Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Biophysical model of synaptic plasticity

Relatore:
Prof. Gastone Castellani

Presentata da:
Ariel Avanzi

Anno Accademico 2019/2020

To my beloved Grandmother

Abstract

In 1982 Elie Bienestock, Leon Cooper and Paul Munro wrote "Theory for the development of neuron selectivity" proposing a synaptic evolution scheme in which incoming patterns rather than converging afferents compete. It briefly became known as BCM theory and it was the springboard for further works on modification of cortial synapses. During the last two decades new formulations of the theory were made, like the IBCM done by Nathan Intrator and Leon Cooper in 1992, and new methods were introduced allowing the creation of more complicated and efficient neural netwoks. The study of these models points out their capability to adapt to different cases in a simple way. Studies has been done rearing animals in a critical period for the development of cortical selectivity and the agreement of the data with the theory has been proved. The whole theory, which is valid for cortical neurons, might be improved with more computing power which could get rid of some approximation.

Keywords: BCM, neuron, selectivity, synapses, plasticity.

Sommario

Nel 1982 Elie Bienestock, Leon Cooper e Paul Munro scrissero "Theory for the development of neuron selectivity" proponendo uno schema di evoluzione sinaptica nel quale competono i modelli in arrivo piuttosto che la convergenza degli afferenti. Brevemente divenne nota come teoria BCM e fu il trampolino di lancio per ulteriori lavori sulla modificazione delle sinapsi corticali. Durante le ultime due decadi sono state fatte nuove formulazioni, come la IBCM di Nathan Intrator and Leon Cooper del 1992, e nuovi metodi sono stati introdotti permettendo la creazione di reti neurali più complicate ed efficienti. Lo studio di questi modelli evidenzia la capacità di adattamento a diverse situazioni in un modo semplice. Sono stati fatti studi allevando animali in un periodo critico per lo sviluppo sella selettività corticale e l'accordo tra i dati e la teoria è stato provato. Tutta la teoria, valida per i neuroni corticali, potrebbe essere migliorata con una maggiore capacià di calcolo che permetterebbe di sbarazzarsi di alcune approssimazioni.

Parole chiave: BCM, neurone, selettività, sinapsi, plasticità.

Acknowledegemets

Prima facie, I would like to express my deepest appreciation to my supervisor Professor Gastone Castellani. Without his invaluable contribution, this paper would have never been accomplished. I would like to thank you very much for your guidance and patience over these months.

I would like to show gratitude to all the Department of Physics and Astronomy for helping me grow. It was such a pleasure studying in a place which is steeped in history and at the same time so inctellectually lively.

Getting through to course of study required more than accademic support, and I have to thank may people for listening to and having to tolerate me over the past years. I cannot being to express my gratitude and appreciation for their friendship. Edoardo Corallo, Alessandro Corona, Marco Ventrella and Beatrice Zanin have been essential during the time spent at University. I also place on record, my sense of gratitue to one and all, who directly or indirectly, have lent their hand in this venture.

I am deeply indebted to Doctor Marco Faustini Fustini, Professor Diego Mazzatenta, Doctor Matteo Zoli. This work would never be done without their help.

Most importantly, none of this could have happened without my family. My grandmother, who offered her encouragement through phone calls and letters. I also thank my parents for the unceasing encouragement, support and attention even during bad times. I am also grateful to every member of my family who played a decisive role during my growth.

Contents

Chapter 1

Introduction

"It has been known for some time that sensory neurons at practically all levels display various forms of stimulus selectivity"[1] which may be regarded as a general property and we might conjecture that the development of such selectivity obeys some general rule (e.g. some of the mechanism by which selectivity develops in embryonic or early postnatal life[1] are sufficiently general to allow a unifying theoretical treatment). The BCM theory of cortical plasticity has been introduced by Bienenstock, Cooper and Munro (BCM)[1] to account for the changes observed in cell response of visual cortex due to changes in visual environment. The theory simplifies the description of the dynamics by choosing as variables the pre- and postsynaptic firing frequencies (i.e. moving time averages of the actual instantaneous variables, where the length of the averaging interval is of the order of magnitude of the membrane time constant, τ) The formal neuron is a device that performs spatial integration (it integrates the signal impinging all over the soma and dendrites) rather than spatiotemporal integration: the output at time t is a function of the input and synaptic efficacies at t, indipendent of the past history.

1.1 Notation

- *Synaptic efficacy* m_j : characterizes the *net effect* of the presynaptic neuton j on the postsynaptic neuron (this effect may be mediated through a complex system). The resulting "ideal synapse"[2] thus may be of either sign, depending on whether the net effect is excitatory or inhibitory; it may also change sign during development)

[1] The experience plays a determining role in the development of selectivity, the precise role is a matter of controversy

- *Integrative power* of the neuron is assumed to be a linear function, that is:

$$c(t) = \sum_j m_j(t) d_j(t)$$

where c(t) denotes the output at time t, $m_j(t)$ is the efficacy of the jth synapse at time t, d_j is the jth compontent of the input at time t (i.e. the firing frequency of the j presynaptic neuron). We can write:

$$m(t) = (m_1(t), m_2(t), ..., m_N(t))$$
$$d(t) = (d_1(t), d_2(t), ..., d_N(t))$$
$$c(t) = m(t) \cdot d(t) \tag{1.1}$$

m(t) and d(t) are real-valued vectors, of the same dimension, N (i.e. the number of ideal synapses onto the neuron). m(t) (i.e.the array of synaptic efficacies at tima t) is called the *state* of the neuron at time t.

- *Selectivity.* It is common usage to estimate the orientation selectivity of a single visual cortical neuron by measuring the half-width at half-height of its orientation tuning curve. The selectivity is measured with respect to a parameter of the stimulation, namely the orientation, wich takes on values over an interval of 180°.

$$Sel_d(N) = 1 - \frac{\text{mean response of } N \text{ with respect to } \mathbf{d}}{\text{maximum response of } N \text{ with respect to } \mathbf{d}} \tag{1.2}$$

The selectivity is estimate *with respect to* or *in* an *environment for the neuron*, that is, a random variable $\mathbf{d}$ that takes on values in the space of inputs to the neuron N. $\mathbf{d}$ represents a random input to the neuron: it is characterized by its probability distribution that may be discrete or continous. This distribution defines an environment, mathematically a random variable $\mathbf{d}$ [2]. Selectivity is estimated (before and after development) with respect to this same envionment. Applied to the formal neuron in state m:

$$Sel_d(m) = 1 - \frac{E[m \cdot \mathbf{d}]}{\text{ess sup}(m \cdot \mathbf{d})}$$

E[...] stands for "expected value of ..." (i.e. the mean value with respect to the distribution of $\mathbf{d}$); ess sup(...) stands for "essential supremum of ...", equivalent to "maximun of..." in most applications. $Sel_d(N)$ always falls between 0 and 1 and the higher the selectivity of N in $\mathbf{d}$ the closer $Sel_d(N)$ is to 1.

[2]The concept that is needed in order to represent the environment, $\mathbf{d}$, during the development period is that of a *stationary stochastic process*, $\mathbf{d}$(t), that is a time-dependent random variable whose distribution is invariant in time

1.2 Modification of Cortical Synapses

The various factors that influence synaptic modification may be divided into two classes:

1. Factors depending on *Global information* in the form of chemical or electrical signaling which presumably influences in the same way most modifiable junctions of a given type in a given area;

2. Factors depending on *Local information* which is available at each modifiable synapse and can influence each junction in a different manner.

An early proposal as to how local information could affect synaptic modification was made by Hebb[3]:

> *When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.*

Thus, the increase of the synaptic strength connecting A to B depends upon the correlated firing of A and B. In order to be used one must state conditions for synaptic decrease (to avoid synaptic saturation with no information stored and no selectivity developed). To do so is required a complementary princilpe as the one proposed by Stent[4]:

> *When the presynaptic axon of a cell A repeatedly and persistently fails to excite the postsynaptic cell B while cell B is firing under the influence of other presynaptic axons, metabolic changes take place in one or both cells such that A's efficiency, as one of the cells firing B, is decreased.*

Thus, the increase of the strength of certain synapses onto neuron B is accompanied by simultaneous decrease of the strength of other synapses onto the same neuron. There thus occurs a *spatial competition between convergent afferents*. Rather than that, the BCM theory proposes a mechanism of synaptic modification that results in a *temporal competition between input patterns* whether synaptic strength increases or decreases depends upon the magnitude of the postsynaptic response as compared with a variable threshold; the change of the jth synapse's strength at the time t obeys the rule:

$$\dot{\mathbf{m}}_j(t) = \phi(\mathbf{c}(t))\mathbf{d}_j(t) - \epsilon\mathbf{m}_j(t) \tag{1.3}$$

where $\phi(\mathbf{c})$ is a scalar function of the postsynaptic activity, $\mathbf{c}$(t), that changes sign at a value, θ_M, of the output called the *modification threshold*:

$$\phi(\mathbf{c}) < 0 \text{ for } \mathbf{c} < \theta_M;\ \phi(\mathbf{c}) > 0 \text{ for } \mathbf{c} > \theta_M$$

The term, $-\epsilon \mathbf{m}_j(t)$, produces a uniform decay of all junctions[3]. Furthermore, $\mathbf{m}$ is driven in the direction of the input $\mathbf{d}$ if the output is large (above θ_M) or opposite to the direction of the input if the output is small. When $\mathbf{d}_j > 0$ and $\mathbf{c}$ is large enough, $\mathbf{m}_j$ increases (as required by Hebb's principle); but if $\mathbf{c}$ is *not large enough*, $\mathbf{m}_j$ decreases. We may regard this as a form of *temporal competitions between incoming patterns.*The idea of this modification scheme was introduced by Cooper et al.[5] using a constant threshold where the response could slip below θ_M and decrease to zero, leading to stable states with a maximal response to more than one pattern. In a threshold modification scheme, namely $\theta_M(t)$, the change of the jth synapse's strength is written as a product of $\mathbf{d}_j(t)$, the presynaptic activity, and $\phi(\mathbf{c}(t),\ \bar{\mathbf{c}}(t))$, a function of the postsynaptic variables, the output $\mathbf{c}$(t) and its average $\bar{\mathbf{c}}(t)$ [4]. Neglecting the uniform decay term ($\epsilon = 0$), together with equation (1.1) yields:

$$\dot{\mathbf{m}}_j(t) = \phi(\mathbf{m}(t) \cdot \mathbf{d}(t), \mathbf{m}(t) \cdot \bar{\mathbf{d}})\mathbf{d}_j(t) \tag{1.4}$$

A candidate for θ_M (i.e. the value of c at which $\phi(c,\ \bar{c})$ changes sign) is $\bar{\mathbf{c}}(t)$ where the time average is meant to be taken over a period T preceding t much longer than τ, the membrane time constant, so that $\bar{\mathbf{c}}(t)$ evolves on a much slower time scale than $\mathbf{c}$(t). This can be approximated [5] by averaging over the distribution of inputs for a given state $\mathbf{m}$(t):

$$\bar{\mathbf{c}}(t) = \mathbf{m}(t) \cdot \bar{\mathbf{d}}$$

This results in the *instability of low selectivity points* and if the state is *bounded from the origin and from infinity* then the *stable equilibrium points are of high selectivity.* These conditions are fulfilled by a single function $\phi(c,\ \bar{c})$ if we define $\theta_M(t)$ to be a *nonlinear function of c(t).* The rquirement on $\phi(c,\ \bar{c})$ thus reads:

$$\begin{cases} \text{sign } \phi(c,\ \bar{c}) \ =\ \text{sign } (\ c - (\frac{\bar{c}}{c_0})^p \bar{c}) \ \text{ for } \ c > 0 \\ \phi(0,\ \bar{c}) \ =\ 0 \ \text{ for all } \ \bar{c} \end{cases} \tag{1.5}$$

$$\theta_M \ =\ (\frac{\bar{c}}{c_0})^p \bar{c}$$

[3]In most cases it does not affect the behavior of the system if ϵ is small enough

[4]The use of $\bar{\mathbf{c}}(t)$ is a new and essential feature introduced by Bienenstock et al. [1]. It is necessary in order to allow both boundedness of the state and efficient threshold modification.

[5]Replacing the time average by an average over the distribution of $\mathbf{d}$ is allowed provided that (1) the process $\mathbf{d}$(t) is stationary, (2) the interval, T, of time integration is short with respect to the process of synaptic evolution (i.e. $\mathbf{m}$(t) changes very little during an interval of length T), (3) T is long compared to the mixing rate of the process $\mathbf{d}$ (i.e. during a period of length T, the relative time spent by the process $\mathbf{d}$(t) at any point d in the input space is nearly proportional to the weight of the distribution of $\mathbf{d}$ at d). Synaptic modification of the type involved in changes of selectivity is a slow process (order of minutes or hours) to be significant, whereas elementary sensory patterns are faster (order of 1 minute or less).

where c_0 and p are two fixed positive constants. The sign of $\phi(c, \bar{c})$ for $c < 0$ is not crucial since c is a positive quantity [6]. The threshold above serves two purposes: allowing its own modification when $\bar{c} \simeq c_0$ as well as driving the state from region such that $\bar{c} \ll c_0$ or $\bar{c} \gg c_0$.

1.2.1 Synaptic growth

The process of synaptic growth, starting near zero to, eventually, end in a stable selective state, may be described as follows. Initially, $\bar{c} \ll c_0$; hence, $\phi(c, \bar{c}) > 0$ for all inputs in the environment: the responses to all inputs grow. With this growth, $\bar{c}$ iincreases, thus increasing θ_M. Now some inputs result in postsynaptic responses that exceed θ_M, while others, those whose direction is far away from (close to orthogonal to) the favored inputs, give a response less than θ_M. The response to the former continues to grow, while the response to the latter decays. This results in a form of *competitions between incoming patterns* rather than competition between synapses. The response to unfavored patterns decays until it reaches zero, where it stabilizes for $\phi(0, \bar{c}) = 0$ for any $\bar{c}$ (1.5). The response to favored patterns grows until the mean response $\bar{c}$ is high enough, and the state stabilizes.

[6] For the sake of mathematical completeness, one may define $\phi(c, \bar{c}) > 0$ for $c < 0$

1.3 Mathematical results

The behavior of equation (1.4) depends on the environment,that is, on the distribution of the stationary stochastic process, **d**. *Discrete distributions* include K possible inputs $d^1, \ldots, d^K$. These will generally be assumed to occur with the same probability *1/K*. The process **d** is therefore a jump process which randomly assumes new values at each time increment; the vector **m** is, roughly, a Markov process [7]. Results obtained only for certain discrete distributions are of two types:

1. Equilibrium point are locally stable if and only if they are of the highest available selectivity with respect to the given distribution of **d**;

2. Given any initial value of **m** in the state space, the probability that **m**(t) converges to one of the maximum selectivity fixed points as t goes to infinity is 1.

If **d** takes on K values, then:

> *Lemma 1. Let* $d^1, \ldots, d^K$ *be linearly indipendent and* $\boldsymbol{d}$ *satisfy* $P[\boldsymbol{d} = d^1] = \ldots = P[\boldsymbol{d} = d^K] = 1/K$. *Then, for any function* ϕ *satisfying equation (1.5), equation (1.4) admits exactly* 2^K *fixed points with selectivities* $0, 1/K, 2/K, , \ldots, (K-1)/K$. *There are K fixed points* $m^1, \ldots, m^K$ *of selectivity (K - 1)/K.*

(K - 1)/K is the maximum possible selectivity with respect to **d** which means a positive response for one and only one of the inputs. The following theorem holds:

> *Theorem 1. Assume, in addition to the conditions of Lemma 1, that* $d^1, \ldots, d^K$ *are mutually orthogonal or close to orthogonal. Then the K fixed points of maximum selectivity are stable, and whatever its initial valus, the state of the system converges almost surely to one of them.*

The proof of Theorem 1 is based on the existence of *trap regions* around each of the selective fixed points.

> *Theorem 2. Under the same conditions as in theorem 1, there exists around* m^i, *stable point, a region* F^i, *the trap region, such that, once the state enters in, it converges almost surely to* m^i

The meaning of theorem 2 is the following: once **m**(t) has reached a certain selectivity, it cannot "switch" to another selective region[8].

[7] A stochastic process that satisfies Markov property (memorylessness): the conditional probability distribution of future states of the process (conditional on both past and present states) depends only upon the present state, not on the history of the process.

[8] Applied to cortical cells in a patterned visual environment, this means that, once they become sufficiently committed to certain orientations, they will remain committed to those orientation (provided that the visual environment does not change), becoming more selective as they stabilize to some maximal selectivity

Simulation suggestion

- For a fairly broad range of environment if $d^i \cdot d^j \geq 0$, even if d^1, ..., d^K are far from being mutually orthogonal, the K fixed points of maximum selectivity are stable;
- Even if the d^1, ..., d^K are not linearly indipendent and are far from being mutually orthogonal, the asymptotic selectivity is close to its maximum value with respect to **d**.

1.3.1 Example (K = 2)

In this simple case **d** takes on values on two possible input vectros, d^1 and d^2, that occur with the same probability $P[\mathbf{d} = d^1] = P[\mathbf{d} = d^2] = 1/2$. Whatever the actual dimension N of the system, it reduces to two dimensions. (Any component of **m** outside of the linear subspace spanned by d^1 and d^2 will eventually decay to zero due to the uniform decay term.) By definition it follows that the maximum value of $Sel_d(m)$ in the state space is 1/2. It is reached for states m which give a null response when d^1 comes in (i.e. are orthogonal to d^1) but a positive response for d^2, or vice versa. Minimum selectivity, namely zero, is obtained for states m such that $m \cdot d^1 = m \cdot d^2$. Then for any value of ϕ satisfying equation (1.5), equation (1.4) admits exactly $2^K = 4$ fixed points (Lemma 1), m^0, m^1, m^2, $m^{1,2}$ with $Sel_d(m^0) = Sel_d(m^{1,2}) = 0$ and$Sel_d(m^1) = Sel_d(m^2) = 1/2$ (the superscript indicate which of the d^i are not orthogonal to m. The behavior depends on the geometry of the inputs, in the present case, on $cos(d^1, d^2)$. If $cos(d^1, d^2) \geq 0$[9], then m^0 and $m^{1,2}$ are unstable, m^1 and m^2 are stable, and wehatever its initial value, the state of the system converges almost surely either to m^1 or to m^2. This characterizes evolution schemes based on *competition between patterns* and states that the state eventually reaches maximal selectivity even when the two input vectors are very close to one another. It requires that some of the synaptic strengths be negative since the neuron has linear integrative power.

[9]When $cos(d^1, d^2) < 0$ the situation is much more complicated: trap regions do not necessarily exist and periodic asymptotic behavior may occur, bifurcating from the stable fixed points when $cos(d^1, d^2)$ becomes too negative[6]

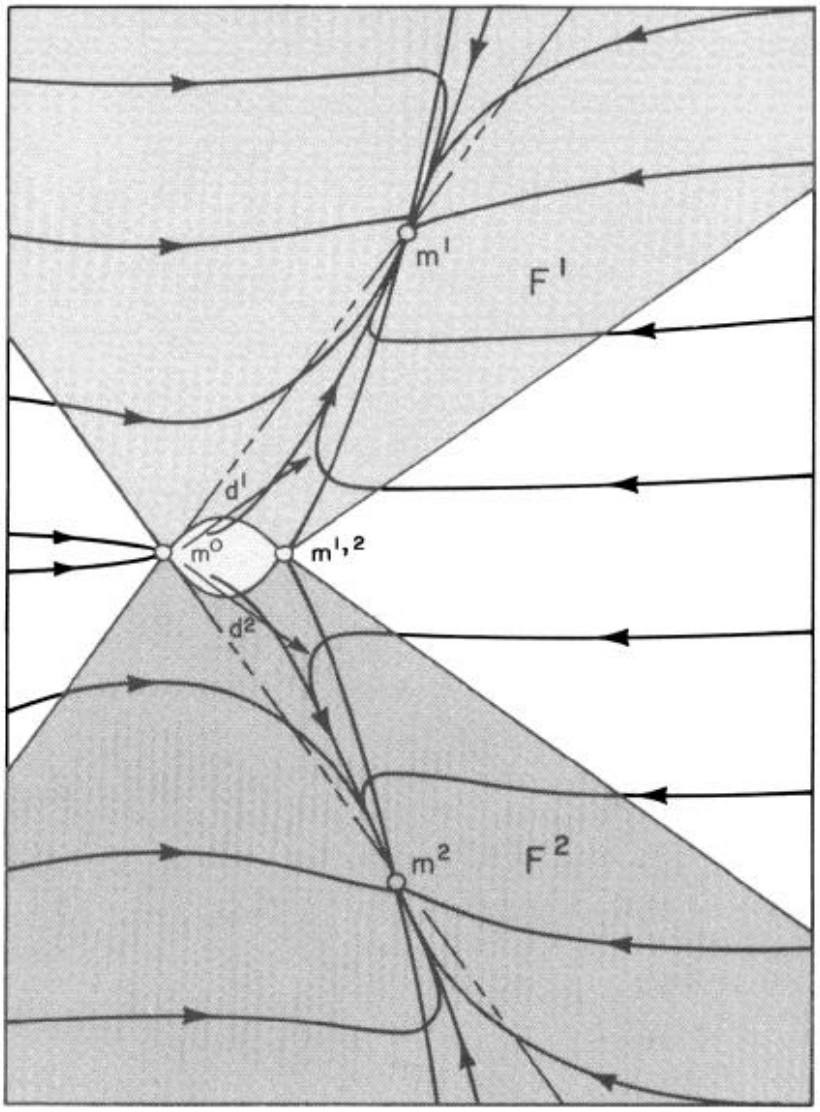

Figure 1.1 – The phase portrait of equation (1.4) in an environment consisting of two inputs. The diagram shows the trajectories of the state of the system, starting from different initial points. This is a simulation with a ϕ that satisfies condition (1.5), using a differen function may slightly change the shape of the trajectories without any essential change in the behavior. The system is a *stochastic* one, which means that the trajectories depend on the precise sequence of inputs. As long as the state is in the *unshaded region*, it is not yet known whether it will eventually be attracted to m^1 or m^2. This is determined as the state enters one of the trap (*shaded*) regions F^1 or F^2. The trajectories shown here are deterministic ones, obtained by alternating $\mathbf{d}$ regularly between d^1 and d^2; in fact they are averaged ones and displays a much more regular and smooth behavior than the actual stochastic ones.

Chapter 2

Models of synaptic plasticity

2.1 Mathematical forms of BCM

A theoretical solution to the problem of visual cortical plasticity, was presented by Cooper, Liberman and Oja[7] (1979). According to this theory, the synaptic efficacy of active inputs increases when the postsynaptic target is cuncurrently depolarized beyond a *modification threshold* (θ_M). However, when the level of postsynaptic activity falls below θ_M, then the strength of active synapses decreases. An important feature was added to this theory in 1982 by Bienenstock, Cooper and Munro (BCM). They proposed that the value of the modification threshold is not fixed, but instead varies as a nonlinear function of the average output of postsynaprtic neuron[1]. The BCM theory exposed so far could be summarized as follows:

1. The change in synaptic weights ($dm_j(t)/dt = \dot{m}_j(t)$) is proportional to presynaptic activity (d_j).
2. The change in synaptic weights is proportional to a non-monotonic function (ϕ) of the postsynaptic activity (c):
   - for low c, the synaptic weight decreases ($dm_j(t)/dt < 0$),
   - for larger c, it increases ($dm_j(t)/dt > 0$),

   The cross over point between those is called the modification threshold (θ_M).
3. The modification threshold is itself a super-linear function of the history of postsynaptic activity.

There are many mathematical forms which satisfy those conditions. Traditionally the approach was to use the simplest form that is still consistent with experiments.

[1]This provided stability properties and explains why the low level of postsynaptic activity during binocular deprivation does not drive the strengths of all cortical synapses to zero

2.1.1 Bienenstock et al. 1982

In the original form, the neuron is assumed to be linear, a uniform weight decay ($-\epsilon m_j$) is present, and the modification threshold is calculated as the power of the mean of the neuron output (i.e. scaled by a constant c_0).

$$\begin{aligned} c(t) &= m(t) \cdot d(t) = \sum_j m_j(t) d_j(t) \\ \dot{\mathbf{m}}_j(t) &= \phi(\mathbf{c}(t)) \mathbf{d}_j(t) - \epsilon \mathbf{m}_j(t) \\ \phi(\mathbf{c}(t)) &= c(c - \theta_M) \\ \theta_M &= E^p[(c/c_0)] \end{aligned}$$

2.1.2 Intrator and Cooper, 1992

Intrator and Cooper[8] presented an objective function formulation for the theory, also called IBCM rule, which indicates what the neuronal goal is and enables simple analysis of the dynamics. This formulation allows to interpret the biological neuron's behavoir from a statistical point of view.

$$\begin{aligned} c(t) &= \sigma(\sum_j m_j(t) d_j(t)) \\ \dot{\mathbf{m}}_j(t) &= \phi(\mathbf{c}(t)) \mathbf{d}_j(t) \sigma'(c) \\ \phi(\mathbf{c}(t)) &= c(c - \theta_M) \\ \theta_M &= E[c^2] \end{aligned}$$

This BCM form can be derived by minimizing the loss function (i.e. objectiove function):

$$R = -\frac{1}{3} E[c^3] + \frac{1}{4} E^2[c^2] \tag{2.1}$$

where E[...] denotes the expectation value with respect to the input environment. The function itself measures the sparseness or bi-modality of the output distribution. It is bounded from below, and it thus has minima which can be obtained by gradient descent $\dot{\mathbf{m}}_j = -\nabla R$. This leads to an approximate solution via the stochastic differential equation $\dot{\mathbf{m}}_j(t) = \phi(\mathbf{c}(t), \theta_M) \mathbf{d}_j(t)$. In order to have stable fixed points, the average used for the modification threshold is calculated with the *square* of the output.

Features Extraction

When a classification of high dimensional vectors is sought, the *curse of dimensionality*[2][9] becomes the main factor affecting the classification performance. In those cases in which important structure in the data actually lies in a much smaller dimensional space, it becomes reasonable to try to reduce the dimensionality before attempting the classification[3]. Unsupervised method use local objective function which may lead to less sensitivity to the number of parameters in the estimation, and therefore have the potential to avoid the curse of dimensionality. A general class of unsupervised dimensionality reduction methods, called *exploratory projection pursuit*, is based on seeking *interesting projections* of high dimensional data points. The notion of interesting projections is motivated by an observation made by Diaconis and Freedman[10], that for most high dimensional clouds, most low dimensional projections are approximately normal. This finding suggests that the important information in the data is conveyed in those directions whose single dimension projected distribution is far from Gaussian. Various projection indices differ on the asumptions about the nature of deviation from normality, and in their computational efficiency. Friedman [11] argues that the most computationally efficient measures are based on polynomial moments, but projection indices based on that are not directly applicable, since they very heavily emphasize departure from normality in the tails of the distribution[12]. The IBCM theory address the problem by applying a sigmoidal (σ) function to the projections, and then applying an objective function based on polynomial moments. The IBCM rule has some nice mathematical properties:

- It is an exploratory projection index that emphasizes deviation from a Gaussian distribution at the center of the distribution, in the form of multi-modality.
- The formulation naturally extends to a lateral inhibition network (with a non-linear saturation transfer function), which can find several projections at once.
- The number of calculations of the gradient grows linearly with the number of projections sought, thus it is very efficient in high dimensional feature extraction.
- The search is constrained by seeking projections that are orthogonal to all but one of the clusters (in the original space). Thus, there are at most K optimal projections and not $K(K-1)/2$ separating hyper-planes as in discriminant analysis methods. This property is very important as it suggests why the "curse of dimensionality" is less problematic with this learning rule (every minima is an optimal one).

[2]The curse of dimensionality is due to the inherent sparsity of high dimensional spaces. The amount of training data needed to get reasonably low variance estimators becomes very high.

[3]This approach can be successful if the dimensionality reduction/feature extraction method loses as little information as possible in the transformation from the high dimensional space to the low dimensional one.

- Most importantly, the neuronal output (or the projection) of an input x (or a cluster of inputs) is proportional to $1/P(x)$, where P(x) is the a-priori probability of the input x. This property, which directly results from the analysis is essential for creating coincidence detectors, and it also indicates the optimality of the learning rule in terms of energy (or code) conservation. If a biologically plausible log saturation transfer function is used as the neuronal non-linearity, it follows that the amplitude or code length associated with the input x is proportional to $-log(P(x))$, which is optimal from information theoretic considerations.

2.1.3 Law and Cooper, 1994

The Law and Cooper form[13] has all of the same fixed points as the Intrator and Cooper form, but the speed of synaptic modification increases when the threshold is small, and decreases as θ_M increases.

$$c(t) = \sigma(\sum_j m_j(t)d_j(t))$$
$$\dot{\mathbf{m}}_j(t) = \phi(\mathbf{c}(t))\mathbf{d}_j(t)/\theta_M$$
$$\phi(\mathbf{c}(t)) = c(c-\theta_M)$$
$$\theta_M = E[c^2]$$

The practical result is that the simulation can be run with artificially high learning rates, and wild oscillations are reduced. This form has been used primarly when running simulations of networks, where the run-time of the simulation can be prohibitive[14].

2.2 Selectivity and Tuning curve

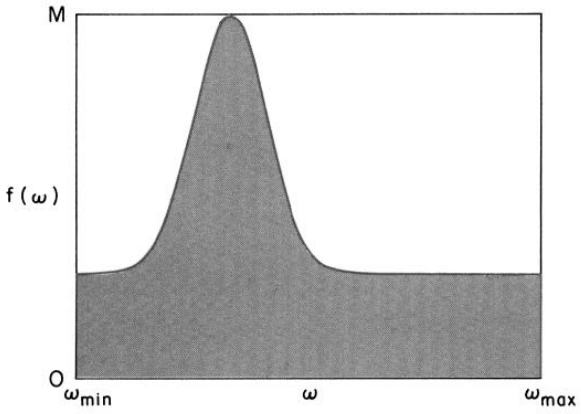

Figure 2.1 – An example of tuning curve: the abscissa displays a parameter of the stimulus ω and the ordinate the neuron's response.The figure is taken from the original BCM paper[1]

The index of selectivity as defined in equation (1.5) becomes

$$Sel_d(m) = 1 - \frac{E[m \cdot \mathbf{d}]}{\text{ess sup}(m \cdot \mathbf{d})}$$

In the figure 2.1 there is an example of tuning curve from which is possible to compute the selectivity with respect to an environment uniformly distributed between ω_{min} and ω_{max}[4]. The neuron's response 0 is the level of the average spontaneous activity; M is the maximum response. It is a simple measure of the breadth of the peak: curves of same selectivity have approximately the same half-width at half-height. Typical values for orientation selectivity of adult cortical cells vary between 0.7 and 0.85 ("specific" cells). Selectivity of broadly tuned but still unimodal cells lies between 0.5 and 0.7 (Buisseret and Imbret[15] and Frégnac and Imbert[16]) Obviously, 0 is the selectivity of an absolutely flat curve, whereas 1 is the selectivity of a Dirac δ function. The selectivity of the neuron then is given:

$$Sel_d(N) \;=\; 1 \;-\; \frac{1}{M(\omega_{max} - \omega_{min})} \int_{\omega_{min}}^{\omega_{max}} f(\omega)\, d\omega \;=\; \frac{\text{light area}}{\text{total box area}}$$

Applying what has been discussed to a concrete example, we obtain orientation selectivity and binocular interaction in the primary visual cortex. Consider first a classical test environment used to construct the tuning curve of cortical neurons: It consists of an elongated light bar successively presented or moved in all orientations (preferably in a random sequence) in the neuron's receptive field. Thus, all of the parameters of the stimulus are constants except the orientation, which is distributed unifomly on a circularly symmetric closed path[5]. The tipical theoretical environment that will be used for constructing the formal neuron's tuning curve is a random variable $\mathbf{d}$ unifomly distributed on a circularly symmetric closed one-parameter family of points in the space R^N.

[4] If the parameter of the stimulus is the orientation then $\omega_{max} - \omega_{min} = 180$.

[5] The assumption is that the retinocortical pathway maps this family of stimuli to the cortical neuron's space of inputs in such a way as to preserve the circular symmetry.

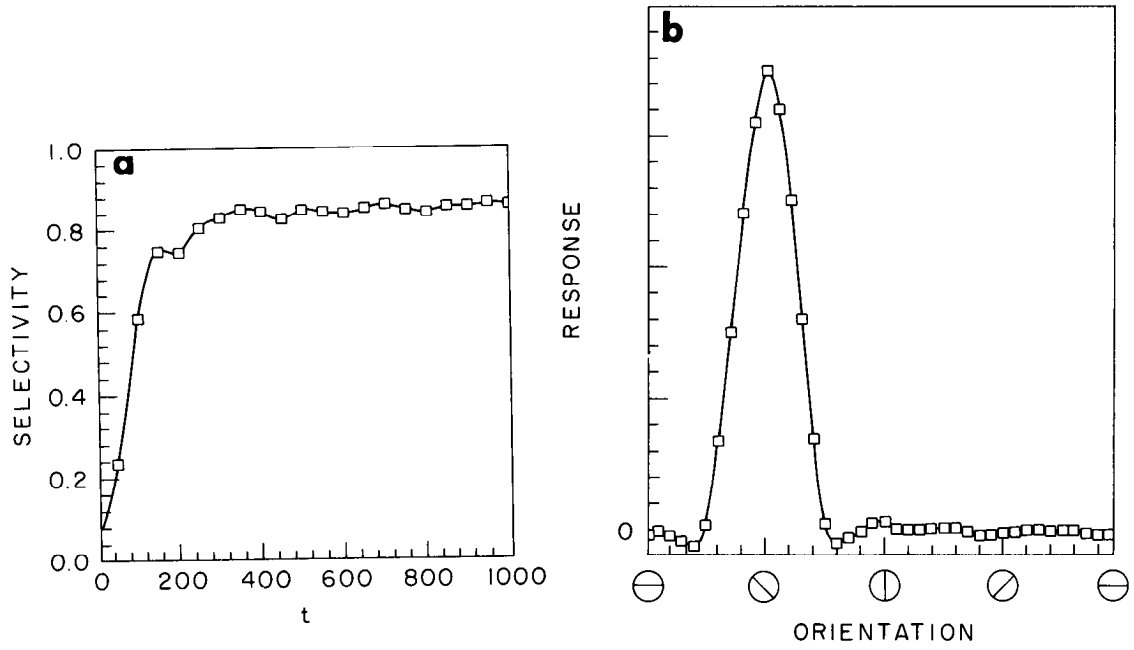

Figure 2.2 – The evolution of a synaptic system in a circular environment: here K = 40 and N = 37 so that the vectors are linearly dependent.The value of the maximum selectivity with respect to **d** is therefore not precisely calculable. The asymptotic selectivity is approximately 0.9.The figure is taken from the original BCM paper[1]
Fig.a demonstrates the progressive buildup of the selectivity;
Fig.b shows the resulting tuning curve.

The parameter coding orientation in the receptive field is, in principle, continous. However, for numerical simulation's sake, the distribution is made discrete so that **d** takes on values on the points $d^1, ..., d^K$[6]. To specify the stationary stochastic process that represents the time sequence of inputs to the neuron it is possible giving to it exactly the same distribution as the circular **d**. In this case the assumption is that the development of orientation selectivity is to a large extent independent of other parameters of the stimulus. The elementary stimulus for a cortical neuron is a rectilinear contrast edge or bar. Any additional pattern present at the same time in the receptive field is regarded as random noise.

[6]The requirement of circular symmetry is expressed mathematically as follows: the matrix of inner products of the vector $d^1, ..., d^K$ is circular (i.e. each row is obtained from its nearest upper neighbor by shifting it one column to the right) and the rows of the matrix are unimodal. A random vaiable, **d**, uniformly distributed on such a set of points will be, hereafter, called a *circular environment*. Such a **d** may be roughly characterized by three parameters: N, K and the d^i vectors.

Chapter 3

Applications

Statistically a neuron is considered as capable of "deciding" whether to fire or not for a given input and vector of synaptic weights. A loss function is attached to each decision and the neuron's task is to choose the decision that minimize that loss. It is natural, then, to seek a synaptic weight vector that will minimize the sum of the losses associated with every input which is the average loss (also called risk). The search for such a vector, which yields an optimal synaptic weight vector, can be viewed as learning or parameter estimation. In those cases where the risk is a smooth function, its minimization can be accomplished by gradient descent. Different loss functions will yield different learning procedures.

3.1 Linear Neuron

Considering a linear neuron with n-stimuli and some useful functions:

- input vector $\mathbf{d} = (d_1, ..., d_N)$,
- synaptic weight vector $\mathbf{m} = (m_1, ..., m_N)$,
- synaptic activity c $= \mathbf{m} \cdot \mathbf{d}$,
- learning rate μ,
- threshold $\theta_M = E[(\mathbf{m} \cdot \mathbf{d})^2]$,
- $\phi(\mathbf{c}, \theta_M) = c(c - \theta_M)$,
- $\hat{\phi}(\mathbf{c}, \theta_M) = c(c - \frac{1}{2}\theta_M)$.

Vectors are in R^N and time dependency is allowed only in the presentation of the training patterns; by requiring that $\mathbf{d}$ is of Type II mixing[1]. These assumption are plausible, since they represent the closest continous approximation to the usual training algorithms, in which training patterns are presented at random. The *learning rate* has to decay in time so that the approximation is valid. The projection index (loss function) is aimed at finding directions for which the projected distribution is far from Gaussian; since high dimensional clusters have a multimodal projected distribution, the aim is to find the function/index that emphasizes that multimodality[2]. The index should exhibit the fact that bimodal distribution is already interesting, and any additional mode should make the distribution even more interesting.

3.1.1 Loss function

Consider the following family of loss functions that depends on the synaptic weight and on the input:

$$\begin{aligned} L_m(\mathbf{d}) &= -\mu \int_0^{\mathbf{m}\cdot\mathbf{d}} \hat{\phi}(\mathbf{s}, \theta_M)\, ds \\ &= -\mu\{\frac{1}{3}(\mathbf{m}\cdot\mathbf{d})^3 - \frac{1}{4}E[(\mathbf{m}\cdot\mathbf{d})^2](\mathbf{m}\cdot\mathbf{d})^2\} \end{aligned} \tag{3.1}$$

For any fixed $\mathbf{m}$ and θ_M, the loss is small for a given input $\mathbf{d}$, when either $c = \mathbf{m}\cdot\mathbf{d}$ is close to zero, or when it is larger than θ_M. Moreover, the loss remains negative for $\mathbf{m}\cdot\mathbf{d} > \theta_M$, therefore any kind of distribution at the right hand side of the threshold is possible, and the preferred ones are those which are concentrated further from θ_M. It is not possible that a minimizer of the average loss will be such that all the mass of the distribution will be concentrated to one side of θ_M because the threshold is dynamic and depends on the projections in a nonlinear way. This implies that θ_M will always move itself to a position such that the distribution will never be concentrated at only one of its sides.

Risk

The risk, which is the expected value of the loss, is given by:

$$\begin{aligned} R_m = E[L_m(\mathbf{d})] &= -\mu E\{\frac{1}{3}(\mathbf{m}\cdot\mathbf{d})^3 - \frac{1}{4}E[(\mathbf{m}\cdot\mathbf{d})^2](\mathbf{m}\cdot\mathbf{d})^2\} \\ &= -\mu\{\frac{1}{3}E[(\mathbf{m}\cdot\mathbf{d})^3] - \frac{1}{4}E^2[(\mathbf{m}\cdot\mathbf{d})^2]\} \end{aligned} \tag{3.2}$$

[1]The mixing property specifies the dependency of the future of the process on its past.

[2]For computational efficiency it is possible to base projection index on plynomial moments of low degree (more than 2).

Since the risk is continuously differentiable, its minimization can be achived via a gradient descent method ($\dot{\mathbf{m}}(t) = -\nabla R_m$) with respect to $\mathbf{m}$, namely:

$$\begin{aligned} \frac{dm_i}{dt} = -\frac{\partial}{\partial m_i} R_m &= \mu\{E[(\mathbf{m}\cdot\mathbf{d})^2 x_i] - E[(\mathbf{m}\cdot\mathbf{d})^2]E[(\mathbf{m}\cdot\mathbf{d})x_i]\} \\ &= \mu E[\phi(\mathbf{m}\cdot\mathbf{d}, \theta_M)x_i] \end{aligned} \tag{3.3}$$

The resulting differential equations give somewhat different version of the law governing synaptic weight modification of the BCM theory:

$$\dot{\mathbf{m}}_j(t) = \phi(\mathbf{c}(t))\mathbf{d}_j(t) - \epsilon\mathbf{m}_j(t)$$

The difference lies in the way the threshold is determined. In the original form it was $E^p[c]$ for $p > 1$, while in the current form is $\theta_M = E[c^p]$ for $p > 1$.The latter takes into account the variance of the activity (for p = 2) and therefore is always positive, this ensures stability even when the average of the inputs is zero[3]. Moreover the original form requires that s history of activity be stored and then via a non-linear process produces the modification threshold. The latter form instead requires that the non-linear process occurs first. The averaged version of the previous equation is, as described in[17]:

$$\dot{\mathbf{m}}(t) = P\,D^T\,\Phi(c, \theta) \tag{3.4}$$

where $\theta = E[c^2] = \sum_{j=1}^{n} p_j(\mathbf{m}\cdot\mathbf{d}_j)^2$ $.\Phi = (\phi_1, ..., \phi_n)$ and p_i, which is the i-th element of the diagonal matrix of the probabilities P, represents the probability of choosing vector d_{ii} from the data set. The matrix of inputs D is composed of different input vectors and its determinant is non-zero since the inputs are linearly indipendent. The dynamics is given by:

$$\begin{bmatrix} \dot{\mathbf{m}}_1 \\ \dot{\mathbf{m}}_2 \\ . \\ . \\ . \\ \dot{\mathbf{m}}_n \end{bmatrix} = \begin{bmatrix} p_1 & 0 & . & . & . & 0 \\ 0 & p_2 & . & . & . & 0 \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ 0 & 0 & . & . & . & p_n \end{bmatrix} \begin{bmatrix} d_{11} & d_{21} & . & . & . & d_{n1} \\ d_{12} & d_{22} & . & . & . & d_{n2} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ d_{1n} & d_{2n} & . & . & . & d_{nn} \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ . \\ . \\ . \\ \phi_n \end{bmatrix} \tag{3.5}$$

[3]The original theory assumed that the inputs were positive, whereas the present relaxes this assumption yieliding stabilty for a lareìger class of bounded inputs.

3.2 Stability Analysis

The analysis of the theory starts finding the stationary states of equation (3.4). Mathematically the condition $\dot{\mathbf{m}}(t) = 0$ implies that $P\,D^T\,\Phi(c,\theta)$ must be zero, and this is possible if and only if $\Phi = 0$, because it is required that the input vectors are linearly indipendent (i.e. $\det D \neq 0$). Then:

$$c_i(c_i - (\sum_{j=1}^{n} p_j c_j^2)) = 0 \quad \text{for } i = 1, ..., n \tag{3.6}$$

Whose solutions are in the following equivalence class:

$$S = \begin{cases} (0,0,...,0,...,0) \\ (0,0,...,0,\frac{1}{p_i},0,...,0) \\ (0,0,...,0,\frac{1}{p_i+p_j},0,...,0,\frac{1}{p_i+p_j},0,...,0) \\ (0,0,...,0,\frac{1}{p_i+p_j+p_k},0,...,0,\frac{1}{p_i+p_j+p_k},0,...,0,\frac{1}{p_i+p_j+p_k},0,...,0) \\ \vdots \\ (1,1,...,1,...,1) \end{cases} \tag{3.7}$$

The corresponding $\mathbf{m}$ solutions are given by $\mathbf{m} = D^{-1}c$.
The next step is to examine the Jacobian matrix :

$$J_n = \begin{bmatrix} d_{11} & d_{21} & \cdots & d_{n1} \\ d_{12} & d_{22} & \cdots & d_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1n} & d_{2n} & \cdots & d_{nn} \end{bmatrix} \begin{bmatrix} \frac{\partial\phi_1}{\partial c_1} & \frac{\partial\phi_1}{\partial c_2} & \cdots & \frac{\partial\phi_1}{\partial c_n} \\ \frac{\partial\phi_2}{\partial c_1} & \frac{\partial\phi_2}{\partial c_2} & \cdots & \frac{\partial\phi_2}{\partial c_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\phi_n}{\partial c_1} & \frac{\partial\phi_n}{\partial c_2} & \cdots & \frac{\partial\phi_n}{\partial c_n} \end{bmatrix} \begin{bmatrix} \frac{\partial c_1}{\partial m_1} & \frac{\partial c_1}{\partial m_2} & \cdots & \frac{\partial c_1}{\partial m_n} \\ \frac{\partial c_2}{\partial m_1} & \frac{\partial c_2}{\partial m_2} & \cdots & \frac{\partial c_2}{\partial m_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial c_n}{\partial m_1} & \frac{\partial c_n}{\partial m_2} & \cdots & \frac{\partial c_n}{\partial m_n} \end{bmatrix} \tag{3.8}$$

Note that the third matrix is the input matrix D since $c = Dm$.

$$J_n = D^T \begin{bmatrix} 2c_1 - \theta - 2p_1c_1^2 & -2p_2c_2c_1 & \cdots & -2p_nc_nc_1 \\ -2p_1c_1c_2 & 2c_2 - \theta - 2p_2c_2^2 & \cdots & -2p_nc_nc_2 \\ \vdots & \vdots & \ddots & \vdots \\ -2p_1c_1c_n & -2p_2c_2c_n & \cdots & 2c_n - \theta - 2p_nc_n^2 \end{bmatrix} D \tag{3.9}$$

From all the above consideration follow that:

- *Stable points* are those with one non-zero coordinate which set all the off-diagonal terms to zero obtaining a diagonal matrix with diagonal elements $(-\frac{1}{p_i})$ for $i = 1, ..., n$. Thus, the Jacobian is negative define[4].
- (0, 0, ..., 0, ..., 0) is *unstable* in a Lyapunov sense (i.e. it is neutrally stable).
- *Instability* of all other points due the Jacobian $D^T(\frac{\partial \Phi}{\partial c})D$ is a quadratic form non-negative define in the case of linearly independent vectors and $\sum_{j=1}^{n} p_j = 1$.

3.3 Non-Linear Neuron

The BCM could be extent to a non-linear neuron due to the fact that the distribution has part of its mass on both sides of the threshold θ_M; this allows a projection index that seeks multi-modalities. However, this projection index will be more general if the loss is insensitive to outliers and if any projected distribution is allowed to be shifted so that the part of the distribution that satisfies $c < \theta_M$ will have its mode at zero. The oversensitivity to outliers is addressed by considering a nonlinear neuron in which the neuron's activity is defined to be $c = \sigma(\mathbf{m} \cdot \mathbf{d})$, where σ represent a smooth sigmoidal function. The ability to shift the projected distribution so that one of its modes is at zero is achieved by introducing a threshold β[5] so that the projection is defined to be $c = \sigma(\mathbf{m} \cdot \mathbf{d} + \beta)$[6] For a nonlinear neuron, $\theta_M = E[\sigma^2(\mathbf{m} \cdot \mathbf{d})]$. The loss function is given by:

$$\begin{aligned} L_m(\mathbf{d}) &= -\mu \int_0^{\sigma(\mathbf{m}\cdot\mathbf{d})} \hat{\phi}(\mathbf{s}, \theta_M)\, ds \\ &= -\mu\{\frac{1}{3}\sigma^3(\mathbf{m} \cdot \mathbf{d}) - \frac{1}{4}E[\sigma^2(\mathbf{m} \cdot \mathbf{d})]\sigma^2(\mathbf{m} \cdot \mathbf{d})\} \end{aligned} \tag{3.10}$$

The gradient of the risk becomes:

$$\begin{aligned} -\nabla_m R_m &= \mu\{E[\sigma^2(\mathbf{m} \cdot \mathbf{d})\sigma'\mathbf{d}] - E[\sigma^2(\mathbf{m} \cdot \mathbf{d})]E[\sigma(\mathbf{m} \cdot \mathbf{d})\sigma'\mathbf{d}\} \\ &= \mu E[\phi(\sigma(\mathbf{m} \cdot \mathbf{d})\mathbf{m} \cdot \mathbf{d}),\, \theta_M)\sigma'\mathbf{d}] \end{aligned} \tag{3.11}$$

[4]The eigenvalues of the matrix $D^T D$ are all positive and real because the matrix is symmetric and positive define and the product of diagonal matrices is commutative.

[5]From the biological point of view β can be considered as spontaneous activity.

[6]The modification equation for finding optimal threshold are obtained observing that it effectively adds one dimension to the input vector and vector of synaptic weight, namely $\mathbf{d} = (d_1, ..., d_n, 1)$; $\mathbf{m} = (m_1, ..., m_n, \beta)$. Therefore β can be found by using the same synaptic modification equation. Hereafter β will be absorbed in the ordinary form.

Where σ' represents the derivative of the σ at the point $(\mathbf{m} \cdot \mathbf{d})$[7]. The analogoue of equation (3.4) is:

$$\dot{\mathbf{m}}(t) \;=\; \Sigma\, P\, D^T\, \Phi(c, \theta) \tag{3.12}$$

Where, as before, D is the input matrix, Φ is the vector of ϕ calculated at the points $(\sigma(\mathbf{m} \cdot \mathbf{d}_1),\, \sigma(\mathbf{m} \cdot \mathbf{d}_2), ..., \sigma(\mathbf{m} \cdot \mathbf{d}_n))$, and Σ is the matrix containing σ':

$$\Sigma \;=\; \begin{bmatrix} \sigma'_1 & 0 & \cdots & 0 \\ 0 & \sigma'_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma'_n \end{bmatrix} \tag{3.13}$$

The matrix Σ is positive definite as σ is smooth and monotonic, thus, the search for stationary states leads to $\Phi = 0$.For convenience it is possible to fidefine the variable ζ such that $\zeta = \sigma(\mathbf{m} \cdot \mathbf{d}) = (\sigma_1, \sigma_2, ..., \sigma_n)$. Thus the fixed points solutions in terms of ζ are equivalent to the solutions of equation (3.6). It follows that the solutions for $\mathbf{m}$ result from solving an equation of the form $\mathbf{m} = D^{-1}\, \sigma^{-1}(\zeta)$; with $\zeta \in S$[8].

[7]The multiplication by σ' reduces sensitivity to outliers of the differential equation since for outliers σ' is close to zero. The gradient descent procedure is valid, provided that the risk is bounded from below.

[8]S is the equivalence class of solutions (3.7)

3.4 Lateral interacting network

When a neuron is in a network, the incoming inputs can arise from the thalamus and another set can arise from other cortical neurons.Then the vector of synaptic weights $\mathbf{m}$ for a single cell becomes a matrix M for all the network neurons. At the same time, the vectori of neuronal activities $\mathbf{c}$ (due to the matrix of inputs D) becomes a matrix[9]. An extension of the single cell BCM neuron to such a network was presented by Scotfiel and Cooper in 1985[18] and a mean field approximation of this network by Cooper and Scofield in 1988[19][10]. For a network with a single input $\mathbf{d}$, the activity of the neuron is affected also by the adjacent neurons in the netwok other than the input itself; namely:

$$\mathbf{c} \;=\; M\mathbf{d} \;+\; L\mathbf{c} \tag{3.14}$$

where L is the cortico-cortical connectivity matrix in which l_{ij} is the interaction between neuron i (the target) and neuron j (the source) and M is the matrix of the feedforward (thalamocortical) synapses; m_{ij} represents the feedforward connections to cell i arising from input channel j. For a consistency condition follows that:

$$\mathbf{c} \;=\; (I \,-\, L)^{-1} M\mathbf{d} \tag{3.15}$$

Which represents the network activity due to a single input vector. For a network of n neurons that recive n input vectors the modification of the weights described by equation (3.4) has almost the same form:

$$\dot{\mathbf{m}}(t) \;=\; \mathcal{P}_n \, \mathcal{D}_n^T \, \Phi \tag{3.16}$$

where, in this case, $\mathcal{D}_n$ and $\mathcal{P}_n$ are the direct product of the input and probability matrices respectively. Φ now is the vector ofthe neuronal activation function:

$$\begin{aligned} \Phi \;&=\; (\phi_{11}, ..., \phi_{ij}, ..., \phi_{nn}) \\ \mathcal{P}_n \;&=\; \bigotimes_{k=1}^{n} P_j \\ \mathcal{D}_n \;&=\; \bigotimes_{k=1}^{n} D_j \end{aligned} \tag{3.17}$$

[9]It should be treated as a supe-vector

[10]The mean field approximation is obtained by repalcing the inhibitory contribution of cell j,c_j in

$$c_i \;=\; m_i \cdot \mathbf{d} \;+\; \sum_j L_{ij} c_j$$

by its average value $\bar{c} \;=\; \frac{1}{N} \sum_i c_i$ so that c_i becomes:

$$c_i \;=\; m_i \cdot \mathbf{d} \;+\; \bar{c} \sum_j L_{ij}$$

For the fixed-points equation the requirement is again $\Phi = 0$ and the neuronal activity takes the form:

$$\mathbf{c} = \mathcal{L}_n \mathcal{D}_n^T \mathbf{m} \tag{3.18}$$

where:

$$\mathcal{L}_n = \begin{bmatrix} \mathcal{L}_{n11} & \mathcal{L}_{n12} & \cdots & \mathcal{L}_{n1n} \\ \mathcal{L}_{n21} & \mathcal{L}_{n22} & \cdots & \mathcal{L}_{n2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{L}_{nn1} & \mathcal{L}_{nn2} & \cdots & \mathcal{L}_{nnn} \end{bmatrix} \mathcal{D}_n = \begin{bmatrix} D & 0 & \cdots & 0 \\ 0 & D & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D \end{bmatrix} \tag{3.19}$$

$\mathcal{L}_{nij}$ is a diagonal $n \times n$ matrix with diagonal elements:

$$\mathcal{L}_{nij} = (\mathcal{I} - \mathcal{L})^{-1}_{ij} \mathcal{I}$$

where $\mathcal{I}$ is the identity $n \times n$ matrix. $\mathcal{D}_n$ is a diagonal $n \times n$ block matrix, where the bloks are the input matrices D. Duing so allows to decoupled the system in n sub-systems which means that the solutions are the direct product of the solutions of the sub-systems. The solutions for the weights could be found with the inverse of $\mathbf{c}$ if $|\mathcal{D}_n| \neq 0$ and $|\mathcal{L}_n| \neq 0$ [11]. From studies over the stability follows that all the stable solutions in the single-neuron case are also stable in the network case. While network interactions do not change the stability of the possible solutions, they do change the basins of attraction associated with different solutions. The associative solutions can be divided into *completely associative* and *partially associative*, where the first kind refers to those solutions that associate all the neurons to a simple input pattern and the others exhibit an incomplete associativity. For a general n-size network the number of stable solutions is n^n, the number of completely selective solutions is $n!$ and the number of completely associative solutions in n. Thus, the numer of solutions with incomplete associativity is $n^n - n! - n$. If all the lateral connections are negative, different neurons reach different stable states (selective state) with higher probability, while if l_{ij} are positive, different neurons are more likely to reach a similar state (associative state).

[11] Deriving the BCM from the objective function gives a slightly different equation (3.16): $\dot{\mathbf{m}}(t) = \mathcal{L}_n \mathcal{P}_n \mathcal{D}_n^T \Phi$. The study of fixed points of this equation is the same because the matrix $\mathcal{L}_n$ is non-singular[17]

Nonlinear neuron with lateral interactions

The objective function in the case of nonlinear neurons with lateral interaction is given by equation(3.10)

$$R_m \;=\; -\frac{1}{3}E[\sigma^3(\zeta)] \;-\; \frac{1}{4}E^2[\sigma^2(\zeta)] \tag{3.20}$$

where ζ is the inhibited activity of the neurons prior to apply the nonlinearity σ. It is possible to define $\zeta = (\mathcal{I} - \mathcal{L})^{-1}M\mathbf{d}$. This leads to the gradient descent dynamics:

$$\begin{aligned} E[-\nabla_m R_m] \;&=\; \{E[\sigma^2(\zeta)\sigma'\nabla_m\zeta] \;-\; E[\sigma^2(\zeta)]E[\sigma(\zeta)\sigma'\nabla_m\zeta\} \\ &=\; E[\phi(\sigma(\zeta,\theta_m)\sigma'\nabla_m\zeta] \\ &=\; \Sigma\mathcal{P}\mathcal{L}\mathcal{D}\phi(\sigma(\zeta),\theta_m) \end{aligned} \tag{3.21}$$

From this equation follows that the stationary solutions arise from the equation $\Phi(\sigma(\zeta)) = 0$), because the matrices Σ, $\mathcal{D}$, $\mathcal{L}$ are positive definete; hence the solutions are:

$$\mathbf{m} \;=\; \mathcal{D}^{-1}\,\mathcal{L}^{-1}\,\Sigma^{-1}(\zeta)$$

with $\zeta \in S$ equivalence classes of solutions reported in (3.7).

Chapter 4

Conclusions

The aim of the present research was to examine a model of synaptic plasticity which reflects some important biological aspects such neuron's tuning curve and selectivity. The model itself is known as BCM learning rule and during the past decades was modified to fit even better the experimental data. The wok points out the learning mechanism adopted by cortical neurons and how these cells extract feature from high dimensional inputs like those we are subjected to. The understanding of the model exposed above is a good starting point to go deeper inside the neural netwok world we are facing.

This study has shown that synaptic modification occurs when th environment used as input changes. This can be expressed mathematically with a set of stochastic differential equations and the application of gradient descent which allows to find the minimum value for the synaptic weight. The precise form of the modification can change but the main thread is that it is a function of the input. Some normalization factors could occur. The application of the model to some cases shows that exists a set of solution for the stability of the weight.

Overall, this study strengthens the idea that the model is working pretty well under various conditions and future works might be done on the side of computing. The computational power achieved in these years is impressive but the model sill need some approximation which means that there is some space for improvement. The findings of this investigation may be completed with computing and some neural network environment so that another brinck will be placed on our knowlwdge of our brain and hopefully help somebody.

This thesis has provided an overview on the biophysical aspects of the neuron's plasticity that is tightly tied up with the evolution and adaptive power animals have. The research area is quite new and the work done pretended to be a useful summary of nowadays' knowledge that could allow other people, researcher and not, to go deeper inside and not stand on the threshold, which by the way is modicable with time as exposed above.

Bibliography

[1] Bienenstock Elie L., Leon N Cooper, and Paul W. Munro. "Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex". In: *Journal Neuroscience* **2** (1982), pp. 32–48.

[2] Nass M.N. and Cooper L.N. *A theory for the development of feature detecting cells in visual cortex*. 1975.

[3] Hebb D.O. *Organization of Behavior*. New York: John Wiley, 1949.

[4] Stent G.S. "*A physiological mechanism for Hebb's postulate of learning*". In: *Proceedings of the National Academy of Sciences U.S.A.* **70** (1973), pp. 997–1001.

[5] Cooper L.N., F. Lieberman, and E.Oja. "A theory for the acquisition and loss of neuron specificity in visual cortex". In: *Biological Cybernetics* **33** (1979), pp. 9–28.

[6] Bienenstock E. *A theory of development of neuronal selectivity*. Doctoral thesis. Providence: Brown University, 1980.

[7] Cooper L.N., Liberman F., and Oja E. *A theory for the acquisition and loss of neurons specificity in visual cortex*. 1979.

[8] Intrator N. and Cooper L.N. *Objective function formulation of the BCM theory of visual cortical plasticity: statistical connections, stability conditions*. 1992.

[9] Bellman R. E. *Adaptive control process*. Princeton, NJ.: Princeton University Press, 1961.

[10] Diaconis P. and Freedman D. *Asymptotics of graphical projection pursuit*. 1984.

[11] Friedman J. H. *Exploratory projection pursuit*. 1987.

[12] Huber P.J. *Projection pursuit (with discussion)*. 1985.

[13] Law C. and Cooper L. *Formation of receptive fields according to the BCM theory in realistic visual environment*. 1994.

[14] B. S. Blais and L. Cooper. "BCM theory". In: *Scholarpedia* 3.3 (2008). revision #91041, p. 1570. DOI: 10.4249/scholarpedia.1570.

[15] Buisseret P. and Imbert M. *Visual cortical cells. Their developmental properties in normal and dark reared kittens*. London. 1976.

[16] Frégnac Y. and Imbert M. *Early development of visual cortical cells in normal and dark reared kittens. Relationship between orientation selectivity and ocular dominance.* London. 1978.

[17] Castellani G. C. et al. *Solutions of the BCM learning rule in a network of lateral interactiong nonlinear neurons.* 1999.

[18] Scofield C. L. and Cooper L. N. *Development and properties of neural networks.* 1985.

[19] Cooper L. N. and Scofield C. L. *Mean-field theory of a neural networks.* 1988.