Chapter 12

Adaptation and Learning

§ 1. Memory and Learning

The phenomenon of memory has long fascinated scientists and philosophers. Theorizing about memory dates back some twenty-five centuries.

> Memory is, therefore, neither perception nor conception, but a state or affection of these, conditioned by the lapse of time. As already observed, there is no such thing as memory of the present while present; for the present is object only of perception, and the future, of expectation, but the object of memory is the past. All memory, therefore, implies a time elapsed; consequently only those animals which perceive time remember, and the organ whereby they perceive time is also that whereby they remember.
>
> As regards the question, therefore, what memory or remembering is, it has now been shown that it is the having of an image, related as a likeness to that of which it is an image; and as to the question of which of the faculties within us memory is a function, it has been shown that it is a function of the primary faculty of sense-perception, i.e. of that faculty whereby we perceive time [ARIS: 714, 715].

Today we know things are not quite so simple as Aristotle thought, although it would be wrong to conclude he erred in all his particulars. There is a large corpus of literature on the psychology of memory, and this textbook is not the place to attempt a review of this. Our discussion here will be restricted to only a few illustrative remarks. William James wrote,

> Memory proper . . . is the knowledge of a former state of mind after it has already once dropped from consciousness; or rather *it is the knowledge of an event, or fact,* which meantime we have not been thinking, *with the additional consciousness that we have thought or experienced it before*.
>
> The first element which such a knowledge involves would seem to be the revival in the mind of an image or copy of the original event. And it is an assumption made by many writers that the revival of an image is all that is needed to constitute the memory of the original occurrence. But such a revival is obviously not a *memory*, whatever else it may be; it is simply a duplicate, a second event, having absolutely no connection with the first event except that it happens to resemble it. . . No memory is involved in the mere fact of recurrence. The successive editions of a feeling are so many independent events, each snug in its own skin. . . A further condition is required before the present image can be held to stand for a *past original*.
>
> That condition is that the fact imaged be *expressly referred to the past*, thought as *in the past*. But how can we think a thing as in the past, except by thinking of the past together with the thing, and of the relation of the two? . . . [If] we wish to think of a particular past epoch, we must think of a name or other symbol, or else of certain concrete events, associated therewithal. Both must be thought of to think the past epoch adequately. . .
>
> It follows that what we began by calling the 'image,' or 'copy,' of the fact in the mind is really not there at all in that simple shape, as a separate 'idea.' Or at least, if it be there as a separate idea no memory will go with it. What memory goes with is, on the contrary, a very complex representation, that of the fact to be recalled *plus* its associates, the whole forming one 'object' . . . known in one integral pulse of consciousness . . . and demanding probably a vastly more intricate brain-process than that on which any simple sensorial image depends [JAME: 648-651].

A specific "memory" is not some photographic-like image, a "snapshot" snipped from the pages of one's experience as a whole. Science has found, as James had predicted, that there is much more to it than that. There appears to be no one special place in the brain where memories reside. Rather, the situation appears to be that the process of "having a memory" involves the participation of numerous regions of the brain, including parts of the cortex involved with motor functions. This is one of the many facts psychology has unearthed contributing to the refutation of the old idea of the British empiricists (and, for that matter, of Aristotle) that some "copy of reality" mechanism exists that stamps the impress of external objects into the brain (or mind). Piaget et al. found that memory is intimately tied to sensorimotor schemes of actions.

> We can, in fact, establish three major hierarchic types of memory, each with several sub-levels: the recognitory memory (1-3); the reconstructive memory (4-7) and the recollective memory . . .
>
> *Type I: the recognitive memory.* Recognition at all levels is an assimilation of the data to schemes of various kinds, ranging from reflex and elementary habits to the motor schemes of sensory exploration.
>
> (1) Elementary recognition is bound up with the continuation or repetition of a reflex action or a potential habit extending that reflex . . .
>
> (2) Next, there is recognition by assimilation to an existing scheme (in the repetition of the schematized action): this is the recognition of signs as signifiers and is bound up with habits and acts of the sensori-motor intelligence – the fact that the signs are treated as signifiers is due precisely to their links with the schemes.
>
> (3) Recognition at the higher levels is bound up with mobile and differentiated schemes (classifications, etc.). . .
>
> *Type II: mnemonic reconstruction.* In contrast to elementary habits, which involve the reproduction, intentionally or otherwise, of schematized actions or of sensori-motor schemes tending towards generalization, a reconstruction is the intentional reproduction of a particular action and of its results. Hence, it involves the recognition of signs, etc. but goes beyond recognition proper in that it constitutes a form of recall by action: it tends to reconstruct a model no longer available for perception, while recognition occurs in the presence of the model.
>
> (4) We may seek the elementary form of the reconstructive memory in sensori-motor imitation, considered as the intentional reproduction of an action performed by oneself, or by somebody else, and often of the motion of an object. Now, this interpretation of the reconstructive memory bears out the genetic point of view, because sensori-motor imitation heralds recall and already constitutes a kind of recall by actions. Moreover, in its deferred and above all in its internalized forms, it becomes representative recall and constitutes the source of the mental image which plays so important a rôle in the recollective memory.
>
> (5) Next comes the reproduction of an isolated and not fully schematized action and reconstruction of its result: this is the situation . . . where the model was copied before it was reconstructed.
>
> (6) Then there is the reconstruction of an object or a configuration (without prior constructions of an imitative or spontaneous kind): this is the situation examined . . . where the model was recalled by memory-drawings and reconstructions, with the latter producing considerably better responses than the former.
>
> (7) Finally, we have reconstruction of a schematized action. . .
>
> *Type III: mnemonic recall.* There is no need to stress the fact that even the recollective memory depends on actions and action schemes, thus ensuring the complete continuity as between reconstructions by actions (type II) and internalized reconstructions represented by the memory-image as the instrument of recall:
>
> (8) The memory-image of a schematized action: . . . here we found precisely that from the age

> of seven years . . . simple recall produces the same results as reconstruction . . . which proves the complete internalization of reconstructive procedures.
>
> (9) Recall by images of any non-schematized action: this is the direct result of the internalization of imitation by images.
>
> (10) Recall by images of objects or events extraneous to the action: this is the 'pure' memory of classical psychologists, but as we saw . . . these 'ill-assorted' configurations are nevertheless subject to active schematizations, the *sine qua non* of their retention.
>
> From the foregoing remarks, the reader will have gathered by what slight transitions children advance from elementary recognitions, closely bound up with actions, to the higher forms of recall, which, thanks to their links with the operational schemes, cannot be entirely divorced from actions because, as we say, operations spring directly from the latter [PIAG16: 392-395].

Educators have long taken note of the distinction between "active learning" and "passive learning," and have known that the former is effective in producing learning while the latter is not. This is why teachers use repetition and drills to convey lessons. It is why students are assigned homework exercises that go beyond mere textbook-reading, and it is why students who do these exercises learn the material while those who do not are unsuccessful in learning. The development of memory in its various forms is tied to actions and action schemes. ***Learning*** is the word we use to describe the successful development of memory. A learner who does not pass beyond rote imitation and recitation is merely a pupil; one who develops the operational schemes for generalizing beyond the particular to a wider range of usage for his or her memory structure is a student. Many young people just entering college find the transition from high school to college difficult precisely because, while previously it was possible for them to be successful by merely being good pupils, in college they are expected and required to be good students.

§ 2. Synaptic Plasticity

Learning implies change of some sort taking place in the mind-brain system. In the physical dimension of the mind-brain system, this change can only be change of some kind in the neural system. Now, normal full-term human infants are born with, in complete or nearly complete measure, all the neuron cells their brains will ever have. True, these cells will grow in size for awhile, but not in number. What *does* change, and dramatically so, is the growth and establishment of synaptic connections. Synapses form in great profusion and either become established or else disestablish and disappear. Between the ages of two to ten or eleven years, a normal child's brain has on the order of twice the number of synapses as an adult. Furthermore, in many parts of the brain there is found to be a critical time period in development, during which synaptic structures can form and become established through neural activity. After this critical period has passed, these structures no longer will form. Thus, deprivation of particular kinds of sensorimotor experience during the critical period results in permanent disability of the particular neural function involved with that experience.

In established synapses, the ***efficacy*** or ***strength*** of the synaptic connection – by which we

mean the amount of postsynaptic response to presynaptic excitation – increases or decreases ***according to the level of activity at the synapse***. This is termed ***long term synaptic plasticity*** and comes in two general categories: ***long term potentiation*** (LTP) and ***long term depression*** (LTD). The discovery of activity-dependent long term potentiation and long term depression underscored a speculation made much earlier by Cajal (1911) and later by psychologist Donald Hebb (1949) that learning and memory result from changes in the strength of synaptic transmission.

Synapses capable of exhibiting LTP and LTD exist in significant numbers in the hippocampus, neocortex, and amygdala – brain structures heavily implicated in the psychological phenomena of memory and learning. The synaptic plasticity model is generally accepted by neuroscientists as the most likely hypothesis for explaining memory and learning phenomena. Indeed, it is difficult to imagine any other mechanism. Nonetheless, experimentally confirming this hypothesis is very, very challenging [GRIM], and so it is worth bearing in mind that what we are discussing here is hypothesis rather than definitely established fact. The SPM ("synaptic plasticity and memory") model does occasionally face challenges from findings coming out of experimental laboratories. It has so far not been overthrown, but this does not necessarily mean that could not happen some day.

The marriage of neural network theory and psychology is often called "connectionist theory." Not all psychologists are "connectionists," and it is probably accurate to say the majority of connectionist psychologists are found in America. American psychology has long tended to go its own way with disregard for trends in European psychology. Connectionism is strongly associated with what is called "the cognitive revolution" that took place in American psychology in 1960 following American psychology's long and sterile love affair with behaviorism dating from the 1920s. The connectionist paradigm played an important role in coming to think of brain organization in terms of computer and computer-like "information processing" analogies. This shows up when a neural network or an artificial intelligence (AI) theorist speaks of "symbols" and "symbolic processing" in the brain. "Symbol," however, is a nice vague term that can mean almost anything someone wants it to mean. Reber's *Dictionary of Psychology* lists nine different "usages" for the term in psychology, none of which tie the word to neurological entities.

Computer science jargon usually ties the word "symbol" to whatever is stored in a unit of computer memory, this "unit" being conveniently defined to fit whatever context seems the most appropriate for the computer science topic at hand. Because synaptic efficacy is what is represented by the weights in a map model, it is but a short metaphorical step to regard synaptic plasticity as not merely the basic mechanism underlying the phenomenon of memory but, rather, as "the memory" itself. One sees the shadow of this metaphor when a neural network theorist

makes a remark to the effect, "the long term memory is stored in the weights." However useful this metaphor may be, it tends to promote thinking of "memory" in computer terms rather than in the proper psychological context of the word "memory." Grimwood et al. remark,

> Learning and memory are generally divided into a set of constituent processes – encoding, consolidation, storage, and retrieval – that occur at different phases of learning and recall, that may involve different brain areas, and that are very likely to involve distinct activity patterns. When considering the role of synaptic plasticity in learning and memory, we must recognize that this role might be different at these different phases. It seems likely that synaptic plasticity is involved in encoding, storage, and the initial stages of consolidation of information, but not in retrieval of that information.
>
> It is worth mentioning that memories should not be confused with the traces that subserve them. Trace encoding can be thought of as the momentary collective activity of large numbers of neurons whose patterns of firing give rise to increases and decreases of synaptic strength that then outlast these very patterns. Memory retrieval is the process of passing neural activity through the network to create patterns of firing that constitute a "memory." The SPM hypothesis asserts that activity-dependent synaptic plasticity is the fundamental mechanism responsible for creating and storing traces. In this sense, LTP enables memory; it is not equivalent to it [GRIM: 525-526].

This is a cautionary remark worth remembering.

§ 3. Long Term Potentiation and Long Term Depression Mechanisms

Long term potentiation (LTP) is a long-lasting increase in the magnitude of synaptic strength. Long-lasting in this context means hours, days, months, years, a lifetime. This is in sharp distinction from short term potentiation (STP), which in this text is called an *elastic modulation* of synaptic strength because in STP the synaptic strength eventually returns to its original level. The functional opposite of LTP is long term depression (LTD), which is a long-lasting decrease in the magnitude of synaptic strength. As we come up to our discussion of adaptation algorithms used in map and network system models, it is worth noting that *all* the widespread algorithms in use today are LTP/LTD models. There are presently *no* adaptation algorithms in *widespread* use in biological signal processing or computational neuroscience aimed at elastic modulations.

The physiological mechanisms for LTP are metabotropic. This means they involve complex biochemical reactions taking place inside the cell. The effect of this metabotropic signal processing acts on the ionotropic channels by means of which signals are produced, processed, and transmitted in neural networks. Thus, the LTP concept belongs to "control processing" rather than to immediate "data processing" (to use some more computer jargon) in neural systems. To the extent that it is conceptually helpful to think of certain neural structures as being "like" the memory elements in a computer (and to no more extent than this), LTP and LTD can be regarded as mechanisms by which network systems ***build and organize their own 'memory circuits'*** within the overall neural system. This is, of course, a simile for talking about the biological function.

LTP mechanisms can be either presynaptic or postsynaptic. This means that the long term changes in synaptic strength are tied to physiological changes taking place in one or the other side of the synaptic junction. Presynaptic LTP and postsynaptic LTP have qualitatively different physical characteristics in terms of conditions that are present in the initiation of LTP. This means that the basic biochemical mechanisms are different for presynaptic LTP and postsynaptic LTP. Likewise, the sites within the brain where these two types of LTP are found are also different.

§3.1 NMDA-Mediated LTP

Historically, postsynaptic LTP was the first type of LTP discovered experimentally (in 1973). It is this type of LTP that appears to correspond to Hebb's now-famous 1949 conjecture:

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

Although today some evidence suggests it might not be necessary for changes to take place in *both* cells (although this does happen in many cases, i.e., in those cases where new synapse growth occurs), the principal condition of Hebb's speculation is found in postsynaptic LTP. This is to say that both the presynaptic and postsynaptic cell fire action potentials, and the presynaptic cell must fire first and within a short time period of the firing of the postsynaptic cell. This pairing of cell firings with subsequent change in the synaptic strength is today called ***Hebbian learning***.

The region of the hippocampus called CA1 is the most-studied brain structure for postsynaptic LTP, and findings from these studies serve as a model for postsynaptic LTP elsewhere in the brain. Two conditions must be satisfied for this form of LTP to be elicited [NICO]. First, the synapses must be activated at a high frequency; this is to say, the presynaptic cells involved must present a tetanus to the postsynaptic cell, and this tetanus must have a high firing rate. Second, the overall intensity of the stimulus must be above a certain threshold intensity. What this means is that there must be a sufficient number of synaptic inputs involved in volley firing of the tetanus or else the LTP effect does not take place.

A great step forward in understanding this phenomenon was made when the NMDA channel was discovered and this channel was found to be involved in those experimental cases where postsynaptic LTP had been observed [MALE]. As you will recall from the earlier chapters, NMDA channels are glutamate-enabled/voltage-activated Ca^{2+} ionotropic channels. For small EPSPs induced in the target cell, these channels are blocked by Mg^{2+} and the NMDA channel plays little part in ionotropic signaling. However, when the postsynaptic membrane is depolarized, this leads to the ejection of the blocking Mg^{2+} particle and the opening of the Ca^{2+}

channel. Because the amount of depolarization required to open many NMDA channels is typically also large enough to elicit an action potential response by the postsynaptic cell, this explains why pairing of pre- and post-synaptic action potentials is necessary for LTP initiation. It also explains why the volley input must surpass a particular threshold of intensity (because a sufficient EPSP must be produced to open the NMDA channels), and why the presynaptic cells must fire first (because they must fire first in order to *enable* the NMDA channels).

Ca^{2+} is one of the most potent metabotropic agents found in biological systems. Under normal resting conditions, the intracellular concentration of free Ca^{2+} is extremely low, on the order of about 50 to 100 nM, compared to extracellular concentration levels (on the order of about 1.5 mM). Basal intracellular Ca^{2+} concentrations are set by the cell's calcium buffering mechanisms, and free Ca^{2+} is taken up in the endoplasmic reticulum (ER), which serves as the cell's "calcium warehouse." The opening of NMDA channels produces a significant rise in intracellular Ca^{2+} levels which, in sufficient quantity, kick off the metabotropic biochemical cascade reactions now thought to be responsible for NMDA-mediated postsynaptic LTP.

Although today there is no doubt that NMDA-receptor-dependent LTP exists, it is less clear whether there is one or more than one form of NMDA-receptor-dependent LTP process [MALE]. The research of Kandel and his co-workers [CHAI], [KAND3] has shed much light on at least some of the biochemical mechanisms involved. We thus have at least one basic qualitative model for an LTP process, although little has been published in regard to turning this into a quantitative model. Figure 12.1 illustrates the signal processing cascades in the Kandel model.

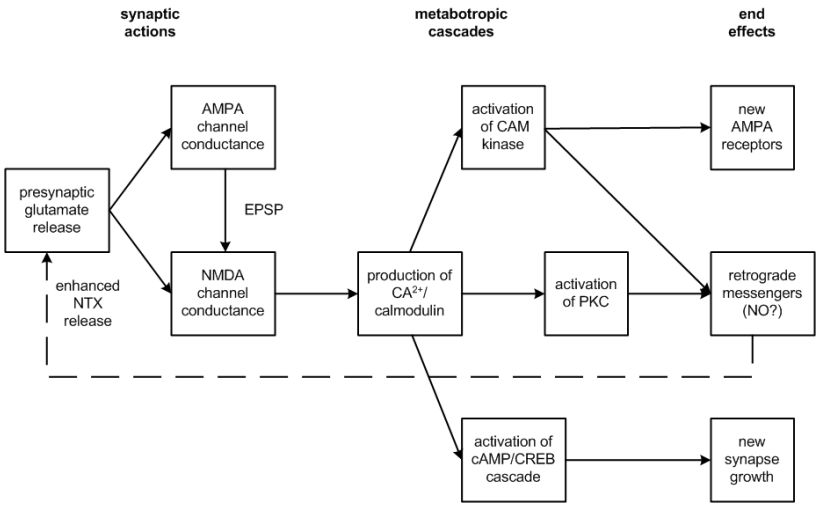


Figure 12.1: Metabotropic cascades thought to be responsible for NMDA-mediated LTP. CAM = Ca^{2+}/calmodulin kinase; PKC = protein kinase C; cAMP = cyclic AMP; CREB = cAMP response element binding protein; NTX = neurotransmitter (glutamate). EPSP = excitatory postsynaptic potential; NO = nitrous oxide.

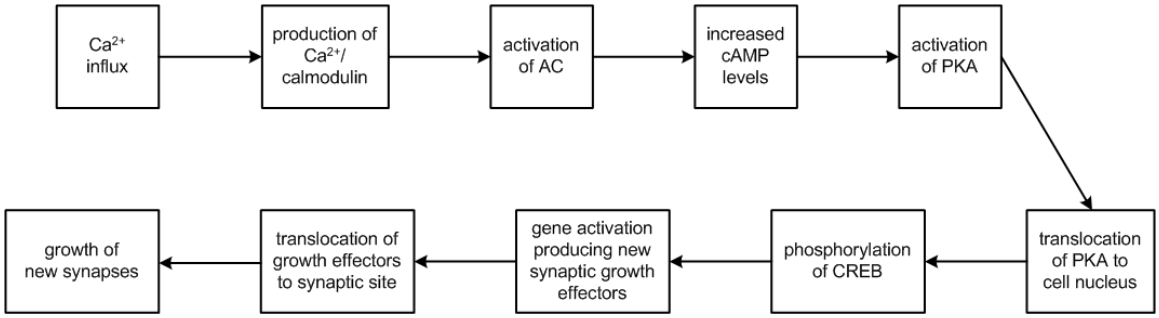


Figure 12.2: The cAMP/CREB metabotropic cascade leading to new synaptic growth in NMDA-mediated LTP. AC = adenylyl cyclase; PKA = protein kinase A.

Free Ca^{2+} introduced into the cell from NMDA channel current binds with calmodulin, a calcium-sensing compound, to form Ca^{2+}/calmodulin. This compound activates a number of different enzymes, called ***protein kinases***, thought to be involved in both short-term synaptic plasticity (elastic modulation) and long-term synaptic plasticity [SCHW2]. Ca^{2+}/calmodulin is called a ***second messenger*** in a metabotropic reaction; the protein kinases are called ***secondary effectors***. The three protein kinases involved in the NMDA-mediated LTP process are the CaM kinase (Ca^{2+}/calmodulin protein kinases) [HANS], PKC (protein kinase C) [TANA], and PKA (protein kinase A) [FRAN].

The end effect of the CaM cascade is a redistribution of AMPA receptors, placing more of these in the synapse [LÜSC]. AMPA receptors have to be recycled in the cell over a relatively short period of time. This has led to the hypothesis that an available pool of AMPA receptors is set up by the cell [MALE], an idea not altogether dissimilar to that of the available pool of vesicles in the model of the presynaptic terminal. Under this model, CaM catalyzes the transport of ready AMPA receptor proteins to (and from) the cell membrane.

The PKC cascade model is a bit more speculative. There is considerable evidence pointing to the involvement of PKC in LTP [NICO], but the precise mechanism by which it acts is not firmly proved. There is experimental evidence for the existence of ***retrograde messengers***, i.e. small molecules, produced by reactions catalyzed by PKC, which diffuse back to the presynaptic terminal [MALE]. There, the hypothesis has it, they stimulate some kind of chemical reaction that leads to enhancement of neurotransmitter release. Nitrous oxide (NO) is postulated to be one possible retrograde messenger. The cell membrane is transparent to NO, meaning that NO produced in the postsynaptic cell can freely diffuse out to other nearby cells. Nonetheless, clear evidence confirming the retrograde messenger hypothesis has not yet been reported, so this part of the model must presently be regarded as merely a probable mechanism for LTP.

Both metabotropic reactions just described are related to what is called ***early LTP***. Early LTP is a transient effect lasting 1 to 3 hours. The third metabotropic cascade in figure 12.1 produces

LTP lasting for at least 24 hours. This consolidated form of LTP is called ***late LTP*** [KAND3]. The cAMP/CREB cascade, illustrated in more detail in figure 12.2, leads to the growth of *new synapses* at the site of LTP induction. This mechanism affects both the pre- and the post-synaptic cells. The presynaptic axon must produce a new presynaptic terminal, and the postsynaptic cell must develop a new postsynaptic density to mate with it [LÜSC].

The cAMP/CREB cascade begins when Ca^{2+}/calmodulin activates adenylyl cyclase (AC), an enzyme that catalyzes the conversion of intracellular ATP (adenosine triphosphate) into the cyclic form of adenosine monophosphate (cAMP). This leads to the activation of the cAMP-dependent protein kinase (PKA). PKA is then translocated to the nucleus of the cell, where it catalyzes a series of reactions leading to phosphorylation of CREB (the cAMP response element binding protein) [CHAI]. CREB activates genes responsible for regulating the synthesis of new proteins. This produces new growth effectors that are, in turn, translocated back to the region of the original synaptic site, where they cause the growth of the new synapse [KAND3].

The model just described is the currently most widely accepted model of NMDA-mediated LTP. It is no doubt clear that this is a qualitative model of the process. More detailed quantitative modeling is, of course, a research subject in progress. One possible approach to modeling this process for augmentation of a Hodgkin-Huxley-like neuron model is to employ the Linvill modeling schema introduced earlier in this textbook.

§3.2 Non-NMDA-Mediated LTP

It is known that NMDA-mediated LTP is not the only form of LTP [NICO], [MALE]. LTP that does not require NMDA receptor activation is known to occur at mossy fiber synapses in the hippocampus, at synapses between parallel fibers and Purkinje cells in the cerebellum, and at corticothalamic synapses in the neocortex. Virtually all investigators agree the site of expression for this form of LTP is presynaptic. This form of LTP is less extensively investigated than the other form just discussed, but evidence suggests that, once again, Ca^{2+}/calmodulin is involved, and that cAMP and PKA are again factors in the process. It appears to be the case that one important action of PKA is to modify some aspect of the synaptic vesicle cycle or perhaps the vesicle release machinery itself [MALE]. To effect long-lasting changes in the efficiency of neurotransmitter release, it would seem necessary for PKA to become ***persistently active*** in the terminal. It is known from the work of Kandel et al. that a metabotropic cascade process similar to the cAMP/CREB signaling cascade discussed above leads to the production of ubiquitin hydrolase (in the cell nucleus) in *Aplysia*, and this compound is capable of persistently phosphorylating (that is, persistently activating) PKA enzymes located back at the presynaptic terminal [KAND3], [CHAI].

§ 3.3 Long Term Depression

It appears to be the case that any synapse capable of expressing LTP is also capable of expressing LTD and vice versa. In other words, it appears likely that there are no synapses that express only LTP or only LTD [BEAR]. It is therefore unsurprising that intracellular Ca^{2+} should again be found a focal point for initiation of LTD. However, because LTD is the precise opposite of LTP, it seems clear that its underlying metabotropic mechanism must in some way be different from those of LTP discussed in the previous section.

The two best-understood forms of LTD are those of the cerebellar parallel fiber-Purkinje cell synapse and the hippocampal Schaffer collateral/commissural-CA1 pyramidal cell synapse. Mature Purkinje cells do not express functional NMDA receptors, but they do express metabotropic glutamate receptors (MGluRs). These MGluRs do not produce ionotropic currents, but they do initiate *two* metabotropic signaling cascades. The first produces the second messenger chemical IP3 (inositol-1,4,5-triphosphate), which binds to receptors located in the endoplasmic reticulum and thereby stimulates release of Ca^{2+} from the cell's "calcium warehouse." The second produces the second messenger DAG (diacylglycerol), which activates the secondary effector protein kinase C (PKC). Thus, so far we have two of the factors present in this cascade that were involved in Kandel's LTP model. Further evidence shows that MGluR activation by itself is not sufficient to produce LTD. Another factor necessary for LTD induction is the opening of voltage-gated calcium channels, which is, of course, an effect that can be brought about by the firing of an action potential by the Purkinje cell. This is something that is, on the face of it, rather strange since a rise in intracellular free Ca^{2+} triggered by the MGluR cascade is certainly sufficient to provide a significant source of calcium by itself. Nonetheless, the fact remains. In addition, it is also found that an influx of Na^{+} current is *also* necessary to evoke LTD in the Purkinje cell. This current is, of course, provided by the ionotropic AMPA channels in the synapse, but it remains unclear why this influx is needed to evoke LTD.

These confusing, and sometimes contradictory, experimental findings illustrate that we have much yet to be discovered before being able to claim an understanding of cerebellar LTD on par with the models for LTP discussed earlier. A more detailed discussion of this topic can be found in [BEAR]. In view of the unsettled nature of non-NMDA-mediated LTD at the present time, we will forego an attempt to summarize this process as we did earlier for LTP. The simple fact is that we currently lack a sufficiently established qualitative model for the biochemical mechanism or mechanisms at work here.

LTD can also be mediated by NMDA receptor channels. Although at first glance it might seem very contradictory that NMDA currents could produce either LTP *or* LTD, the critical

factor here seems to be the *amount* of Ca^{2+} entering the cell via the NMDA channels. In the case of LTP, a *strong* stimulus is required, which produces a large Ca^{2+} current influx. However, some lower level of Ca^{2+} current flow does take place for EPSP levels insufficient to evoke an action potential. The obvious implication is that it is not the mere presence of free Ca^{2+} in the post-synaptic cell that matters, but rather the *concentration* of Ca^{2+}. Experiments have pointed to the existence of at least three different mechanisms for LTD, all of which require Ca^{2+} concentrations to be *above* some critical threshold for LTD induction but *below* some other, higher threshold for LTP induction [BEAR].

§ 4. The BCM Model

While the *mechanistic* modeling picture of long term synaptic plasticity still appears somewhat cloudy, more significant success has been achieved by *functional* modeling. One of the most important of these models was proposed in 1982 by Bienenstock, Cooper, and Munro. It is known as the BCM model [BIEN]. Indeed, the BCM model was an important inspirational idea for researchers investigating LTD [BEAR], which is something to be highly prized in a functional model: function suggestive of mechanism.

In 1982 the existence of LTP/LTD was already known but the physiological mechanisms were not. Bienenstock et al. therefore based their model on phenomenological signaling characteristics. They used an Instar map model, which they called a "neuron model with 'ideal' synapses" – a terminology they explained could imply "a complex system including perhaps several interneurons." The activity levels of the Instar's inputs and outputs were taken to be firing rates relative to the level of average spontaneous activity. The Instar's weights W represented the "efficacy" of "synaptic connections" to the "neuron."

The novel ideas in the BCM weight adaptation rule were: (1) weight modification was a function of average activity levels determined over some time interval much larger than the membrane time constant of a neuron; this type of "average" is elsewhere called a "moving average" because average output activity $\langle y \rangle$ is actually a function of time t; (2) the weight modification function is a function of a weight modification threshold θ; and (3) θ is itself a function of a running average of the activity levels; this has since come to be called a ***sliding threshold***. Property (3) is essential for the operation of the adaptation operation. If θ is held to a fixed constant value, the adaptation is generally unstable.

The BCM rule is actually a rule schema. This is to say Bienenstock et al. did not provide one specific adaptation rule but rather a general form an adaptation rule should obey. This was expressed in [BIEN] through the use of an unspecified adaptation function ϕ and an unspecified

function for $\theta(t)$. Furthermore, they did not specify how temporal averages, $\langle y \rangle$ and $\langle \mathbf{X} \rangle$, were to be determined for the Instar's output and input activities. There are a number of different ways in which a "running average" over some time window can be defined and computed. These include the un-weighted moving average (where each past value is given equal weighting in determining the overall moving average), the weighted moving average (where typically "older" past values are given less weight than "newer" past values in determining the average; this is sometimes called "averaging with a forgetting factor"), and a recurrent form of averaging called the autoregressive moving average or ARMA. The only hard requirement is that $\langle y \rangle$ should change less rapidly than $y(t)$ and $\langle \mathbf{X} \rangle$ should change less rapidly than $\mathbf{X}(t)$. Thus, BCM is a rule schema and there are many possible BCM "rules."

Bienenstock et al. did make one specialization, namely that it was sufficient to determine $\langle y \rangle$ in terms of $\langle \mathbf{X} \rangle$, i.e. $\langle y \rangle = g(W^T(t) \cdot \langle \mathbf{X} \rangle)$ where g is the Instar activation function. The only requirements placed on g are that g should be a continuous and monotonic function. Of the activation functions introduced in chapter 11, this requirement rules out only the discontinuous functions and the radial basis function. The rule for $\langle y \rangle$ requires that the weights W change slowly relative to changes in $\mathbf{X}(t)$ so that $\langle y \rangle$ changes slowly relative to $y(t)$.

The BCM adaptation schema implements what Bienenstock et al. termed "temporal competition between input patterns." This vaguely explained term means, in effect, that weight changes depend on "average" activities compared to the sliding threshold $\theta(t)$. To the extent that one could say the Instar "learns the average activities" in the setting of W, BCM constitutes a form of what is called ***unsupervised adaptation***. As is common in the literature dealing with unsupervised adaptation algorithms, Bienenstock et al. expressed the BCM rule schema in the form of a differential equation,

$$\frac{d}{dt}(W(t)) = \phi\left[y(t), \langle y \rangle, \theta(t)\right] \cdot \mathbf{X}(t) - \varepsilon \cdot W(t)$$

where $\varepsilon > 0$ is a small constant called the "uniform decay term constant." Its presence in the equation is necessary to ensure stability in the differential equation.

Because so many terms in BCM are left unspecified, it is typically unnecessary to go to a great degree of exactness in converting the BCM weight equation into difference equation form for computer implementation. Rather, it is perfectly legitimate to pose a "BCM rule" directly in difference equation form. The general difference equation form for BCM is

$$W(t+1) = (1-\gamma) \cdot W(t) + \phi\left[y(t), \langle y \rangle, \theta(t)\right] \cdot \mathbf{X}(t) \qquad (12.1)$$

where $\gamma > 0$ serves as the "uniform decay term" constant. The adaptation function ϕ is a scalar-valued function multiplying input vector **X**.

The sliding threshold for weight modification is generally some nonlinear function of y, and is typically a direct function of $\langle y \rangle$, the running average of $y(t)$. Most commonly $\langle y \rangle = g(W^T(t) \cdot \langle \mathbf{X} \rangle)$ is used and $\theta(t)$ is described by a difference equation of the form

$$\theta(t+1) = (1-\alpha_2)\cdot\theta(t) + \alpha_2 \cdot (\langle y \rangle)^p \tag{12.2}$$

where $0 < \alpha_2 < 1$ and $p \geq 2$ are fixed constants. The simplest form for (12.1) is

$$W(t+1) = (1-\gamma)\cdot W(t) + \alpha_1 \cdot (y(t) - \theta(t))\cdot \mathbf{X}(t) \tag{12.3}$$

when the activation function g is non-negative (e.g. the unipolar sigmoid function), $0 < \alpha_1 < 1$ and $\alpha_1 < \alpha_2$. The condition $0 < \gamma < \alpha_1 < \alpha_2$ ensures $\theta(t)$ changes more quickly than W and the adaptation is not unduly dominated by γ. Typically α_2 is on the order of about three times larger than α_1 and γ is on the order of about five times smaller than α_1.

When g is a non-negative activation function, the elements of **X** will likewise be non-negative. It might therefore appear that equation (12.3) then will decrease W whenever $y - \theta < 0$, and increase W whenever $y - \theta > 0$. However, this is not the case since the values approached by the elements of W are also functions of the values of the elements of **X**. Figure 12.3 illustrates the dynamic for a two-input example with constant $\mathbf{X} = [0.15 \quad 0.7]^T$. The Instar used a unipolar sigmoid activation function with threshold $\Theta = 0.5$ and slope parameter $\sigma = 3$. The adaptation parameters for the BCM rule were $\alpha_1 = 0.1$, $\alpha_2 = 0.3$, and $\gamma = 0.02$. The initial conditions are as shown in the figure. Although the weight for the weaker input, w_1, was initially larger than w_2, the figure shows that the adaptation produces a redistribution of the weights such that the larger input comes to have the larger weight associated with it.

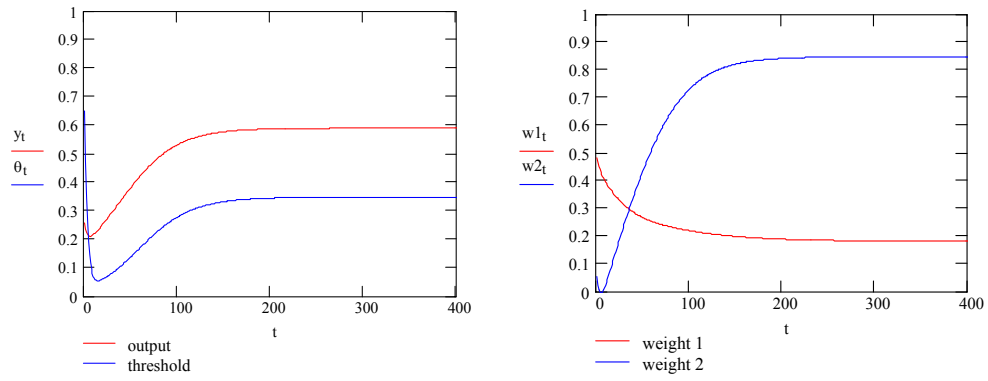


Figure 12.3: Two-input Instar adaptation using the BCM rule with constant inputs. See text for parameters.

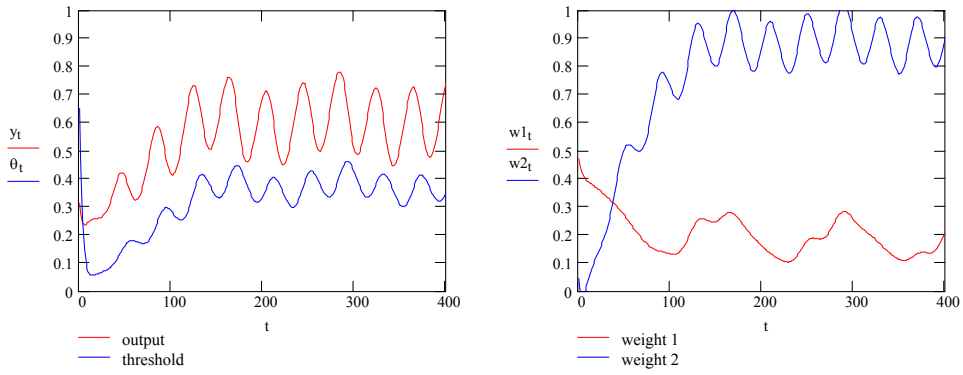


Figure 12.4: BCM rule adaptation of the same Instar when the inputs are time-varying. ⟨**X**⟩ is the same as in the previous example, but the two input signals have sinusoidal variations about the mean value.

The adapted Instar response from figure 12.3 can be said to be "selective" in the sense that the final weight vector W comes to weight the more active input (x_2) more heavily than the lower-activity input x_1. With a sigmoid function threshold of $\Theta = 0.5$, x_1 would be regarded as having lower-than-background activity while x_2 would be said to have higher-than-background activity. Bienenstock et al. defined a statistical selectivity metric based on the assumption **X** was a vector of random variables following some probability distribution. Their metric is

$$S \overset{\Delta}{=} 1 - \frac{\text{mean value of } y \text{ over distribution of } \mathbf{X}}{\text{maximum value of } y \text{ over distribution of } \mathbf{X}} \ .$$

For the example system of figure 12.3, if we assume x_1 and x_2 are statistically independent and each is uniformly distributed over the range from 0 to 1, the selectivity metric function is

$$S = 1 - \frac{g\left(0.5 \cdot \left(w_1 + w_2\right)\right)}{g\left(w_1 + w_2\right)}$$

where g is the unipolar sigmoid function used for the Instar. Using the final weight values for the simulation in figure 12.3, the selectivity in this example is $S = 0.385$.

Bienenstock et al. presented several theorems governing convergence of the algorithm, the selectivity achievable with it, and whether or not the final weight settings would become stable – that is, unresponsive to statistical variations in **X** once the adaptation process had run its course. The theorems are based on several conditions, most notably the assumptions that the probability distributions for **X** are stationary (not changing over time) and that the Instar was adequately exposed to a statistically sufficient number of input cases **X**. Figure 12.4 presents the results of another simulation using the same Instar, adaptation parameters and ⟨**X**⟩ as before. This time, however, x_1 and x_2 varied sinusoidally about their mean values. We see that the parameters of the

system show time-varying responses to these time-varying inputs, although the average of the responses tracks figure 12.3 reasonably well. (The inputs in this case are said to be statistically "cyclostationary"). Using a snapshot value for W at the end of the simulation, the BCM selectivity is $S = 0.369$, which is reasonably close to the first case.

However, the fact that the adaptation *does* track time-varying changes in the input signal presents certain practical considerations in using BCM in a network system model. In [BIEN] the biological system under consideration was the region of the primary visual cortex where neural structures, called ocular dominance columns, form during early post-natal brain development. There is a critical development period in this cortex, during which neural connections develop and become permanent. If a baby is deprived of appropriate stimulus during this period, these crucial structures do not develop and serious and permanent visual impairments results. Mathematically, this situation is equivalent to setting α_1, α_2, and $\gamma = 0$ outside the time span of the critical development period, and setting them to their non-zero values inside this time span. Thus, the ***adaptation stability*** theorems in [BIEN] implicitly assume appropriately time-varying changes in the adaptation parameters such that W is established for a specific environment at a specific stage of development.

In BCM-based adaptation for other network system models, where a "critical development period" is not part of the system phenomena being modeled, non-stationary input statistics for $\mathbf{X}$ will have the consequence that W will track these statistical changes, "forgetting" whatever W settings it "learned" earlier. All unsupervised adaptation algorithms must make some kind of tradeoff between ***plasticity*** (the ability to adaptively respond to input signals) and ***stability*** (the ability to develop robust final weight settings that no longer adapt to changes in input signals). This is because adaptation functions favoring plasticity tend to be detrimental to stability, and vice versa. The BCM rule favors plasticity at the expense of stability in the face of nonstationary input statistics unless the adaptation parameters α_1, α_2, and γ are themselves controlled in some appropriate fashion (i.e., unless they are not strictly *constants*). Control functions added to the network system for this purpose are said to address the ***stability-plasticity dilemma*** (if, that is, they are successful).

Another issue the BCM rule has, in common with many other adaptation algorithm schemes, arises when we are dealing with a network system with multiple Instar nodes. Suppose we have N Instars arranged in a single layer, all receiving the same input vector $\mathbf{X}$, and let us further suppose these Instars do not connect to each other. Let us further assume their initial weight vectors W_n are not identical, nor are their initial values for θ_n (the weight modification thresholds). (If all the Instars have identical initial conditions, they will all respond exactly the same way, assuming

each has the same adaptation parameters). Depending on the statistics of the input signals **X** and the order in which the inputs are presented, at least some of the *N* Instars might achieve different final stable weight settings eventually [BIEN]. On the other hand, it is also possible they might all reach exactly the same final weight setting. Except in very special circumstances, it is usually impossible to predict a priori which of these outcomes will occur.

Bienenstock et al. provided a brief discussion about multiple-Instar networks, and they tell us they have achieved some unreported test cases where the Instars reached different and stable weight settings – and could therefore be said to have succeeded in classifying their input space, analogously to our earlier discussion in this chapter. We are told this was achieved by providing fixed (non-adaptive) feedback weights interconnecting the Instars. However, this discussion in [BIEN] is dialectic, non-quantitative, non-specific, and is insufficient to form a basis for generalization. This problem – that is, **adaptation of a *network system***, as opposed to merely adapting a network *node* – is one of the principal issues of interest in neural network theory. There does not currently exist any general theorem or solution, applicable in all circumstances, for the general problem of network system adaptation.

Equation (12.2) is not in the most general form proposed by Bienenstock et al. An adaptation threshold rule in more general form replaces (12.2) by

$$\theta(t+1) = (1-\alpha_2)\cdot\theta(t) + \alpha_2 \cdot \left(\frac{\langle y \rangle}{c_0}\right)^p \cdot \langle y \rangle \qquad (12.4)$$

where c_0 is a positive constant. Figure 12.5 illustrates the behavior of this adaptation rule for $\theta(t)$ for c_0 = 0.65, p = 2, and the other simulation parameters unchanged from the previous cases. Comparing this to figure 12.4, the main difference made by (12.4) is a reduction in average levels

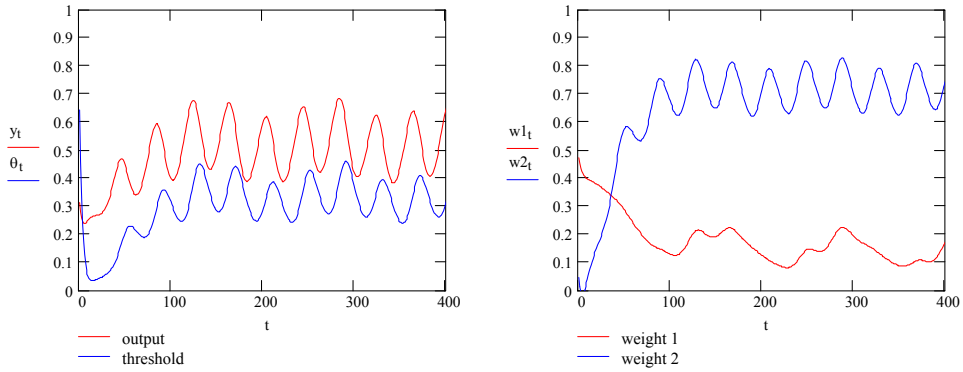


Figure 12.5: Simulation of BCM adaptation under the same conditions as for figure 12.3 using rule (12.4) with c_0 = 0.65 and p = 2.

for $y(t)$ and W. In effect, parameter c_0 determines the point where $y(t) - \theta(t)$ changes sign in the weight adaptation, leaving α_2 to be used exclusively to control the rate of adaptation. For the special case where the second term on the right-hand side of (12.4) is a constant, the asymptotic value for $\theta(t)$ as t becomes very large is easily shown to approach

$$\lim_{t\to\infty} \theta(t) \to \left(\frac{\langle y \rangle}{c_0} \right)^p \langle y \rangle \tag{12.5}$$

independently of the initial condition on $\theta(t)$ and independently of α_2, provided the magnitude of α_2 remains less than 1.

§ 5. The Instar Rule

The BCM rule developed from efforts to find stable adaptation methods capturing the experimentally observed signaling features of LTP and LTD. These efforts were inspired to a large degree by Hebb's postulate. Cooper, Lieberman, and Oja had proposed a rule that was a direct precursor to the BCM rule in 1979 [COOP]. The Cooper-Lieberman-Oja rule tended to lack robustness and had problems with stability and selectivity owing to their use of a fixed threshold for adaptation. The BCM rule, with its sliding threshold mechanism, overcame these early issues. It is, however, a mathematical answer to a mathematical problem. One could justly ask what physiological basis there could be for the BCM adaptation rule.

Bienenstock et al. discussed this briefly in their paper, but it would have to be admitted that this discussion was a plausibility argument rather than a physiological argument. The fact is that the BCM rule is phenomenological and is more "inspired" by biology than "descriptive" of biology. The same can be said for many adaptation rules in use today. Some theorists note a degree of similarity between the BCM rule and other more physiologically-motivated (but still phenomenological) adaptation models. An example of this is a spiking-neuron-model adaptation rule based on what is called the ***calcium control hypothesis*** [GERS1: 362-383]. This adaptation rule has features similar to, and tends to bolster the plausibility argument for, the BCM rule but does not demonstrate the biological fealty of the model.

In this section we will look at an adaptation rule treated by Grossberg several years before the BCM rule. We will call it the ***Instar adaptation rule*** or IAR. Letting $X = [x_1\ x_2\ .\ .\ .\ x_N]^T$ be the vector of inputs to the Instar, $W = [w_1\ w_2\ .\ .\ .\ w_N]^T$ be the weight vector, and $y = g(X,W)$ be its output activity level, the IAR in difference equation form is

$$W(t+\Delta t) = W(t) + \eta \cdot (X(t) - W(t)) \cdot y(t) \tag{12.6}$$

where η is called the adaptation rate parameter and $0 \le \eta < 1$. We easily see from (12.6) that when $X(t) = W(t)$ no change in the weight settings occurs. The IAR acts to produce a set of weights that constitutes a copy of the input pattern. For this reason, the pattern X is said to be "stored in the weights" and the Instar is said to "learn the input pattern." Thus Grossberg refers to the weight settings W_n in a network of Instars as the "long term memory" or LTM of the network. (One should bear in mind that this "memory" terminology is made in a mathematical context and is not to be confused with the psychological meaning of the word "memory").

Relatively few constraining conditions are necessary for (12.6) to have stable fixed-point solutions. The analysis is non-trivial owing to the fact (12.6) is a nonlinear difference equation. Nonetheless, there are a few features of (12.6) we can bring into the light without resort to complicated nonlinear analysis. Rewrite (12.6) as

$$W(t+\Delta t) = \left[1 - \eta \cdot y(t)\right] W(t) + \eta \cdot y(t) X(t). \tag{12.7}$$

The order of the nonlinearity in (12.7) depends on the activation function. For example, if the activation function is the simple linear function $y = X^T W$, (12.7) can be written as

$$W(t+\Delta t) = \left[\mathbf{I} + \eta \cdot X(t) X^T(t)\right] \cdot W(t) - \eta \cdot W(t) W^T(t) X(t)$$

which is a quadratic function of the weights and the input signals. (Here **I** is the identity matrix). Other choices for activation function g can lead to a higher-than-quadratic equation. This makes general analytic solutions of the difference equation (12.7) impossible to obtain in closed form. However, some useful information can be obtained from some simple cases.

Suppose the input stimulus is held constant, i.e. $X(t) = X$ is a constant-valued vector. If there is a fixed point solution for (12.6), it is defined by $W(t + \Delta t) = W(t)$, for which we obtain $W = X$ as noted before. The question is: Is this a *stable* fixed point solution? That is, if W at some time t is given by $W = X + \Delta W$, where ΔW is some perturbation, will (12.6) converge to $W = X$ or will it diverge away from this solution?

To explore this, let us set $W(t) = X + \Delta W(t)$ and $W(t + \Delta t) = X + \Delta W(t + \Delta t)$ and insert these into (12.7). From this we obtain

$$X + \Delta W(t+\Delta t) = X + \Delta W(t) + \eta \cdot \left(X - X - \Delta W(t)\right) \cdot y(t).$$

After some minor algebraic manipulation, this reduces to

$$\Delta W(t+\Delta t) = \left[1 - \eta \cdot y(t)\right] \cdot \Delta W(t).$$

Our interest now focuses on the scalar term on the right-hand side of this expression. A sufficient

condition for this equation to converge to $\Delta W = [\ 0\ ]$ is to have the magnitude of this term strictly less than 1 at all time steps. Stated symbolically, the condition is satisfied if

$$0 < \eta \cdot y(t) < 2\,. \tag{12.8}$$

This condition is guaranteed if η is constrained as $0 < \eta < 1$, $y_{\min}(t) \geq 0$ and $y_{\max}(t) < 2$. (This is a sufficient condition; it is not a necessary condition). At first glance, this constraint might seem very artificial and non-biological. However, it really is not. First, if signal variables x are non-negative, any saturating non-negative activation function will do. Even if g is non-saturating, the condition can be enforced by a suitable *normalization* of vectors X and W. Grossberg has shown that a normalization condition on X can be automatically produced by a relatively simple single-layer Instar network [GROSS5]. Following Grossberg's terminology, we will call such a single-layer network a ***shunting network***. If the initial settings of the Instar weights, $W(0)$, is set up so that $W(0)$ likewise meets the normalization constraint (with W a non-zero vector), then an Instar fed by a shunting network and undergoing IAR adaptation will converge to X when X is held constant.

Is the idea of a shunting network biologically plausible? Grossberg argues that it is. To follow this argument, two things are noteworthy. First, the nodes within a Grossberg shunting network are not simple Instars as we have earlier defined them. We will call them ***shunting node Instars***. The details of this model will be explained later, but suffice it to say for now that they are linear, time-varying Instars in which the excitation variable s is governed by a state variable equation (i.e., it is not merely the sum of inputs; thus, the node is said to "have short-term memory"). Second, an individual biological neuron is not a shunting node Instar. However, Grossberg's networks work on the scale of a map model, not a neuron model. As we have already seen earlier in this textbook, biological structures at this scale are networks made up of large numbers of interconnected excitatory and inhibitory neurons. It is not at all unfeasible to make the hypothesis that such a network can act to limit the total state of excitation in its population of neurons, and this is all that a shunting node Instar model asks.

As for IAR adaptation, this adaptation scheme is likewise true to the phenomenological dynamics of LTP/LTD. In the presence of a persistent input X, the weights of the Instar become a mirror of this input (provided $\eta \neq 0$), and thus large inputs x_i eventually produce a large w_i (LTP), and small x_i produce a small w_i (LTD). The rate at which the Instar adaptation approaches this steady-state solution depends on the adaptation rate factor η and on the total amount of excitation s produced by the input vector. Like the BCM rule, the IAR will attempt to "follow" the changes in a time-varying sequence $X(t)$ of input vectors.

If the adaptation rate is too fast, the Instar weights will not converge to a single, stable fixed-point value. The weights in this case are said to be "too plastic." If the Instar is to have any chance of obtaining relatively stable "learning" (again, "learning" in a mathematical, not a psychological, context), the adaptation rate must be slow relative to the time variations in $X(t)$. This property of adaptation turns out to be a rather general feature of most adaptation methods and it goes to the heart of the stability- plasticity dilemma. With a slow adaptation rate, "learning" is statistical, that is, the weight adaptation tends toward the mean value of the probability distribution of the inputs $X(t)$, assuming this distribution does not itself change over time. (Such a probability distribution is said to be "stationary").

To see the general character of this process, let us briefly examine the statistics of equation (12.6). The statistical ***expectation*** or "expected value" of a random variable u with probability distribution function $p(u)$ is defined as $E\{u\} = \int u \cdot p(u) du$ where the integral is taken over all possible values of u. Note that time does not appear as a parameter in this expression; for this reason, the expectation is also often called an ***ensemble average*** because the statistic is the mean value of everything that *could* happen at any given time t. If the probability distribution function is independent of time (stationary), then the expectation is also independent of time. Applying this operation to both sides of (12.6) gives us

$$E\{W(t+\Delta t)\} = E\{W(t)\} + \eta \cdot E\{(X(t) - W(t)) \cdot y(t)\} .$$

Now, without further assumptions this expression defies closed-form solution. Therefore let us assume the adaptation rate is slow relative to the rate at which $X(t)$ can change, and let us further assume that successive time values of $X(t)$ are statistically independent. Then since the present value of $W(t)$ depends only on *past* values of X, $W(t)$ and $X(t)$ are likewise statistically independent at time t. This is called ***the independence assumption***, and was first introduced into the theory of adaptive systems by Widrow [WIDR5].[1] If u and v are two statistically independent random variables, then $E\{uv\} = E\{u\} \cdot E\{v\}$. Applying this to the expression above and noting that $E\{W(t + \Delta t)\} = E\{W(t)\}$ if the statistics of the overall system are stationary, we obtain

$$E\{W(t) \cdot y(t)\} = E\{X(t) \cdot y(t)\} .$$

A quantity $E\{u \cdot v\}$ is called the ***cross correlation*** of u and v. We can see from the expression above that what is enforced by the IAR is equality between the cross correlations of the products of W and X with the output activity. Because $y(t)$ is expressly *not* independent of W and X, we

[1] Strictly speaking, the independence assumption is rarely true of an adaptive system. However, under the condition of slow adaptation under stationary probabilities, many years of research in adaptive systems has shown it to lead to results that, while approximate, are very accurate.

cannot merely separate the terms as $E\{W\}\cdot E\{y\} = E\{X\}\cdot E\{y\}$. To obtain an approximate understanding of the result, let $y = W^T X$. Then invoking the independence assumption gives us

$$E\left\{XX^T\right\}\cdot E\{W\} = E\left\{WW^T\right\}\cdot E\{X\} \Rightarrow E\{W\} = \left[E\left\{XX^T\right\}\right]^{-1}\cdot E\left\{WW^T\right\}\cdot E\{X\}$$

provided the indicated matrix inverse exists (which will be the case in practical situations).

Under the conditions stated for this derivation, the product of the two terms multiplying $E\{X\}$ will approximate the identity matrix and $E\{W\} \cong E\{X\}$ in the steady-state. In most practical circumstances, the cancellation of the two autocorrelation matrices will not be exact and so W will only approximately equal $E\{X\}$, but in such cases the error is usually quite small. In somewhat more technical language, the *bias* in the weight vector solution is said to be *bounded* [McCA]. Often, in practical circumstances, this bound can be quite tight.

§ 6. Network Adaptation

The remarks made earlier concerning network adaptation in the case of the BCM rule apply also to the IAR and, indeed, to adaptation in all network systems. Consider a single layer of non-interconnected Instars all receiving the same input vector X. If all the Instars are allowed to adapt in response to all inputs $X(t)$, then, as the statistical analysis presented above shows, they will all adapt to the same steady-state solution. To a limited degree this may be biologically useful; redundant networks protect the organism from suffering disastrous consequences from neural cell death (up to a point). However, it is clearly undesirable – and biologically unrealistic – to have complete redundancy in the adaptive solutions of all Instars in a network system. Something must be present in the system to prevent this.

It is clear merely from the multiplicity of different network adaptation schemes neural network theory has brought out since the late 1980s that there is more than one possible approach for this problem. We will not attempt to survey the entire field in this textbook. That would be an undertaking that would either be staggeringly huge (if in-depth coverage were to be given to each major method) or disappointingly vague (if we were merely to mention each with a skeletal outline of the method used). What we can and will discuss is some of the general reasonings by which different network adaptation methods are guided. Beginning in chapter 13, we will also go into a select few network methods that appear to be biologically pertinent.

When one considers a particular biological network system, from a broad perspective it can be said that such a system receives a particular set of input tract signals (usually originating from some other network system within the central nervous system as a whole) and produces a set of output tract signals (usually sent to other network systems in the CNS). To focus our discussion,

we will restrict our attention to the network system and postpone consideration of interactions among different network systems. We can then regard the set of input tract signals as constituting a mathematical ***input space***, and regard the signals within the network system, including also its output signal tracts, as constituting a ***solution space*** in the abstract sense that one can regard the maps within the network system has "solving" some signal processing task, i.e. responding "appropriately" to the input signals. When we consider just those signal tracts constituting the outputs of the network system, we can say these output signals constitute an ***output space***. If a particular map within the network system responds with strong activity to some particular set of inputs constituting a subset within the input space, and responds weakly or not at all to signals not belonging to this subset, we can call the subset of signals that evoke strong activation responses from the map the ***response space*** of the map.

Figure 12.6 is a conceptual representation of this idea of partitioning spaces within an input space. Let us assume the dashed lines in the figure delimit regions of the input space where some network mechanism has determined that a particular Instar will respond by adapting its weights when an input vector X falls within this region. We will suppose the symbols **x** represent actual instances where a particular input vector has occurred. We will further suppose the symbols **o** are

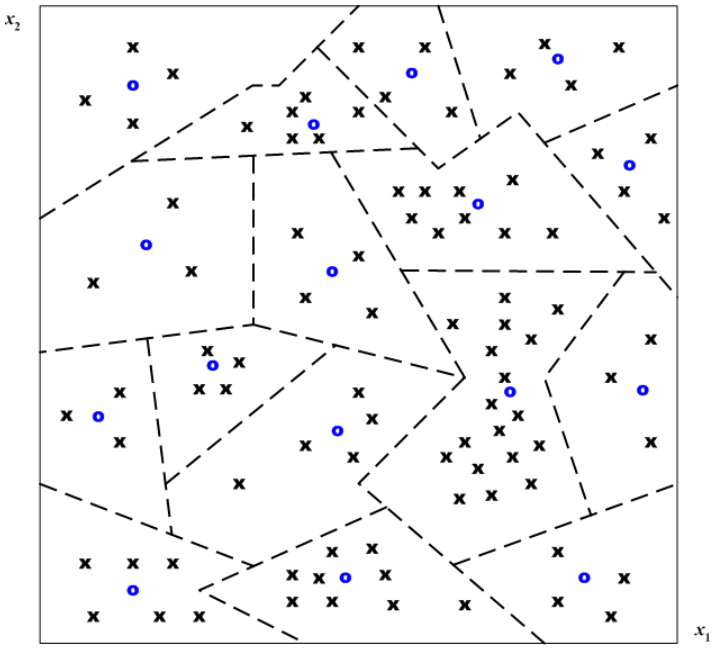


Figure 12.6: Conceptual diagram of space partitioning in a network system of Instars. The dashed boundary lines denote regions for which some network decision mechanism has selected a particular Instar to respond to an input by adapting its weight values. **x** denotes an actual input vector applied to the network. **o** denotes an Instar weight vector, which is presumed to have taken on a value representative of the estimated mean value of the input vectors to which the Instar has adapted. For discussion purposes, we assume this figure represents one planar projection of a multi-dimensional input space with more than 2 dimensions.

representations of the Instars' weight vectors W. Because of the hypothetical selection mechanism determining which Instar will adapt in any given case, the adaptation rule for a specific Instar does not "see" the entire probability distribution of X. Rather, it "sees" only some subset of this probability distribution, namely the one corresponding to the adaptation region "belonging" to that Instar. Given enough adaptation epochs, the Instar will eventually adapt its W vector to the expected value of the X vectors lying within the Instar's designated adaptation region.

The resulting steady-state W vector for a particular Instar is said to be a ***prototype vector*** describing its region of the input space. In effect, and under certain reasonably benign conditions, the W vector can be said to "represent" *all* the input vectors in its region in a statistical sense. In principle, then, when the next X vector is received, the Instar with the closest W vector, i.e. the W vector for which $||W - X||^2$ is the least, will respond with the strongest positive activation if certain conditions apply to the vectors X.

To see this we need some concepts from analytic geometry. Let us assume figure 12.6 is a two-dimensional plane within some $N >> 2$ input space, and let us further assume the **x** and **o** symbols in the figure are projections onto this space from vectors lying almost in this plane. Let us further assume that all actual input vectors are normalized so that $||X||^2 = 1$. Assuming the adaptation process has reached steady-state for all the Instars, this likewise implies $||W||^2 = 1$ for each Instar. The scalar quantity $W^T X$ is called the ***dot product*** or ***scalar product*** of W and X in the language of the mathematicians. Let us use the notation $\cos(W, X)$ to denote the cosine of the solid angle between vectors W and X. $\cos(W, X) = 1$ implies the angle is zero and $W = X$. It is a basic property of vectors that $s = W^T X = ||W|| \cdot ||X|| \cdot \cos(W, X)$. Since we have assumed the vectors are of unit length, the largest scalar product s (the excitation variable of the Instar) has the largest magnitude for that Instar for which $\cos(W, X)$ is closest to zero. This means the Instar response will be the strongest for that Instar with weight vector most nearly equal to X. The Instar is said to have the capacity for ***generalization*** in the sense that it most strongly responds to *any* X that is closest to (in the Euclidean sense) its weight vector than to that of any other Instar.

What we have just discussed is, of course, a special case because we have put some restrictive assumptions on the nature of the input space of X vectors. In the terminology of neural network theory, this special case is called the ***classification problem***. You should be able to easily see the similarity of this example in comparison to the discussion of separating boundaries from chapter 11. Classification is one thing artificial neural networks do very well. We can, however, ask if this problem is pertinent to biological signal processing.

To this question, the answer is "yes." There is very solid experimental evidence telling us that neuronal organization in the primary sensory cortices of the brain during the critical development

phase of an organism's life adapts in such a way that specific neural network systems develop a strong response to very particular and restricted kinds of afferent sensory signals, and develop weak or no response to others. The same has been found to be true for some neuronal structures that adaptively form in the cerebellum. While it would be naive and more than a bit rash to over-generalize this and say the same is true for *all* neuronal structures in the CNS, it is nonetheless clear that one important functional task the brain organizes itself to carry out is signal classification. In the language of Piagetian developmental psychology, neural network systems which develop to carry out solutions to the classification problem are said to ***assimilate*** the afferents they classify. Classification theory is arguably the most highly developed topic in neural network theory.

Still, our discussion leaves hanging the question of what hypothetical mechanism was responsible for *selecting* which Instar was to adapt in response to a given afferent signal X. Selection mechanisms constitute an important part of the core of the theory of network adaptation in neural network theory. Indeed, it was lack of progress in precisely this area of the theory that led, in part, to Minsky's and Papert's scathing 1968 critique of the field. Equally, it was the discovery of workable mathematical mechanisms for addressing this problem that led in large part to the widespread "rebirth" of the field in the mid- to late- 1980s. The groundwork for this was set during the "dark age" of neural network research from the end of the sixties to the eighties by pioneers such as Grossberg, Kohonen, Malsburg, and others. chapter 13 begins our discussion of network adaptation theory.

§ 7. The LMS Algorithm

This chapter would not be complete without a discussion of the best-known and most widely used adaptation algorithm, the least-mean-squared or LMS algorithm. Although its biological significance has been called into question by some, most notably by Grossberg [GROSS10], the LMS algorithm has long been a mainstay in the engineering world of artificial neural network theory and adaptive signal processing. It was developed, simultaneously and independently along side Rosenblatt's perceptron rule, from statistical adaptation theory [WIDR6] and is the algorithm used by the special case version of the Instar known as the Adaline [WIDR1].

Adaline was a more or less direct outgrowth of von Neumann's work with the McCulloch-Pitts model; as Widrow and Hoff put it, "This element [the Adaline] bears some resemblance to a 'neuron' model introduced by von Neumann, whence the name." An Adaline is characterized by both its use of the LMS algorithm and its requirement for the signals in the network to be bipolar, a signal property necessary for the proper operation of the LMS algorithm.

The LMS algorithm belongs to the class of – and, indeed, in many ways can be regarded as the

father of[2] – ***supervised adaptation*** rules. In supervised adaptation, some signal variable in the map model (usually the excitation variable s) is compared with a ***desired response*** signal. The numerical difference between the two is called the ***error signal***, and the adaptation rule seeks to minimize some measure of this error signal. In the case of the LMS algorithm, it minimizes the mean-squared value of the error signal.

Figure 12.7 illustrates the Adaline map. As is obvious, it is a version of Instar with the explicit addition of the adaptation algorithm (LMS), a desired response signal d_n, and the generated error signal ε. There are in fact two versions of the LMS algorithm, today called μ-LMS and α-LMS, as well as a more involved version called LMS/Newton [WIDR7: 142-147] and several other variations on the general theme. Historically, μ-LMS was the first of the LMS family to be developed, and when someone refers to "*the* LMS algorithm," this is the one usually meant. The μ-LMS and α-LMS algorithms are similar in many ways, although in some circumstances one or the other may show superior performance behaviors [WIDR3]. Because μ-LMS is the simpler and easier to understand version, our discussion in this textbook will be confined to it, and when "LMS algorithm" is used here, it will mean μ-LMS.

The presence of the desired response signal d_n is what makes LMS a *supervised* adaptation algorithm. In neural network contexts, the presence of signal d_n is usually taken as equivalent to a requirement that the Adaline must be "trained" and, therefore, requires a "teacher." The "teacher"

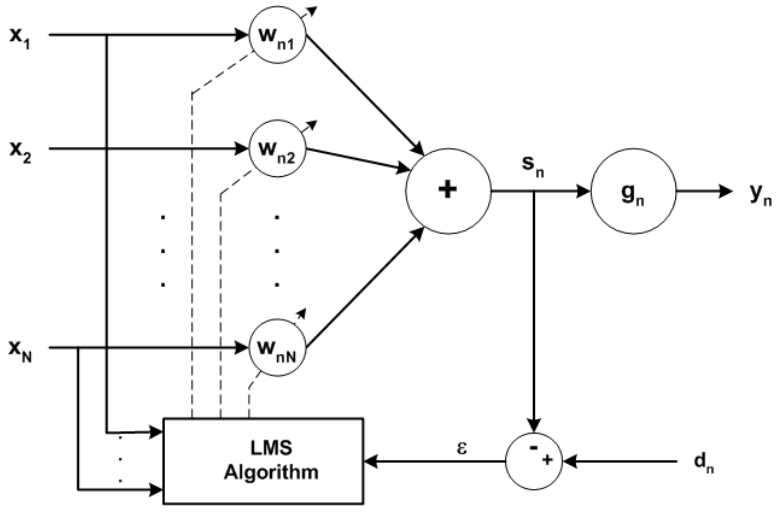


Figure 12.7: The Adaline map model.

[2] R.W. Lucky of Bell Telephone Laboratories also developed supervised adaptation rules for application in modems in the early 1960s. Lucky's algorithms are somewhat similar to, but not the same as, the LMS algorithm, and his papers on the subject make no mention of Widrow's work. Rather, Lucky's work was carried out independently, and so it is not accurate to call his algorithms "children" of LMS.

requirement is what leads Grossberg and others to conclude the LMS algorithm, and network adaptation algorithms employing it, "does not model a brain process . . . This shortcoming does not limit the model's possible value in technological applications, which can benefit from a steepest descent algorithm, but it undermines the model's usefulness in explaining behavioral or neural data" [GROSS10].

As true as Grossberg's remark is when the network system directly employs the agency of an external "teacher" or "supervisor" in "training" the system, its truth is less obvious if the "teacher" is *another* network system within the overall system. Widrow et al. also introduced, in 1973, a system structure by which an Adaline network (called a "Madaline network") can be "trained" by another subsystem within the overall system [WIDR8]. Such systems are today called ***actor-critic models***. It can be argued that the "critic" in an actor-critic system fulfills the role of affective psychological phenomena ("feelings", "values", etc.), and, to the extent this might be true, it might also be true the LMS algorithm is not so unfaithful to biological reality as a straight-up external teacher method is. It is a psychological fact that "emotions", "values", "interests", "motivations", etc. are heavily implicated in learning. Nonetheless, the overall issue is likely to remain controversial for the foreseeable future.

The LMS algorithm is a ***gradient descent*** or ***steepest descent*** algorithm, so we need to discuss what this means. We are all familiar with what the "grade" of a highway going up and down a hill means. The steeper the grade, the more rapid is the rate of change in the height of the hill. The ***gradient*** of a scalar function f of variables $w_1, w_2, \ldots w_N$ is a vector of derivatives of the function with respect to these variables. Mathematically,

$$\nabla f(W) \equiv \left[\partial f/\partial w_1 \quad \partial f/\partial w_2 \cdots \partial f/\partial w_N\right]^T \tag{12.9}$$

In the case of the LMS algorithm, the scalar function is the square of the error signal and the variables are the weights of the Adaline map. Widrow has shown that the quantity

$$-\frac{1}{2}\cdot\varepsilon\cdot X$$

is an unbiased estimate of the gradient of the mean-squared value of ε as a function of W (where X is, of course, the input signal vector). Accordingly, the LMS algorithm is

$$W(t+1) = W(t) + 2\cdot\mu\cdot\varepsilon(t)\cdot X(t) \tag{12.10}$$

where $\mu > 0$ is the adaptation rate constant. Its value determines the stability or instability of the adaptation process. Provided the adaptation is stable, (12.10) converges in the mean to the set of weights that minimizes the mean-squared error.

It is clear from the language just used that LMS is a statistical adaptation process, and so it is not surprising that the stability of the adaptation process should depend on the statistics of the input vector X. Let $R = E\{XX^T\}$, i.e. R is the correlation matrix for X. A sufficient condition for the adaptation to be stable is [WIDR7]

$$0 < \mu < \mathrm{Tr}(R) \tag{12.11}$$

where Tr(R) is the sum of the diagonal elements of R. If $||X||^2 = 1$ then Tr(R) = N.

A great deal has been written on the LMS algorithm and its properties, and we will not repeat all this here. The interested reader can consult [WIDR3] and [WIDR7] as excellent starting points to probe further. However, a couple general comments are in order here. First, like other adaptation algorithms, the LMS algorithm works best when the adaptation process is slow. As a rule of thumb, most researchers have reported best results when μ is one or two orders of magnitude smaller than the upper limit permitted by (12.11). Because R is a statistical entity, it is oftentimes the case that its trace, Tr(R) is unknown a priori unless X is *normalized* beforehand. A common normalization restricts X such that $X^TX = 1$. (This is one of the things accomplished automatically by the α-LMS algorithm; refer to [WIDR3] for details).

The second important remark has to do with the network environment in which LMS is used. The success of the algorithm relies upon the squared-error signal being a *quadratic* function of the weights. Because $\varepsilon = d_n - s_n = d_n - X^TW$, it is easily shown that

$$\varepsilon^2 = d_n^2 - 2 \cdot d_n \cdot X^T W + W^T X X^T W$$

which is a quadratic function of the weights if $X(t)$ is independent of $s_n(t)$.[3] However, if X contains as one of its elements a signal x_i that is either a direct or the indirect function of y_n, then it can also be shown that the embedding of this term in the expression above leads to a function that is a higher-than-quadratic function of W. (An additional W term "hides" within the expression for X). This has the undesirable result that the gradient function has multiple maxima and minima, and in this case no general rule has been found that can guarantee the convergence of the LMS algorithm to a global minimum-mean-squared-error solution.

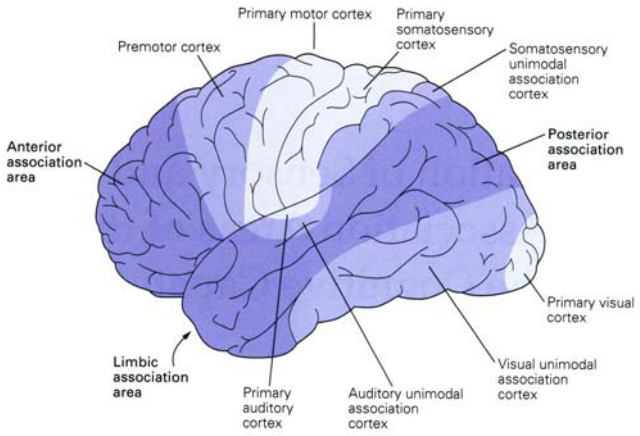
This property of the LMS algorithm has largely limited its application to feedforward network systems, i.e. the LMS algorithm is not very successful in handling recurrent neural network systems. A great deal of research effort has been expended over the years trying to extend the LMS algorithm to the case of recurrent network systems. Some limited successes, applicable in certain very restricted special cases, have been reported. However, an honest appraisal must, in

[3] Obviously s_n is not independent of X. But X can still nonetheless be independent of s_n.

the opinion of your author, conclude that these proposed approaches are not very generally applicable to *widespread* practice. Even the special case approaches seem to sacrifice the elegant simplicity that is one of the LMS algorithm's most attractive features. None can foresee when some breakthrough might change all this, but until and unless this happens, the restriction of the practical usefulness of LMS to feedforward network systems is a much more important "un-biological model" issue confronting it than are arguments against its biological plausibility based on its use of a desired response signal d_n.

Finally, it is important to comment that coming up with an appropriate desired response d_n is usually the trickiest and most crucial step in constructing an adaptive network model based on any supervised adaptation algorithm. The general rule of thumb is that d_n must represent some truly desirable response property of the system so that ε does in fact represent ***performance feedback*** to the adaptation process. Coming up with an appropriate scheme for *generating or supplying* d_n is the keystone for applying the LMS algorithm in a network system model.

Exercises



1. According to Aristotle's theory, what general regions of the brain would constitute the "faculty" of memory?
2. Explain what James' idea of "memory" adds, in terms of brain structures, in addition to what Aristotle thought a "memory" is?
3. What brain structures do Piaget's findings add to the constitution of memory beyond what James' theory calls for?
4. The simplest mathematical model for Hebbian learning is $\Delta w_{ij} = \eta \cdot x_j \cdot y_i$ where x_j is the presynaptic activity level, y_i is the output activity of the i^{th} map node and η is a positive constant called the "learning rate." This simple model has a number of fatal flaws in the "learning dynamics" it produces. Using an Instar map with unipolar sigmoid activation function, propose an adaptation algorithm based on his Hebbian learning rule and simulate the weight adaptation performance of your model. What problems do you find with this adaptation algorithm? Repeat this exercise for the case where the Instar uses a

bipolar sigmoid function.

5. Propose a Linvill network model for the metabotropic process leading to production of Ca^{2+}/calmodulin in figure 12.1. Do not omit the calcium buffering mechanism. You do not need to find quantitative parameter values for your model.

6. Explain how the CaM kinase cascade in figure 12.1 leads to change in synaptic efficacy. Do not be hesitant about using mathematics to make your explanation specific and clear.

7. The simplest form of moving average of a signal y is the M-sample moving average

$$\langle y(t) \rangle = \frac{1}{M} \sum_{m=t-M+1}^{t} y(m).$$

By using delay elements (a "shift register") to provide the inputs, this computation is easily performed by an Instar with its output equal to s. Illustrate this using an Instar diagram and simulate the response of this map to a unit step input for $M = 2$, 5, and 10.

8. Write a computer simulation program for the BCM rule according to equations (12.2) and (12.3) and verify its correctness by reproducing the plots given in the text.

9. Repeat exercise 8 using (12.4) as the adaptation rule. Verify your simulator against the figures given in the text. Then examine the adaptation performance of the Instar for the case where **X** is a sequence of uniformly distributed random variables with the same mean values as in the test simulation. (MATLAB and MATHCAD both provide random number generators you can use for this). Do the weights reach a steady-state mean value, and how does this value compare to the steady state values reach if **X** is a constant vector with elements equal to the mean value?

10. Derive (12.5).

11. Create a computer simulation of the Instar adaptation rule and compare its behavior to the BCM rule.