this covariance matrix. In figure 6.21 we show the covariance matrix C calculated with a bank of real sounds recorded in natural environments. As it can be seen, we can consider it as essentially diagonal. This is reasonable since in a real environment there is an enormous variety of sound sources each one with its corresponding complex spectral pattern. Since these sources can occur in very different combinations we would not expect strong correlations between different frequency bands.

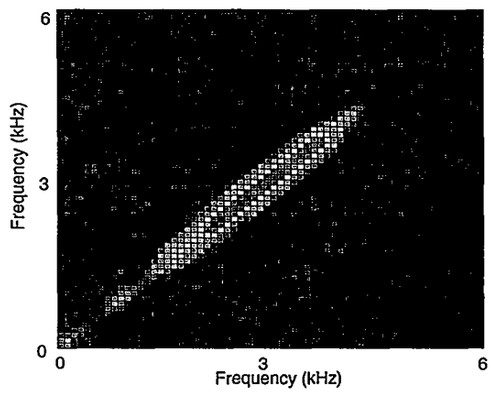

Figure 6.21: Correlation matrix of the frequency components of sounds in natural environments. Red: high correlation. Blue: low correlation. The sounds were obtained from demo sounds at the Macaulay Library of Natural Sounds (Cornell university, http://birds.cornell.edu/lns/), the Earthear catalog (www.earthear.com) and the CD "Amazon Rainforest" (Hugues, Carlton Home Entertainment, 1995). The different samples represent a variety of sounds. The sounds were resampled at 11 kHz and then the FFT of moving windows of 128 samples was computed (this corresponds to a block length of 11.6 ms), representing an estimation of the spectral content at that instant. The correlation matrix is then estimated using the spectrograms of the different sounds.

Therefore the eigenvectors of C are the axes in the frequency domain, and so will be the eigenvectors of Φ. However the knowledge of C does not completely

fix the solution since there is a family of different optimal solutions. Which of the configurations out of the family should be chosen by the biological system ? The concrete election of the orthogonal transformation now depends on the biological constraints of the system. If the system is trying to minimize the consumed energy and the individual statistical independency between the neurons then it should choose a factorial code [Barlow, 1989, Olshausen & Field, 1996, Baddeley, 1996]. Then the optimal configuration is equal to the simple one since then the pdf of the population activity is factorial [Campa *et al.*, 1995]. Thus the greatest eigenvectors are each one represented by only one neuron, and the rest of the population does not respond at anything at all.

However the required gains of the neurons, $(\frac{\sigma_y}{\sigma_x}\sqrt{\beta\frac{\sigma_x}{\Delta_x}})$, can be too large respect to the maximum gain of the neuron which we will call γ. Then an exactly factorial solution can not be implemented in the biological system. How can the system implement a nearly factorial code ? The solution to this situation is to code an eigenvector by a group of neurons so that their receptive fields are proportional to the corresponding eigenvector and on the other hand the square sum of the individual gains are equal to the square of $\frac{\sigma_y}{\sigma_x}\sqrt{\beta\frac{\sigma_x}{\Delta_x}}$. This configuration can be easily proved to be a rotation of the simple solution and therefore an optimal configuration. We want the group of neurons to be as small as possible in order to not affect much the factoriality of the code. Then each of these groups should be composed by $\frac{\sigma_y}{\gamma\sigma_x}\sqrt{\beta\frac{\sigma_x}{\Delta_{x_i}}}$ neurons each with maximum gain γ. The exact number of neurons which code each frequency band is thus inversely proportional to the square root of the required precision for that zone Δ_{x_i}.

In order to illustrate the results we will consider a population of 100 neurons with 10 frequency bands. These numbers are arbitrary and are not critical for our results. Figure 6.22a shows the optimal configuration when $\frac{\sigma_y}{\gamma\sigma_x}\sqrt{\beta\frac{\sigma_x}{\Delta_{x_i}}} = 5$ for all the frequency bands. That is, all the frequency bands carry a similar amount of information about the task. We see that all the frequency bands are represented by the same number of neurons. Importantly, this number of neurons depends on the required precision for estimating the frequency band but not on the statistics of the stimuli. On the other hand, the rest of neurons in the model are not necessary to solve

the task with the desired precision under the given noise conditions, and therefore are not included in the configuration. These can be seen as neurons which are reserved for possible novel situations in the environment (section 3.4.1). All these results are in accord with section 3.4.1 where we analyzed our realistic model of auditory cortex.

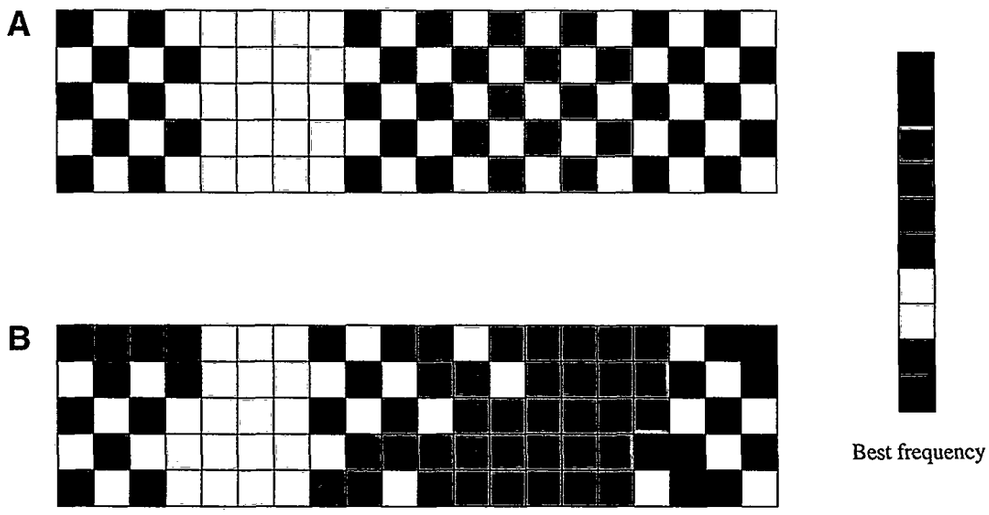

Figure 6.22: Configuration that maximizes ΔP in the auditory cortex. **A:** The desired precision for each frequency band is the same for each one. **B:** The *8th* frequency band is required to be estimated with higher precision respect to the others.

In figure 6.22 we show the optimal configuration when the system is required to estimate the *8th* frequency band more precisely. All other frequencies are represented as before but now the 8th frequency band is represented by more neurons. Therefore the system assigns to this frequency band more resources than to the others. Again this is exactly the same properties shown by the biological system [Kilgard & Merzenich, 1998] and by our realistic network. Therefore we can conclude that the auditory cortex is maximizing the effective processing measure ΔP with the additional ingredient of being factorizing the code.

Part IV

Conclusions and Future Work

Chapter 7

Conclusions and Future Work

7.1 Summary of the thesis

This thesis presents a new formal framework based on a new information processing measure that allows both the analysis and design of adaptive complex systems. The framework can be used to analyze how different biological perceptual systems construct optimal internal representations of the environment. Additionally the proposed framework allows new predictions and may shed a new light on how the different perceptual systems perform the analysis of their environment. From a more analytical point of view the proposed framework allows the construction of a map of several of the existing algorithms in neural computation and machine learning, for instance, PCA, Fisher discrimination analysis, C4.5, etc. and can help to elucidate their common analogies and different functionalities. Finally the proposed framework gives rise to the construction of new powerful algorithms for the design of adaptive complex systems. For instance we present a new learning method for decision tree construction derived from our framework which contains several features of existing methods. We also present a new algorithm for constructing optimal nonlinear representations (nonlinear feature extraction) in classification problems.

The thesis starts with a theoretical study of biological systems since they are very efficient in their interaction with their environment. In order to perform optimally

they must construct adequate internal representations of the complex sensory information they receive. Therefore a study of principles of organization in those systems will provide us with useful insights about what the principles of organization of an optimal adaptive system should be. Specifically, we show that the neural configuration which maximizes the information transfer in a simple model of the olfactory epithelium has very similar properties to the real system: the receptive fields show bipolarity, unspecificity, and homogeneous distribution. Thus we can say that the olfactory epithelium represents the information optimally from the point of view of information theory. However this representation is genetically encoded and does not result as a consequence of self-organization or experience. Then we proceed to study the representation of the information in auditory and visual cortices where plasticity and learning occurs and the internal representations are more elaborated. As we show the internal representations formed in the computer models are very similar to the real ones, and emerge as a consequence of the cooperation and competition at several levels. The receptive fields formed by these mechanisms give rise to neurons with more specific receptive fields than those seen in the olfactory model. Near neurons in space tend to code the same feature while far neurons tend to code different aspects of the information. Thus the information is represented by *functional groups of neurons* (for related concepts see for example [Hebb, 1949, Abeles, 1991, Tononi & Edelman, 1998]). Due to this specificity, the internal representations in the auditory and visual cortex are *sparse codings*, that is, the stimulus is represented by only a few active cells out of a potentially much higher number [Olshausen & Field, 1996, Baddeley, 1996] as in contrast with the olfactory system. This strategy is efficient in the sense that it minimizes the complexity and energetic cost of the code while maximizing the representational accuracy for natural images [Olshausen & Field, 1996, Baddeley, 1996].

The above considerations suggest that if a general framework exists then it should be valid for describing any complex system with an arbitrary internal code and arbitrary implementation details. This suggests that the general theory should be expressed in terms of the internal states of the system and not on physical parameters related to the specific implementation of the system. On the other hand the notion of *low complexity of the representation* is a key ingredient in the cortical representations

of the visual and auditory cortices. Thus the biological systems reduce the intrinsic complexity of the sensory input by constructing higher level representations of the information (e.g. *border detectors* in the visual cortex versus *pixel detectors* in the array of photoreceptors in the eye [Bear *et al.*, 1996]). This is intuitive since the ultimate goal of the animal is to solve the tasks the environment imposes and thus an efficient representation of the information which captures the regularities of the environment is critical [Barlow, 1989]. Thus the concept of *complexity reduction* seems to be another basic ingredient in our desired theory. Finally, the experimental observations in the auditory cortex and our theoretical analysis show that the representation of information in that structure is modulated by experience as the behavioral importance of the stimuli change in time. Thus the internal representation is biased to behaviorally important stimuli. These experiments reveal the important principle that internal representations, even at primary levels of processing, are influenced by the tasks the environment imposes on the animal. This seems very reasonable since the resources that a biological system has are limited, and due to the high complexity of the environment and the large amount of different stimuli it receives, the animal should focus on that part of the information which is really relevant for the task. Apart from these considerations about the limitation of resources, to focus on the relevant part of the information neglecting its spurious and noisy aspects is critical for a proper generalization. These considerations conducts us to the principle that *the concept of task should be a crucial element in a general theory of information processing.*

Given these conclusions we derive a general mathematical framework which contains the essence of these principles. In order to make it as general as possible we use the notions of *agent* and *environment* valid for any system which interacts with its environment. We introduce the concept of *amount of effective information processing* (ΔP) performed by a part $\mathcal{W}$ of the agent which has to solve the task imposed by the environment. This notion does not depend on the specific implementation but on the global statistical relations between that part of the agent and the environment. Then the crucial notion of distance to the task emerges, which depends on both the level of uncertainty and complexity that the information processed by $\mathcal{W}$ has with respect to the task. Since we express this theory using the general notions of agent

and environment and the mathematical tools the theory uses are implementation-independent, the framework can be used to analyze the interaction of any complex adaptive system with its environment (whether biological or artificial) as well as to obtain new optimal learning strategies for artificial systems.

Since biological systems are very efficient interacting with their environment we would expect them to build internal representations which maximize ΔP. As we will show the theoretical configuration which maximizes ΔP in the auditory cortex has very similar properties to the biological system. Moreover the specific properties of this optimal configuration are very similar to those studied with the realistic model. This is the same when we calculate the theoretical configuration which maximizes ΔP in the olfactory and visual cortex. Therefore we conclude that ΔP seems to be maximized in biological systems, which constitutes a validation of our theory and demonstrates the potentiality of our framework to study and understand biological systems.

On the other hand we demonstrate how our framework can be used to obtain the optimal learning algorithm for an autonomous artificial system in different conditions. The optimal algorithm is then the strategy which maximizes ΔP in that system. For instance if a noisy linear system processes a gaussian signal in order to transmit as much information as possible about it then principal component analysis emerges as one of the solutions which maximize ΔP. On the other hand, if the task is to classify the signal in different classes then Fisher discriminant analysis [Duda & Hart, 1973] arises as the optimal solution when there is high overlapping between classes and the statistics are well represented by gaussians. Classical learning algorithms for tree construction are also obtained by maximization of ΔP in classification problems. For example the basic algorithm of C4.5 [Mitchell, 1997] is obtained as a special case when the complexity of the internal representation is not taken into account.

Finally we demonstrate the utility of the framework for developing new optimal learning schemes. For example we show that the principle of ΔP applied to decision trees construction induces a learning algorithm which combines the good features of the known methods of *information gain* and *gain ratio* [Mitchell, 1997]. It also shows a natural ability of *early stopping* and as a particular case contains the

[LópezdeMántaras, 1991] distance for attribute selection. On the other hand when the principle of ΔP is used in a nonlinear layer of neurons for a classification task, a nonlinear discriminant analysis algorithm emerges with very interesting features. The amount of resources used by the system is automatically adjusted to the required precision and the complexity of the problem, providing a useful strategy for avoiding overfitting. Thus the algorithm selects the more efficient representation for a given accuracy. The algorithm also shows a natural tendency to maximize the margins of the decision frontiers which provides a natural link with support vector machines [Vapnik, 1998].

Thus the proposed theory constitutes a unified general framework which allows us to describe, analyze and compare different complex adaptive systems independently on their physical implementation and whether they are biological or artificial. On the other hand, the theory allows us to develop new optimal learning schemes for artificial systems for different problems and implementations.

7.2 Analysis of adaptive complex systems with the new framework

In this section we will analyze the initial biological models in the light of the proposed framework. On the other hand, we will summarize how the proposed framework allows the construction of a map of several of the existing algorithms in neural computation and machine learning.

7.2.1 The olfactory, visual and auditory systems in the context of the new framework

As we have shown in different examples along the thesis the maximization of ΔP is a well-defined problem without the need of additional constraints. In general it leads to not a single solution but a family of optimal configurations. However biological systems have physical constraints which limit their possible configurations. In section 6.4 we demonstrated that the combination of the principle of maximization of ΔP

together with the specific biological limitations of that system conducts to properties very similar to the biological systems.

The olfactory epithelium

In section 6.4.1 we have derived the configuration which maximizes ΔP for the model of the olfactory epithelium. We considered a restriction in the gene pool size, limited individual sensitivities in the neurons and high desired accuracy. The optimal configuration with very similar properties to the real system: bipolarity in the sensitivities, mixing of narrow and broadly tuned receptive fields, and maximum number of different receptive fields.

The primary auditory cortex

In section 6.4.2 we used our framework to analyze the auditory cortex. The maximization of ΔP leads to a family of solutions and assumed that the concrete election done by the system was based in biological constraints. If we assume this constraint to be energy minimization and limited gaining in the neurons, which are reasonable constraints in this biological system, the concrete election of the optimal configuration has very similar features to the real system. On one hand, the receptive fields of the neurons are very specific, each one focusing in a single frequency band. The number of neurons responding to the band depends on the behaviorally importance of that band and not on its probability, which corresponds to the biological experiments [Kilgard & Merzenich, 1998] and what we observed in the realistic computer model (section 3.4.1). Finally, our model predicts that in some situations there can be neurons which do not respond to anything, being thus recruited when new stimuli appear. This is interesting since in the realistic model this also occurs (section 3.4.1).

The primary visual cortex

Olshausen and Field showed that the basic properties of receptive fields of cells in primary visual cortex (spatial location, orientation selectivity and structure to different scales) can be derived by a principle that maximizes the information transfer

while maximizing the spareseness of the code [Olshausen & Field, 1996]. The maximization of our measure ΔP maximizes the information transfer about the task while minimizing the global complexity of the code. It is important to mention that this is not equivalent to maximizing the spareness of the code since this is a term which applies to individual neurons and therefore to implementation specificities, whereas our notion of complexity depends on the global states of the network no matter what the specific implementation is.

As for the auditory cortex we expect that the optimization of ΔP in a visual cortex model using natural images will conduct to a family of optimal solutions. Again, the choice of the concrete solution will depend on the physical constraints in the system. We would expect in analogy with the auditory cortex that the configuration in the family which minimizes the consumed energy will have properties analog to the real system as in the model of Olshausen and Field occurs [Olshausen & Field, 1996].

The retina

Redlich and Atick demonstrated that the basic properties of ganglion cells at the retina can be derived by minimization of a measure of redundancy they proposed [Atick & Redlich, 1990]. Importantly, the minimization was performed imposing a given level of information transmission in the system and the translation invariance of image statistics. In preliminary work we obtained similar results when a maximization of ΔP is performed under the assumption of translation invariance in image statistics.

7.2.2 Our theory as a framework that explains known machine learning algorithms

We have shown that the principle of optimization of the effective information processing ΔP in artificial systems conducts to well known existing algorithms. We do not need to impose additional constraints in contrast with other methods. The optimization of ΔP is thus a well defined mathematical problem. Thus, our theory allows a unified framework that allows the interpretation and contextualization of learning algorithms which were developed in different areas. For example, we show that PCA

is the solution which maximizes ΔP in a linear system which processes a gaussian signal in order to transmit it optimally. As in PCA, a natural order of the eigenvectors appears and our theory shows how many of them we should choose given the desired level of accuracy in the transmission. But PCA is not the only optimal solution since any orthogonal transformation of this solution is also an optimal solution.

On the other hand, if the role of the linear system is to obtain an optimal internal representation for classification, then the Fisher discriminator emerges when the data can be described using second-order statistics and the clusters of data belonging to different classes are highly overlapped.

Finally when our framework is applied to decision tree construction, the basic algorithm of C4.5 [Quinlan, 1993] and the entropic distance for attribute selection introduced by [LópezdeMántaras, 1991] emerge as special cases of the general solution that maximizes ΔP.

7.3 Design of artificial adaptive systems with the new framework

The theory we propose also allows the development of new learning algorithms. We have showed that when applied to decision tree construction a new learning algorithm emerges which combines features of different well-known algorithms such as *information gain* and *gain ratio* [Mitchell, 1997]. It also shows a natural ability of *early stopping* and as a particular case contains the [LópezdeMántaras, 1991] distance for attribute selection. We have also presented a learning algorithm of optimal nonlinear internal representations for a classification problem (section 6.3.3). As we showed, the representations learned by the system have interesting properties such as minimization of complexity while maximization of the quality of the representation. These are important properties for a learning system since the power of generalization is strongly related with the structural complexity of the learning system [Vapnik, 1998].

We expect that when we apply our theory to other problems such as clustering or source separation similar algorithms to known optimal algorithms will emerge. For

example, we showed that PCA emerges in a linear network which tries to maximize the information transfer. Interestingly, PCA leads to a representation of the information which has a factorial pdf. The search of factorial codes is precisely the key feature of independent component analysis. In future work we will study ICA in the context of the framework.

7.4 Comparison with the Information Bottleneck Method

Next we will describe the Information Bottleneck Method [Tishby *et al.*, 1999] since it shares several similarities with the general theoretical framework exposed in this thesis.

The information Bottleneck Method [Tishby *et al.*, 1999] (IB from now on) has several commonalities with the framework presented in this paper in that it also allows for the construction of learning systems by searching for an optimal internal representation. This framework is derived from an interpretation of rate distortion theory [Cover & Thomas, 1991]. Following the notation we use in this thesis, the Information Bottleneck Method attempts to minimize the functional

$$L = I(Y;X) - \alpha I(Y;G) \tag{7.1}$$

with α a constant, the previous equation can be rewritten as that is, to minimize

$$L = \alpha H(Y|G) - (\alpha - 1)H(Y) - H(Y|X) \tag{7.2}$$

On the other hand, with the new information processing measure proposed in this paper the following expression must be minimized,

$$d(X,Y) = H(Y|G) + \beta H(G|Y) = (1+\beta)H(Y|G) - \beta H(Y) + \beta H(G) \tag{7.3}$$

where the last term does not take part in the optimization since it is constant for

a given problem. The first difference is that $H(Y|X)$ plays a role in L being zero only when $\alpha \to \infty$. That is, in IB the introduction of noise is not penalized but quite on the contrary whereas in the framework here proposed the introduction of noise is always penalized due to the introduction of spurious states. Since $\beta > 0$ (otherwise it would not be a distance) this means that $\beta = \alpha - 1$ when $H(Y|X) = 0$. However β must be > 0 (distance), so they are equivalent only when $\alpha > 1$.

When we analyze a continuous system and discretize it we have

$$H_d(a) = H(a) - H(q_a) \tag{7.4}$$

where $H_d(a)$ is the entropy of the discretized variable, $H(a)$ is its differential entropy and $H(q_a)$ is the distortion induced by the quantization. Then:

$$d(X_d, Y_d) = H(Y_d|G_d) + \beta H(G_d|Y_d) =$$

$$= (1+\beta)H(Y|G) - (\beta - 1)H(q_y) - \beta H(Y) + \beta(H(G) - H(q_g)) \tag{7.5}$$

so only when the quantization in Y is equal to its total noise variance $H(Y|X)$ and $\alpha \geq 1$ then the two methods are equivalent.

7.5 Future work

Future work includes the application of our theory to specific problems where temporal dynamics are relevant, for instance problems related with sequential pattern recognition. When using the formalism of Hidden Markov Models [Rabiner & Juang, 1986] we expect to obtain optimal algorithms with links to existing ones such as the Viterbi algorithm [Viterbi, 1967].

The examples showed in this thesis were focused on optimal internal representations in an agent whose optimal actions are communicated by the environment for a set of examples. However, for many real problems it is the agent itself which has to interact with the environment in order to learn what the optimal actions are (cf. *active learning*, [Shen, 1994]). Thus, the agent should learn optimal representations as well as the optimal actions (note that optimal actions depend on the concrete specificities

of the problem and not on the global statistics of the agent-environment system). Both concepts are very strongly related and should be learned simultaneously. Since there is a whole theory about learning optimal actions though the interaction with the environment (*reinforcement learning*) we will address the integration of our theory with these concepts.

On the biological side we plan to study the application of our theory to the understanding and the design of concrete biological experiments. For example, the derived methodology can be used to detect which ensembles of neurons contribute to solve a given task. This can be useful in psychophysical experiments where the animal is trained to solve a given task (for example, to discriminate between different stimuli) and the activity of different neural populations is recorded simultaneously while the animal is performing this task [Georgopoulos *et al.*, 1982, Britten *et al.*, 1992].

Part V

Appendices

Appendix A

Introducción

A.0.1 Los sistemas biológicos vistos como sistemas adaptativos eficientes

La tesis comienza con un estudio teórico de los sistemas biológicos ya que son muy eficientes en la interacción con su entorno. Para alcanzar dicha eficiencia, es crucial que dichos sistemas formen una adecuada representación interna de la información sensorial compleja que reciben [Barlow, 1961, Atick, 1992]. La construcción de dicha representación eficiente de la información tiene varias ventajas. Primero, *un menor gasto computacional y energético:* Los estímulos naturales llegan de una manera muy ineficiente ya que tienden a tener regularidades estadísticas. Por ejemplo, los píxeles en imágenes naturales tienen bastante correlación espacial, temporal y en color [Ruderman, 1994]. De esta forma, la representación de la imagen en la actividad global de los fotorreceptores es muy ineficiente. Como ejemplo claro de dicha ineficiencia basta considerar que el grado de compresión alcanzado normalmente por el sistema MPEG en archivos audio-visuales es de 30:1 [Furht, 1998]. De esta forma, una recodificación de estas señales utilizando un código menos redundante hace que el procesamiento posterior de dicha información sea más simple y menos costoso energéticamente [Attneave, 1954, Barlow, 1961, Atick, 1992, Baddeley, 1996]. Segundo, la representación interna puede tener consecuencias enormes en la *capacidad del animal de aprender las relaciones entre los objetos* del entorno [Barlow, 1989]. Por

ejemplo, las propiedades de una imagen son el resultado de los objetos presentes en la escena. La capacidad del animal de aprender relaciones funcionales entre los objetos depende crucialmente de su capacidad para representar los objetos como entidades independientes. Tercero, la representación interna influye directamente en la capacidad de generalización del sistema. Además, dicha representación interna debe depender de la tarea que se está llevando a cabo. Por ejemplo algunas variaciones como la distancia al objeto (y de esta forma el tamaño de su proyección en la retina) no son importantes para un tipo de tareas (por ejemplo, reconocer la identidad del objeto), con lo que en realidad representan ruido. En cambio, dichas variaciones pueden ser decisivas para otro tipo de tareas (la distancia a un objeto es crítica para poder cogerlo). El sistema debe ser entonces capaz de representar la información de manera apropiada según sea la tarea que se esté realizando. De esta forma el estudio de los principios de organización en dichos sistemas nos proporcionará información valiosa acerca de cuáles deben ser los principios de organización en un sistema adaptativo eficiente.

A lo largo de las últimas décadas la cantidad de datos experimentales obtenidos acerca del sistema nervioso ha ido creciendo enormemente. Esto ha permitido el desarrollo de una nueva rama de la neurociencia, la neurociencia teórica, que estudia el funcionamiento de los sistemas neuronales mediante modelos teóricos y simulaciones computacionales. Aunque hay muchos aspectos de los sistemas nerviosos que aún no han sido abordados experimentalmente, los modelos teóricos pueden ser útiles y dan lugar a predicciones concretas si el nivel de descripción del modelo se corresponde con el problema que se trata de explicar. Para una revisión de los diferentes tipos de modelos de sistemas neuronales ver [Koch & Segev, 1998] y [Arbib, 1998].

La tesis comienza analizando la representación de información en el epitelio olfativo. Esta estructura está compuesta por millones de neuronas receptoras olfativas, que representan la primera etapa de procesamiento en el sistema olfativo [Kandel *et al.*, 1991]. De esta forma la representación de la información en esta estructura es crítica para un funcionamiento óptimo del sistema. Esto es especialmente importante en el sistema olfativo, ya que los olores están compuestos por miles de sustancias químicas sencillas que pueden aparecer en multitud de combinaciones

diferentes [Pearce *et al.*, 2002]. En cambio, casi todos las neuronas receptoras olfativas tienen un campo receptivo inespecífico, respondiendo a una enorme variedad de compuestos químicos [Sicard & Holley, 1984]. La cuestión de si la causa de esta inespecificidad es alguna limitación física en los procesos químicos de transducción, o si por el contrario esto es benificioso para el sistema, no está clara.

Mediante el estudio de un modelo sencillo del epitelio olfativo mostraremos que la configuración neuronal que maximiza la transferencia de información tiene propiedades muy similares a la que se observa en el sistema real. De esta forma podemos decir que la inespecificidad de las neuronas receptoras olfativas es beneficiosa para el funcionamiento del sistema, optimizando la transferencia de información a etapas posteriores. Sin embargo esta representación está codificada genéticamente y no resulta de un fenómeno de aprendizaje o autoorganización. De esta forma procederemos a estudiar la representación de la información en las cortezas visual y auditiva, donde sí ocurren fenómenos de aprendizaje y plasticidad y las representaciones internas son más elaboradas.

En los últimos años los neurocientíficos han profundizado en los mecanismos responsables del aprendizaje y adaptación en los sistemas biológicos (para una revisión ver por ejemplo [Alkon *et al.*, 1991, Buonomano & Merzenich, 1998]. Basándose en ellos, diferentes autores han ido proponiendo diferentes modelos de aprendizaje en estos sistemas [Sejnowski, 1977, Stent, 1973, Bienenstock *et al.*, 1982, Brown & Chattarji, 1998, Fregnac, 1998]. Sin embargo, observaciones fisiológicas recientes en neuronas de la corteza revelan la existencia de propiedades no descritas previamente. Por ejemplo, las relaciones temporales a escala de milisegundos entre las señales que recibe una neurona son cruciales para el tipo de plasticidad que se activa [Markram *et al.*, 1997, Zhang *et al.*, 1998, Bi & Poo, 1998]. Dado que estos mecanismos pueden ser críticos para entender cómo se crean las representaciones internas en la corteza, introduciremos dichos mecanismos en modelos realistas de corteza.

Como se mostrará, las representaciones internas formadas por los modelos computacionales son muy similares a las de los sistemas biológicos, y emergen como consecuencia de la competición y cooperación neuronal en diferentes niveles. Los campos

receptivos formados por estos mecanismos dan lugar a neuronas más específicas que las neuronas receptoras olfativas. Neuronas cercanas en el espacio tienden a codificar las mismas características, mientras que las neuronas lejanas codifican aspectos diferentes de la información. De esta forma la información está representada por *grupos funcionales de neuronas.* Debido a esta especificidad las representaciones internas de las cortezas visual y auditiva son *códigos esparcidos*, esto es, el estímulo se representa por sólo unas células activas de todas las que potencialmente se podrían activar en otro tipo de codificación [Olshausen & Field, 1996, Baddeley, 1996], en contraste con las neuronas receptivas del sistema olfativo. Esta estrategia es eficiente en el sentido de que minimiza la complejidad y el costo energético del código, a la vez que maximiza la precisión en la representación interna de las imágenes reales [Olshausen & Field, 1996, Baddeley, 1996].

A.0.2 Derivación de principios generales del análisis de los sistemas biológicos

Las consideraciones anteriores sugieren que si existe un marco teórico general entonces debería ser válido para describir cualquier sistema complejo cuya codificación interna de la información y detalles de implementación sean arbitrarios. Esto sugiere que el marco teórico debería ser expresado en términos de los estados internos del sistema y no de parámetros físicos relacionados con la implementación específica del sistema. De esta forma nuestro marco debería ser *independiente de la implementación.* Por otra parte la noción de *baja complejidad en la representación* es un ingrediente esencial en las representaciones de las cortezas visual y auditivas. De esta forma los sistemas biológicos reducen la complejidad de la entrada sensorial construyendo representaciones más elaboradas de la información (por ejemplo "detectores de bordes" en la corteza visual versus "detectores de píxeles" en el conjunto de fotorreceptores de la retina [Bear *et al.*, 1996]). Esto es intuitivo ya que el objetivo final del animal es solucionar las tareas que el entorno le impone, siendo para ello crítica una representación eficiente de la información que capture las regularidades del entorno [Barlow, 1989]. De esta forma el concepto de *reducción de la complejidad* es otro ingrediente básico

en la teoría que deseamos. Finalmente, las observaciones experimentales realizadas por otros autores en la corteza autitiva y nuestro análisis teórico muestran que la representación de la información en esa estructura es modulada a través de la experiencia, según va cambiando en el tiempo la importancia de los estímulos. De esta forma la representación interna está sesgada hacia los estímulos importantes. Estos experimentos revelan el importante concepto de que las representaciones internas, aun en etapas primarias de procesamiento, están influenciadas por las tareas que el entorno le impone al animal. Esto parece muy razonable ya que los recursos de los que dispone un sistema biológico son limitados, y debido al alto grado de complejidad del entorno y a la gran cantidad de estímulos diferentes que recibe, el animal se debería enfocar en la parte de la información que es realmente relevante para su tarea.

Aparte de estas consideraciones acerca de la limitación de recursos, el enfoque de las representaciones en los aspectos importantes de la tarea es crítico para una generalización apropiada. Estas consideraciones nos conducen a la noción de que el concepto de tarea debería ser un elemento crucial en una teoría del procesamiento de información. Esto es, la nueva medida de procesamiento de información debería ser *dependiente de la tarea.*

A.0.3 El marco formal general

El objetivo de esta tesis es proporcionar un nuevo marco teórico que permita tanto el análisis de de sistemas complejos adaptativos existentes, como el diseño de nuevos sistemas artificiales complejos. En este contexto la teoría de la información estudia dichos sistemas en función de sus propiedades estadísticas globales pero no de sus detalles de implementación [Cover & Thomas, 1991]. De esta forma la maximización de la información mutua parece muy apropiada para describir las propiedades globales de los sistemas sensoriales, siendo muy exitosa en la descripción de algunos sistemas neuronales [Atick, 1992, Borst & Theunissen, 1999, Dayan & Abbott, 2001]. Sin embargo mostraremos que el principio de maximización de la información mutua no es válido para describir las propiedades de las representaciones internas en etapas de procesamiento más avanzadas como la corteza auditiva.

De estas ideas y de las conclusiones mostradas en la sección A.0.2 derivamos un marco matemático general que contiene la esencia de estas ideas. Para hacerlo lo más general posible, se usan las nociones de *agente* y *entorno* válidas para cualquier sistema que interacciona con el entorno. Se introduce el concepto de *cantidad efectiva de procesamiento de información* (ΔP) realizado por una parte $\mathcal{W}$ del agente que debe solucionar la tarea impuesta por el entorno. Esta noción no depende de los detalles de la implementación sino en las relaciones globales estadísticas entre esa parte del agente y su entorno. Entonces veremos que emerge la noción crucial de *distancia a la tarea*, que depende de los niveles de incertidumbre y complejidad que la información procesada por $\mathcal{W}$ tiene respecto a la tarea. Dado que la teoría está expresada usando las nociones generales de agente y entorno y las herramientas matemáticas usadas son independientes de la implementación, el marco teórico puede ser usado para analizar la interacción de cualquier tipo de sistema adaptativo, tanto biológico como artificial, con su entorno. La teoría propuesta servirá también para obtener nuevas estrategias de aprendizaje para sistemas artificiales.

A.0.4 Validación de nuestra teoría en los modelos biológicos

Dado que los sistemas biológicos son muy eficientes interaccionando con su entorno, esperaríamos que construyeran representaciones internas que maximicen ΔP. Como mostraremos la configuración teórica que maximiza ΔP en el modelo de corteza auditiva tiene propiedades muy parecidas a las del sistema biológico. Por una parte predice que las neuronas responden a bandas específicas de frecuencia. Por otra parte predice que los estímulos se representan internamente por una cantidad de recursos proporcional a su importancia pero no a su probabilidad de aparición. Estas propiedades derivadas de la maximización de ΔP son muy similares a las observaciones experimentales [Weinberger, 1993, Kilgard & Merzenich, 1998]-

La configuración teórica que maximiza ΔP en el epitelio olfativo está compuesta por un repertorio de neuronas que muestran máxima diversidad en su patrón de respuestas, sensibilidades bipolares y campos receptivos inespecíficos. Todas estas propiedades han sido descritas en trabajos experimentales [Sicard & Holley, 1984,

Schild & Restrepo, 1998, Sanhueza *et al.*, 2000].

Finalmente, en trabajo preliminar hemos obtenido que la maximización de ΔP en un modelo simplificado de la retina [Atick & Redlich, 1990] obtenemos propiedades similares a las de las células ganglionares de la retina. De esta forma concluimos que ΔP parece ser maximizado en los sistemas biológicos, lo cual constituye una validación de nuestra teoría y demuestra el potencial de nuestro marco teórico para estudiar y comprender los sistemas biológicos.

A.0.5 Algoritmos conocidos de aprendizaje en sistemas artificiales vistos como soluciones particulares en nuestra teoría

Hay una enorme variedad de técnicas de aprendizaje para sistemas artificiales, útiles para diferentes problemas (para una revisión ver por ejemplo [Mitchell, 1997]). Por ejemplo, una clasificación usual de estos algoritmos los divide en algoritmos supervisados, algoritmos no supervisados, y algoritmos de aprendizaje por refuerzo [Mitchell, 1997, Sutton & Barto, 1998]. Esta tesis muestra cómo emergen del mismo marco teórico tanto algoritmos de aprendizaje supervisados como no supervisados. Adicionalmente se puede obtener una representación óptima de la información para técnicas de aprendizaje por refuerzo.

Demostraremos cómo nuestro marco teórico puede ser usado para obtener el algortitmo de aprendizaje óptimo para un sistema autónomo artificial en condiciones diferentes. Veremos que si un sistema lineal ruidoso procesa una señal gaussiana con el objetivo de transmitir tanta información cono sea posible de ella, entonces la técnica de análisis de componentes principales (PCA) emerge como una de las soluciones que maximiza ΔP. Por otra parte, si la tarea es clasificar la señal en diferentes clases entonces la técnica del análisis discriminante de Fisher [Duda & Hart, 1973] es la solución que maximiza ΔP cuando hay un alto grado de solapamiento entre las clases y las estadísticas están bien representadas por gaussianas. Además se muestra que algunas técnicas clásicas de construcción de árboles de decisión se obtienen por maximización de ΔP en problemas de clasificación. Por ejemplo el algoritmo básico

de C4.5 [Mitchell, 1997] es obtenido como un caso especial cuando la complejidad de la representación interna no es tenida en cuenta.

De esta forma nuestra teoría es una marco unificador que permite crear un mapa con diferentes técnicas de aprendizaje automático, pudiendo ayudarnos en comprender sus analogías y diferencias fundamentales.

A.0.6 Utilidad del marco teórico para obtener nuevos esquemas de aprendizaje para sistemas artificiales

Finalmente demostraremos la utilidad del marco teórico para el desarrollo de nuevos esquemas óptimos de aprendizaje en sistemas artificiales. Por ejemplo mostraremos cómo el principio de maximización de ΔP aplicado a árboles de decisión induce un algoritmo de aprendizaje que combina las propiedades de métodos conocidos como el *information gain* y el *gain ratio* [Mitchell, 1997]. Dicho algoritmo muestra una capacidad natural de detención de la expansión y como caso particular contiene la distancia para selección de atributos propuesta por [LópezdeMántaras, 1991]. Por otra parte cuando el principio de maximización de ΔP es usado en una capa de neuronas no lineales para una tarea de clasificación, emerge un algoritmo de extracción de características no lineales con propiedades muy interesantes. Por una parte, la cantidad de recursos usada por el sistema se ajusta automáticamente a la precisión requerida y a la complejidad del problema, proporcionando una estrategia útil para evitar el sobreajuste. De esta forma el algoritmo selecciona la representación más eficiente para una precisión dada. Por otra parte el algoritmo también muestra una tendencia natural a maximizar los márgenes de las fronteras de decisión, lo que proporciona un vínculo natural con la técnica de *support vector machines* [Vapnik, 1998].

A.0.7 Esquema general de la tesis y metodología

La tesis comienza con el estudio de la representación interna de infromación en sistemas sensoriales biológicos. Por una parte, usamos modelos que tratan de capturar las propiedades globales del sistema. Estos modelos se basan en medidas teóricas de información que describen la eficacia del sistema basándose en sus propiedades

estadísticas globales y no en los detalles de implementación. De esta forma estos modelos son matemáticamente sencillos pero nos permiten realizar predicciones concretas acerca del sistema. Por otra parte, usaremos modelos más realistas que incluyen detalles específicos acerca de la dinámica de las neuronas. Estos modelos nos posibilitarán entender cuestiones más específicas como por ejemplo cómo se construyen dichas representaciones internas.

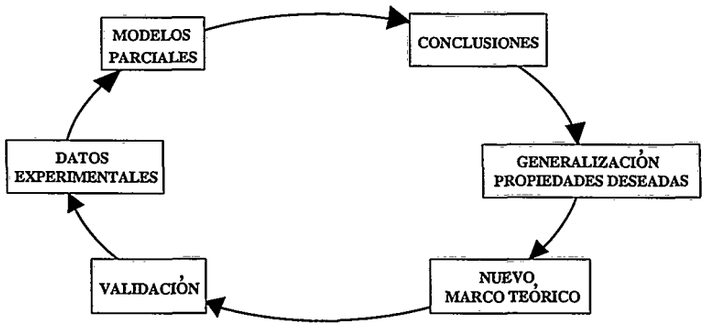

Figure A.1: Esquema general de la tesis

Del análisis de ambos tipos de modelos obtendremos una lista de propiedades deseables que nuestro deseado marco teórico general debería tener. Entonces derivaremos un formalismo matemático que satisface dichas propiedades y es expresado usando herramientas matemáticas independientes de la implementación. Esto permite la generalización de la teoría y su aplicación en otros contextos como aprendizaje automático en sistemas artificiales.

Finalmente, el marco teórico es validado en dos direcciones diferentes. Primero, mostraremos que la aplicación de la teoría en problemas artificiales nos lleva a la obtención de algoritmos muy conocidos de aprendizaje automático, demostrando la validez de la teoría en estos problemas. Adicionalmente el marco teórico nos permite obtener nuevos algoritmos de aprendizaje óptimos de los que damos dos ejemplos en esta tesis. Segundo, mostraremos que cuando aplicamos la teoría a los sistemas biológicos, predice configuraciones con propiedades muy parecidas a las del sistema biológico. Esto nos proporciona otra validación de nuestra teoría. Finalmente discutiremos cómo este marco teórico puede ser usado en el diseño y análisis de nuevos experimentos biológicos.

Appendix B

Conclusiones

B.1 Conclusiones obtenidas del estudio de los sistemas biológicos

La tesis empieza abordando el estudio del procesamiento de información en sistemas biológicos. Ya que dichos sistemas son muy eficientes en su interacción con el entorno, el estudio de principios de organización en dichos sistemas nos puede proporcionar pistas sobre las propiedades generales que debe tener un sistema adaptativo para ser eficiente. En concreto, queremos desarrollar un marco teórico que exprese dichas propiedades.

De los estudios teóricos realizados en el epitelio olfativo, la corteza visual y la corteza auditiva concluimos que:

- El marco teórico no debe depender de los detalles específicos del sistema. Por ello, debe depender de las propiedades estadísticas de la interacción del sistema con su entorno.

- El concepto de tarea es crucial: incluso en etapas primarias de procesamiento, la representación de información en los sistemas neuronales está sesgada hacia los estímulos que llevan información acerca de la tarea.

- El concepto de reducción de la complejidad es clave: el sistema debe procesar la

información construyendo representaciones internas que eliminen la parte que no lleva información relacionada con la tarea a realizar. De esta forma las representaciones internas son más elaboradas y más cercanas a la tarea según se va procesando la información en el animal.

B.2 Desarrollo de un nuevo marco teórico para el estudio del procesamiento de información en sistemas adaptativos

Basándonos en las conclusiones anteriores desarrollamos un marco teórico general que usa las nociones de *agente* y *entorno*. Se obtiene una medida efectiva de procesamiento de información (ΔP) que realiza una parte del agente dada la tarea global del sistema. Dicha medida depende de la complejidad de la representación de la información en esa parte del agente y en la tarea global que debe realizar el agente. La medida no depende de detalles de implementación del sistema sino de propiedades estadísticas. De esta forma, el marco teórico obtenido es aplicable tanto para analizar sistemas adaptativos complejos (biológicos y artificiales) como para diseñar nuevos algoritmos de aprendizaje automático.

B.3 Análisis de sistemas adaptativos complejos con el nuevo marco teórico

B.3.1 Análisis de sistemas biológicos

El epitelio olfativo

Se maximiza ΔP para un modelo del epitelio olfativo que incluye restricciones realistas: tamaño máximo en el número de genes que puede expresar y limitación en la respuesta máxima de una neurona a un olor. Además se supone que el sistema está interesado en discriminar los estímulos con gran precisión. Entonces la solución

que maximiza ΔP tiene propiedades muy similares al sistema real: bipolaridad en la respuesta a los estímulos, mezcla de campos receptivos específicos con no específicos, y máxima diversidad en el número de campos receptivos diferentes.

La corteza auditiva primaria

La maximización de ΔP en un modelo de la corteza auditiva primaria da lugar a propiedades muy parecidas a las del sistema real. Los campos receptivos de las neuronas se enfocan en una banda específica de frecuencias y el número de neuronas asociadas a cada banda depende del grado de asociación de dicha banda con estímulos aversivos ("dolor").

La corteza visual primaria y la retina

Se ha realizado trabajo preliminar cuyos resultados parecen indicar que la maximización de ΔP en estos sistemas también conduce a las propiedades observadas experimentalmente.

Conclusión general

Los sistemas biológicos parecen maximizar la medida efectiva de procesamiento de información propuesta, ΔP. Esto constituye una validación importante para nuestra teoría.

B.3.2 Nuestra teoría como marco para el estudio de algoritmos de aprendizaje automático ya existentes

En la tesis mostramos que el principio de maximización de ΔP en sistemas artificiales conduce a algoritmos de aprendizaje automático ya conocidos. Por ejemplo, el análisis de componentes principales (PCA) aparece como una de las soluciones que maximiza ΔP en un sistema lineal con entradas gaussianas y cuyo objetivo es maximizar la tasa de transferencia de información de esas entradas.

Si por otra parte el objetivo del sistema lineal es construir una representación óptima para la clasificación, el análisis del discriminante de Fisher es la solución que maximiza ΔP cuando las estadísticas de las entradas son gaussianas y hay mucho solapamiento entre clases. Finalmente, cuando el marco teórico se aplica a la construcción de árboles de decisión, aparecen técnicas ya conocidas como C4.5 [Quinlan, 1993] y la distancia entrópica entre atributos [LópezdeMántaras, 1991] como las soluciones óptimas en situaciones especiales.

B.4 Diseño de sistemas artificiales adaptativos con el nuevo marco teórico

Finalmente mostramos cómo el marco teórico presentado permite el desarrollo de nuevos algoritmos de aprendizaje para sistemas artificiales. Dichos algoritmos presentan propiedades muy interesantes y tienen aplicaciones prácticas concretas tales como la extracción de información y la predicción en bases de datos reales.

Por ejemplo, presentamos un nuevo método de inducción de árboles de decisión que reúne las características de otros métodos ya conocidos tales como el *information gain* y el *gain ratio* [Mitchell, 1997]. El nuevo método de inducción obtenido muestra también una capacidad natural de parar la expansión automáticamente (*early stopping*) y como caso particular contiene a la distancia entrópica propuesta por [LópezdeMántaras, 1991].

Por otra parte, presentamos un algoritmo nuevo de construcción de características no lineales para problemas de clasificación. El algoritmo muestra unas propiedades muy interesantes tales como la minimización de la complejidad para una precisión dada y la maximización de los márgenes en las fronteras de decisión.

Appendix C

Technical appendices

C.1 Learning with noise in the auditory model: mathematical analysis

The noise continuously excites the input neurons corresponding to the lowest frequencies (fig. 3.14 A). This in turn drives the thalamic neurons tuned to low frequencies leading to a response of all cortical excitatory neurons at the first presentation due to their initially diffuse receptive fields. After a few seconds, however, the efficacy of the synapses from the input population to the thalamic population, which transduce the presented frequencies, diminishes due to their short-term depression. This prevents continuously present harmonics from further activating the thalamic and cortical populations. This can be analyzed by calculating the expected value of the changes of the depressing synapses (eq. 3.3):

$$\langle \dot{\Gamma}_i(t) \rangle = \tau_d^{-1}(1 - \langle \Gamma_i(t) \rangle) - f \langle \Gamma_i(t) \delta(t - t_{spike_i}) \rangle \tag{C.1}$$

The temporal dynamics of Γ (time constant of 4 Sec) is much slower than the temporal dynamics of the input neuron (time constant of 19 ms) so we can write:

$$\langle \Gamma_i(t) \delta(t - t_{spike_i}) \rangle = \langle \Gamma_i(t) \rangle \langle \delta(t - t_{spike_i}) \rangle = \langle \Gamma_i(t) \rangle F_N \tag{C.2}$$

where F_N is the firing rate of the input neuron which responds to a continuous stimulus. Using this in (eq. C.1) we have that $\langle \Gamma_i(t) \rangle$ converges to:

$$\langle \Gamma_i \rangle = \frac{1}{1 + \tau_d f F_N} \tag{C.3}$$

Thus $\langle \Gamma_i \rangle$ goes to zero as the firing rate of the input neuron increases. Hence, continuously presented audio signals are filtered out through rapid synaptic depression. The main harmonics of the continuously presented signal induce high activation in the corresponding input neuron leading to a continuous high excitation of the thalamic neurons they project to. The efficacies of the synapses connecting these input neurons with the thalamic neurons would have a very low $\langle \Gamma_i \rangle$ strongly attenuating the strength of the signal they transduce.

However, as shown in Fig. 3.14 B not all aspects of the continuously presented stimulus are filtered out. This is due to fluctuations in harmonics which have a small contribution to the signal and are not filtered out by the short-term depressing synapses. The weak contribution of these harmonics makes the corresponding input neuron fire at a low firing rate. From (eq. C.3) we see that its connection to the thalamic neuron is not strongly affected. However, in our system a thalamic neuron needs to receive 2-3 effective spikes in a short period of time in order to fire. This means that those input neurons firing at a low frequency are not able to trigger a spike in thalamic neurons, even when the connection has a high $\langle \Gamma_i \rangle$. However, a momentary increment in the harmonic contribution would increase the firing rate of the input neuron, thus having the possibility to fire 2-3 spikes in a short period of time with a high $\langle \Gamma_i \rangle$, making the corresponding thalamic neuron fire. Therefore, we see that short-term depressing synapses are not able to completely filter out the continuous noise. As a result fluctuations in the harmonics of the noise are processed by the cortical network, mixed with the information about the tones presented to the system. Furthermore, these fluctuations in the noise can activate those input neurons that are activated when the 0.74 kHz tone is presented (fig. 3.14). Therefore we see that the noise overlaps with the signal both temporally and spatially.

Hence, one would expect that the continuously presented noise would interfere

with the development of receptive fields specific to the tones. However, those thalamocortical synapses that transduce information about the noise tend to get weaker. The learning rule decorrelates signals that are independent, in this case the fluctuations in the spectrum of the noise and the tones played by the synthesizer. This effect can be understood by calculating the expected increment of synaptic strength per postsynaptic spike, $\Delta W_{i,j}$, from equations (3.4, 3.5, 3.6). First, we introduce the notation:

$$\Pi_{i,j} \equiv \langle \frac{t_0}{t_0 + \|t_i - t_j\|} | BPAP_j \wedge (\|t_i - t_j\| < W) \rangle \tag{C.4}$$

that is, the expected value of the quantity $\frac{t_0}{t_0+|t_i-t_j|}$ given that the backpropagating action potential in the postsynaptic neuron is not attenuated by the inhibition ("$BPAP_j$"), and that this event and the presynaptic spike fall within the temporal association window W (see Learning Dynamics); t_i and t_j are the times when the presynaptic and postsynaptic neurons spike, respectively. Analogously:

$$\Phi_{i,j} \equiv \langle \frac{t_0}{t_0 + \|t_i - t_j\|} | BPAPm_j \wedge (\|t_i - t_j\| < W) \rangle \tag{C.5}$$

that is, the expected value of the quantity $\frac{t_0}{t_0+|t_i-t_j|}$ given that the backpropagating action potential in the postsynaptic neuron is attenuated by the inhibition ("$BPAPm_j$"), and that this event and the presynaptic spike occur within the temporal association window W. Given these definitions the expected increment of synaptic strength per postsynaptic spike can be calculated from eqs. (3.4, 3.5, 3.6) resulting in:

$$\langle \Delta W_{i,j} \rangle = \alpha\, \Pi_{i,j}\, p(BPAP_j \wedge (\|t_i - t_j\| < W) | sp_j) - \beta\, \Phi_{i,j}\, p(BPAPm_j \wedge (\|t_i - t_j\| < W) | sp_j) -$$

$$- \eta\, p(BPAP_j \wedge (\|t_i - t_j\| \geq W) | sp_j) \tag{C.6}$$

where "sp_j" means "there is a spike in the jth postsynaptic neuron". In case the ith thalamic cell encodes just fluctuations in the noise while another signal is making the cortical excitatory population fire, the activity of this cell is uncorrelated with

activity of cortical excitatory cell j, that is:

$$p(BPAP_j \wedge (\|t_i - t_j\| < W)|sp_j) = p(BPAP_j|sp_j)\, p(\|t_i - t_j\| < W|sp_j) \quad \text{(C.7)}$$

$$p(BPAP_j \wedge (\|t_i - t_j\| \geq W)|sp_j) = p(BPAP_j|sp_j)\, p(\|t_i - t_j\| \geq W|sp_j) \quad \text{(C.8)}$$

Because both β and $\Phi_{i,j}$ are positive,

$$\langle \Delta W_{i,j} \rangle \leq \alpha\, \Pi_{i,j}\, p(BPAP_j \wedge (\|t_i - t_j\| < W)|sp_j) - \eta\, p(BPAP_j \wedge (\|t_i - t_j\| \geq W)|sp_j) \quad \text{(C.9)}$$

Noting that $\Pi_{i,j} \leq 1$, and using equations (C.4, C.5):

$$\langle \Delta W_{i,j} \rangle \leq \alpha\, p(BPAP_j|sp_j)\, p(\|t_i - t_j\| < W|sp_j) - \eta\, p(BPAP_j|sp_j)\, p(\|t_i - t_j\| \geq W|sp_j) \quad \text{(C.10)}$$

Therefore, $\Delta\langle W_{i,j} \rangle$ is guaranteed to be negative (forcing the final value of the synaptic strength towards 0) if

$$\eta > \alpha \frac{p(\|t_i - t_j\| < W|sp_j)}{1 - p(\|t_i - t_j\| < W|sp_j)} \quad \text{(C.11)}$$

If the noisy output of thalamic neuron i can be described as a Poisson process with rate F, then:

$$p(\|t_i - t_j\| < W|sp_j) = \int_0^{2W} f e^{-tF} = 1 - e^{-2FW} \quad \text{(C.12)}$$

Using this in (eq. C.11) we obtain

$$\eta > \alpha \left(e^{2FW} - 1\right) \quad \text{(C.13)}$$

This equation shows that the smaller the association window W is, the easier it is for the learning mechanism to prune the synapses that carry noisy information. In addition, the smaller the Poisson noise rate is, the easier it is for the learning mechanism to prune the synapse which transduces it.

Therefore, this learning mechanism decorrelates the noise from the receptive fields of the cortical excitatory cells sensitive to tones. Effectively, we see in (fig. 3.15 A) that the receptive fields of the neurons that fire to the tones are decorrelated from noise.

A few neurons develop receptive fields specific to frequencies that are part of the noise: two are finally selective to frequencies lower than 0.7 kHz and one is selective to 0.90 kHz (fig. 3.16). These neurons, however, do not respond to any of the tones (fig. 3.15 B). Finally, the remaining neurons do not respond to either the tones or the noise, remaining "unspecific" (fig. 3.15 C).

C.2 Useful properties of matrices and gaussian distributions

C.2.1 Compact notation for the gaussian distributions

We introduce the notation:

$$\mathcal{G}(A, \vec{b}) \equiv \frac{1}{(2\pi)^{\frac{N}{2}} (\det A)^{\frac{1}{2}}} e^{-\frac{1}{2}\vec{b}^T A^{-1} \vec{b}} \tag{C.14}$$

where N is the number of dimensions of the vector $\vec{b}$.

C.2.2 Useful properties of the compact notation

- $$\mathcal{G}(A, \vec{b}) = \mathcal{G}(A, -\vec{b}) \tag{C.15}$$

- For a scalar variable

$$\mathcal{G}(a^2, b) = \frac{1}{a} \mathcal{G}(1, \frac{b}{a}) \tag{C.16}$$

- In general

$$\mathcal{G}(A,\vec{b}) = \frac{1}{(\det A)^{\frac{1}{2}}}\,\mathcal{G}(I, A^{-\frac{1}{2}}\cdot\vec{b}) \tag{C.17}$$

where $A^{-\frac{1}{2}}$ is the inverse of the square root matrix of A.

C.2.3 Basic properties of gaussian distributions

- **Normalization:** $\int_{\mathcal{R}^N}\mathcal{G}(A,\vec{x}-\vec{m})\,d^N x = 1$ with N being the number of dimensions of $\vec{x}$.
- **Mean value:** $\int_{\mathcal{R}^N}\vec{x}\cdot\mathcal{G}(A,\vec{x}-\vec{m})\,d^N x = \vec{m}$.
- **Covariance matrix:** $\int_{\mathcal{R}^N}(\vec{x}-\vec{m})(\vec{x}-\vec{m})^T\mathcal{G}(A,\vec{x}-\vec{m})\,d^N x = A$

C.2.4 Useful properties of determinants

If A is a square matrix,

- $\det A = \det(A^T)$

If A and B are two square matrices with the same dimensions,

- Multiplication of determinants:

$$\det(A\cdot B) = (\det A)\cdot(\det B)$$

- The order of the factors in the determinant is not important:

$$\det(A\cdot B) = \det(B\cdot A)$$

Proof:

$$\det(A\cdot B) = (\det A)\cdot(\det B) = (\det B)\cdot(\det A) = \det(B\cdot A)$$

where we have used property C.2.4.

- Invariance under multiplication order:

$$\det(I + AB) = \det(I + BA)$$

Proof:

$$\det(I + AB) = \det\left((B^{-1} + A)B\right) = \det\left(B \cdot (B^{-1} + A)\right)$$

where we have used property C.2.4.

Then:

$$\det\left(B \cdot (B^{-1} + A)\right) = \det(I + BA)$$

C.2.5 Inversion of matrices

For any square and invertible matrices A and B (no matter if they are defined positive or not, symmetric or not):

$$(A+B)^{-1} = A^{-1}(A^{-1} + B^{-1})^{-1}B^{-1} = B^{-1}(A^{-1} + B^{-1})^{-1}A^{-1} \tag{C.18}$$

Proof:

$$(A+B)^{-1} = (A(I + A^{-1}B))^{-1} = (A(B^{-1} + A^{-1})B)^{-1} = B^{-1}(B^{-1} + A^{-1})^{-1}A^{-1}$$

On the other hand,

$$(A+B)^{-1} = (B(B^{-1}A + I))^{-1} = (B(B^{-1} + A^{-1})A)^{-1} = A^{-1}(B^{-1} + A^{-1})^{-1}B^{-1}$$

C.2.6 Multiplication of two gaussians depending on the same variable

$$\mathcal{G}(A,\vec{x}-\vec{b})\cdot\mathcal{G}(C,\vec{x}-\vec{d})=\mathcal{G}((A^{-1}+C^{-1})^{-1},\vec{x}-\vec{f})\cdot\mathcal{G}(A+C,\vec{b}-\vec{d}) \qquad \text{(C.19)}$$

where $\vec{f}=(A^{-1}+C^{-1})^{-1}(A^{-1}\vec{b}+C^{-1}\vec{d})$.

Proof:

$$\mathcal{G}(A,\vec{x}-\vec{b})\cdot\mathcal{G}(C,\vec{x}-\vec{d})=$$

$$\frac{1}{(2\pi)^{dim_x}\,(\det\,A)^{1/2}\,(\det\,C)^{1/2}}\exp\frac{-1}{2}\left[(\vec{x}-\vec{b})^TA^{-1}(\vec{x}-\vec{b})+(\vec{x}-\vec{d})^TC^{-1}(\vec{x}-\vec{d})\right]$$

Now we rewrite the exponent as:

$$(\vec{x}-\vec{b})^TA^{-1}(\vec{x}-\vec{b})+(\vec{x}-\vec{d})^TC^{-1}(\vec{x}-\vec{d})=(\vec{x}-\vec{f})^TE^{-1}(\vec{x}-\vec{f})+g$$

Developing this expression we get:

$$\vec{x}^T(A^{-1}+C^{-1})\vec{x}-2\vec{x}^T(A^{-1}\vec{b}+C^{-1}\vec{d})+\vec{b}^TA^{-1}\vec{b}+\vec{d}^TC^{-1}\vec{d}=$$

$$=\vec{x}^TE^{-1}\vec{x}-2\vec{x}^TE^{-1}\vec{f}+\vec{f}^TE^{-1}\vec{f}+g$$

from where we get:

- $E=(A^{-1}+C^{-1})^{-1}$
- $A^{-1}\vec{b}+C^{-1}\vec{d}=E^{-1}\vec{f}\rightarrow\vec{f}=(A^{-1}+C^{-1})^{-1}(A^{-1}\vec{b}+C^{-1}\vec{d})$ which can be rewritten equivalently as

  $\vec{f}=\vec{b}-(A^{-1}+C^{-1})^{-1}C^{-1}(\vec{b}-\vec{d})$ or $\vec{f}=\vec{d}+(A^{-1}+C^{-1})^{-1}A^{-1}(\vec{b}-\vec{d})$.
- $\vec{b}^TA^{-1}\vec{b}+\vec{d}^TC^{-1}\vec{d}=\vec{f}^TE^{-1}\vec{f}+g\rightarrow g=\vec{b}^TA^{-1}\vec{b}+\vec{d}^TC^{-1}\vec{d}-\vec{f}^TE^{-1}\vec{f}$

  Using the two equivalent expressions for $\vec{f}$ we get:

$$g = \vec{b}^T A^{-1}\vec{b} +$$

$$+\vec{d}^T C^{-1}\vec{d} + \left[(\vec{b}-\vec{d})^T C^{-1}(A^{-1}+C^{-1})^{-1} - \vec{b}^T\right](A^{-1}+C^{-1})\left[\vec{d} + (A^{-1}+C^{-1})^{-1}A^{-1}(\vec{b}-\vec{d})\right]$$

expanding:

$$g = \vec{b}^T A^{-1}\vec{b} + \vec{d}^T C^{-1}\vec{d} + (\vec{b}-\vec{d})^T C^{-1}\vec{d} - \vec{b}^T(A^{-1}+C^{-1})\vec{d} - \vec{b}^T A^{-1}(\vec{b}-\vec{d}) +$$

$$+(\vec{b}-\vec{d})^T C^{-1}(A^{-1}+C^{-1})^{-1}A^{-1}(\vec{b}-\vec{d})$$

which after simplifying gets:

$$g = (\vec{b}-\vec{d})^T C^{-1}(A^{-1}+C^{-1})^{-1}A^{-1}(\vec{b}-\vec{d})$$

For every invertible matrices A and B it holds $A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1} = B^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} = (A+B)^{-1}$ (property C.18), so we have the simple expression

$$g = (\vec{b}-\vec{d})^T(A+C)^{-1}(\vec{b}-\vec{d})$$

Finally, noting that $A + C = A(I + A^{-1}C) = A(C^{-1} + A^{-1})C$ we obtain

$$\det\left(A^{-1}+C^{-1}\right)^{-1}\cdot\det(A+C) = \det\left(A^{-1}+C^{-1}\right)^{-1}\cdot\det\left(A(C^{-1}+A^{-1})C\right) = \det(A)\cdot\det(C)$$

which allows us to write the final expression

$$\mathcal{G}(A,\vec{x}-\vec{b})\cdot\mathcal{G}(C,\vec{x}-\vec{d}) = \mathcal{G}((A^{-1}+C^{-1})^{-1},\vec{x}-\vec{f})\cdot\mathcal{G}(A+C,\vec{b}-\vec{d}).$$

C.2.7 "And" of two gaussian variables

Using the previous result together with the normalization property of gaussians (section C.2.3) it is straightforward to obtain

$$\int_{\mathcal{R}^N}\mathcal{G}(A,\vec{x}-\vec{b})\cdot\mathcal{G}(C,\vec{x}-\vec{d})\,d^N x = \mathcal{G}(A+C,\vec{b}-\vec{d}) \qquad \text{(C.20)}$$

C.2.8 Convolution of gaussian variables

$$\int \mathcal{G}(A, \vec{s} - B\,\vec{u}) \cdot \mathcal{G}(C, \vec{u} - \vec{t}\,)\, d^N u = \mathcal{G}(A + BCB^T, \vec{s} - B\,\vec{t}\,) \tag{C.21}$$

Proof:

$$\mathcal{G}(A, \vec{s} - B\,\vec{u}) \cdot \mathcal{G}(C, \vec{u} - \vec{t}) =$$

$$= \frac{1}{(2\pi)^{\frac{dim_s + dim_u}{2}} (\det A)^{1/2} (\det C)^{1/2}} \exp \frac{-1}{2} \left[(\vec{s} - B\vec{u})^T A^{-1} (\vec{s} - B\vec{u}) + (\vec{u} - \vec{t})^T C^{-1} (\vec{u} - \vec{t}) \right]$$

After simple algebraic manipulations and Woodsbury's formula $(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1} U)^{-1} V^T A^{-1}$ [Golub & van Loan, 1996] we can express the power of the exponential as:

$$-\frac{1}{2} \left[(\vec{s} - B\vec{u})^T A^{-1} (\vec{s} - B\vec{u}) + (\vec{u} - \vec{t})^T C^{-1} (\vec{u} - \vec{t}) \right] =$$

$$= -\frac{1}{2} (\vec{u} - \vec{m})^T D^{-1} (\vec{u} - \vec{m}) - \frac{1}{2} (\vec{s} - B\vec{t})^T E^{-1} (\vec{s} - B\vec{t})$$

with $D = (B^T A^{-1} B + C^{-1})^{-1}$, $E = A + BCB^T$ and $\vec{m} = D(B^T A^{-1} \vec{s} + C^{-1} \vec{t})$.

Therefore,

$$\mathcal{G}(A, \vec{s} - B\,\vec{u}) \cdot \mathcal{G}(C, \vec{u} - \vec{t}) = \frac{(\det D)^{1/2} (\det E)^{1/2}}{(\det A)^{1/2} (\det C)^{1/2}} \cdot \mathcal{G}(D, \vec{u} - \vec{m}) \cdot \mathcal{G}(E, \vec{s} - B\vec{t})$$

If now we integrate in $\vec{u}$:

$$\int \mathcal{G}(A, \vec{s} - B\,\vec{u}) \cdot \mathcal{G}(C, \vec{u} - \vec{t}\,)\, d^N u = \frac{(\det D)^{1/2} (\det E)^{1/2}}{(\det A)^{1/2} (\det C)^{1/2}} \cdot \mathcal{G}(E, \vec{s} - B\vec{t}) \int \mathcal{G}(D, \vec{u} - \vec{m})\, d^N u =$$

$$= \frac{(\det D)^{1/2} (\det E)^{1/2}}{(\det A)^{1/2} (\det C)^{1/2}} \cdot \mathcal{G}(E, \vec{s} - B\vec{t}) \tag{C.22}$$

Let us integrate both sides of this expression in $\vec{s}$:

$$\int\int \mathcal{G}(A, \vec{s} - B\,\vec{u}) \cdot \mathcal{G}(C, \vec{u} - \vec{t})\, d^N\vec{u}\; d^N\vec{s} = \frac{(\det\, D)^{1/2}\,(\det\, E)^{1/2}}{(\det\, A)^{1/2}\,(\det\, C)^{1/2}} \int \mathcal{G}(E, \vec{s} - B\vec{t})\, \vec{s}$$

The double integral of the left term is 1, since

$$\int\int \mathcal{G}(A, \vec{s}-B\,\vec{u})\cdot\mathcal{G}(C, \vec{u}-\vec{t})\, d^N u\, d^N\,\vec{s} = \int \mathcal{G}(A, \vec{s}-B\,\vec{u}) \left(\int \mathcal{G}(C, \vec{u} - \vec{t})\, d^N\,\vec{s}\right)\, d^N\, u =$$

$$= \int \mathcal{G}(A, \vec{s} - B\,\vec{u}) \cdot 1 \cdot d^N\, u = 1$$

by the normalization property of the gaussians. For the same reason the single integral of the right term is also 1. Therefore,

$$\frac{(\det\, D)^{1/2}\,(\det\, E)^{1/2}}{(\det\, A)^{1/2}\,(\det\, C)^{1/2}} = 1$$

Finally, if we use this result in eq. C.22 we get the property we were looking for:

$$\int \mathcal{G}(A, \vec{s} - B\,\vec{u}) \cdot \mathcal{G}(C, \vec{u} - \vec{t}\,)\, d^N\, u = \mathcal{G}(A + BCB^T, \vec{s} - B\,\vec{t}\,)$$

C.2.9 Derivative of the determinant of a matrix

Simple derivative

Let us consider Y, a non-singular square matrix of $c \times c$ components. We are interested in calculating the derivative of $\det Y$ respect to its $l\,m$ entry, that is, $\frac{\partial}{\partial Y_{lm}} \det Y$. First we will express the determinant using the Laplacian expansion by minors [Muir, 1960]:

$$\det Y = \sum_{i=1}^{c} (-1)^{i+j}\, Y_{ij}\, M_{ij} \tag{C.23}$$

where j can be arbitrarely chosen in the $1\,..\,c$ range, and M_{ij} is a so-called minor of Y, obtained by taking the determinant of Y with row i and column j "crossed out".

If we choose $j = m$ none of the minors in the expansion depends on Y_{lm}, and we can write:

$$\frac{\partial}{\partial Y_{lm}} \det Y = \sum_{i=1}^{c} (-1)^{i+m} \frac{\partial Y_{im}}{\partial Y_{lm}} M_{im} = \sum_{i=1}^{c} (-1)^{i+m} \delta_{il} M_{im} = (-1)^{l+m} M_{lm} \quad \text{(C.24)}$$

where δ_{il} is the Kronecker delta. But we know that the lm component of the inverse of Y, $[Y^{-1}]_{lm}$, can be expressed as [Muir, 1960]:

$$\left[Y^{-1}\right]_{lm} = \frac{1}{\det Y} (-1)^{l+m} M_{lm} \quad \text{(C.25)}$$

Therefore, we get $\frac{\partial}{\partial Y_{lm}} \det Y = \det Y \, [Y^{-1}]_{lm}$. In abbreviated form:

$$\frac{d}{dY} \det Y = (\det Y) \, Y^{-T} \quad \text{(C.26)}$$

where Y^{-T} is the inverse transposed of Y, and we define $\frac{d}{dY} \det Y$ as a $c \times c$ matrix with entries defined by $\left[\frac{d}{dY} \det Y\right]_{lm} \equiv \frac{\partial}{\partial Y_{lm}} \det Y$. In case Y is symmetrical:

$$\frac{d}{dY} \det Y = \det Y \, Y^{-1} \quad \text{(C.27)}$$

Complex derivative

Let us consider the determinant $\det\left(A + B\,U\,\Phi\,U^T\,B^T\right)$, where A is a $t \times t$ matrix, B is a $t \times c$ matrix, U is a $c \times d$ matrix and Φ is a symmetric $d \times d$ matrix. Its derivative respect to a given entry of U can be expressed using the chain rule of derivatives as:

$$\frac{\partial \det\left(A + B\,U\,\Phi\,U^T\,B^T\right)}{\partial U_{lm}} = \sum_{i=1}^{t} \sum_{j=1}^{t} \frac{\partial \det Y}{\partial Y_{ij}} \frac{\partial Y_{ij}}{\partial U_{lm}} \quad \text{(C.28)}$$

with $Y = A + B\,U\,\Phi\,U^T\,B^T$. Using the result of the previous section we get:

$$\frac{\partial \det\left(A + B\,U\,\Phi\,U^T\,B^T\right)}{\partial U_{lm}} = \sum_{i=1}^{t}\sum_{j=1}^{t}\left[\left(A + B\,U\,\Phi\,U^T\,B^T\right)^{-1}\right]_{ij}\frac{\partial\left[B\,U\,\Phi\,U^T\,B^T\right]_{ij}}{\partial U_{lm}} \tag{C.29}$$

Note that A does not appear in the last term since it does not depend on U_{lm}. The term $\left[B\,U\,\Phi\,U^T\,B^T\right]_{ij}$ can be expressed as:

$$\left[B\,U\,\Phi\,U^T\,B^T\right]_{ij} = \sum_{\alpha=1}^{c} B_{i\alpha}\sum_{\beta=1}^{d} U_{\alpha\beta}\sum_{\gamma=1}^{d}\Phi_{\beta\gamma}\sum_{\lambda=1}^{c}\left[U^T\right]_{\gamma\lambda}\left[B^T\right]_{\lambda j} =$$

$$= \sum_{\alpha=1}^{c} B_{i\alpha}\sum_{\beta=1}^{d} U_{\alpha\beta}\sum_{\gamma=1}^{d}\Phi_{\beta\gamma}\sum_{\lambda=1}^{c} U_{\lambda\gamma}B_{j\lambda} \tag{C.30}$$

If we derivate this quantity respect to U_{lm} we obtain:

$$\frac{\partial\left[B\,U\,\Phi\,U^T\,B^T\right]_{ij}}{\partial U_{lm}} = \sum_{\alpha=1}^{c} B_{i\alpha}\sum_{\beta=1}^{d}\delta_{\alpha l}\delta_{\beta m}\sum_{\gamma=1}^{d}\Phi_{\beta\gamma}\sum_{\lambda=1}^{c} U_{\lambda\gamma}B_{j\lambda} + \sum_{\alpha=1}^{c} B_{i\alpha}\sum_{\beta=1}^{d} U_{\alpha\beta}\sum_{\gamma=1}^{d}\Phi_{\beta\gamma}\sum_{\lambda=1}^{c}\delta_{\lambda l}\delta_{\gamma m}B_{j\lambda} =$$

$$= B_{il}\sum_{\gamma=1}^{d}\Phi_{m\gamma}\sum_{\lambda=1}^{c} U_{\lambda\gamma}B_{j\lambda} + \sum_{\alpha=1}^{c} B_{i\alpha}\sum_{\beta=1}^{d} U_{\alpha\beta}\Phi_{\beta m}B_{jl} = B_{il}\left[\Phi\,U^T\,B^T\right]_{mj} + B_{jl}\left[B\,U\,\Phi\right]_{mj} \tag{C.31}$$

If we use this in eq. C.29 we get:

$$\frac{\partial\left(A + B\,U\,\Phi\,U^T\,B^T\right)}{\partial U_{lm}} =$$

$$\sum_{i=1}^{t} B_{il}\sum_{j=1}^{t}\left[\left(A + B\,U\,\Phi\,U^T\,B^T\right)^{-1}\right]_{ij}\left[\Phi\,U^T\,B^T\right]_{mj} + \sum_{i=1}^{t}\left[B\,U\,\Phi\right]_{im}\sum_{j=1}^{t}\left[\left(A + B\,U\,\Phi\,U^T\,B^T\right)^{-1}\right]_{ij}B_{jl} =$$

$$= \sum_{i=1}^{t} B_{il}\left[\left(A + B\,U\,\Phi\,U^T\,B^T\right)^{-1}B\,U\,\Phi\right]_{im} + \sum_{i=1}^{t}\left[B\,U\,\Phi\right]_{im}\left[\left(A + B\,U\,\Phi\,U^T\,B^T\right)^{-1}B\right]_{il} =$$

$$\left[B^T\left(A + B\,U\,\Phi\,U^T\,B^T\right)^{-1}B\,U\,\Phi\right]_{lm} + \left[\Phi\,U^T\,B^T\left(A + B\,U\,\Phi\,U^T\,B^T\right)^{-1}B\right]_{ml} =$$

$$2\left[B^T\left(A+BU\Phi U^T B^T\right)^{-1} BU\Phi\right]_{lm} \tag{C.32}$$

In abbreviated form:

$$\frac{d\left(A+BU\Phi U^T B^T\right)}{dU} = 2\left[B^T\left(A+BU\Phi U^T B^T\right)^{-1} BU\Phi\right] \tag{C.33}$$

C.3 Property used in the derivation of ΔP

Let us consider the set $\mathcal{A} = \{\vec{x}_1, \vec{x}_2, ...\}$. If $\vec{x}^*$ is the element of $\mathcal{A}$ which minimizes $f(\vec{x}) + \beta g(\vec{x})$ with $\beta > 0$, then there does not exist any $\vec{x} \in \mathcal{A}$ with neither $f(\vec{x}) < f(\vec{x}^*)$ and $g(\vec{x}) \leq g(\vec{x}^*)$, nor $f(\vec{x}) \leq f(\vec{x}^*)$ and $g(\vec{x}) < g(\vec{x}^*)$.

Proof: Suppose there exists such a $\vec{x}$ satisfying $f(\vec{x}) < f(\vec{x}^*)$ and $g(\vec{x}) \leq g(\vec{x}^*)$, or $f(\vec{x}) \leq f(\vec{x}^*)$ and $g(\vec{x}) < g(\vec{x}^*)$. Then any of these two situations imply $f(\vec{x})+\beta g(\vec{x}) < f(\vec{x}^*) + \beta g(\vec{x}^*)$ which contradicts the hypothesis.

C.4 Shannon's conditioned entropy satisfies the requirements for an uncertainty measure

Next we will demonstrate that Shannon's conditioned entropy satisfies the requirements of section 5.2.2 for a proper uncertainty measure.

Independency on the specific implementation

The conditioned entropy of A given B, $H(A|B)$, is defined as:

$$H(A|B) = -\sum_{i,j} p(a_i \wedge b_j) \cdot \log(p(a_i|b_j)) \tag{C.34}$$

Since it depends only on statistical relations between the states this measure satisfies the requirement about independency on the specific implementation. Moreover, if C and D are relabellings of A and B respectively, then their statistical relations would be conserved and therefore $H(A|B) = H(C|D)$.

Positiveness and zero uncertainty

$H(A|B)$ satisfies these requirements (see [Cover & Thomas, 1991] for details).

Triangular inequality

Using the equalities [Cover & Thomas, 1991]:

$$I(A;C|B) = H(A|B) - H(A|B,C) = H(C|B) - H(C|A,B) \tag{C.35}$$

valid for any A, B, C we get $H(A|B) = H(A|B,C) + H(C|B) - H(C|A,B)$. On the other hand, $H(A|B,C) \leq H(A|C)$ [Cover & Thomas, 1991] and $H(C|A,B)$ $ge0$ in general. Therefore,

$$H(A|B) = H(A|B,C) + H(C|B) - H(C|A,B) \leq H(A|C) + H(C|B) \tag{C.36}$$

which is the triangular inequality we were looking for: $H(A|B) \leq H(A|C) + H(C|B)$.

Uncertainty about an external variable is never reduced in a closed system (Generalized data processing inequality theorem)

Let us consider the Markov chain in figure C.1.

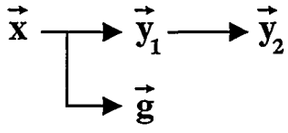

Figure C.1: Flow of information in a closed system. The second processing step is statistically independent of g given y_1.

$\vec{y}_1$ and $\vec{g}$ are (possibly) stochastic functions of $\vec{x}$, and $\vec{y}_2$ is a (possibly) stochastic function of $\vec{y}_1$. That is, $\vec{y}_2$ and $\vec{g}$ are conditionally independent given $\vec{y}_1$. The joint probability function of $\vec{y}_1$, $\vec{y}_2$ and $\vec{g}$ can be then described as:

$$p(\vec{y}_1, \vec{y}_2, \vec{g}) = p(\vec{y}_1)\, p(\vec{y}_1|\vec{g})\, p(\vec{y}_2|\vec{y}_1) \tag{C.37}$$

Then it is easy to prove that $I(\vec{y}_2; \vec{g}|\vec{y}_1) = 0$ (Cover & Thomas 1991). On the other hand, $I(\vec{g}; \vec{y}_1, \vec{y}_2)$ can be described in two equivalent manners (Cover & Thomas 1991):

$$I(\vec{g}; \vec{y}_1, \vec{y}_2) = I(\vec{y}_1; \vec{g}) + I(\vec{y}_2; \vec{g}|\vec{y}_1) = I(\vec{y}_2; \vec{g}) + I(\vec{y}_1; \vec{g}|\vec{y}_2) \tag{C.38}$$

where, using the fact that $I(\vec{g}; \vec{y}_2|\vec{y}_1) = 0$ it follows

$$I(\vec{y}_2; \vec{g}) \leq I(\vec{y}_1; \vec{g}) \tag{C.39}$$

that is, no any processing can increase the information of a closed system about the objective. Since $I(\vec{y}_1; \vec{g}) = H(g) - H(g|\vec{y}_1)$ and $I(\vec{y}_2; \vec{g}) = H(g) - H(g|\vec{y}_2)$ [Cover & Thomas, 1991] this inequality is equivalent to:

$$H(\vec{g}|\vec{y}_2) \geq H(\vec{g}|\vec{y}_1) \tag{C.40}$$

that is, no any processing performed by a closed system can decrease the uncertainty about the objective.

C.5 Equivalent expressions for ΔP when Shannon's entropy is chosen as the uncertainty measure

The original definition of ΔP can be rewritten in an useful equivalent manner which makes it useful for comparison with other techniques.

$$\Delta P(\vec{x} \to \vec{y}|\vec{g}) = d(\vec{x}, \vec{g}) - d(\vec{y}, \vec{g}) \tag{C.41}$$

If we choose Shannon's conditioned entropy as a measure of uncertainty we get:

$$\Delta P(\vec{x} \to \vec{y}|\vec{g}) = H(\vec{x}|\vec{g}) - H(\vec{y}|\vec{g}) + \beta(H(\vec{g}|\vec{y}) - H(\vec{g}|\vec{x})) \tag{C.42}$$

Using the equalities $I(a;b) = H(a) - H(a|b) = H(b) - H(b|a)$ [Cover & Thomas, 1991] it is possible to rewrite this equation as:

$$\Delta P(\vec{x} \to \vec{y}|\vec{g}) = H(\vec{x}) - H(\vec{y}) + (\beta+1)(H(g|x) - H(g|y)) \tag{C.43}$$

or, analogously,

$$\Delta P(\vec{x} \to \vec{y}|\vec{g}) = H(\vec{x}) - H(\vec{y}) + (\beta+1)(I(y;g) - I(x;g)) \tag{C.44}$$

C.6 Calculation of the entropy of a multidimensional gaussian variable

Here we consider the quantization of a gaussian variable of N dimensions with pdf:

$$p(\vec{y}) = \mathcal{G}(E, \vec{y}) \tag{C.45}$$

with E being an $N \times N$ covariance matrix. Now the quantization is defined by N independent directions, each one with its own quantization bin Δ_α. The quantization is thus defined by a matrix Δ where its columns $\vec{q}_\alpha$ are the different quantization directions and the norms of these columns are the corresponding quantization bins Δ_α.

Now we consider a stochastic quantization of the variable $\vec{y}$. Thus each quantization direction α has an associated dispersion σ_α. The activation functions defining the symbols are then multidimensional gaussian distributions with covariance matrix $Q^T \Phi Q$, with Φ being a diagonal matrix with $\Phi_{\alpha\alpha} = \frac{\sigma_\alpha^2}{\Delta_\alpha^2}$. These functions can be expressed as $\mathcal{G}(Q^T \Phi Q, \vec{y} - \vec{c}_{i_1,i_2,\ldots,i_N})$, where $\vec{c}_{i_1,i_2,\ldots,i_N}$ is the point where the symbol $y^{\Delta}_{i_1,i_2,\ldots,i_N}$ is centered:

$$\vec{c}_{i_1,i_2,\ldots,i_N} = \sum_{\alpha=1}^{N} i_\alpha \, \vec{q}_\alpha = Q \cdot \begin{bmatrix} i_1 \\ i_2 \\ \cdot \\ \cdot \\ i_N \end{bmatrix} = Q \cdot [i_1 \, i_2 \ldots i_N]^T \tag{C.46}$$

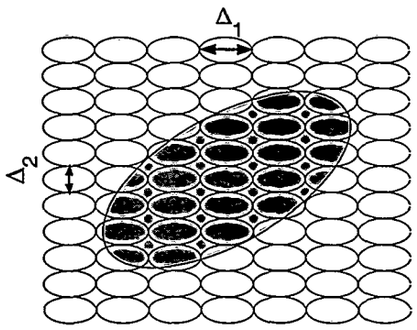

Figure C.2: Stochastic quantization of a multidimensional gaussian variable. Analogously to the scalar case, each discretization symbol has its own activation function, given by a gaussian distribution. Their dispersion along the quantization directions are represented by the lengh of their principal axis.

As in the previous section we have to normalize the activation functions in order to calculate the probabilities:

$$p(y^{\Delta}_{i_1,i_2,\ldots,i_N}|\vec{y}) = \frac{\mathcal{G}(Q^T \Phi Q, \, \vec{y} - \vec{c}_{i_1,i_2,\ldots,i_N})}{\sum_{j_1,j_2,\ldots,j_N=-\infty}^{\infty} \mathcal{G}(Q^T \Phi Q, \, \vec{y} - \vec{c}_{j_1,j_2,\ldots,j_N})} \tag{C.47}$$

Now we consider the equality:

$$\lim_{\alpha \to 0} \left(\sum_{n_1,n_2,\ldots,n_N=-\infty}^{\infty} \det(\alpha A) \cdot \mathcal{G}(I, \alpha \, A \cdot [n_1 \, n_2 \ldots n_N]^T - \vec{b}) \right) =$$

$$= \lim_{\alpha \to 0} \left(\sum_{n_1,n_2,\ldots,n_N=-\infty}^{\infty} \alpha^N \det(A) \cdot \mathcal{G}(I, A \cdot [\alpha n_1 \, \alpha n_2 \ldots \alpha n_N]^T - \vec{b}) \right) =$$

$$= \int_{\mathcal{R}^N} (\det A) \cdot \mathcal{G}(I, \, A \cdot \vec{x} - \vec{b}) \, d^N x =$$

$$= \int_{\mathcal{R}^N} \mathcal{G}(A^{-2}, \, \vec{x} - A^{-1}\vec{b}) \, d^N x = 1 \tag{C.48}$$

for A definite positive, where we have used the property C.17.

This implies:

$$\lim_{tr(A)\to 0}\left(\sum_{n_1,n_2,\ldots,n_N=-\infty}^{\infty}\det(A)\cdot\mathcal{G}(I,A\cdot[n_1\,n_2\ldots n_N]^T-\vec{b})\right)=1 \tag{C.49}$$

with A definite positive, where $tr(A)\to 0$ forces that all the elements of A tend to 0. This lets us make the following approximation for small $|A|$:

$$\sum_{n_1,n_2,\ldots,n_N=-\infty}^{\infty}\det(A)\cdot\mathcal{G}(I,\,A\cdot[n_1\,n_2\ldots n_N]^T-\vec{b})\simeq 1 \tag{C.50}$$

which is the multidimensional version of the approximation exposed in the previous section (eq. 6.13) for scalar gaussian variables. Analogously, this series can be proved to be always convergent as long as $\det A\neq 0$.

This together with property C.17 lets us write:

$$\sum_{j_1,j_2,\ldots,j_N=-\infty}^{\infty}\mathcal{G}(Q^T\Phi Q,\,\vec{y}-\vec{c}_{j_1,j_2,\ldots,j_N})\;=\sum_{j_1,j_2,\ldots,j_N=-\infty}^{\infty}\mathcal{G}(Q^T\Phi Q,\,Q\cdot[j_1\,j_2\ldots,j_N]^T-\vec{y})$$

$$=\sum_{j_1,j_2,\ldots,j_N=-\infty}^{\infty}(\det(Q^T\Phi Q))^{-\frac{1}{2}}\cdot\mathcal{G}\left(I,\,(Q^T\Phi Q)^{-\frac{1}{2}}\cdot\left(Q\cdot[j_1\,j_2\ldots,j_N]^T-\vec{y}\right)\right)\simeq(\det Q)^{-1} \tag{C.51}$$

which lets us simplify eq. C.47 as:

$$p(y^{\Delta}_{i_1,i_2,\ldots,i_N}|\vec{y})\simeq(\det Q)\cdot\mathcal{G}(Q^T\Phi Q,\,\vec{y}-\vec{c}_{i_1,i_2,\ldots,i_N}) \tag{C.52}$$

Now we calculate the probability of the discretized symbols:

$$\begin{aligned}p(y^{\Delta}_{i_1,i_2,\ldots,i_N})&=\int_{\mathcal{R}^N}p(y^{\Delta}_{i_1,i_2,\ldots,i_N}|\vec{y})\cdot p(\vec{y})\,d^N y=\\&=(\det Q)\int_{\mathcal{R}^N}\mathcal{G}(Q^T\Phi Q,\,\vec{y}-\vec{c}_{i_1,i_2,\ldots,i_N})\;\mathcal{G}(E,\vec{y})=\\&=(\det Q)\cdot\mathcal{G}(Q^T\Phi Q+E,\,\vec{c}_{i_1,i_2,\ldots,i_N})\end{aligned} \tag{C.53}$$

where we have made use of the theorem of gaussians convolution (eq. C.21).

The negative logarithm of this expression is:

$$-\log p(y^{\Delta}_{i_1,i_2,\ldots,i_N}) = \frac{1}{2}\left[\log\frac{(2\pi)^N \det(Q^T\Phi Q+E)}{\det Q} + \vec{c}^{\,T}_{i_1,i_2,\ldots,i_N}\,(Q^T\Phi Q+E)\,\vec{c}_{i_1,i_2,\ldots,i_N}\right] \tag{C.54}$$

Therefore we can calculate the entropy as:

$$H(y^{\Delta}) = -\sum_{i_1,i_2,\ldots,i_N=-\infty}^{\infty} p(y^{\Delta}_{i_1,i_2,\ldots,i_N})\,\log\,p(y^{\Delta}_{i_1,i_2,\ldots,i_N}) =$$

$$= \frac{1}{2}\left[\log\frac{(2\pi)^N \det(Q^T\Phi Q+E)}{\det Q}\right]\sum_{i_1,i_2,\ldots,i_N=-\infty}^{\infty} \det Q\cdot\mathcal{G}(Q^T\Phi Q+E,\vec{c}_{i_1,i_2,\ldots,i_N}) +$$

$$+\frac{1}{2}\sum_{i_1,i_2,\ldots,i_N=-\infty}^{\infty} \det Q\cdot\left(\vec{c}^{\,T}_{i_1,i_2,\ldots,i_N}\cdot(Q^T\Phi Q+E)\cdot\vec{c}_{i_1,i_2,\ldots,i_N}\right)\cdot\mathcal{G}(Q^T\Phi Q+E,\vec{c}_{i_1,i_2,\ldots,i_N}) \tag{C.55}$$

Using eq. C.50 the first summatory of this equation is 1. In order to calculate the second summatory let us consider the equation C.50:

$$\sum_{n_1,n_2,\ldots,n_N=-\infty}^{\infty} \det(A)\cdot\mathcal{G}(I,\,A\cdot[n_1\,n_2\ldots n_N]^T) \simeq 1$$

Notice we have taken $\vec{b}=0$ since we will not need it for our further considerations. Suppose that the above approximation holds well for A (A is small enough). Since the second term in the equation does not depend on A, this means that the left term is constant over perturbations in A. Then, if we make the change $A \to \lambda A$ with λ close to 1 then the summatory remains constant. Therefore:

$$\frac{d}{d\lambda}\left(\sum_{n_1,n_2,\ldots,n_N=-\infty}^{\infty} \det(\lambda A)\cdot\mathcal{G}\left(I,\,\lambda A\cdot[n_1\,n_2\ldots n_N]^T\right)\right)_{\lambda=1} \simeq 0$$

Computing the derivatives in $\lambda=1$ and rearranging terms we obtain:

$$\sum_{n_1,n_2,\ldots,n_N=-\infty}^{\infty} \det(A) \cdot [n_1\, n_2 \ldots n_N] \cdot A^2 \cdot [n_1\, n_2 \ldots n_N]^T \cdot \mathcal{G}(I,\, A \cdot [n_1\, n_2 \ldots n_N]^T) \simeq N \tag{C.56}$$

Remember that $\vec{c}_{i_1,i_2,\ldots,i_N} = Q \cdot [i_1\, i_2 \ldots i_N]^T$ and the property $\mathcal{G}(Q^T\Phi Q + E, \vec{c}_{i_1,i_2,\ldots,i_N}) = (\det(Q^T\Phi Q+E)^{\frac{-1}{2}})\ \mathcal{G}(I, (Q^T\Phi Q+E)^{-\frac{1}{2}} \cdot \vec{c}_{i_1,i_2,\ldots,i_N})$ (eq. C.17). Using these facts together with eq. C.56, where we choose $A = (Q^T\Phi Q + E)^{-\frac{1}{2}}$, we obtain that the second summatory in C.55 is $\simeq 1$.

Then:

$$H(y^\Delta) \simeq \frac{1}{2}\log\frac{(2\pi)^N \det(Q^T\Phi Q + E)}{\det Q} + \frac{N}{2} \tag{C.57}$$

This can be simplified as:

$$H(y^\Delta) \simeq \frac{1}{2}\log(2\pi e)^N \det(\Phi + Q^{-T} E\, Q^{-1}) \tag{C.58}$$

As in the previous section, we will determine the optimal values of the diagonal entries of Φ in order to obtain a good approximation of the deterministic quantization $H(y^\Delta)_d$. Remember that Φ was defined as a diagonal matrix with $\Phi_{ii} = \frac{\sigma_i^2}{\Delta_i^2}$. Since the deterministic entropy is invariant under arbitrary elections of the coordinate axis, we have $\frac{\sigma_i}{\sigma} = c$ with c being a constant. Finally, the entropy of the deterministic entropy goes to zero when the discretization bins of all the quantization directions go to ∞ (that is, all the components of Q^{-1} go to 0). This determines $2\pi e c^2 = 1$. Using these properties we obtain:

$$H(y^\Delta) \simeq \frac{1}{2}\log\det(I + 2\pi e\, Q^{-T} E\, Q^{-1}) \tag{C.59}$$

As in the previous section, it can be demonstrated that in the other limit (the quantization bins go to 0) this entropy coincides with the differential entropy plus the term due to the quantization $\frac{1}{2}\log\det Q^2$. Therefore this approximation is valid in both limits.

These properties can be also shown to be satisfied independently for each different

quantization direction. For example, if we make one of the quantization bins in $\vec{y}$ go to ∞ the projections of $\vec{y}$ on that quantization direction will not contribute to the total entropy. Notice that this equation for multidimensional variables C.59 contains as a special case the scalar equation 6.21.

C.7 Maximization of ΔP in a linear system with linear objectives

C.7.1 Derivation of the expression of ΔP

In order to quantify the information processing measure in our system we need to calculate:

$$\Delta P(\vec{x}^{\Delta} \rightarrow \vec{y}^{\Delta}|\vec{g}^{\Delta}) = d(\vec{x}^{\Delta}, \vec{g}^{\Delta}) - d(\vec{y}^{\Delta}, \vec{g}^{\Delta}) =$$

$$= H(\vec{x}^{\Delta}|\vec{g}^{\Delta}) + \beta H(\vec{g}^{\Delta}|\vec{x}^{\Delta}) - H(\vec{y}^{\Delta}|\vec{g}^{\Delta}) - \beta H(\vec{g}^{\Delta}|\vec{y}^{\Delta})$$

Since $H(a) - H(a|b) = H(b) - H(b|a)$ we can rewrite this equation and rearrange it as:

$$\Delta P = (1+\beta)H(\vec{x}^{\Delta}|\vec{g}^{\Delta}) - \beta H(\vec{x}^{\Delta}) - (1+\beta)H(\vec{y}^{\Delta}|\vec{g}^{\Delta}) + \beta H(\vec{y}^{\Delta}) \tag{C.60}$$

This rearrangement will simplify our calculations.

Now we need to calculate the entropies. Since $p(\vec{x}) = \mathcal{G}(C, \vec{x})$ and $p(\vec{x}'|\vec{x}) = \mathcal{G}(N_x, \vec{x}')$, we calculate the entropy of the quantized version of $\vec{x}$ as (eq. 6.27)

$$H(\vec{x}^{\Delta}) = \frac{1}{2}\log\det(I + N_x^{-1}\, C) \tag{C.61}$$

If we do the same with $\vec{y}$ we can calculate the entropy of its quantized version as:

$$H(\vec{y}'^{\Delta}) = \frac{1}{2}\log\det(I + N_y^{-1}\, A(C + N_x)A^T) \tag{C.62}$$

In order to calculate the conditioned entropy $H(\vec{x}^{\Delta}|\vec{g}^{\Delta})$ we will first calculate the joint entropy $H(\vec{x}'^{\Delta}, \vec{g}^{\Delta})$ and then using the property $H(a|b) = H(a,b) - H(b)$ we will obtain $H(\vec{x}^{\Delta}|\vec{g}^{\Delta})$. In order to calculate the joint entropy we need the pdf $p(\vec{x}, \vec{y})$. Since the input statistics and input noise are gaussian, and the goal is linear, the pdf of $\begin{pmatrix} \vec{x} \\ \vec{g} \end{pmatrix}$ is defined by a multidimensional gaussian. Its covariance matrix will be:

$$\begin{pmatrix} \left\langle (\vec{x}-\overline{\vec{x}})(\vec{x}-\overline{\vec{x}})^T \right\rangle & \left\langle (\vec{x}-\overline{\vec{x}})(\vec{g}-\overline{\vec{g}})^T \right\rangle \\ \left\langle (\vec{g}-\overline{\vec{g}})(\vec{x}-\overline{\vec{x}})^T \right\rangle & \left\langle (\vec{g}-\overline{\vec{g}})(\vec{g}-\overline{\vec{g}})^T \right\rangle \end{pmatrix} = \begin{pmatrix} C & CW^T \\ WC & WCW^T \end{pmatrix}$$

The total quantization matrix is

$$\begin{pmatrix} Q_x & 0 \\ 0 & Q_g \end{pmatrix}$$

Using these equations in 6.24 we have:

$$H(\vec{x}^{\Delta}, \vec{g}^{\Delta}) = \frac{1}{2} \log \det \begin{pmatrix} I + 2\pi e Q_x^{-T} C Q_x^{-1} & 2\pi e Q_x^{-T} C W^T Q_g^{-1} \\ 2\pi e Q_g^{-T} W C Q_x^{-1} & I + 2\pi e Q_g^{-T} W C W^T Q_g^{-1} \end{pmatrix}$$

which can be rearranged as:

$$H(\vec{x}^{\Delta}, \vec{g}^{\Delta}) = \frac{1}{2} \log \det (I + 2\pi e Q_g^{-T} W C W^T Q_g^{-1}) +$$

$$\frac{1}{2} \log \det \left((I + 2\pi e Q_x^{-T} C Q_x^{-1}) - (2\pi e)^2 (Q_x^{-T} C W^T Q_g^{-1})(I + 2\pi e Q_g^{-T} W C W^T Q_g^{-1})^{-1} (Q_g^{-T} W C Q_x^{-1}) \right)$$

The second determinant can be simplified to

$$\det \left(I + Q_x^{-T} \left(\frac{1}{2\pi e} C^{-1} + W^T Q_g^{-1} Q_g^{-T} W \right)^{-1} Q_x^{-1} \right) =$$

$$= \det \left(I + 2\pi e Q_x^{-T} \left(C^{-1} + 2\pi e W^T Q_g^{-1} Q_g^{-T} W \right)^{-1} Q_x^{-1} \right)$$

Therefore, we obtain the conditioned entropy as:

$$H(\vec{x}^{\Delta}|\vec{g}^{\Delta}) = \frac{1}{2}\log\det\left(I + 2\pi e Q_x^{-T}\left(C^{-1} + 2\pi e W^T Q_g^{-1} Q_g^{-T} W\right)^{-1} Q_x^{-1}\right) \quad \text{(C.63)}$$

Remembering the noise-quantization duality we have to assign $Q_x = \frac{1}{\sqrt{2\pi e}} N_x^{1/2}$ so that:

$$H(\vec{x}^{\Delta}|\vec{g}^{\Delta}) = \frac{1}{2}\log\det\left(I + N_x^{-1}\left(C^{-1} + 2\pi e W^T Q_g^{-1} Q_g^{-T} W\right)^{-1}\right) \quad \text{(C.64)}$$

We will calculate the other entropy we need, $H(\vec{y}^{\Delta}|\vec{g}^{\Delta})$, using the same strategy. As in the previous case, the pdf of $\begin{pmatrix} \vec{y} \\ \vec{g} \end{pmatrix}$ is also a multidimensional gaussian, now with covariance matrix:

$$\begin{pmatrix} \left\langle (\vec{y}-\overline{\vec{y}})(\vec{y}-\overline{\vec{y}})^T \right\rangle & \left\langle (\vec{y}-\overline{\vec{y}})(\vec{g}-\overline{\vec{g}})^T \right\rangle \\ \left\langle (\vec{g}-\overline{\vec{g}})(\vec{y}-\overline{\vec{y}})^T \right\rangle & \left\langle (\vec{g}-\overline{\vec{g}})(\vec{g}-\overline{\vec{g}})^T \right\rangle \end{pmatrix} = \begin{pmatrix} A(C+N_x)A^T & ACW^T \\ WCA & WCW^T \end{pmatrix}$$

The total quantization matrix is in this case:

$$\begin{pmatrix} Q_y & 0 \\ 0 & Q_g \end{pmatrix}$$

Using these equations in 6.24 we have:

$$H(\vec{y}^{\Delta}, \vec{g}^{\Delta}) = \frac{1}{2}\log\det\begin{pmatrix} I + 2\pi e Q_y^{-T} A(C+N_x)A^T Q_y^{-1} & 2\pi e Q_y^{-T} ACW^T Q_g^{-1} \\ 2\pi e Q_g^{-T} WCA^T Q_y^{-1} & I + 2\pi e Q_g^{-T} WCW^T Q_g^{-1} \end{pmatrix}$$

which can be rearranged as:

$$H(\vec{y}^{\Delta}, \vec{g}^{\Delta}) = \frac{1}{2}\log\det(I + 2\pi e Q_g^{-T} WCW^T Q_g^{-1}) +$$

$$\frac{1}{2}\log\det\left(I+2\pi eQ_y^{-T}\left(AN_xA^T+A\left(C^{-1}+2\pi eW^TQ_g^{-1}Q_g^{-T}W\right)^{-1}A^T\right)Q_y^{-1}\right) \tag{C.65}$$

Then we obtain the conditioned entropy as:

$$H(\vec{y}^{\Delta}|\vec{g}^{\Delta})=\frac{1}{2}\log\det\left(I+2\pi eQ_y^{-T}A\left(N_x+\left(C^{-1}+2\pi eW^TQ_g^{-1}Q_g^{-T}W\right)^{-1}\right)A^TQ_y^{-1}\right) \tag{C.66}$$

Using the noise-quantization duality we have $Q_y=\frac{1}{\sqrt{2\pi e}}N_y^{1/2}$ so:

$$H(\vec{y}^{\Delta}|\vec{g}^{\Delta})=\frac{1}{2}\log\det\left(I+N_y^{-1}\left(AN_xA^T+A\left(C^{-1}+2\pi eW^TQ_g^{-1}Q_g^{-T}W\right)^{-1}A^T\right)\right) \tag{C.67}$$

This entropy, together with the previously calculated for $H(\vec{x}^{\Delta})$ (eq. C.61), $H(\vec{y}^{\Delta})$ (eq. C.62), and $H(\vec{x}^{\Delta},\vec{y}^{\Delta})$ (eq. C.63) are all the terms we need in order to calculate ΔP.

Then we can express ΔP as:

$$\Delta P(x^{\Delta}\rightarrow y^{\Delta}|g^{\Delta})=\frac{1}{2}\ln\frac{(\det(I+D))^{1+\beta}}{(\det(I+S))^{\beta}}+\frac{1}{2}\ln\frac{(\det(I+V\Phi V^T))^{\beta}}{\det(I+VV^T)^{\beta+1}} \tag{C.68}$$

where $D=N_x^{-1}(C^{-1}+2\pi eW^TN_g^{-1}W)^{-1}$, $S=N_x^{-1}C$, $\Phi=(N_x+N_xD)^{-1/2}(C+N_x)(N_x+N_xD)^{-1/2}$ and $V=N_y^{-1/2}A(N_x+N_xD)^{1/2}$.

The second term in the summation C.68 is the one which determines the maximization of Δ since the other does not depend on the responses of the neurons (matrix A).

It can be easily proved that if R is an orthogonal transformation in the space of neurons, then the solution $\hat{V}=RV$ has exactly the same ΔP than V. Thus there does not exist a unique optimal configuration but a family of optimal solutions.

C.7.2 General properties of the optimal solution

The functional to maximize is:

$$\frac{1}{2}\ln\frac{\det(I+V\Phi V^T)^\beta}{\det(I+VV^T)^{\beta+1}} \tag{C.69}$$

where Φ is a definite positive matrix and $\beta > 0$, and the free parameters are the entries of the receptive field matrix V of $ny \times M$ components. The *ith* row of V defines the receptive field of neuron i. We are interested in obtaining the matrix V which maximizes globally the functional. Note that this maxima occurs in a finite configuration, since the functional tends to $-\infty$ if the absolute value of one or more of the components of V tends to ∞.

Therefore we have to look at the fixed points of the functional. Thus the gradient of the functional must be null at the global maximum $\hat{V}$:

$$\beta(I+\hat{V}\Phi\hat{V}^T)^{-1}\hat{V}\Phi - (\beta+1)(I+\hat{V}\hat{V}^T)^{-1}\hat{V} = 0 \tag{C.70}$$

where we have used the formula for the gradient of a determinant developed in appendix C.2.9. It is easy to verify that, if R_y is a rotation matrix of $ny \times ny$ components (rotation in the space of neurons), then the change $\hat{V} \to R_y\hat{V}$ in C.69 keeps this functional unaltered. Moreover, $R_y\hat{V}$ is also a solution of C.70. Thus, if $\hat{V}$ maximizes C.69, then $R_y\hat{V}$ is also a maximum, for any rotation matrix R_y. Therefore we do not have a unique maximum but a *family* of optimal configurations.

Since VV^T is a symmetric matrix, there exists a rotation R_y such that $R_yVV^TR_y^T$ is diagonal. That is, the change $\hat{V} \to R_y\hat{V} \equiv \tilde{V}$ makes $\tilde{V}\tilde{V}^T$ diagonal. Since this point is also a fixed point as argued before, we can write:

$$\beta(I+\tilde{V}\Phi\tilde{V}^T)^{-1}\tilde{V}\Phi - (\beta+1)(I+\tilde{V}\tilde{V}^T)^{-1}\tilde{V} = 0 \tag{C.71}$$

Multiplying to the right by $\tilde{V}^T$ and rearranging we get:

$$\beta(I+\tilde{V}\Phi\tilde{V}^T)^{-1}\tilde{V}\Phi\tilde{V}^T = (\beta+1)(I+\tilde{V}\tilde{V}^T)^{-1}\tilde{V}\tilde{V}^T \tag{C.72}$$

Since $\tilde{V}\tilde{V}^T \equiv D$ is diagonal, $\beta > 0$, and Φ is definite positive, it is easy to see that $\tilde{V}\Phi\tilde{V}^T \equiv D_2$ must also be diagonal. Then we can rearrange equation C.71 as

$$\Phi\tilde{V}^T = (1+\frac{1}{\beta})(I+D)^{-1}(I+D_2)\,\tilde{V}^T \tag{C.73}$$

which is equivalent to $\Phi\vec{\tilde{v}}_i = \gamma_i\vec{\tilde{v}}_i$ with $\gamma_i \equiv (1+\frac{1}{\beta})\,[(I+D)^{-1}(I+D_2)]_{ii}$ and $i = 1..n_y$. That is, the columns of the global maximum $\tilde{V}^T$, $\vec{\tilde{v}}_i$, are either eigenvectors of Φ or null vectors. Note that since $\tilde{V}\tilde{V}^T$ is diagonal, then $\vec{\tilde{v}}_i^T\vec{\tilde{v}}_j = 0$ for any pair $i \neq j$. Therefore, the set of non null rows of $\tilde{V}$ forms an orthogonal set of vectors. This eliminates the possibility of repeated rows. Note also that if we permute the rows of $\tilde{V}$ we obtain another equivalent optimal solution.

C.7.3 Specific properties of the optimal configuration

Since both $\tilde{V}\tilde{V}^T$ and $\tilde{V}\Phi\tilde{V}^T$ are diagonal, the value of the functional C.69 evaluated in $\tilde{V}$ can be rewritten as:

$$\frac{1}{2}\sum_{i=1}^{n_y}\ln\frac{(1+\vec{\tilde{v}}_i^T\Phi\vec{\tilde{v}}_i)^\beta}{(1+\vec{\tilde{v}}_i^T\vec{\tilde{v}}_i)^{\beta+1}} \tag{C.74}$$

which is the summation of the separate contributions of the rows of $\tilde{V}$. Since $\log(1) = 0$, the contribution of the null rows of $\tilde{V}$ is null. Now we will find out which eigenvectors of Φ should be included in the solution, and with which norm.

Suppose $\vec{a}$ is an eigenvector of Φ with eigenvalue equal to λ, and squared norm equal to t (therefore $t \geq 0$). Is it included in the optimal configuration ? If it were, it would contribute to the functional with:

$$\frac{1}{2}\ln\frac{(1+t\,\lambda)^\beta}{(1+t)^{\beta+1}} \tag{C.75}$$

and the squared norm t should acquire the value which maximizes this contribution. If we calculate the derivative of this expression respect to t we obtain:

$$\frac{1}{2}\frac{\lambda\beta(1+t)-(\beta+1)(1+t\lambda)}{(1+t\,\lambda)(1+t)} \tag{C.76}$$

The sign of this derivative is positive as long as $t < \beta - \frac{\beta+1}{\lambda}$, and negative if $t > \beta - \frac{\beta+1}{\lambda}$.

Since t must be greater or equal to 0, this lends two possible situations (figure C.3):

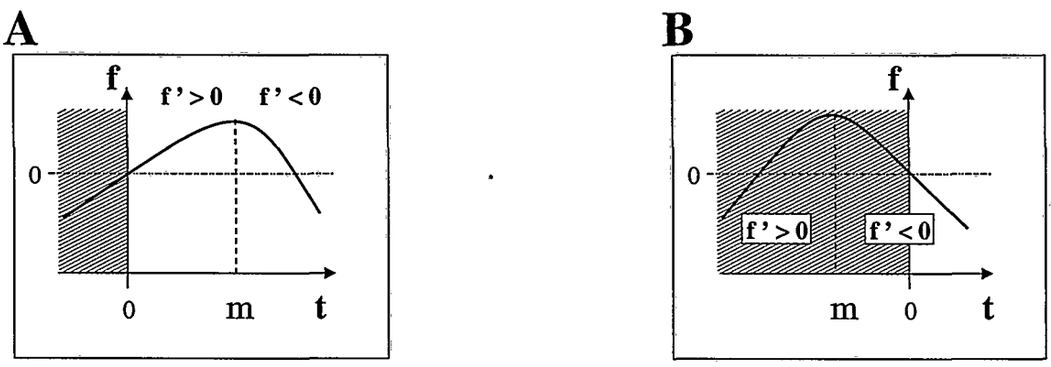

Figure C.3: Maximum of a function $f(t)$ which satisfies $f(0) = 0$, $f'(t) > 0$ for $t < m$, and $f'(t) < 0$ for $t > m$. The maximization must be done with the restriction $t \geq 0$. **A**: $m > 0$. The maximum occurs at $t = m$. **B**: $m < 0$. The maximum occurs at $t = 0$ since the region $t < 0$ is not allowed (dashed style).

1. $\beta - \frac{\beta+1}{\lambda} \leq 0$: the optimal value for t is 0 and therefore $\vec{a}$ is null. Thus this eigenvector can not exist in the optimal solution (its contribution is less than 0, therefore a null vector is better).

2. $\beta - \frac{\beta+1}{\lambda} > 0$: in case this eigenvector exists in the optimal configuration, it has squared norm equal to $\beta - \frac{\beta+1}{\lambda}$ and its contribution to the functional is greater than zero.

Thus we have determined which eigenvectors can not exist in the optimal configuration, and which others are candidates to be included. Let us call n_c the number of these candidates. In case $n_c = n_y$, the optimal configuration would be then formed by all these candidates. If $n_c < n_y$, the optimal configuration is formed by all these candidates plus $n_y - n_c$ null vectors. Finally, if $n_c > n_y$, we have to choose the n_y candidates with greatest contribution.

Let us prove that if $\vec{a}$ and $\vec{b}$ are two eigenvectors of Φ having eigenvectors λ_a and λ_b respectively, which satisfy $\lambda_a > 1 + \frac{1}{\beta}$, $\lambda_b > 1 + \frac{1}{\beta}$ (they are candidates to be in the optimal configuration), and $\lambda_a > \lambda_b$, then the optimal contribution of $\vec{a}$ is greater

than that of $\vec{b}$. Remember that the optimal contribution of a candidate eigenvector occurs at $t = \beta - \frac{\beta+1}{\lambda}$. Then its contribution to ΔP is:

$$\frac{1}{2}\ln\frac{(1+t\lambda)^\beta}{(1+t)^{\beta+1}} = \frac{1}{2}\ln\frac{(1+\lambda\beta-\beta-1)^\beta}{(1+\beta-\frac{\beta+1}{\lambda})^{\beta+1}} = \frac{1}{2}\ln\frac{\beta^\beta(\lambda-1)^\beta}{(\beta+1)^{\beta+1}(1-\frac{1}{\lambda})^{\beta+1}} =$$

$$\frac{1}{2}\ln\frac{\beta^\beta}{(\beta+1)^{\beta+1}}\frac{(\lambda-1)^\beta}{(\lambda-1)^{\beta+1}\left(\frac{1}{\lambda}\right)^{\beta+1}} = \frac{1}{2}\ln\frac{\beta^\beta}{(\beta+1)^{\beta+1}} + \frac{1}{2}\ln\frac{\lambda^{\beta+1}}{(\lambda-1)} \qquad \text{(C.77)}$$

Note that all the terms are well defined since $\beta - \frac{\beta+1}{\lambda} > 0$ implies $\lambda > 1$.

Finally let us calculate the derivative of the eigenvector contribution respect to λ:

$$\frac{d}{d\lambda}\left(\frac{1}{2}\ln\frac{\beta^\beta}{(\beta+1)^{\beta+1}} + \frac{1}{2}\ln\frac{\lambda^{\beta+1}}{(\lambda-1)}\right) = \frac{\beta+1}{\lambda} - \frac{1}{\lambda-1} = \frac{(\beta+1)(\lambda-1)-\lambda}{\lambda(\lambda-1)} = \frac{\beta\lambda-\beta-1}{\lambda(\lambda-1)} \qquad \text{(C.78)}$$

which, if $\lambda > 1 + \frac{1}{\beta}$, is always positive. Therefore, if $\lambda_a > \lambda_b$, then the optimal contribution of $\vec{a}$ to ΔP is strictly greater than that of $\vec{b}$. As a conclusion, in case that $n_y < n_c$, we should choose the candidates with greatest eigenvalues.

C.7.4 Derivation of the specific equations for the autoencoder

In this specific case, the objective if the system is to reconstruct the original signal x with precision given by Δ_g. Therefore, $W = I$, and for clarity reasons we define the symbol $\Delta_x \equiv \Delta_g$ for the required reconstruction precision. Then

$$\Phi = (N_x + (C^{-1} + 2\pi e N_g^{-1})^{-1})^{-1/2}(C + N_x)(N_x + (C^{-1} + 2\pi e N_g^{-1})^{-1})^{-1/2} \qquad \text{(C.79)}$$

For simplicity, let us assume that $N_x = \sigma_x^2 I$, $N_y = \sigma_y^2 I$ and $N_g = \Delta_g^2 I$. Then the eigenvectors of Φ coincide with the eigenvectors of C. It is easy to show that the eigenvalues of Φ are:

$$\lambda_i = \frac{\sigma_x^2 + \sigma_{c_i}^2}{\sigma_x^2 + (2\pi e \Delta_g^{-2} + \sigma_{c_i}^{-2})^{-1}} \tag{C.80}$$

where $\sigma_{c_i}^2$ are the eigenvalues of C. If we define $a_i \equiv \frac{\sigma_{Ci}^2}{\sigma_x^2}$ and $b \equiv \frac{\Delta_x^2}{2\pi e \sigma_x^2}$, we can rewrite this expression as:

$$\lambda_i = \frac{1 + a_i}{1 + (b^{-1} + a_i^{-1})^{-1}} = \frac{(a_i + b)(1 + a_i)}{a_i + b + a_i b} \tag{C.81}$$

Then the constraint $\lambda_i > 1 + \frac{1}{\beta}$ can be written as

$$\lambda_i - 1 = \frac{(a_i + b)(1 + a_i)}{a_i + b + a_i b} - 1 = \frac{a_i^2}{a_i + b + a_i b} > \frac{1}{\beta} \tag{C.82}$$

which leads to

$$\beta a_i^2 - (1 + b)a_i - b > 0 \tag{C.83}$$

with solutions

$$a_i = \frac{1 + b \pm \sqrt{(1 + b)^2 + 4\beta b}}{2\beta} \tag{C.84}$$

which roots have opposite sign. Since $a_i \geq 0$ the positive root is the only value where eq. C.83 collapses to zero. Then it is easy to see that

$$\lambda_i > 1 + \frac{1}{\beta} \leftrightarrow a_i > \frac{1 + b + \sqrt{(1 + b)^2 + 4\beta b}}{2\beta} \tag{C.85}$$

C.8 Gradient calculation for the Non Linear Feature Extraction algorithm

The functional to maximize is:

$$\Delta P = d(\vec{x},\vec{g}) - d(\vec{x},\vec{g}) = d(\vec{x},\vec{g}) + \beta\frac{1}{2}\log\det(I+\tilde{Q}_y^{-2}S_c) - (\beta+1)\sum_{i=1}^{N_c} p_{c_i}\frac{1}{2}\log\det(I+\tilde{Q}_y^{-2}S_{w_i}) \tag{C.86}$$

Where $S_c = \frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i - \vec{\mu})(\vec{y}_j^i - \vec{\mu})^T$, $S_{w_i} = \frac{1}{N_i}\sum_{j=1}^{N_i}(\vec{y}_j^i - \vec{\mu}_i)(\vec{y}_j^i - \vec{\mu}_i)^T$, $\vec{\mu} = \frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\vec{x}_j^i$ and $\vec{\mu}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\vec{x}_j^i$.

The response of the neuron k to the j th example in class i, $\vec{x}_j^i$, is a nonlinear function $f_k(\vec{x}_j^i, \vec{\alpha}_i)$ where $\vec{\alpha}_i$ is the set of internal parameters of that unit. Therefore we will compute the gradient of our functional respect to the internal parameters in order to find the ones which optimize the performance of our network.

The derivative of eq. C.86 respect to α_{kz} is:

$$\frac{\partial\Delta P}{\partial\alpha_{kz}} = \frac{1}{2}\beta\sum_{l,m=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{lm}\frac{\partial\left[S_c\right]_{lm}}{\partial\alpha_{kz}} - \frac{\beta+1}{2}\sum_{i=1}^{N_c}p_{c_i}\sum_{l,m=1}^{R}(\tilde{Q}_y^2 + S_{w_i})^{-1}\frac{\partial\left[S_{w_i}\right]_{lm}}{\partial\alpha_{kz}} \tag{C.87}$$

For clarity we will define the two terms:

$$A_{kz} \equiv \frac{1}{2}\sum_{l,m=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{lm}\frac{\partial\left[S_c\right]_{lm}}{\partial\alpha_{kz}} \tag{C.88}$$

and

$$B_{kz}^i \equiv \frac{1}{2}\sum_{l,m=1}^{R}(\tilde{Q}_y^2 + S_{w_i})^{-1}\frac{\partial\left[S_{w_i}\right]_{lm}}{\partial\alpha_{kz}} \tag{C.89}$$

Therefore,

$$\frac{\partial\Delta P}{\partial\alpha_{kz}} = \beta A_{kz} - (\beta+1)\sum_{i=1}^{N_c}p_{c_i}B_{kz}^i \tag{C.90}$$

We will first calculate A_{kz}. We need to calculate the derivative of S_c:

$$\left[\frac{\partial S_c}{\partial\alpha_{kz}}\right]_{lm} = \left[\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial(\vec{y}_j^i - \vec{\mu})}{\partial\alpha_{kz}}(\vec{y}_j^i - \vec{\mu})^T + \frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i - \vec{\mu})\frac{\partial(\vec{y}_j^i - \vec{\mu})}{\partial\alpha_{kz}}\right]_{lm}$$

If we split the term $\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial(\vec{y}_j^i-\vec{\mu})}{\partial\alpha_{kz}}(\vec{y}_j^i-\vec{\mu})^T$ in two summatories:

$$\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial(\vec{y}_j^i-\vec{\mu})}{\partial\alpha_{kz}}(\vec{y}_j^i-\vec{\mu})^T=\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\vec{y}_j^i}{\partial\alpha_{kz}}(\vec{y}_j^i-\vec{\mu})^T-\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\vec{\mu}}{\partial\alpha_{kz}}(\vec{y}_j^i-\vec{\mu})^T=$$

$$=\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\vec{y}_j^i}{\partial\alpha_{kz}}(\vec{y}_j^i-\vec{\mu})^T-\frac{\partial\vec{\mu}}{\partial\alpha_{kz}}\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i-\vec{\mu})^T$$

The last term is zero since $\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i-\vec{\mu})^T=\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i)^T-\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\mu^T=\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i)^T-\mu^T=0$ because $\vec{\mu}=\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\vec{y}_j^i$ (eq. 6.3.3). Therefore,

$$\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial(\vec{y}_j^i-\vec{\mu})}{\partial\alpha_{kz}}(\vec{y}_j^i-\vec{\mu})^T=\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\vec{y}_j^i}{\partial\alpha_{kz}}(\vec{y}_j^i-\vec{\mu})^T$$

The transpose of this equation is the other term we need:

$$\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i-\vec{\mu})\frac{\partial(\vec{y}_j^i-\vec{\mu})^T}{\partial\alpha_{kz}}=\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i-\vec{\mu})\frac{\partial(\vec{y}_j^i)^T}{\partial\alpha_{kz}}$$

Therefore we can write:

$$\left[\frac{\partial S_c}{\partial\alpha_{kz}}\right]_{lm}=\left[\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\vec{y}_j^i}{\partial\alpha_{kz}}(\vec{y}_j^i-\vec{\mu})^T+\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\vec{y}_j^i-\vec{\mu})\frac{\partial(\vec{y}_j^i)^T}{\partial\alpha_{kz}}\right]_{lm}$$

If $\vec{a}$ and $\vec{b}$ are two column vectors with the same number of components, then $[\vec{a}\vec{b}^T]_{lm}=[\vec{a}]_l[\vec{b}]_m$. Therefore,

$$\left[\frac{\partial S_c}{\partial\alpha_{kz}}\right]_{lm}=\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_l}{\partial\alpha_{kz}}(\left[\vec{y}_j^i\right]_m-[\vec{\mu}]_m)+\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\left[\vec{y}_j^i\right]_l-[\vec{\mu}]_l)\frac{\partial\left[\vec{y}_j^i\right]_m}{\partial\alpha_{kz}}$$

Note that since the neuron k only depends on $\vec{\alpha}_k$, the derivative $\frac{\partial[\vec{y}_j^i]_l}{\partial\alpha_{kz}}$ is 0 for $l\neq k$. Then:

$$\left[\frac{\partial S_c}{\partial \alpha_{kz}}\right]_{lm} = \delta_{lk}\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_m - [\vec{\mu}]_m\right) + \delta_{mk}\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_l - [\vec{\mu}]_l\right)$$

where δ_{lk} is the Kronecker delta.

Substituting into C.88:

$$A_{kz} =$$

$$= \frac{1}{2}\sum_{l,m=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{lm}\left(\delta_{lk}\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_m - [\vec{\mu}]_m\right) + \delta_{mk}\frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_l - [\vec{\mu}]_l\right)\right) =$$

$$= \frac{1}{2N}\sum_{l,m=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{lm}\delta_{lk}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_m - [\vec{\mu}]_m\right) +$$

$$+\frac{1}{2N}\sum_{l,m=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{lm}\delta_{mk}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_l - [\vec{\mu}]_l\right) =$$

$$= \frac{1}{2N}\sum_{m=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{km}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_m - [\vec{\mu}]_m\right) +$$

$$+\frac{1}{2N}\sum_{l=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{lk}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_l - [\vec{\mu}]_l\right)$$

Since $(\tilde{Q}_y^2 + S_c)$ is symmetric, we can write:

$$A_{kz} = \frac{1}{N}\sum_{m=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{km}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left(\left[\vec{y}_j^i\right]_m - [\vec{\mu}]_m\right)$$

Changing the order in the summatories:

$$A_{kz} = \frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\sum_{m=1}^{R}\left[(\tilde{Q}_y^2 + S_c)^{-1}\right]_{km}\left(\left[\vec{y}_j^i\right]_m - [\vec{\mu}]_m\right)$$

which can be written in a compact form:

$$A_{kz} = \frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left[(\tilde{Q}_y^2 + S_c)^{-1}(\vec{y}_j^i - \vec{\mu})\right]_k$$

We can calculate B_{kz}^i using the same strategy, yielding:

$$B_{kz}^i = \frac{1}{N_i}\sum_{j=1}^{N_i}\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left[(\tilde{Q}_y^2 + S_{w_i})^{-1}(\vec{y}_j^i - \vec{\mu}_i)\right]_k$$

Therefore, eq. C.90 is calculated as:

$$\frac{\partial\Delta P}{\partial\alpha_{kz}} = \sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\left(\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left[\frac{\beta}{N}(\tilde{Q}_y^2 + S_c)^{-1}(\vec{y}_j^i - \vec{\mu}) - \frac{(\beta+1)}{N_i}p_{c_i}(\tilde{Q}_y^2 + S_{w_i})^{-1}(\vec{y}_j^i - \vec{\mu}_i)\right]_k\right)$$

Finally, since $p_{c_i} = \frac{N_i}{N}$ we have:

$$\frac{\partial\Delta P}{\partial\alpha_{kz}} = \frac{1}{N}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}\left(\frac{\partial\left[\vec{y}_j^i\right]_k}{\partial\alpha_{kz}}\left[\beta(\tilde{Q}_y^2 + S_c)^{-1}(\vec{y}_j^i - \vec{\mu}) - (\beta+1)(\tilde{Q}_y^2 + S_{w_i})^{-1}(\vec{y}_j^i - \vec{\mu}_i)\right]_k\right) \tag{C.91}$$

C.9 Construction of decision trees

The output of a tree for a pattern is the terminal node that classifies it. Then $d(Y,G) = H(Y|G) + \beta H(G|Y)$. Consider the tree in figure C.4 A where the subtree I is a child of the root node, and J are the rest of the children of the root. Because a choice can be broken down into several successive choices, the global entropy is the weighted sum of the individual values of H:

$$H(Y) = p(I)H_I(Y) + (1-p(I))H_J(Y) - p(I)\log p(I) - (1-p(I))\log(1-p(I)) \tag{C.92}$$

where $H_I(Y)$ and $H_J(Y)$ are the entropies calculated with the local statistics respectively. Note that $p(I)H_I(Y) - p(I)\log p(I) - (1-p(I))\log(1-p(I))$ is just the

entropy of the same tree replacing J by a single terminal node ($H_{no\,J}(Y)$) (figure C.4).

Thus $H(Y) = H_{no\,J}(Y) + p(J)\,H_J(Y)$, where $p(J)$ is the probability of a pattern to reach J. Analogously, $H(Y|G) = H_{no\,J}(Y|G) + p(J)\,H_J(Y|G)$.

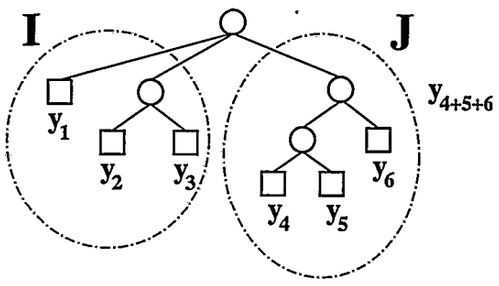

Figure C.4: General schema of a decision subtree. Decision nodes are drawn as circles whereas classification nodes are squares. Each terminal node corresponds to a different state of the system. We call y_{14+5+6} to the node resulting of replacing J by a single terminal node.

Then $d(\vec{Y}, \vec{G})$ can be written as:

$$d(\vec{Y}, \vec{G}) = d(\vec{Y}_{no\,J}, G) + p(J)\,(H_J(Y|G) - \beta I_J(Y;G)) \tag{C.93}$$

where $d(\vec{Y}_{no\,J}, \vec{G})$ is the distance of the tree without the subtree J, and $H_J(Y)$ and $I_J(Y;G)$ are computed using the local statistics in J.

References

Bibliography

[Abbott, 1990] Abbott, L. (1990) Learning in neural network memories. *Network,* **1**, 105–122.

[Abbott & Dayan, 1999] Abbott, L. & Dayan, P. (1999) The effect of correlated variability on the accuracy of a population code. *Neural Comput.,* **11**, 91–101.

[Abbott & Sejnowski, 1999] Abbott, L. & Sejnowski, T. (1999) *Neural codes and distributed representations: foundations of neural computation.* MIT Press.

[Abeles, 1991] Abeles, M. (1991) *Corticonics : Neural Circuits of the Cerebral Cortex.* Cambridge University Press.

[Albus & Wolf, 1984] Albus, K. & Wolf, W. (1984) Early post-natal development of neuronal function in the kitten's visual cortex: a laminar analysis. *J. Physiol. (Lond.),* **348**, 153–85.

[Alkon *et al.*, 1991] Alkon, D., Amaral, D., Bear, M., Black, J., Carew, T., Cohen, N., Disterhoft, J., Eichenbaum, H., Golski, S., Gorman, L., Lynch, G., Mcnaughton, B., Mishkin, M., Moyer, J., Olds, J., Olton, D., Otto, T., Squire, L., Staubli, U., Thompson, L. & Wible, C. (1991) Learning and memory. *Brain Res. Rev.,* **16**, 193–220.

[Amari *et al.*, 1996] Amari, S., Cichochi, A. & Yang, H. H. (1996) *A new learning algorithm for blind signal separation,* vol. 8, of *Advances in Neural Information Processing Systems.* MIT Press.

[Andrade & Morán, 1996] Andrade, M. & Morán, F. (1996) Structural study of the development of ocularity domains using a neural network model. *Biol. Cybern.,* **74**, 243–54.

[Arbib, 1998] Arbib, M., ed. (1998) *The handbook of brain theory and neural networks.* MIT Press, Cambridge, Massachusetts; London, England.

[Arfken, 1985] Arfken, G. (1985) *Mathematical Methods for Physicists.* 3th edition,, Orlando, FL: Academic Press.

[Atick, 1992] Atick, J. (1992) Could information theory provide an ecological theory of sensory processing ? *Network,* **3**, 213–51.

[Atick & Redlich, 1990] Atick, J. & Redlich, A. (1990) Towards a theory of early visual processing. *Neural Comp.,* **2**, 308–320.

[Attneave, 1954] Attneave, F. (1954) Some informational aspects of visual perception. *Psychol. Rev.,* **61**, 183–93.

[Baddeley, 1996] Baddeley, R. (1996) Visual perception. an efficient code in v1? *Nature,* **381**, 560–561.

[Bakin & Weinberger, 1996] Bakin, J. & Weinberger, N. (1996) Induction of a physiological memory in the cerebral cortex by stimulation of the nucleus basalis. *Proc. Natl. Acad. Sci. USA,* **93**, 11219–11224.

[Barlow, 1961] Barlow, H. (1961) *Sensory Communication.* Cambridge, Massachusetts: MIT Press.

[Barlow, 1989] Barlow, H. (1989) Unsupervised learning. *Neur. Comput.,* **1**, 295–311.

[Bear *et al.*, 1996] Bear, M., Connors, B. & Paradiso, M. (1996) *Neuroscience: Exploring the Brain.* Lippincott, Williams & Wilkins.

[Bell & Sejnowski, 1995] Bell, A. & Sejnowski, T. (1995) An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.,* **7** (6), 1129–59.

[Bi & Poo, 1998] Bi, G. & Poo, M. (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.,* **18**, 10464–72.

[Bienenstock *et al.*, 1982] Bienenstock, E., Cooper, L. & Munro, P. (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.,* **2** (1), 32–48.

[Blakemore & van Sluyters, 1975] Blakemore, C. & van Sluyters, R. (1975) Innate and environmental factors in the development of the kitten's visual cortex. *J. Physiol. (Lond.),* **248**, 663–716.

[Bliss & Collingridge, 1993] Bliss, T. & Collingridge, G. (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature,* **361**, 31–9.

[Borst & Theunissen, 1999] Borst, A. & Theunissen, F. (1999) Information theory and neural coding. *Nat. Neuroscience,* **2** (11), 947–57.

[Braastadt & Heggelund, 1985] Braastadt, B. & Heggelund, P. (1985) Development of spatial receptive-field organization and orientation selectivity in kitten striate cortex. *J. Neurophysiol.,* **53**, 1158–78.

[Britten *et al.*, 1992] Britten, K., Shadlen, M., Newsome, W. & Movshon, J. (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience,* **12**, 4745–4765.

[Brown & Chattarji, 1998] Brown, T. & Chattarji, S. (1998) *The handbook of brain theory and neural networks.* Cambridge, Massachusetts; London, England: MIT Press.

[Bruce *et al.*, 1981] Bruce, C., Desimone, R. & Gross, C. (1981) Visual properties of neurons in a polysensory area in the superior temporal sulcus of the macaque. *J. Neurophysiol.,* **46**, 369–84.

[Buck & Axel, 1991] Buck, L. & Axel, R. (1991) A novel multigene family may encode odorant receptors - a molecular-basis for odor recognition. *Cell,* **65**, 175–187.

[Buonomano & Merzenich, 1998] Buonomano, D. & Merzenich, M. (1998) Cortical plasticity: from synapses to maps. *Annu. Rev. Neurosci.,* **21**, 149–86.

[Buzsaki & Kandel, 1998] Buzsaki, X. & Kandel, E. (1998) Somadendritic backpropagation of action potentials in cortical pyramidal cells of the awake rat. *J. Neurophysiol.,* **79** (3), 1587–91.

[Campa *et al.*, 1995] Campa, A., Giudice, P. D., Parga, N. & Nadal, J.-P. (1995) Maximization of mutual information in a linear noisy network: a detailed study. *Network: computation in neural systems,* **6**, 449–468.

[Chapman & Stryker, 1993] Chapman, B. & Stryker, M. (1993) Development of orientation selectivity in ferret visual cortex and effects of deprivation. *Journal of Neuroscience,* **13**, 5251–62.

[Chess *et al.*, 1994] Chess, A., Simon, I., Cedar, H. & Axel, R. (1994) Allelic activation regulates olfactory receptor gene expression. *Cell,* **78**, 823–834.

[Cover & Thomas, 1991] Cover, T. & Thomas, J. (1991) *Elements of Information Theory.* John Wiley, New York.

[Cruz & Dorronsoro, 1998] Cruz, C. S. & Dorronsoro, J. (1998) A nonlinear discriminant algorithm for data projection and feature extraction. *IEEE Trans. on Neural Networks,* **9**, 1370–6.

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

[Deco & Obradovic, 1996] Deco, G. & Obradovic, D. (1996) *An Information–Theoretic Approach to Neural Computing.* Springer.

[Deneve *et al.*, 1999] Deneve, S., Latham, P. & Pouget, A. (1999) Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience,* **2**, 740–745.

[Douglas & Martin, 1991] Douglas, R. & Martin, K. (1991) A functional microcircuit for cat visual cortex. *J. Physiol.,* **440**, 735–769.

[Duda & Hart, 1973] Duda, R. O. & Hart, P. E. (1973) *Pattern Classification and Scene Analysis.* J. Wiley and Sons.

[Fregnac, 1998] Fregnac, Y. (1998) *The handbook of brain theory and neural networks.* Cambridge, Massachusetts; London, England: MIT Press.

[Freund & Gulyas, 1991] Freund, T. & Gulyas, A. (1991) Gabaergic interneurons containing calbindin d28k or somatostatin are major targets of gabaergic basal forebrain afferents in the rat neocortex. *J. Comp. Neurol.,* **314**, 187–99.

[Freund & Meskenaite, 1992] Freund, T. & Meskenaite, V. (1992) Gamma-aminobutyric acid-containing basal forebrain neurons innervate inhibitory interneurons in the neocortex. *Proc. Natl. Acad. Sci. USA,* **89**, 738–742.

[Frigo & Johnson, 1998] Frigo, M. & Johnson, S. (Seattle, WA, 1998) Fftw: an adaptive software architecture for the fft. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* vol. 3, pp. 1381–1384.

[Furht, 1998] Furht, B., ed. (1998) *Handbook of Multimedia Computing.* CRC Press, Cambridge, Massachusetts; London, England.

[Georgopoulos *et al.*, 1982] Georgopoulos, A., Kalaska, J., Caminiti, R. & Massey, J. (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.,* **2** (11), 1527–37.

[Gerstner *et al.*, 1993] Gerstner, W., Ritz, R. & van Hemmen, J. (1993) Why spikes? hebbian learning and retrieval of time-resolved excitation patterns. *Biol. Cybern.*, **69**, 503–15.

[Golub & van Loan, 1996] Golub, G. H. & van Loan, C. F. (1996) *Matrix Computations.* 3th edition,, Jonhs Hopkins Series in the Mathematical Sciences, Baltimore, MD.

[Hasselmo, 1993] Hasselmo, M. (1993) Acetylcholine and learning in a cortical associative memory. *Neural Comp.,* **5**, 32–44.

[Hebb, 1949] Hebb, D. O. (1949) *The Organization of Behavior.* John Wiley & Sons, New York.

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

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

[Hubel & Wiesel, 1963] Hubel, D. & Wiesel, T. (1963) Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *J. Neurophysiol.,* **26**, 994–1002.

[Humphrey & Saul, 1995] Humphrey, A. & Saul, A. (1995) Strobe rearing alters the spatiotemporal structure of simple cell receptive fields in cat area 17. In *Soc. Neurosci. Abstr.* vol. 21, p. 1648.

[K. Obermayer, 1995] K. Obermayer, T. Sejnowski, G. B. (1995) Neural pattern formation via a competitive hebbian mechanism. *Behav. Brain Res.,* **66** (1–2), 161–7.

[Kandel *et al.*, 1991] Kandel, E. R., Schwartz, J. M. & Jessell, T. M. (1991) *Principles of Neural Science.* 3rd ed. edition,, New York: Elsevier.

[Kandel *et al.*, 2000] Kandel, E. R., Schwartz, J. M. & Jessell, T. M., eds (2000) *Principles of Neural Science.* 4th ed. edition,, McGraw-Hill / Appleton & Lange.

[Kelly, 1985] Kelly, J. (1985) Auditory system. In *Principles of Neural Science*, (Kandel, E. & Schwartz, J., eds), pp. 396–408 Elsevier.

[Kilgard & Merzenich, 1998] Kilgard, M. & Merzenich, M. (1998) Cortical map reorganization enabled by nucleus basalis activity. *Science,* **279**, 1714–8.

[Koch & Segev, 1998] Koch, C. & Segev, I., eds (1998) *Methods in Neuronal Modeling. From Synapses to Networks.* second edition,, MIT Press: Cambridge, Massachusetts.

[König *et al.*, 1995] König, P., Engel, A., Roelfsema, P. & Singer, W. (1995) How precise is neuronal synchronization? *Neural Comput.,* **7**, 469–85.

[Körding & König, 2000] Körding, K. & König, P. (2000) A learning rule for dynamic recruitment and decorrelation. *Neural Networks,* **13**, 1–9.

[Köster & Sakmann, 1998] Köster, H. & Sakmann, B. (1998) Calcium dynamics in single spines during coincident pre- and postsynaptic activity depend on relative timing of back-propagating action potentials and subthreshold excitatory postsynaptic potentials. *Proc. Natl. Acad. Sci. U.S.A.,* **95** (16), 9596–601.

[Kratz *et al.*, 2002] Kratz, E., Dugas, J. C. & Ngai, J. (2002) Odorant receptor gene regulation: implications from genomic organization. *Trends in Genetics,* **18** (1), 29–34.

[Larkum *et al.*, 1999] Larkum, M., Zhu, J. & Sakmann, B. (1999) A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature,* **398**, 338–341.

[Laurent, 1999] Laurent, G. (1999) A systems perspective on early olfactory coding. *Science,* **286** (5440), 723–8.

[Levine, 1998] Levine, D. (1998) *PGAPack Parallel Genetic Algorithm Library.*

[LópezdeMántaras, 1991] LópezdeMántaras, R. (1991) A distance-based attribute selection measure for decision tree induction. *Machine Learning Journal,* **6**, 81–92.

[Maes, 1994] Maes, P. (1994) Modeling adaptive autonomous agents. *Artificial Life Journal,* **1** (1 and 2), 135–62.

[Magee *et al.*, 1998] Magee, J., Hoffman, D., Colbert, C. & Johnston, D. (1998) Electrical and calcium signaling in dendrites of hippocampal pyramidal neurons. *Annu. Rev. Physiol.,* **60**, 327–46.

[Markram *et al.*, 1997] Markram, H., Lubke, J., Frotscher, M. & Sakmann, B. (1997) Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science,* **275**, 213–215.

[Martin *et al.*, 2000] Martin, S., Grimwood, P. & Morris, R. (2000) Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu. Rev. Neurosci.,* **23**, 649–711.

[Meister *et al.*, 1991] Meister, M., Wong, R., Baylor, D. & Shatz, C. (1991) Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science,* **252**, 939–43.

[Miller *et al.*, 1999] Miller, K., Erwin, E. & Kayser, A. (1999) Is the devlopment of orientation selectivity instructed by activity ? *J. Neurobiology,* **41**, 44–57.

[Miller *et al.*, 1989] Miller, K., Keller, J. & Stryker, M. (1989) Ocular dominance column development: analysis and simulation. *Science,* **245**, 605–15.

[Mitchell, 1997] Mitchell, T. M. (1997) *Machine learning.* McGraw-Hill Series in Computer Science.

[Miyashita *et al.*, 1997] Miyashita, M., Kim, D.-S. & Tanaka, S. (1997) Cortical direction selectivity without directional experience. *Neuroreport,* **8** (5), 1187–91.

[Mombaerts, 2001] Mombaerts, P. (2001) How smell develops. *Nature Neuroscience,* **4**, 1192–1198.

[Movshon & van Sluyters, 1981] Movshon, J. & van Sluyters, R. (1981) Visual neural development. *Annu. Rev. Psychol.,* **32**, 477–522.

[Muir, 1960] Muir, T. (1960) *A Treatise on the Theory of Determinants.* New York: Dover.

[Oja, 1982] Oja, E. (1982) A simplified neuron model as a principal component analyzer. *J. of Mathematical Biology,* **15**, 267–273.

[Olshausen & Field, 1996] Olshausen, B. & Field, D. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature,* **381**, 607–609.

[Pearce *et al.*, 2002] Pearce, T., Schiffman, S., Nagle, H. & Gardner, J., eds (2002) *Handbook of Machine Olfaction: Electronic Nose Technology.* Wiley-VCH, Weinheim, Germany.

[Pearce & Sánchez-Montanés, 2002] Pearce, T. C. & Sánchez-Montanés, M. A. (2002) Chemical sensor array optimization: geometric and information -theoretic approaches. In *Handbook of Machine Olfaction: Electronic Nose Technology*, (Pearce, T. C., Schiffman, S. S., Nagle, H. T. & Gardner, J. W., eds), Wiley-VCH, Weinheim, Germany.

[Pouget *et al.*, 1999] Pouget, A., Deneve, S., Ducom, J. & Latham, P. (1999) Narrow versus wide tuning curves: what's best for a population code? *Neural Comput.,* **11**, 85–90.

[Pouget *et al.*, 1998] Pouget, A., Zhang, Z., Deneve, S. & Latham, P. (1998) Statistically efficient estimation using population coding. *Neural Comput.,* **10** (2), 373–401.

[Prechelt, 1994] Prechelt, L. (1994) *PROBEN1: A Set of Neural Network Benchmark Problems and Benchmarking Rules.* Technical Report 21/94, Fakultat fur Informatik, Universitat Karlsruhe, Germany Internal Report, Max–Planck–Institute of Biophysical Chemistry Gottingen, West Germany.

[Press *et al.*, 1992] Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992) *Numerical Recipes in C, Second Edition.* Cambridge University Press.

[Quinlan, 1986] Quinlan, J. R. (1986) Induction of decision trees. *Machine Learning,* **1(1)**, 81–106.

[Quinlan, 1993] Quinlan, J. R. (1993) *C4.5: Programs for machine learning*, vol. 58,. Morgan Kauffman.

[Rabiner & Juang, 1986] Rabiner, L. & Juang, B. (1986) An introduction to hidden markov models. *IEEE Acoustics Speech and Signal Processing (ASSP) Magazine,* **3** (1), 4–16.

[Rao & Sejnowski, 2000] Rao, R. & Sejnowski, T. (2000) Predictive sequence learning in recurrent neocortical circuits. In *Advances in Neural Information Processing Systems 12, (S.A. Solla T.K. Leen K.-R. Muller, eds.)* pp. 164–70 MIT Press.

[Reed, 2000] Reed, R. (2000) Regulating olfactory receptor expression: controlling globally, acting locally. *Nature Neuroscience,* **3** (7), 638–9.

[Rolls, 2000] Rolls, E. (2000) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron,* **27**, 205–18.

[Ruderman, 1994] Ruderman, D. L. (1994) The statistics of natural images. *Network,* **5**, 517–548.

[Sánchez-Montanes, 2001] Sánchez-Montanes, M. (2001) Strategies for the optimization of large scale networks of integrate and fire neurons. *Lecture Notes in Computer Science,* **2084**, 117–125.

[Sánchez-Montanés & Corbacho, 2002] Sánchez-Montanés, M. & Corbacho, F. (2002) Towards a new information processing measure for neural computation. In *International Conference on Artificial Neural Networks (ICANN 02)* vol. 2415, p. 637.

[Sánchez-Montanés & Corbacho, 2003] Sánchez-Montanés, M. & Corbacho, F. (2003) A new information processing measure for adaptive complex systems. *IEEE Transactions on Neural Networks (submitted),* .

[Sánchez-Montanés *et al.*, 1999] Sánchez-Montanés, M., Corbacho, F. & Sigüenza, J. (1999) Development of directionally selective microcircuits in striate cortex. In *Foundations and Tools for Neural Modeling*, (Mira, J. & Sánchez-Andrés, J. V., eds), vol. 1606, of *Lecture Notes in Computer Science* pp. 53–65 Springer Verlag.

[Sánchez-Montanés *et al.*, 2001] Sánchez-Montanés, M., König, P. & Verschure, P. F. M. J. (2001) Learning in a neural network model in real time using real world stimuli. *Neurocomputing,* **38–40**, 859–865.

[Sánchez-Montanés *et al.*, 2002] Sánchez-Montanés, M., König, P. & Verschure, P. F. M. J. (2002) Learning sensory maps with real-world stimuli in real time using a biophysically realistic learning rule. *IEEE Transactions on Neural Networks,* **13** (3), 619–632.

[Sánchez-Montanés & Pearce, 2001] Sánchez-Montanés, M. & Pearce, T. (2001) Fisher information and optimal odor sensors. *Neurocomputing,* **38–40**, 335–341.

[Sánchez-Montanés & Pearce, 2002] Sánchez-Montanés, M. & Pearce, T. (2002) Why do olfactory neurons have unspecific receptive fields ? *Biosystems,* **67**, 229–238.

[Sánchez-Montanés *et al.*, 2000] Sánchez-Montanés, M., Verschure, P. & König, P. (2000) Local and global gating of synaptic plasticity. *Neural Comput.,* **12** (3), 519–529.

[Sanhueza *et al.*, 2000] Sanhueza, M., Schmachtenberg, O. & Bacigalupo, J. (2000) Excitation, inhibition, and suppresion by odors in isolated toad and rat olfactory receptor neurons. *Am. J. Physiol. Cell Physiol.,* **279**, C31–9.

[Schild & Restrepo, 1998] Schild, D. & Restrepo, D. (1998) Transduction mechanisms in vertebrate olfactory receptor cells. *Physiol. Rev.,* **78** (2), 429–466.

[Sejnowski, 1977] Sejnowski, T. (1977) Storing covariance with nonlinearly interacting neurons. *J. Math. Biology,* **4**, 303–321.

[Serizawa *et al.*, 2000] Serizawa, S., Ishii, T., Nakatani, H., Tsuboi, A., Nagawa, F., Asano, M., Sudo, K., Sakagami, J., Sakano, H., Ijiri, T., Matsuda, Y., Suzuki, M.,

Yamamori, T., Iwakura, Y. & Sakano, H. (2000) Mutually exlusive expression of odorant receptor transgenes. *Nature Neuroscience,* **3** (7), 687–693.

[Seung & Sompolinsky, 1993] Seung, H. S. & Sompolinsky, H. (1993) Simple models for reading neuronal population codes. *Proc. Natl. Acad. Sci. USA,* **90**, 10749–53.

[Shannon & Weaver, 1949] Shannon, C. E. & Weaver, W. (1949) *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

[Shen, 1994] Shen, W. (1994) *Autonomous Learning from the Environment.* W. H. Freeman & Co.

[Sherman & Spear, 1982] Sherman, S. & Spear, P. (1982) Organization of visual pathways in normal and visually deprived cats. *Physiol. Rev.,* **62**, 738–855.

[Sicard & Holley, 1984] Sicard, G. & Holley, A. (1984) Receptor cell responses to odorants: similarities and differences among odorants. *Brain Research,* **292**, 283–296.

[Singer & Rauschecker, 1982] Singer, W. & Rauschecker, J. (1982) Central core control of development plasticity in the kitten visual cortex. ii: electrical activation of mesencephalic and diencephalic projections. *Exp. Brain Res.,* **47**, 223–233.

[Singer *et al.*, 1979] Singer, W., von Grünau, M. & Rauschecker, J. (1979) Requirements for the disruption of binocularity in the visual cortex of strabismic kittens. *Brain Res.,* **171**, 536–540.

[Spruston *et al.*, 1995] Spruston, N., Schiller, Y., Stuart, G. & Sakmann, B. (1995) Activity-dependent action potential invasion and calcium influx into hippocampal ca1 dendrites. *Science,* **268**, 297–300.

[Stent, 1973] Stent, G. (1973) A physiological mechanism for hebb's postulate of learning. *Proc. Natl. Acad. Sci. USA,* **70**, 997–1001.

[Stuart & Sakmann, 1994] Stuart, G. & Sakmann, B. (1994) Active propagation of somatic action potentials into neocortical pyramidal cell dendrites. *Nature,* **367**, 69–72.

[Sutton & Barto, 1998] Sutton, R. & Barto, A. (1998) *Reinforcement Learning.* MIT Press.

[Tishby *et al.*, 1999] Tishby, N., Pereira, F. & Bialek, W. (1999) The information bottleneck method. In *Proceedings of the 37-th Allerton Conference on Communication and Computation.*

[Tononi & Edelman, 1998] Tononi, G. & Edelman, G. (1998) Consciousness and complexity. *Science,* **282**, 1846–51.

[Tsien, 2000] Tsien, J. (2000) Linking hebb's coincidence-detection to memory formation. *Curr. Opin. Neurobiol.,* **10**, 266–273.

[Tsubokawa & Ross, 1996] Tsubokawa, H. & Ross, W. (1996) Ipsps modulate spike backpropagation and associated ca_i^{2+} changes in the dendrites of hippocampal ca1 pyramidal neurons. *J. Neurophysiol.,* **76**, 2896–906.

[Vapnik, 1998] Vapnik, V. (1998) *Statistical Learning Theory.* Wiley-Interscience.

[Varela *et al.*, 1997] Varela, J., Sen, K., Gibson, J., Fost, J., Abbott, L. & Nelson, S. (1997) A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. *J.Neurosci.,* **17**, 7926–7940.

[Verschure, 1997] Verschure, P. (1997) *Xmorph.* Internal Report, Institute of Neuroinformatics, ETH-UZ.

[Viterbi, 1967] Viterbi, A. (1967) Error bounds for convulational codes and asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory,* **13**.

[von der Malsburg, 1973] von der Malsburg, C. (1973) Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik,* **14**, 85–100.

[Weinberger, 1993] Weinberger, N. (1993) Learning induced changes of auditory receptive fields. *Cur. Opin. Neurobiol.,* **3**, 570–577.

[Weinberger *et al.*, 1993] Weinberger, N., Javid, R. & Lepan, B. (1993) Long-term retention of learning-induced receptive-field plasticity in the auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.,* **90**, 2394–8.

[White *et al.*, 2001] White, L., Coppola, D. & Fitzpatrick, D. (2001) The contribution of sensory experience to the maturation of orientation selectivity in ferret visual cortex. *Nature,* **411** (6841), 1049–52.

[Wilke & Eurich, 2002] Wilke, S. & Eurich, C. (2002) Representational accuracy of stochastic neural populations. *Neural Comp.,* **14**, 155–89.

[Wimbauer *et al.*, 1997] Wimbauer, S., Wenisch, O., Miller, K. & van Hemmen, J. (1997) Development of spatiotemporal receptive fields of simple cells: i. model formulation. *Biol. Cybern.,* **77** (6), 453–61.

[Zhang & Sejnowski, 1999] Zhang, K. & Sejnowski, T. (1999) Neuronal tuning: to sharpen or broaden ? *Neural Comput.,* **11**, 75–84.

[Zhang *et al.*, 1998] Zhang, L., Tao, H., Holt, C., Harris, W. & Poo, M. (1998) A critical window for cooperation and competition among developing retinotectal synapses. *Nature,* **395**, 37–44.

[Zozulya *et al.*, 2001] Zozulya, S., Echeverri, F. & Nguyen, T. (2001) The human olfactory receptor repertoire. *Genome Biology,* **2** (6), 1–12.

Reunido el tribunal que suscribe en el día
de la fecha, acordó calificar la presente Tesis
doctoral con Sobres. Cum Laudem
Madrid, 26 de Septiembre de 2003

J. DONOSO

(G. DECO)

/ E. Korutcheva /

(P. VERSCHURE)

R. López de Mántaras