

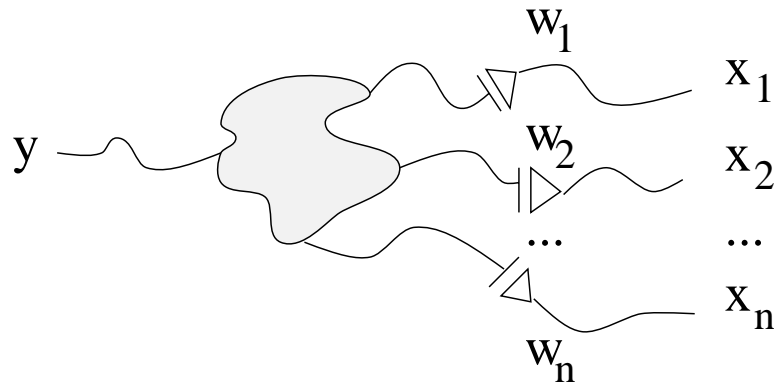
Brian Blais' Homemade Guide to the BCM Theory of Synaptic Plasticity

This guide is based on notes I took during a talk Professor Cooper gave for the IBNS Research Seminar.

1 The Model of the Neuron

A Simple Picture

We'd like to understand how the brain works but that is, at least immediately, an impossible task. In the standard physics fashion we introduce a simplified model of the smallest workable unit of the big problem and make that the system we'd like to understand, namely the neuron. It is a well established idea in neuroscience that signals between neurons is electro-chemical in nature. The behavior of a neuron depends on these signals coming into it from other neurons. It also depends manifestly on the history of these inputs, making the overall behavior of the neurons quite complicated. The model we introduce exhibits these basic characteristics without the extreme detail which would make the modeling impossible. The simple picture of the neuron is



where \mathbf{x} is the vector of inputs into the neuron, y is the output value, and \mathbf{w} is the vector of *synaptic weights*, or efficiencies of the connections from each neuron to the target neuron.

What are the physical mechanisms behind \mathbf{x}, \mathbf{w} , and y ? What is the relation between the simplified model and reality? These are questions that we must face if we are to say we have any understanding of the neuron.

An Argument from Evolution

The claim has been made that much of *memory* and *learning* takes place in the modification of the chemical properties at the synapse. The electrical signal impinging on the synapse is changed to a chemical signal sent across the synapse, and then changed back to an electrical signal. The chemical signal is affected by the properties of the receptors at the synapse, so that the electrical signal actually “seen” by the cell is a filtered version of the incoming signal. The change in this filter is called *synaptic modification*. One can justify the development of the nervous system as a network of many neurons communicating with electrical signals filtered through a chemical signaling process, using a simple argument from evolution. This argument is intended to loosely motivate the development of such a system, and to highlight the important aspects of neural systems, not to get all the details correct or suggest that this is the only way for things to have happened.

As life on this planet developed, it started off small. This is, of course, the only possible way it could have developed. Later, however, it became evolutionarily favorable to have larger creatures. A larger creature is harder to eat, it has a larger chance of surviving because larger creatures are often more robust to damage, and larger creatures allow for the possibility of redundant parts making it more resilient to change. Unfortunately larger creatures require more resources, and they necessitate more efficient communication and transport systems between their body parts. One possible way to achieve this is to have some kind of electrical communication.

The first types of electrical communication were most likely quite simple, perhaps akin to a single neuron between one end of the creature and the other. Such a design would have advantages over non-electrical communication systems, due to speed, but it would be even more advantageous to have *more than one* neuron. A multiple neuron system would be far less susceptible to damage. The existence of more than one neuron implies connections between them. The communication across the connections in actual neurons is chemical in nature, as we have mentioned before. This begs the question *why chemically based connections, rather than electrical ones?*

Passing an electrical signal through a chemical-signal filter has a number of advantages. Its biggest advantage is its ability to achieve fine and long lasting change. The dynamic communication system which would result from this is well suited to a creature needing to adapt its behavior to changing environmental conditions. The chemical synapse, then, serves the dual purpose of an adaptable communication system and a method for information storage. Our understanding of the behavior of the synapse is tantamount to our understand of learning and memory in general, so we ask the question *how does the synapse change?*

Back to the Model

We return now to the model to begin to answer some of these questions. For now we will say that the inputs to the neuron, $\mathbf{x} \equiv (x_1, \dots, x_n)$, are related to the magnitude of electrical activity along the connections to the neuron. Likewise the output, y , is related to the magnitude of electrical activity of the neuron itself. We will need to make this more precise later but for now we will not pin down the definitions in order to avoid making premature assumptions.

We have made the further approximation that the synaptic efficiency, despite its apparent complexity, can be thought of as a single number called a *weight*. We do this both because it is the easiest thing to write down, and we presume that the full complexity of a neuron is not completely necessary for its general behavior. When we find that this approximation is not adequate to understanding the neuron, we can then modify it. If it does work, then we have verified that the neuron can be understood from simple fundamental principles. In this same vein we can state that the output of the neuron is given simply by

$$y = w_1x_1 + \dots + w_nx_n = \mathbf{w} \cdot \mathbf{x}$$

If we wanted to be slightly more biologically plausible, we could rule out arbitrarily large output values, using a squashing (or sigmoid) function:

$y = \sigma(\mathbf{w} \cdot \mathbf{x})$, where $\sigma(\cdot)$ looks like



Now that we have defined what the output of a neuron is, we face the question asked above: *how do the weights change?*, or *what is the learning rule?* In mathematical language we wish to specify

$$\frac{d\mathbf{w}(t)}{dt} \equiv \dot{\mathbf{w}}(t)$$

The learning rule, then, is the foundation of the entire model. We will discuss in Section 2 the details of the learning rules we want to consider. For now all we need to do is specify the possible constraints on our choice for a learning rule. The constraints are

- **Simplicity:** We require that the rules we consider are *simple*. We don't want to run into the problem that there are too many parameters to fit to our system. At the same time we don't want to lose all information by oversimplification. So, as a rule of thumb, we will introduce the model as simple as possible and let only poor predictions force us into making it more complicated.
- **Locality:** We require that (almost) all the information the neuron has access to is localized to that neuron. These would include the inputs to the neuron, the synaptic weights, the output of the neuron, and perhaps

some parameters intrinsic to the neuron itself. They would *not* include the activities of neurons that are not directly connected to it, global information about the input patterns being presented to the network of which the neuron is a part, etc. There still could be some limited global information, but there is no evidence to motivate a model of it at this point.

The question we want to treat next is *given a particular learning rule, how do we verify it experimentally?* It becomes necessary to explore some of the macroscopic consequences of the learning rules, because direct verification of a particular rule is quite difficult to do.

Macroscopic Consequences

The most notable characteristic of neurons in the brain is that most of them respond preferentially to some small subset of the input patterns presented to them. One can find many examples of this in any elementary neurobiology text: cells in the visual cortex which are selective to the orientation of a visual stimulus, cells in the auditory cortex which are selective to frequency bands, cells in the motor cortex which are selective to movement direction, etc. Though neurons are often classified by the things for which they are selective, selectivity is not a static property: it can be changed with the input environment.

The most striking example of this is the set of visual deprivation experiments (Wiesel and Hubel, 1963; Wiesel and Hubel, 1965; Singer, 1977; Gu et al., 1989; Daniels and Saul, 1985). Simply put, these experiments probe the change in orientation selectivity in the visual cortex due to visual environment changes. We will begin with the crudest of these experiments to initially justify the model. The model will then have more specific consequences that can be measured. Along the way we will be forced define more specifically the parameters in the model, and then look at attempts for their direct measurement.

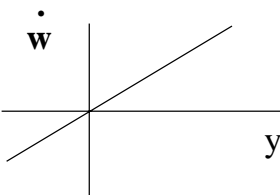
2 The BCM Learning Rule

The Hebb Learning Rule

The cornerstone of all learning rules is the Hebb rule (Hebb, 1949). The original Hebb rule states how synapse efficacies are strengthened:

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

Mathematically this is expressed

$$\dot{\mathbf{w}} = y\mathbf{x}$$


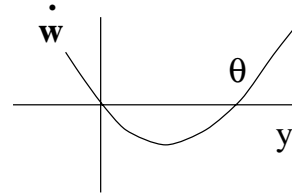
This rule has the obvious problem that repeated inputs force the neuron to the largest possible value (infinity, if we don't use the $\sigma(\cdot)$ function). A variant of this Hebb rule, called the Oja rule (Oja, 1982), solves this problem by introducing a decay factor into the learning rule proportional to y^2 . The properties of this learning rule are explored in many sources (Shouval et al., 1996; Liu and Shouval, 1994), and will not be discussed here.

Modeling Selectivity: BCM

Hubel and Wiesel demonstrated that normally reared animals have orientation selective visual cortical neurons. These neurons responded preferentially to bars of light of a particular orientation. Dark reared animals, the went on to show, do *not* have orientation selective cells. *What kind of modification can account for this?* We need a

rule which can become selective to a subset of the input and is stable. The proposed rule(Beiensenstock et al., 1982) looks like

$$\dot{\mathbf{w}} = \phi(t, \dots)\mathbf{x} \equiv y(y - \theta)\mathbf{x}$$



Initially patterns can yield activity both above and below the threshold, θ . Patterns which yield activity above θ increase the weights, driving the activity up. Patterns which yield activity below θ decrease the weights, driving the activity to zero. In the end, the neuron will respond strongly to a subset of the inputs. Of course this is all true only if some of the patterns initially fall both above and below the threshold. If they all fell below, then the response them would be driven to zero yielding a totally unselective cell. If they all fall above then we have the same problem we ran into in the Hebb case: the weights will saturate. This would be a problem for *any* patterns whose response fell above the threshold, so we need to cope with this problem. This is solved by allowing the threshold to slide, as a non-linear function of the activity of the neuron, say $\theta \sim y^2$. It would have to be a non-linear function for the threshold to “catch up” with the activity.

The easiest way to understand the BCM rule, and the role of the sliding threshold, is to see a few simple examples worked out.

3 BCM Examples

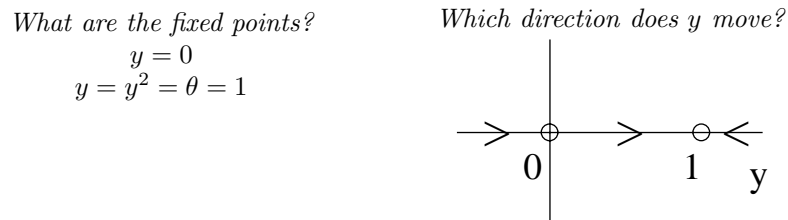
BCM example 1: 1D with constant input

Before we can write down the one dimensional BCM equations, we must realize that we have some freedom to choose the form of θ (see the section of **Summary of Assumptions** on page 6). We usually take θ to be $E[y^2]$, where $E[\cdot]$ is either an average over the *entire* input space, or an average over a *temporal window*, both of which can be considered to be equivalent in the proper limits(Iterator, 92). Regardless, in many examples the specific form *does not* matter significantly, so we will usually take whichever form is most convenient.

Having said that, the one dimensional BCM equations are

$$\begin{aligned} y &= wx && \text{(Activation equation)} \\ \dot{w} &= \eta y(y - \theta)x && \text{(BCM learning equation with learning rate } \eta) \\ \theta &= E[y^2] = y^2 && \text{(sliding threshold using average over input space)} \end{aligned}$$

assuming a constant input x . The fixed points of the learning equation occur when $y = 0$ and $y = \theta$. The following diagram demonstrates that, for the one dimensional case, only the $y = \theta$ fixed point is stable.



Thus, the threshold slides to the value of y^2 , the only stable fixed point. The one dimensional case, however, is too simple to demonstrate selectivity, so we will explore that issue with the two dimensional case. The 1D case does give us an intuition for the motion of the sliding threshold, and the stability of the model.

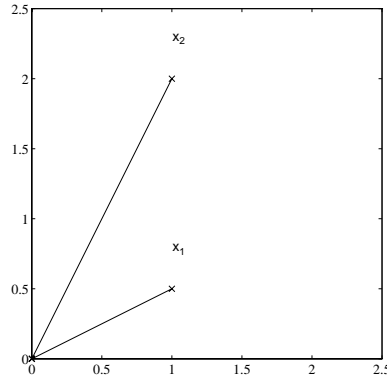
BCM example 2: 2D with 2 linearly independent inputs

As before, we can write down the BCM equations, and find the fixed points.

$$\begin{aligned}
y &= \mathbf{w} \cdot \mathbf{x} && \text{(Activation equation)} \\
\dot{\mathbf{w}} &= \eta y (y - \theta) \mathbf{x} && \text{(BCM learning equation with learning rate } \eta) \\
\theta &= E[y^2] && \text{(sliding threshold using average over input space)}
\end{aligned}$$

with fixed points $y = 0$ and $y = \theta$.

In the two dimensional case, each input is a 2D vector. One possible input set is



there are 4 possible weight vectors satisfying the fixed point conditions.

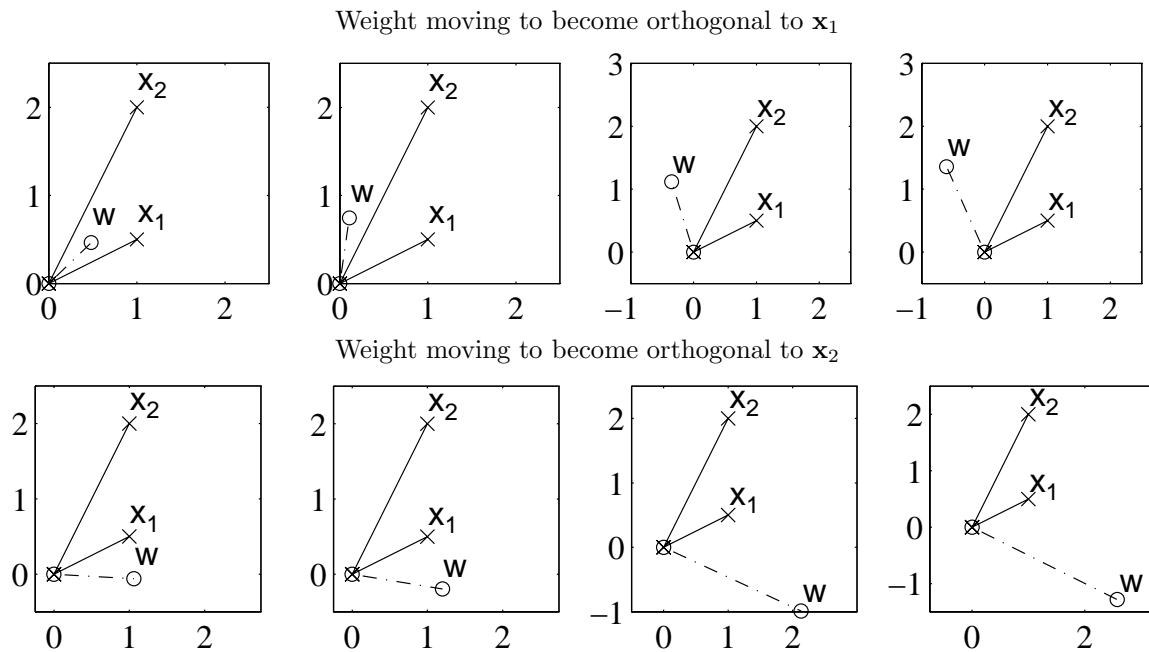
- $(\mathbf{w} \cdot \mathbf{x}_1) = 0$ and $(\mathbf{w} \cdot \mathbf{x}_2) = 0$: zero weight vector
- $(\mathbf{w} \cdot \mathbf{x}_1) = 0$ and $(\mathbf{w} \cdot \mathbf{x}_2) = \theta \neq 0$: weight orthogonal to \mathbf{x}_1 and has projection θ on \mathbf{x}_2
- $(\mathbf{w} \cdot \mathbf{x}_1) = \theta \neq 0$ and $(\mathbf{w} \cdot \mathbf{x}_2) = 0$: weight orthogonal to \mathbf{x}_2 and has projection θ on \mathbf{x}_1
- $(\mathbf{w} \cdot \mathbf{x}_1) = \theta \neq 0$ and $(\mathbf{w} \cdot \mathbf{x}_2) = \theta \neq 0$: weight has equal projections, θ , on both \mathbf{x}_1 and \mathbf{x}_2

It can be shown (Interator, 92) that the only two *stable* fixed points are the ones where the weight vector is orthogonal to just one of the inputs. If each input is presented with probabilities, $P(\mathbf{x}_1)$ and $P(\mathbf{x}_2)$, then it follows, without loss of generality, for the case where the weight is orthogonal to \mathbf{x}_2

$$\begin{aligned}
\theta &= E[y^2] = \sum_{i=1}^2 P(\mathbf{x}_i) (\mathbf{w} \cdot \mathbf{x}_i)^2 \\
&= P(\mathbf{x}_1) (\mathbf{w} \cdot \mathbf{x}_1)^2 \\
&= P(\mathbf{x}_1) \theta^2 \\
\theta &= \frac{1}{P(\mathbf{x}_1)}
\end{aligned}$$

Thus the threshold becomes a measure of the inverse probability of the input the neuron is selective to, and it determines how strongly it responds to that particular input.

Examples of the evolution of a weight vector, for each stable fixed point, can be seen in the following diagram:



It is easy to see that, given slightly noisy input, the neuron will converge to about the same fixed points, and thus will be selective. Given *only* noise as input, the neuron will not become selective at all. So initially this gives agreement to the roughest experimental results:

patterned input \longrightarrow selectivity
 non-patterned input \longrightarrow no selectivity

Summary of Assumptions

It is necessary to outline all of the assumptions of the model, so that we are aware of both the freedoms and restrictions we have in modelling. The assumptions are as follows.

Assumptions not specific to BCM

- The output (or activity) of the neuron, $y(t)$, is a scalar function of time. It is somehow related to the integrated current output by the neuron, and will thus increase with increasing firing rate and also with increasing output potential.
- The description of the synapse involves only a scalar function of time, $w(t)$, called the weight.
- The output is given by a function of the dot product of the weights with the inputs, $y = f(\mathbf{w} \cdot \mathbf{x})$. Usually $f(\cdot)$ is a squashing, or sigmoid, function.
- Plasticity is governed by the change in the weights as a function of time. Both learning and memory can be accomplished this way.

Assumptions specific to BCM

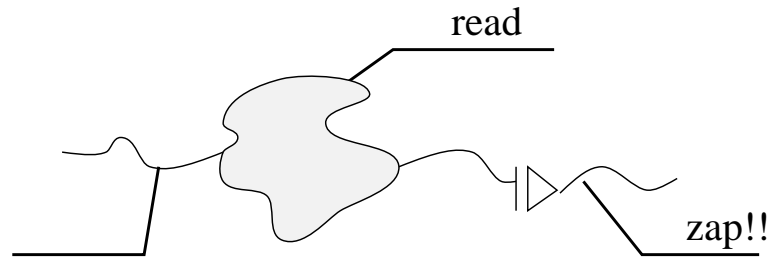
- The *learning rule*, $d\mathbf{w}/dt$, is a function of both the current firing rate y , and the input vector \mathbf{x} .
- $d\mathbf{w}/dt = \phi(t, \dots)\mathbf{x}$ has a threshold in the activity, above which the weights are increased (e.g. Hebb, LTP) and below which the weights are decreased (e.g. Anti-Hebb, LTD). The usual form is $d\mathbf{w}/dt = \eta y(y - \theta)\mathbf{x}$, where η is a parameter determining how quickly the weights are changed.
- The threshold, θ , is a non-linear function of the output of the neuron, and is thus a whole-cell parameter. The usual form is $d\theta/dt = \frac{1}{\tau}(y^2 - \theta)$, where τ is a parameter determining how quickly θ slides.

4 Experimental Verification

We have a theoretical structure which has consequences that can be measured. This would *suggest* the validity of the theory, but we'd like to have a direct verification of the theory.

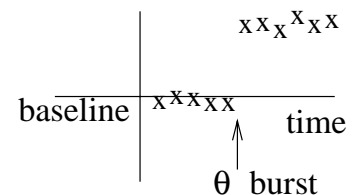
Measuring Long Term Potentiation (LTP)

In order to measure LTP, one sets up an electrode to stimulate the neuron and then read its response. A simplistic picture of the setup is



The procedure for achieving LTP is then

- Establish a baseline response using a low frequency input, like 0.1 Hz, until the neural activity stabilizes.
- Give a high frequency burst (called a Theta burst) of 100 Hz for a short time.
- Reestablish response to baseline stimulus of 0.1 Hz.



Aside: In a very handwaving fashion, one can think of LTP as the long term storage of dramatic events to the neuron.

The properties of LTP have been explored in great detail. Differences have been found in different parts of the brain, in vivo versus in vitro, in different animals, etc. As physicists we prefer to assume global rules unless qualitative differences arise. We will then assume that LTP is basically the same everywhere in the brain, and explore the function of it in the brain.

In our learning rule, LTP would be the part of the $\dot{w}(t)$ function *above* the threshold, where modification tends to increase the response to stimulus.

Measuring Long Term Depression (LTD): The Rest of the Learning Rule

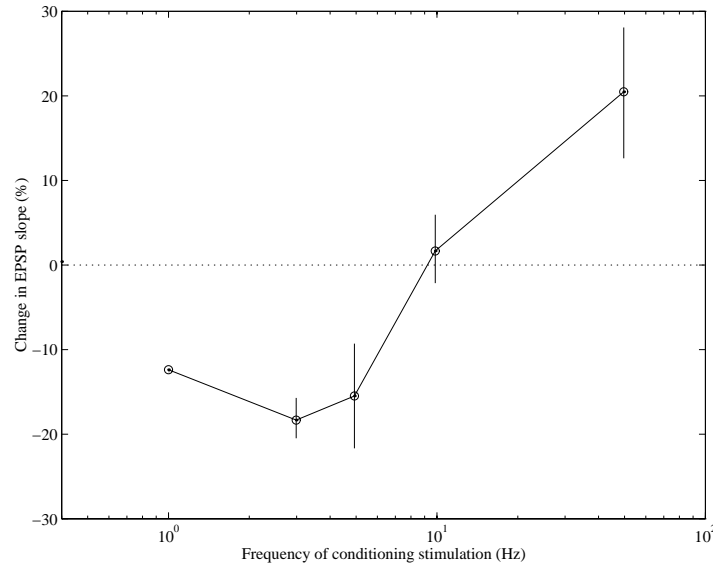
Verifying the rest of the learning rule requires us to define the variables in the theory more precisely. If we have $\Delta w(t) = \phi(t, \dots)x(t)$, and we know that the modification function $\phi(\cdot)$ is related to the postsynaptic depolarization, then x must occur over this time scale. For instance, if we give one electric pulse every minute for 15 minutes, are all the pulses included in determining the value of $x(t)$? No. x is some small grouping, or local time average, of the incoming pulses. Unfortunately we don't know the true time scale over which this average is performed, so a direct number for x given the firing rate or magnitude of a neuron is not feasible. One can, however, find a clever workaround to this problem.

If one sends in N pulses $x(t_1), x(t_2), \dots, x(t_N)$ over a time period where the modification function $\phi(\cdot)$ does not change significantly (ie. the threshold θ doesn't slide much), then the change in w is given by

$$\begin{aligned} \Delta w(t_1) &= \phi(t_1, \dots)x(t_1) \\ \Delta w(t_2) &= \phi(t_2, \dots)x(t_2) \\ &\vdots \end{aligned}$$

$$\begin{aligned}\Delta w(t_N) &= \phi(t_N, \dots)x(t_N) \\ \Delta w &= \sum_{i=1}^N \phi(t_i, \dots)x(t_i) = \phi(\cdot) \sum_{i=1}^N x(t_i) \\ &= \phi(\cdot) \cdot (\text{total input})\end{aligned}$$

So, if one can keep the total input the same, and change only the input frequency, then for each frequency one is probing the value of the $\phi(\cdot)$ function. The result of the experiment (Dudek and Bear, 1992) is



References

- Beienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal Neuroscience*, 2:32–48.
- Daniels, J. D. and Saul, A. B. (1985). Ocular dominance, selectivity, and responsiveness in kitten area 17 neurons, after dark rearing plus brief monocular experience. In *Society for Neuroscience*.
- Dudek, S. M. and Bear, M. F. (1992). Homosynaptic long-term depression in area CA1 of hippocampus and the effects on NMDA receptor blockade. *Proc. Natl. Acad. Sci.*, 89:4363–4367.
- Gu, Q., Bear, M. F., and Singer, W. (1989). Blockade of NMDA-receptors prevents ocularity changes in kitten visual cortex after reversed monocular deprivation. *Developmental Brain Research*, 47:281–288.
- Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.
- Intrator, N. (92). Feature extraction using an unsupervised neural network. *Neural Computation*, 4:98–108.
- Liu, Y. and Shouval, H. Z. (1994). Localized principal components of natural images - an analytic solution. *Network*, 5:317–324.
- Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273.
- Shouval, H., Intrator, N., Law, C. C., and N Cooper, L. (1996). Effect of eye misalignment on ocular dominance according to BCM and PCA synaptic modification. *Neural Computation*, 8(5):1021–1040.
- Singer, W. (1977). Effects of monocular deprivation on excitatory and inhibitory pathways in cat striate cortex. *Experimental Brain Research*, 134:508–518.
- Wiesel, T. N. and Hubel, D. H. (1963). Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Neurophysiology*, 26:1003–1017.
- Wiesel, T. N. and Hubel, D. H. (1965). Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens. *J. Neurophysiol.*, 28:1029–1040.