Chapter 14

Simple Adaptive Competitive Networks

§ 1. Winner-Take-All Partitioning

The most basic and fundamental task carried out by adaptive competitive networks is solving a partitioning problem. In the standard terminology of neural network theory this is often referred to as segmentation, and WTA competitive networks are often called classifiers or feature detectors. Nonetheless, in order to accomplish either of these things, the first and most fundamental job the adaptation schema of the network system must perform is to determine which map will respond to any given input vector with an adaptation response. This is the partitioning problem.

As Grossberg never tires of reminding us, most of the classical network solutions for competitive classifiers suffer from the stability-plasticity problem. This will be true for the simple networks presented in this chapter. Our objective here is two-fold: (1) to become familiar with the basic ideas of competitive classification networks; (2) to understand what causes the stability vs. plasticity problem. This will set the groundwork for the transition to ART networks.

The most basic form of classical WTA partitioning is a two-layer network. The first layer is the adaptive layer that performs the actual classification task. The second layer is a non-adaptive layer where the competition is actually carried out. The second layer identifies the "winner" in the first layer that will be allowed to adapt in response to the input vector signal. The first layer will have $N > 1$ nodes (each a map model), and to each node in the first layer there will be a corresponding node in the second layer receiving its output signal.

When we regard this network system as a classifier, N nodes in the first layer implies N different possible classifications of the input vector X. This at once raises a fundamental question with which computational neuroscience must concern itself, namely: What determines N? If we regard each input vector X as a point in an input space Ξ, N nodes implies partitioning this input space into N subspace $\Xi_1, \Xi_2, \ldots, \Xi_N$ so that Ξ is the union, $\Xi = \Xi_1 \cup \Xi_2 \cup \ldots \cup \Xi_N$. Why should Ξ be divided up into N subspaces rather than some other number L subspaces? In artificial neural networks this is rarely a problem because the designer of the artificial network knows what problem he is trying to solve and the problem defines N. But for a model of brain function we have no nicely predefined "problem" the network system is to solve.

Empirical psychophysics can partially help us here, as can physiology studies. For example,

we know that neocortical functional columns outnumber anatomical columns in the visual cortex, which implies each anatomical column (to the extent these can be identified) carries out some usually small plurality of signal processing tasks. If we had reliable statistics for the ratio of the number of functional columns to anatomical columns, we could use this ratio to guide the selection or selections of N in our network systems models. Likewise, if we observe through PET or fMRI scans the involvement of a particular region of cortex in several different kinds of psychological phenomena, we again could perhaps use this to take a guess at an appropriate value for N by assuming different psychological situations imply differences in the signals X converging on that region of the brain. The point, though, is that we generally have no a priori empirical reason to prefer one value of N over some other value. This is one of the many reasons theoretical neuroscience must stay connected to experimental neuroscience. If computational neuroscience is to be explanatory and predictive rather than merely descriptive, this connection must be reciprocal.

For our purposes here, we will take it for granted that by some means or another we have made our choice for N (while understanding that this is in fact taking *a lot* for granted). Several other system-level issues must still be addressed. By our selection of N we have put a cap on the maximum number of subspaces that can emerge from adaptive partitioning.[1] But this does not tell us what the topology of this partitioning will be. Should all the Ξ_n be the same "size"? Should they all have the same "shape"? Should all the Ξ_n be disjoint or should some of them overlap? Should some of them initially overlap but later, through adaptation, become disjoint? Again, there is no single a priori answer to these questions. However, the choice of map model we make, the initial conditions we give to their weights W_n, the adaptation rate (or rates) of the nodes – all these decisions will affect the qualitative nature of the solution the network system will develop. Part of the task of theory in neural network theory is to understand how different model structures and algorithms affect outcomes in the solution space.

Another question concerns what to do about "ties." It is always possible, in principle, for a winner-take-all competition to result in a tie. What does the occurrence of a tie implicate so far as neuroscience is concerned? Should we regard a tie as a mechanism setting up one of Piaget's cycle ruptures preventing the equilibration of the system (in which case it would be a "disturbance" too large for the central process of equilibration to deal with)? Should it be a telltale indicator for the "focusing of attention" (a "spot light" for a critic structure to weigh in on) or

[1] If a map node never "wins" a competition, it cannot be said with a straight face that this map is actually classifying or representing anything. If every input X is "covered" by another map node there is no distinct Ξ_n this "loser" map node represents.

should it be a condition for just the opposite response? Briefly put, what is the neurological or psychological significance of a tie? One presumption we probably should *not* make is that a tie is a mere "mathematical inconvenience" without either action implication or meaning implication.

Should the competition require some minimum "margin of victory"? If the activity of one first layer node in response to X is $y_i = 0.9995$ and another is $y_j = 0.9994$, did i "beat" j by "enough" to declare classification i is "correct" and classification j is "incorrect"? This question goes to the statistical issue of ***covariance***, $E\{(X - E\{X\}) \cdot (X - E\{X\})^T\}$, if inputs X are regarded as "noisy." (And, of course, this leads in turn back to the question of "when is a 'feature' of X a 'significant' feature and when is it 'noise'?", as we saw Grossberg point out in chapter 13). One way to phrase this "margin of victory" question is: Should a classifier make "crisp" classifications or should it make "fuzzy" classifications? Should every X belong to exactly one Ξ_n (implying subspaces should not overlap) or should X have a "degree of membership" in more than one Ξ_n?

As soon as we pose a network system model, we are making a *de facto* decision regarding all these issues. This is something the modeler must always understand. It is also why it is important to understand how different classical network systems behave in regard to these issues. With this prolegomenon in place, let us now get down to some actual cases.

§ 2. The Instar-MAXNET Network

Our first example network is illustrated in figure 14.1. It is composed of two layers: an Instar layer and a MAXNET competitive layer. The input vector X is comprised of M input signals with

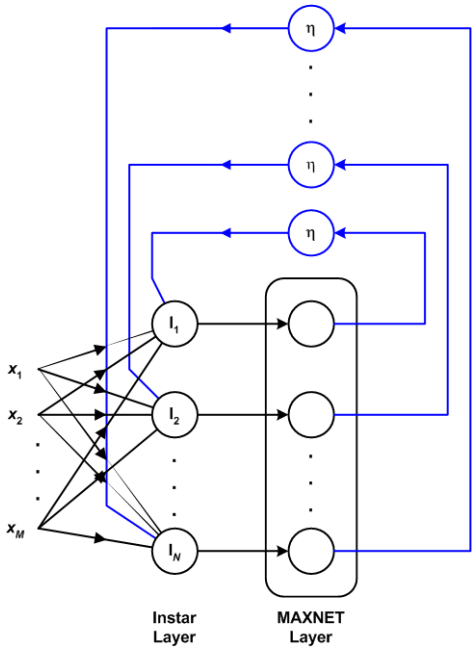

Figure 14.1: The Instar-MAXNET Network.

N nodes in both the Instar and MAXNET layers. The first-layer Instars use the unipolar sigmoid activation function g(s) and adapt according to the Instar adaptation rule (IAR)

$$W(t+1) = W(t) + \eta \cdot (X(t) - W(t)) \cdot g[s(t)] \tag{14.1}$$

where t is the time index and s is the Instar's excitation variable, $s = X^T W$. The weights laterally connecting the Instars in the MAXNET are fixed and we will assume it uses activation function

$$y = g(s) = \begin{cases} 1, & s - q > 1 \\ s - q, & s \geq q \\ 0, & s < q \end{cases} \tag{14.2}$$

where the fixed constant $0 \leq q < 1$ is called a ***quenching threshold*** and s is the excitation variable for the MAXNET Instars.

The adaptation constant η for the IAR is determined by feedback from the MAXNET. There are two simple choices for how η is determined. We assume $\eta = 0$ during the MAXNET competition in both cases. At the end of the competition, the first and simplest case is to set the adaptation constant equal to η for any first-layer Instar for which the corresponding MAXNET output is not equal to zero, and to set $\eta = 0$ for all others. We will call this the Case I method. The second and next simplest case is to set $\eta = \eta_0 \cdot y_n$ where y_n is the output of the n^{th} node of the MAXNET layer and $0 < \eta_0 \leq 1$ is a scaling factor. We will call this the Case II method.

We will assume some X is applied at time step t and the first-layer Instar outputs are calculated. The MAXNET then carries out its competition. Provided at least one y_n from the first layer exceeds the quenching threshold, the MAXNET will select the first-layer Instar with the highest activity (or, in the case of a tie, it will select no winner). Assuming the IAR stability condition (11.8) from chapter 12 is satisfied, the winner will adapt and move its W vector in the direction of X. This completes the processing at time step t.

This seems at first glance to achieve the desired end result, namely the partitioning of Ξ according to which first-layer Instars respond with the most activity to input X. But what kind of partitioning of Ξ does this network actually produce? Does it, for example, assign the winner according to which Instar weight W is nearest in Euclidean distance to X, i.e. to the Instar for which $||X - W||^2$ is least?

In general the answer to this question is no. To see this, let us look at the case where we have two input signals, $X^T = [x_1 \;\; x_2]$, and two Instars. It will be convenient to express the X and W vectors in polar coordinates so that $X = [R \cdot \cos(\theta) \;\; R \cdot \sin(\theta)]^T$ and $W_i = [r_i \cdot \cos(\phi_i) \;\; r_i \cdot \sin(\phi_i)]^T$, $i = 1$ or 2. Some simple trigonometry gives us

$$s_i = W_i^T X = R \cdot r_i \cdot \cos(\phi_i - \theta).$$

Assuming both first-layer Instars produce outputs above the quenching threshold, the ratio of their respective excitation variables is

$$\frac{s_1}{s_2} = \frac{r_1 \cdot \cos(\phi_1 - \theta)}{r_2 \cdot \cos(\phi_2 - \theta)}.$$

There are several things to note about this result. First, since the sigmoid is a monotonic function, whichever Instar has the greatest s_i will also have the greatest y_i and will be declared the "winner." This is equivalent to saying Instar I_1 will be the winner if the ratio above is greater than one, and vice versa if it is less than one. Note that this ratio is *independent* of $R = ||X||$. The winner decision is a function only of $\theta = \arctan(x_2/x_1)$ insofar as X enters into it.

Now suppose W_1 and W_2 both have the same orientation in W-space, i.e. $\phi_1 = \phi_2$. In this case, *X plays no role whatsoever* in determining the winner. The winner will always be the Instar with the larger $||W||$. Most likely, this is not what a modeler would have in mind for the behavior of this network. With two co-aligned Instars, one would be "dead" insofar as adaptation is concerned (although it would still produce output activity, which would merely be a scaled version of the other's output activity).

Next suppose $||W_1|| = ||W_2||$, i.e. $r_1 = r_2$. For this case we will assume the weight vectors are not aligned (since this would imply $W_1 = W_2$). Now both $||W||$ and $||X||$ are irrelevant to the winner selection, and the winning Instar will be the one for which the angular difference between W and X is the least. (Because all the x and w variables are non-negative, our "space" is merely a quarter-circle in the first quadrant of coordinate system). Finally we at least have a situation where "closeness" (in this case, angular "closeness") determines the winner. But note that if $||X|| \neq ||W||$, the condition $||W_1|| = ||W_2||$ will not be maintained after the adaptation. This is because the winning W will move toward X under IAR adaptation. The only way to maintain the "closeness" property for the network would be to *renormalize* the weight vector so that every weight vector always maintained the same value for $||W||$. This, of course, reduces the number of actual dimensions in our "W-space" and requires a modification of the IAR.

The general case for this example, where no special properties are assigned to the magnitudes or angles of X and W, is complicated to analyze. We can at least say this about it: The network will "learn" something, but what that "something" is lacks any easy interpretation. Furthermore, it is not at all clear what, if any, conditions on the system will result in a stable learned configuration where each Instar "stakes out" some piece of Ξ to "call its own" despite the fact that

the adaptation dynamics of each individual Instar are very well behaved. This is an important lesson: *The network adaptation, not the map node adaptation, determines what will happen in a network system*. Our first example is not a very good classifier. The situation described above is not improved at all if we replace the MAXNET with a Mexican Hat competitive layer, so we will move on to our next example.

§ 3. The Radial Basis Function-MAXNET Network

The reason the Instar-MAXNET network of the previous section does not work well is because the winner selected by the MAXNET has no necessary relationship to the distance between the input vector and the weight settings. If W is to be a prototype vector representing some subspace of Ξ, the winner of the competition should be that Instar for which

$$\Delta^2 \overset{def}{=} \|X - W\|^2 = (X - W)^T (X - W) \tag{14.3}$$

is the smallest. This is not guaranteed if the excitation variable is $s = X^T W$.

Expanding the right-hand side of (14.3), we obtain

$$\Delta^2 = X^T X + W^T W - 2X^T W \Rightarrow X^T W - \frac{1}{2}\left(X^T X + W^T W\right) = -\frac{\Delta^2}{2}.$$

We can define the quantity $\Theta = (X^T X + W^T W)/2$ as a ***sliding threshold***, an idea we encountered with the BCM adaptation rule. Setting $s = X^T W - \Theta$ we have $s = -\Delta^2/2$ and our new excitation variable is now directly proportional to the squared Euclidean distance between X and W.

Any activation function $g(s)$ having the property $g(s) = 1$ if $\Delta = 0$ and $g(s) \to 0$ as $|\Delta| \to \infty$ is called a ***radial basis function*** (RBF). For $s = X^T W - \Theta$, $\Theta = (X^T X + W^T W)/2$, let us set

$$g(s) = \exp(\alpha \cdot s) = \exp\left(-\alpha \cdot \Delta^2/2\right) \tag{14.4}$$

where α is any positive constant. If the first-layer Instars of figure 14.1 are modified to use these definitions for s_n and g_n we will call the resulting network a radial basis function-MAXNET or RBF-MAXNET. Since $y_n = g_n(s_n)$, the MAXNET competition between two RBF Instars gives us

$$\frac{y_1}{y_2} = \frac{\exp\left(-\alpha \cdot \Delta_1^2/2\right)}{\exp\left(-\alpha \cdot \Delta_2^2/2\right)}$$

and the winner will be the RBF Instar for which $|\Delta|$ is the least.

The coverage of input vectors X by RBF Instars in Ξ is illustrated for a two-dimensional case in figure 14.2. The output activation of an RBF Instar decreases as the distance between X and W

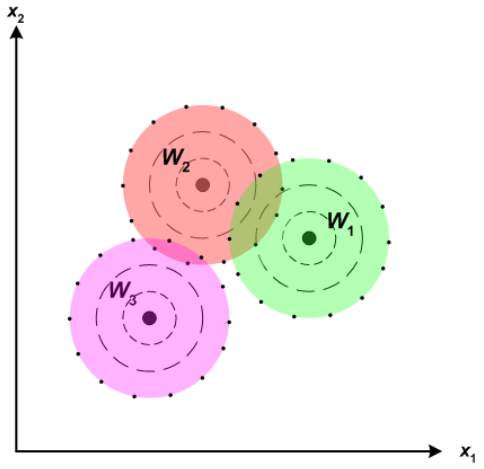

Figure 14.2: RBF-defined coverage regions in Ξ.

increases as a Gaussian function and has circular symmetry[2] about W. The rate of decrease is controlled by the α parameter in the activation function. As α is increased, the region around W for which the activation is significant rapidly decreases.

There is an interaction between α and the quenching threshold q of the MAXNET in the dynamics of the adaptation since any y_n below q will fail to excite its corresponding MAXNET Instar. There can thereby develop "gaps" in the spatial coverage of the RBF Instars. Note, however, that as the W "centers" move due to the IAR adaptation, a point X that formerly was in a "gap" can come to be covered later if one of the W vectors moves in such a way as to bring that region into its cover. However, with fixed α and q, the motion of a W point covering a gap may well also uncover a new "gap" in the region it leaves behind. Specifically, an RBF Instar participates in the MAXNET competition only if

$$y_n = \exp\left(-\alpha \cdot \Delta_n^2/2\right) > q \Rightarrow |\Delta_n| < \sqrt{2 \cdot \ln(1/q)/\alpha}$$

or $\|X - W_n\| < \sqrt{2 \cdot \ln(1/q)/\alpha}$. This expression bounds the coverage region of the RBF Instar.

§3.1 Attentional Functions

There are two mechanisms by which the MAXNET competition can fail to produce a winner. The first is when all RBF Instars in the first layer produce activations falling below the quenching threshold of the MAXNET. This case corresponds to the situation where X falls into a gap in the RBF coverage of Ξ. How should the network system respond to this situation? The answer to this question depends, or should be made to depend, on what interpretation is given to the biological

[2] In three dimensions this would be spherical symmetry. In higher than three dimensions, it is called hyper-spherical symmetry.

function of the network. This is in contrast to how one might decide the issue as merely an exercise in mathematics or in artificial neural network engineering. First we may note that when the quenching threshold $q = 0$, there are no gaps in coverage so far as adaptation is concerned. (The RBF Instars always produce some activation response regardless of the value of q). We may further note that IAR adaptation can be regarded in some sense as a form of Hebbian "learning" inasmuch as a small average value of input x produces a correspondingly small value of weight w under IAR and a large one produces a correspondingly large weight.[3] This is because $W \rightarrow E\{X\}$ in the statistical adaptation process implemented by an IAR. To the extent that we regard the adaptation process as Hebbian or Hebb-like, a small output activation for the Instar implies NMDA-mediated synaptic plasticity should not take place in the neurons said to be represented within the RBF Instar map. Thus, from the perspective of the ***data path*** from X to Y, there appears to be little obvious biological justification for doing anything other than allowing a quenched competitor to stay quenched and preventing adaptation from taking place. This data path argument, admittedly dialectic, favors ***learning stability*** in the network system model.

However, a strict and narrow interpretation of this argument loses sight of the fact that every network system is part of the larger central nervous system as a whole. Data paths are not the only pathways in the CNS. There are also modulatory ***control pathways*** of many kinds, such as those implemented by the metabotropic signaling projections from various brain stem nuclei. The data path argument favoring stability tends to also favor the network counterpart of Piagetian ***cycle rupture*** in adaptation (chapter 13) when, for fixed q in the MAXNET layer, an input X cannot be "assimilated" by one of the RBF Instars. In chapter 13 we spoke of the central process of equilibration and defined adaptation as the equilibrium between assimilation and accommodation. What about accommodation at the level of the total system?

This consideration leads us into the psychologically murky waters where hypotheses dealing with the psychological phenomena of ***attention*** and ***consciousness*** are found. While it is true enough that universally accepted theories of attention and consciousness have not yet been achieved, there is general agreement that both are psychological factors for which the underlying brain activity is wide-scale and involves many, many different brain regions. Damasio presents one interesting picture of this in [DAMA5]. Some, e.g. Damasio, describe attention as "a spot lighting mechanism." William James provided a bit more lengthy description:

> Millions of items of the outward order are present to my senses which never properly enter into my experience. Why? Because they have no *interest* for me. *My experience is what I agree to*

[3] This Hebbian interpretation is a bit more cloudy for RBF Instars located in a region of Ξ where all the x values in an X are relatively large.

> *attend to*. Only those items which I *notice* shape my mind – without selective interest, experience is an utter chaos. Interest alone gives accent and emphasis, light and shade, background and foreground – intelligible properties, in a word. It varies in every creature, but without it the consciousness of every creature would be a gray chaotic indiscriminateness, impossible for us to even conceive.
>
> Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence [JAME: 402-404].

ART networks contain a subsystem Grossberg dubs the "attentional subsystem," the function of which is to make the network system respond to "significant novelties." This is a most important part of an ART network insofar as resolving the stability-plasticity dilemma is concerned. But if the "essence" of attention is "focalization and concentration of consciousness," this raises the hardly-any-less-difficult issue of what "consciousness" is. There is no shortage of interesting speculations regarding this issue, all of which are colored – either explicitly or, more often, implicitly – by one or another brand of metaphysical outlook. But a proper perspective of science, a proper approach to this issue, should be one that focuses on the practical and observable consequences of the otherwise-hidden-from-observation-by-other-people object called "consciousness." Noting poorly correlated brain activities that appear to accompany presentations of behaviors we say are *indicative* of consciousness is not the same thing as observing "consciousness itself." If we seek a working and *practical* definition of the term, we can hardly improve upon the definition given by Kant: ***consciousness is the representation that another representation is in me***.

Now *this* definition gives us something to grasp in looking at the non-adaptation issue arising from failure of any RBF Instars to participate in the MAXNET competition. Even if all y_n fall below quenching threshold q, there are still representations (signals) present in the network system, namely X and Y. The issue at hand is merely that the MAXNET does not "see" X and is not responding to Y. Three distinct types of representations – X, Y, and the MAXNET output – are present in the network; all it lacks is a function for "representing that a representation is in me" and a function for either acting or not acting in response to this second-order control signal representation. We will call this function an ***attentional subsystem***. It has two tasks. First, it must provide a signal indicating that other signals are present in the data pathway system. Second, it must provide an appropriate response action to this signal. For the case of the non-adaptation issue with the RBF-MAXNET network, this response action is to *enforce plasticity* in the adaptation response. The easiest means for doing so is to lower the quenching threshold q, noting that if $q = 0$ a competition will ensue because now every RBF Instar covers *all* of the input space Ξ so far as adaptation response is concerned.

Figure 14.3 illustrates these modifications to the RBF-MAXNET network system. We will call the control line leading back from the attentional subsystem to the MAXNET the ***vigilance control***.[4] The vigilance control determines the level of the quenching threshold q for the MAXNET Instars, and thereby controls the span of coverage of the input space Ξ. For sake of completeness, we will assume it will also be possible and desirable to interconnect the attentional subsystem with other subsystems in the overall neural system, but for the time being we will not be concerned with this aspect of the RBF-MAXNET network system.

Let Z be the vector of outputs from the MAXNET. If $||Z|| = 0$, $||X|| \neq 0$, and if every $y_n < q$ in Y, this condition means a competition could have taken place but did not because no RBF Instar had a close enough match to X represented in its weights W_n. Should a competition have occurred? That is, should X have stimulated a more vigorous reaction Y from the RBF layer? Or was X merely "normal background noise" in the overall system for which a non-response by the RBF layer is appropriate? Here we encounter face-on the signal-vs.-noise issue discussed in chapter 13. In PET and fMRI studies, the scan data is typically evaluated in terms of activity being "significantly above normal background" levels, "significantly below normal background" levels, and "normal background" levels. Activities significantly above and significantly below background levels are generally regarded as important for correlating brain activity with whatever

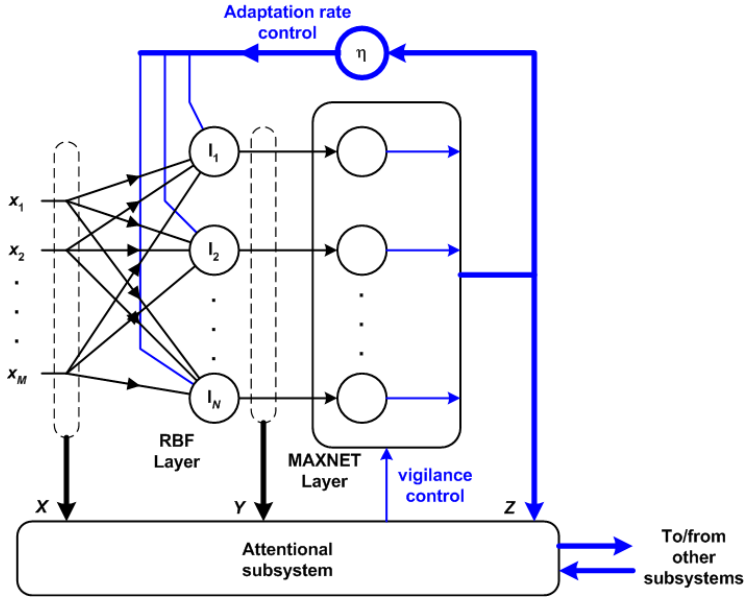

Figure 14.3: RBF-MAXNET with attentional subsystem. Bold lines represent signal vectors X (input signal vector), Y (RBF layer vector), and Z (MAXNET signal vector). The diagram simplifies the representation of the adaptation rate feedback pathway by representing all the individual rate controls as contained in a single functional block in the diagram.

[4] The terminology here differs from the use Carpenter and Grossberg make of the terms vigilance and attentional subsystem in ARTMAPS.

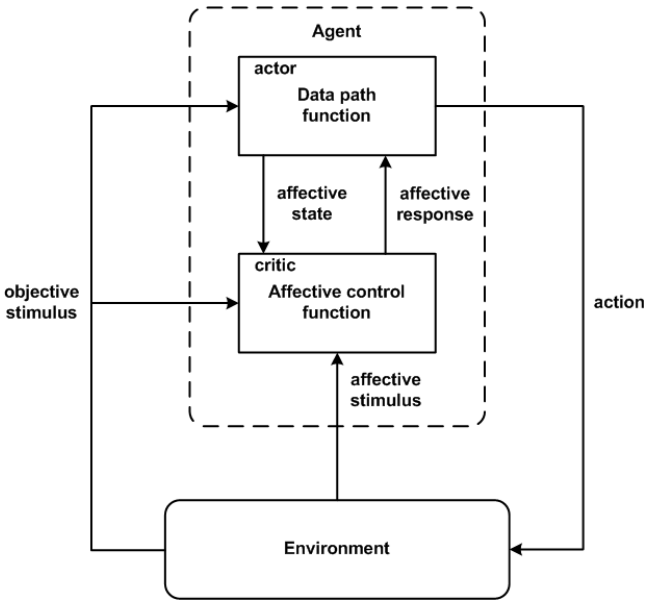

Figure 14.4: Improved actor-critic model when individual attentional subsystems of the various network system models in the overall agent model are regarded as belonging to the agent's overall affectivity-representing system. In this model, environmental effects are not the only source of information for the critic function. Internal conditions, such as those imputed by the attentional subsystem of figure 14.3, should also play a role in coordinating the overall functioning of the system.

psycho-physical object is being studied. "Normal background" levels are assumed to be activity levels that would being taking place as a mere part of the brain's normal commerce in maintaining life. Activity significantly below normal background is regarded as important because the implication is that some other brain function depending on a level of activity from the region showing below-normal activity is being altered by this *lack* of activity. All this is, of course, presupposition on the part of neuroscientists, but it seems to be a very reasonable presupposition and we presently have no compelling evidence to say it is wrong.

If "normal background" activity level is in some sense a "neutral gear" representative of stable overall system function in a state of equilibrium, and if higher-than-normal and lower-than-normal activity levels indicate some process of equilibration is in progress, this suggests two things. Viewed on the large scale, the fact that brain function (as represented by activity levels) is highly distributed suggests that what we are calling the attentional subsystem in figure 14.3 is part of a larger organization of affective control and synchronization of the brain's various specialized systems. This, in turn, suggests that the actor-critic model of chapter 13 – which, we recall, was developed in the context of artificial neural networks – should be modified, perhaps as depicted in figure 14.4 above, to reflect the probable role internal states in the agent play in determining the overall functioning of the system.

The second thing this suggests is that below-normal-background activity should *not* evoke

responses from the network system modules to which it projects. This is tantamount to declaring that some region of the Ξ-space illustrated in figure 14.2 should be treated as a region where the *proper* response of the network system is to make little or *no* response to X, and where the network system should *not* adapt in response to a stimulus. The simplest guess for how to model such a region is to suppose it is described by the quarter-circle defined by $||X|| < \Omega$, where Ω is some threshold, possibly set or controlled by some other function within the larger overall system.

Assuming this, then $||X|| < \Omega$, $\max(y_n \in Y) < q$, and $||Z|| = 0$ indicates the system of figure 14.3 has responded properly, but $||X|| > \Omega$ indicates the system response is improper. In the latter case, the attentional subsystem should respond by means of its "vigilance control" and lower threshold q to evoke a competition.

§3.2 Active Ties and Narrow Margins

The second way in which an all-zero output Z is produced by the MAXNET occurs when two or more of the RBF Instars produce the same output activity $y > q$. This, too, is a detectable condition defined by $||X|| > \Omega$, $\max(y_n \in Y) > q$, and $||Z|| = 0$. What should the network system of figure 14.3 do in this case?

Two RBF Instars, m and n, will tie at any point X for which $||X - W_m|| = ||X - W_n||$. The locus of points $\{X\}$ for which this condition holds can be regarded as a decision boundary between Instars m and n. If the system were modeled as having unlimited arithmetic precision (and if it were possible for a computer to represent quantities with unlimited precision), this case might be regarded as a merely formal difficulty since the probability of an X precisely meeting this condition would be zero (presuming X is regarded as a random variable following some continuous probability distribution function) unless $W_m = W_n$. (This condition of identical weights is usually regarded as disastrous for the functioning of the RBF-MAXNET system, although in some circumstances a useful interpretation for "coincident Instars" might be possible). However, real neural systems do not have unlimited precision (and real computers round and/or truncate their calculations to some finite precision) and so the issue is worth discussing. Let us say that an ***active tie*** exists between Instars m and n at some X if

$$\left| \|X - W_m\| - \|X - W_n\| \right| < \varepsilon \tag{14.5}$$

where ε is some small non-negative constant. What should the network system do? Although it might not be immediately apparent, this question takes us back to an issue we briefly touched upon earlier in this text, namely the issue of so-called grandmother cell encoding.

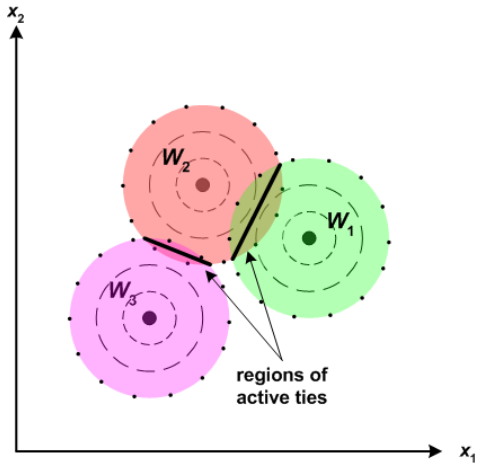

Figure 14.5: RBF coverage regions with active tie decision boundaries. The two dark lines are equally distant from the centers of the regions for which they form an adaptation decision boundary. The width of the active tie region is determined by ε.

If we say that every point X in Ξ must "belong" to exactly one Instar then one or the other (but not both) Instars involved in an active tie should adapt in response to X. This is a generalization of the original, simpler idea of the grandmother cell that arose from feedforward neural network theory for Adalines and perceptrons. In our present case, we would say X should enter in to determining the expected value of X ***in*** the region $\Xi_m \subset \Xi$ "covered" by Instar m (if this is the Instar chosen to adapt in response to X). There would then be a need to "break the tie."

However, let us recall that all the RBF Instars in the network will be producing outputs regardless of whether or not adaptation takes place. The output ***vector*** Y ***is a valid representation*** insofar as "coding" the response to X is concerned. Figure 14.5 repeats our earlier illustration of the coverage of Ξ but with the addition of sub-regions in which we define active ties. Suppose we regard an X falling within an active tie region as belonging to *both* Instars for which the active tie region forms an adaptation decision boundary. We must then decide if the appropriate action is to adapt *both* Instars or *neither*.

Our *mathematical* purpose – to learn an encoding for X – is already served if we do nothing. We therefore ask if there is any *biological* requirement for us to adapt either or both Instars. It is sometimes argued that, because signals X have enough activity to stimulate responses above the threshold q, this implies under Hebb's principle that both Instars should be adapted. This argument would carry some significant force *if* the synaptic weights of the Instar maps were direct correspondents of NMDA synapses. However, let us remember that a map model is a model of the collective actions of thousands of individual neurons. For this reason, a synapse-level argument of this kind cannot be applied with validity. For one thing, the map model "hides" the majority of biological synapses from our view. X is only an input tract and we do not know

the details of what internal interconnections may exist within the neuron system represented by a map model. We must therefore conclude that we have no biological justification *requiring* the adaptation of either RBF Instar.

There *is* a *functional* reason why it may be inadvisable to adapt either Instar. If both are made to adapt in the active tie reason, the "centers" represented by W_m and W_n will move towards each other. This is because IAR adaptation moves W in the direction of X. Now, there are a number of practical and functional reasons why we wish to avoid the possibility that two Instars could ever move to the same point W in Ξ. If our adaptation model calls for us to do nothing in cases of active ties, then occurrences of active ties cannot become a potential mechanism for causing two RBF Instars to become coincident. There may be (and there are) other mechanisms by which some form of undesirable coincidence of Instars could develop, but active ties will not be one of them. Inasmuch as we would have to add special modifications to the system of figure 14.3 to *force* adaptation in the case of an active tie, and because we do have a functional reason *not* to adapt active ties, the best adaptation policy for this case appears to be: ***do nothing***; do not adapt active ties.

How can active tie zones be defined in the network system of figure 14.3? Here we recall that output levels from a MAXNET are generally low-valued when two or more MAXNET inputs are close to each other in numerical value. That will be the case in the event of active ties. Therefore an active tie zone can be defined by adding an ***adaptation threshold*** κ to the η function generators of figure 14.3, i.e.

$$\textbf{case I:} \quad \eta = \begin{cases} \eta_0, & z \geq \kappa \\ 0, & \text{otherwise} \end{cases}; \tag{14.6a}$$

$$\textbf{case II:} \quad \eta = \begin{cases} \eta_0 \cdot (z - \kappa), & z \geq \kappa \\ 0, & \text{otherwise} \end{cases} \tag{14.6b}$$

where κ is some small positive constant. The relationship between κ and ε in (14.5) will be a function of the inhibitory weight settings in the MAXNET, with more rapid competitions tending to produce smaller winner outputs z. As pointed out earlier, having a finite active tie zone is more biologically realistic because of the limited precision of neuronal parameters.

§3.3 Stability and Plasticity in the RBF-MAXNET

Like other simple adaptive competitive networks that have been developed over the years, the RBF-MAXNET has stability vs. plasticity issues. Stability in a "learning" network generally refers to the ability of the network to retain past successful classifications in the face of later

adaptations. Plasticity generally refers to the ability of the network to quickly learn new classifications. The extreme example of this is the psychological phenomenon usually called ***one-time-event learning***. Like most people, you have most likely had at least one experience in life where something happened just one time that you never forgot. Often such events are accompanied by some kind of strong affective reaction. Perhaps it was the first time you met your future wife or husband; perhaps it was something a teacher said to you; perhaps it was an unpleasant first encounter with a wasp. There are many, many different examples of one-time events that seem to "stamp themselves" indelibly into a person's experience.

Generally speaking, the ability for a network system to exhibit one-time-event "learning" calls for an ability to make rapid adaptations. However, rapid adaptations also tend to work against the network's ability to retain old "lessons" – i.e. rapid adaptation tends to oppose stability in network systems. Let us examine how the RBF-MAXNET behaves in regard to stability-vs.-plasticity performance. Without loss of generality, we will consider a simple example of two Instars and consider X values that fall on a straight line connecting the Instar weight locations as shown in figure 14.6 below.

Let us assume for the sake of discussion that inputs X are uniformly distributed over the x-axis scale range in figure 14.6 from 0 to 1. Given initial weight locations for the Instars (0.2 and 0.8) it

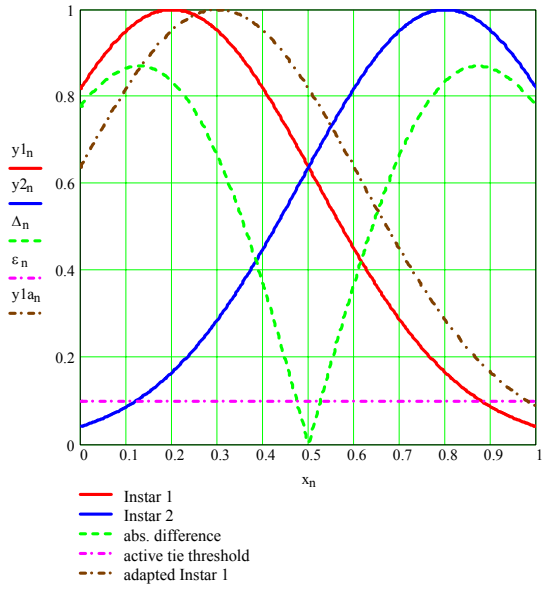

Figure 14.6: An example of network system adaptation dynamics between two RBF Instars. For both Instars the α parameter in the activation function is 10. Instar 1 is assumed to be initially located at $w_1 = 0.2$, and Instar 2 is assumed to be initially located at $w_2 = 0.8$. The dashed green line depicts the absolute value of the difference between the two Instar activations as a function of x. The dash-dotted magenta line represents the active tie threshold; points where the green dashed line fall below it are in the active tie region. The brown dash-dotted line represents a possible new center for Instar 1 following an adaptation.

would not be unreasonable to suppose at first glance that the adaptation process would result in the two Instars dividing up their respective Ξ regions along either side of the center of the initial active tie region at $x = 0.5$; if this was what in fact would happen, it would imply Instar 1 would center itself at $w_1 = 0.25$ and Instar 2 would center itself at $w_2 = 0.75$.

However, this is not what will generally happen. Given the initial situation depicted in the figure, suppose the next input happens to fall at, say, $x = 0.4$; this is in the "winning sphere" of Instar 1, and it falls outside the active tie region, so Instar 1 will adapt. Let us further suppose that Instar 1 ends up with a new weight center $w_1 = 0.3$ after adaptation (either because adaptation is fast or because x happens to dwell for a long time at its $x = 0.4$ value[5]). Its activity vs. x curve is then depicted by the dash-dotted brown curve in figure 14.6. We can see that the center of the active tie region has now also shifted to about $x = 0.55$. This points out the first important qualitative character of adaptation in the RBF-MAXNET system: the space-partitioning decision boundaries set by active tie bands move as the weights of the Instars adapt. For this reason, our reasonable initial intuition stated above is wrong; we cannot tell a priori where the Instar locations will end up merely from a knowledge of their initial positions and knowledge of the probability distribution function of inputs $\{X\}$. By that line of reasoning, after the first adaptation we would then have to say we would expect Instar 1 to end up centered at $w_1 = 0.275$ and Instar 2 to end up at $w_2 = 0.775$, and this, too, will generally be wrong.

With inputs x obeying a uniform probability distribution, if we assume the next x is statistically independent of the previous x (an assumption likely to be false for biological signals), there is a 50% chance the next x input will also fall in between the two Instar locations. If it does, and if it does not fall within the active tie band (which, you should notice, has gotten smaller in width after the first adaptation), the Instars will move still closer to one another. If it does not, they will move apart again. If the adaptation rate is very slow, then it is more likely than not that the statistical properties of the time sequence of x will be able to "balance out" the movements of the two Instar centers and lead to the development of a reasonably stable average location for each. (In which case our initial intuitive guess might turn out to be true after all). Such a system could exhibit stability, but it would come at the price of not being able to respond to one-time-event learning situations.

But, on the other hand, if the adaptation rate is fast, or if the input sequence is time-correlated in such a way that the next x is likely to be near the previous value, the Instar locations could

[5] It is a common experimental observance that large-scale activity patterns do tend to build up and dwell for long time periods, oftentimes on the order of 100 ms or longer, when the subject is given an external sensory stimulus. For an example, see [BRUN].

approach each other very closely. Since we are not allowing both Instars to adapt at the same time, by limiting η_0 to values $\eta_0 < 1$ no Instar can move, in one step, a distance greater than the difference between its present w and x; therefore, the Instars cannot "pass" each other in their travels along the x axis in the figure, nor can they "land on top" of one another. This, at least, is a good thing because were it not true the system would be terribly *unstable*.

However, the Instar weights would tend to move around quite a bit as the system developed over time (owing to the greater plasticity attending a rapid adaptation rate). This, in turn, would lead to significant variance in the Y output values, and it would also mean that a particular X input would sometimes lead to a winning MAXNET activity first for one of the Instars, and then for the other. The variance in the W vectors is commonly called ***weight noise***, and this weight noise leads to the variance (noise) in the output vectors Y. What this means is that even if the statistics of X were stationary (not changing over time), the statistics of the "coded" output vectors Y ***will*** be time dependent, owing to the significant time variation of the W vectors and the nonlinear function relating X to Y. If the statistics of X are also non-stationary (as they are likely to be if X is the output of another adaptive network system), the situation becomes even more complicated.

This is one of the things Grossberg was getting at in his remarks we quoted earlier on the general instability of simple competitive networks. In more Piagetian-like terminology, the accommodations of the system (adaptations) do not preserve the system's previous assimilation characteristics when the adaptation rate is rapid or when the input statistics are significantly non-stationary over time. If, as was argued in chapter 13, a psychologically realistic model must be one in which accommodation preserves prior assimilations, the RBF-MAXNET cannot accomplish this unless the adaptation rate is slow (and, often, not even then). This, by the way, also tends to argue in favor of the case II method for determining η since this method tends to enforce slower adaptation dynamics than does case I.

§3.4 Is the RBF Activation Function 'Biological'?

There is one more question we should deal with before ending our discussion of the RBF-MAXNET network system. Is our use of the radial basis function for the activation function mere mathematical chicanery, or is it a biologically reasonable description?

It would clearly be biologically unrealistic if an Instar was supposed to represent a neuron rather than a neuronal ***map*** modeling the collective input-output responses of many thousands or tens of thousands or hundreds of thousands of neurons. There is absolutely nothing in a biological neuron that says the neuron's response is in any way more active merely because all its synaptic inputs in some way "match" its synaptic strengths (the EPSPs and IPSPs produced in response to

synaptic excitation).

If we recognize a map model for what it is and ask the same question, the honest answer is "no one really knows." Maybe it is; maybe it isn't; maybe sometimes it is and sometimes it isn't. This applies not only to RBF Instar models but to *all* map models. Most neuroscience papers that employ map models maintain a gentlemanly silence in regard not only to arguments why the model is biologically realistic but oftentimes even in regard to what map model is being used to produce the outcomes that make up the paper's topic. Personally, your author does not approve of this because it tends to make it at the least inconvenient, and at the most difficult or impossible, to replicate the authors' work and check their claims. As Claude Bernard famously pointed out, a scientist should never be ashamed to say, "I don't know." Indeed, healthy science demands no less than this of us.

Still, applied mathematics in science serves us as a very, very, very precise language and so it is always a good idea to listen to one's equations to hear what they are saying to us. What does the mathematical representation of the RBF Instar say? Let us start with X. At the map model level, each element x in X represents a ***tract*** of signals and the information it conveys is merely an abstract measure of the "level of activity" at the source. As we discussed earlier, at the psycho-physical level of experimental science our main tools (PET, fMRI, etc.) convey measures of phenomena we know (or think we know) to be directly related to the amount of neuron firing taking place in a specific region. The abstract variables x tell us nothing directly about firing rates, temporal sequencing of action potential patterns, nor even the extent to which the overall activity is merely local (does not project to other regions). It is reasonable to say these measurements are related to the size of the neuron population that *is* active within the region, but this does not necessarily mean it is proportional to projected output activities from the region. PET or fMRI data might result from merely a higher metabolic rate brought on by some sort of local metabotropic signaling making direct contribution only to local signaling activity levels, whether this signaling activity is projected from the region or not.

For any given numerical value of x in a map model, this value might correspond to a large number of slowly-firing projection neurons, or it might correspond to a smaller number of rapidly-firing projection neurons. Or the same number x might sometimes correspond to the first case, and other times might correspond to the second case. Unless one has detailed anatomical and physiological data to say otherwise, there is simply no way to be sure. One can hope the metaphors one uses in thinking about the research problem are accurate, but to paraphrase the English philosopher John Locke, "Hope is like desire, and desire causes pain." It is better and more desirable to know when you don't know than to think you know something that isn't true.

What do we actually know about x? We know it *represents* some abstract measure of source activity and we *make the hypothesis* that this measure is *functionally related* to what happens in the maps to which it is projected.

Now, everything just said about x and X applies equally and in just the same way to y at the output of a map model and to the vector Y at the output of a network system made up of map models. The mathematical input-to-output transformation, $T(X) = y$, produced by a mathematical map model says to us nothing more and nothing less than "the output activity level of the map is related to the input activity level patterns in this *functional* way." $T(X)$ says *nothing* about biological mechanisms.

In the case of the RBF Instar, this output activity is maximal when X matches W and falls off in a Gaussian fashion as the mismatch between X and W increases. What, then, does W represent? It almost certainly does *not* represent the strength of the synaptic efficacy of neurons in the population. The map is not a neuron. As part of the transformation function $T(X)$ of the map, what W represents is an ***abstract tuning function***. Whatever might be the fine detail of firing rates, temporal patterns, correlations among action potential patterns within the tracts, or whatever else affects neuronal response within the population, W merely says to us, "whatever it is that is going on the signaling X represents, *this* cell population will be most excited when this vector activity measure has the vector numerical value W." This is what W says to us, no more and no less.

The RBF sliding threshold, $0.5 \cdot (X^T X + W^T W)$, says *nothing at all* to us about biology. *This* part of the RBF Instar map is indeed mathematical chicanery; its purpose is to *force* a correspondence between y and X so that our hypothesis about output activity being related to input activity in the functional manner described above is satisfied in the mathematical description of the Instar. The sliding threshold is part of the *functional implementation* of our hypothesis. Inasmuch as the hypothesis is *biologically* reasonable, the sliding threshold is *functionally* reasonable, no more and no less.

Our particular radial basis function has a free parameter, α. What does α say to us? It says the map has some particular degree of selectivity for activity patterns being projected to it. Small values of α say the map is not very selective in its response to specific incoming activity patterns; large values say the map is somehow or other highly "tuned" to respond to whatever specific signal details are hiding in W.

We could go on and discuss what the mathematics of the adaptation method implemented by the RBF-MAXNET network are saying to us. As this discussion is the corollary, it is left to you to think about. The key hint is this: *the adaptation method aims at supporting the hypothesis*. We have already discussed the issue that exists between (1) the stability-plasticity dilemma as it

shows up in this network system and (2) the incompatibility between fast adaptation and the psychological requirement that accommodation preserve prior assimilation capacities. We have seen that fast adaptation by this network system is not compatible with (2), while slow adaptation is not compatible with single-event "learning" phenomena. If you suspect the adaptation algorithm is chicanery in functional service of the basic hypothesis – well, good for you!

§4. The Functions of Competitive Networks

§4.1 Heteroassociation

We have just finished looking in some detail at two types of simple competitive networks. In both cases the aim of the network was to classify input vectors X by forming ***prototype*** vectors W. We saw that our first example, the Instar-MAXNET network, does this very poorly unless certain special steps are taken. Specifically, the input vectors had to be ***normalized*** so that $||X||^2 = X^TX$ had some constant value (say, for convenience, $X^TX = 1$) for every input vector; likewise, the weight vectors also have to be normalized to this same value. Normalization allows this network to function as a prototype vector classifier because, for $X^TX = W^TW = 1$, we have

$$\Delta^2 = \| X - W \|^2 = (X - W)^T (X - W) = X^T X + W^T W - 2X^T W = 2 - 2X^T W .$$

In this case, then, $s = X^TW$ provides a means for measuring the distance between X and W, and the Instar with the largest activation (assuming g is a monotonic function of s) is the one for which the distance between X and W is least. Weight normalization requires a change be made to the weight adaptation algorithm, e.g.,

$$W(t + \Delta t) = \frac{W(t) + \eta \cdot (X(t) - W(t))}{\| W(t) + \eta \cdot (X(t) - W(t)) \|} \quad . \tag{14.7}$$

Adaptation rule (14.7) is commonly called the ***WTA*** or ***winner take all*** rule. In addition, the network system must also have some sort of pre-processing on X to ensure its normalization.

The RBF-MAXNET network is also a prototype vector classifier, but one that does not require the normalization pre- and post-processing of the weights and inputs. While the Instar-MAXNET network is restricted to work on the "hyper-surface" of the hyper-sphere defined by $X^TX = 1$, the RBF-MAXNET will work over the "hyper-volume" Ξ.

Both networks can be regarded as ***nonlinear mapping functions***, i.e. $\mathbb{N}: X \rightarrow Y$, where X and Y can be regarded in the abstract as members of an input space, Ξ, and an output space, Ч. The network function is said to "associate" an input X with an output Y, and so such networks are often called ***heteroassociative networks***. The vectors W are frequently referred to as "categories"

or "concepts," although this is mere romance in the language used by neural network theorists.

In principle at least, a network of heteroassociative network systems could be regarded as a kind of instrument for computing. The problem, of course, is that "association" is the only function they actually provide if network "learning" is unsupervised using the IAR method. This is no trite matter; association is one of the fundamental constitutive psychological functions in Piaget's theory. But if the overall mathematical function of the system is to go beyond mere association in an unsupervised adaptive system, something else has to intervene.

Again in principle, this something else would fall to an evaluating or "critic" type of functional subsystem. But it is difficult to see how any overall system comprised of nothing but heteroassociations could provide a means for "universal" or "general purpose" functions to emerge. Indeed, because Piaget et al. found they needed *four* constitutive functions to describe observable behaviors (the association, repetition, identifier, and permutator functions), we should suspect that simple heteroassociative networks do not provide a complete basis for biological signal processing.

§4.2 Autoassociation and Outstar Nodes

One desirable assimilation function is ***autoassociation***. Suppose we have a network system for which the mapping function is $\mathbb{N}$: $X \rightarrow X$. Now let ΔX be some small perturbation on X. A network is said to be an ***autoassociative network*** if $\mathbb{N}$: $(X + \Delta X) \rightarrow c \cdot X$ where c is some constant. Such a network is said to "filter out the noise ΔX in the signal" or "recognize" X when X has "missing pieces" or "noise" in its representation. In general, a network system of parallel and non-interacting Instars, such as those of our earlier examples, cannot implement auto-association using unsupervised adaptation methods. Something new must be introduced into the network to make unsupervised autoassociation possible.

Several different types of recurrent networks for doing autoassociation have been developed over the years. Probably the most famous of these is the Grossberg-Hopfield network[6]. However, one of the major drawbacks to these networks is the fact that they require training, i.e. their adaptation is supervised rather than unsupervised. A second drawback is that their capacity to "store" autoassociation patterns by means of their W vectors is rather limited. For these reasons and because they are primarily of interest to artificial neural network engineering, we will spend no time on them.

[6] This network is more commonly known as the Hopfield network because it was through the work of John Hopfield in 1982 that it became well known. However, Grossberg and his colleagues had carried out much important analysis work on this network prior to its "rediscovery" in the early 1980s.

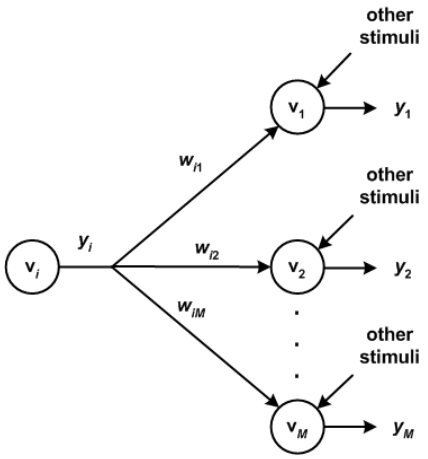

Figure 14.7: The classical Outstar node. Node v_i and the weight connections w_{ij} constitute the Outstar.

A simple extension of the RBF-MAXNET network will allow us to turn this network into an autoassociative network. The method involves another kind of map model network node called an ***Outstar***. Like the Instar, there are various kinds of Outstars; they are distinguished by the form their adaptation law takes. The classical Outstar is illustrated in figure 14.7 [CARP6]. Other variants include Grossberg's Γ-Outstar [GROS11] and the ***facilitated*** or *F*-Outstar introduced below.

Again like the Instar, the Outstar node, v_i in figure 14.7, represents a large population of neurons rather than a single neuron. For Instar adaptation using the IAR method, the weight adaptation is probably best thought of as modeling the function of *postsynaptic* long term potentiation and long term depression. This is to say, the long-lasting change in the weights of the connections *W* is conveniently regarded as "belonging to" the Instar population rather than to the signaling source population. As the RBF-MAXNET network illustrates, *control* of this adaptation process is carried out "locally" in the competitive network without calling upon any special properties of whatever network is the source of the input signals *X*.

In the case of the Outstar, it is convenient to conceptualize the adaptation process as modeling the function of *presynaptic* LTP and LTD. As we discussed in chapter 12, LTP and LTD is found to occur via both postsynaptic and presynaptic mechanisms. The classical Outstar adaptation rule, which we will call the c-OAR, in discrete-time form is

$$W(t+\Delta t) = W(t) + \lambda \cdot (Y(t) - W(t)) \cdot y_i \tag{14.8}$$

where $Y = [y_1 \;\; y_2 \cdots y_M]^T$ from figure 14.7 and y_i is the activation output of Outstar v_i. λ is the adaptation rate constant and *W* is the vector of weights connecting v_i to the destination nodes v_1 to v_M. For the classical Outstar, it is assumed the destination nodes are excited into activity by other

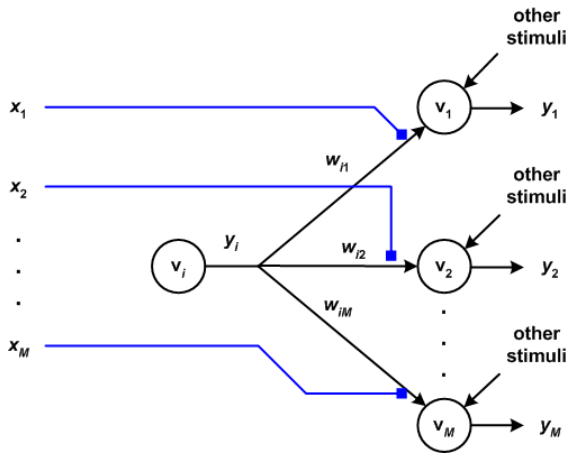

Figure 14.8: The facilitated or *F*-Outstar. In this variant of the Outstar the weight adaptation is controlled by a model of metabotropic presynaptic LTP/LTD due to metabotropic axoaxonal synapses made by a third *facilitating* population of interneurons with output activities *X*. The facilitating population acts as an adaptation control source.

stimulus sources such that their y_m output activations are not dominated by the $w_{im} \cdot y_i$ terms, the Outstar signals coming into nodes v_m. If the Outstar signal were to significantly contribute to determining the activity vector *Y* of the destination nodes, (14.8) will lead to adaptation instability because of the positive feedback inherent in the c-OAR. For this reason, the classical Outstar is almost always found embedded within a recurrent network structure where no "run-away" of *Y* can be induced due to v_i. The most widely known example of this is found in ART networks.

A variation of the Outstar is shown in figure 14.8, which we will call the *F*-Outstar (facilitated Outstar). For the *F*-Outstar adaptation is controlled by a third *facilitating input vector*, *X*, sourced by another set of population nodes. This population is presumed to make axoaxonal connections with the Outstar axon population at the presynaptic terminals of these axons. Furthermore, this axoaxonal connection is regarded as being metabotropic and, therefore, no part of signals *X* are transmitted to the destination nodes v_1 to v_M. The *F*-Outstar's adaptation rule (*f*-OAR) therefore does not suffer from any positive feedback effects such as those with which we must concern ourselves in the case of the c-OAR. The *f*-OAR is

$$W(t+\Delta t) = W(t) + \lambda \cdot (X(t) - W(t)) \cdot y_i \,. \tag{14.9}$$

Biological support for the *F*-Outstar is provided by experimental findings showing that LTP/LTD can be induced by presynaptic connections involving neurotransmitters such as serotonin (5-HT). The textbook example of this is provided by studies of the gill-withdrawal reflex in *Aplysia* [KAND3], the theory of which contributed significantly to Kandel's winning of the Nobel Prize in medicine. Like the IAR, the *f*-OAR adapts the weights *W* to match the expected value of the control inputs *X* provided the Outstar activation y_i is non-zero.

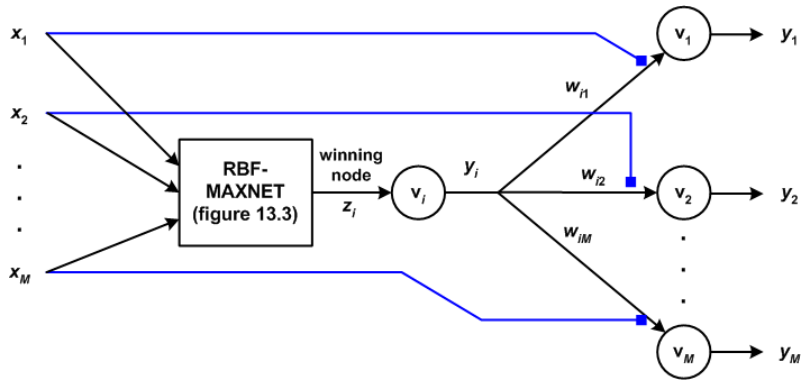

Figure 14.9: Autoassociative network formed by combining an RBF-MAXNET with *F*-Outstar nodes. Each node in the MAXNET projects its post-competition activation z_i to an *F*-Outstar with output y_i. Each Outstar in turn projects to a bank of map nodes (typically Instars) v_1 through v_M. There is one destination node for each input signal x_m, and each v_m receives an input from each *F*-Outstar. For each *F*-Outstar its *m*-th "axon" is facilitated by the corresponding input signal x_m. After competition there will be at most one MAXNET node for which $z_i \neq 0$, and this defines which *F*-Outstar will undergo facilitated adaptation according to the *f*-OAR.

To examine the stability conditions for the *f*-OAR we proceed as before and assume X to be constant and $W = X + \Delta W$. Inserting these into (14.9) and re-arranging terms gives us

$$\Delta W(t + \Delta t) = (1 - \lambda \cdot y_i(t)) \cdot \Delta W(t)$$

which, as before, is guaranteed to converge to zero if $|1 - \lambda \cdot y_i| < 1$. Also as previously shown for the IAR, if $X(t)$ follows a stationary random process the *f*-OAR weights will settle into a steady state defined by $E\{X \cdot y_i\} = E\{W \cdot y_i\}$, leading to $E\{W\} \cong E\{X\}$ by invoking the independence assumption on y_i.

Figure 14.9 illustrates how the *F*-Outstar can be used to convert the RBF-MAXNET of figure 14.3 into an autoassociative network. Each node in the MAXNET projects its post-competition activation z_i to an *F*-Outstar with output y_i. Each Outstar in turn projects to a bank of map nodes (typically Instars) v_1 through v_M. There is one destination node for each input signal x_m, and each v_m receives an input from each *F*-Outstar. For each *F*-Outstar its *m*-th "axon" is facilitated by the corresponding input signal x_m. After competition there will be at most one MAXNET node for which $z_i \neq 0$, and this defines which *F*-Outstar will undergo facilitated adaptation according to the *f*-OAR. After adaptation of the RBF-MAXNET and the *F*-Outstars have reached statistical steady state, the winning RBF Instar, acting through the MAXNET, projects an activation to its Outstar, which in turn projects a weighted signal $y_i \cdot W_i = y_i \cdot E\{X_i\}$ to nodes v_1 through v_m. Here $E\{X_i\}$ is the weight vector that has been "learned" by the RBF Instar. In the simplest case the Outstar activation y_i is linearly proportional to z_i, which would reflect in the intensity of the Outstar activation how "close" the competition between Instars was. Alternatively, we could have $y_i = H(z_i - \gamma)$ where H is the Heaviside function and γ is a quenching threshold for the Outstar. The

destination nodes v_1 through v_m can receive other inputs from other signal tracts $X^{(2)}$ without upsetting the autoassociative learning from the signal tracts X.

§5. General Discussion

The competitive networks presented in this chapter work very well in a statistically stationary input environment, but they are not without their drawbacks. The first and easiest to note is the limited number of partitions in Ξ, the input space, imposed a priori by the number of MAXNET nodes in the network. If we have N Instars and N MAXNET nodes, then we have at most N partitions in Ξ. This is a constraint caused by the basic structure of the network itself, and if more partitions are needed, the size of the network must grow linearly with the number of partitions.

There is, of course, reason to think that specific biological neural networks in the central nervous system do have limited capacity. Furthermore, the growth and development process in brain maturation clearly imposes a numerical constraint on network size. However, what is not constrained by this basic factor is the number of synaptic connections, i.e. the number of tracts in the input vector X and their distribution of connections (called the network system's ***receptive field***). The Instars can "prune" the number of input connections if some particular subset of x input signals is chronically zero or near zero; the network responds to this by driving the corresponding connection weights w to zero. Similarly, if some (but not all!) the connection weights are initially set to zero, the adaptation process can drive them to non-zero settings if the corresponding x inputs are non-zero on the average. Nonetheless, the maximum number of such connections *is* fixed a priori by the structure and cannot change.

There is an a priori constraint that must be applied in the initial setting of the weight vectors W_n of the Instars. Specifically, every Instar must have a *unique* initial weight vector setting. If two Instars should come to have *functionally identical* weight vectors, $W_n = W_k$, $n \neq k$, the result for the network system is disastrous because these two Instars will then forever produce output signals that will tie. This issue is more complicated than it may appear at first glance because functional identity between weight vectors involves the characteristics of the input vector X. To illustrate this, suppose $W_n = [\ w_1\ w_2 \cdots w_a\ ]^T$ and $W_k = [\ w_1\ w_2 \cdots w_b\ ]^T$, $w_a \neq w_b$. Now suppose that all input vectors are of the form $X = [\ x_1\ x_2 \cdots 0\ ]^T$, i.e. the x_M signal turns out to be inactive. Mathematically speaking, this means the *dimension* of input space Ξ is less than what was initially supposed. Instars n and k will forever produce tied outputs and neither will ever adapt. Note that there is no way to distinguish this situation from that of a normal active tie with the structure of the system as presented earlier, i.e. the case where $X \neq W_n \neq W_k$ but $y_n = y_k$.

There is a more serious issue attending these networks, however, and it involves the stability-

plasticity dilemma. Assuming we do not get into the difficulty of coincident weights just discussed, let us revisit the steady-state solutions we used in the theory presentations above. We said that the Instar weight will approach a steady-state solution $E\{W\} \cong E\{X\}$ where E denotes statistical expectation. Now, in deriving this result we had to invoke the assumption that the statistics of X do not change over time. In more formal language, the pairing of a signal space Ξ and a probability distribution function $p(\Xi)$, written $[\Xi, p(\Xi)]$, is called a ***probability space***. If the probability distribution p is not a function of time, written $p(\Xi, t) = p(\Xi)$, the probability space is said to be ***stationary***. In deriving the $E\{W\} \cong E\{X\}$ result, one of the steps in this derivation was the assertion that in the steady-state $E\{W(t + \Delta t\} = E\{W(t)\}$ after some sufficiently large t. The ability to make this assertion completely depends on the assumption $p(\Xi, t) = p(\Xi)$.

One way to say "stationary probability space" mathematically is to say $\partial p(\Xi, t)/\partial t = 0$. Now, in general $\partial p(\Xi, t)/\partial t \neq 0$, i.e., probability spaces are usually *not* stationary. Suppose you were to measuring the average number of cars passing by some particular spot on the main street of your hometown at different hours of the day. Most likely you will find the average number of passing cars at 5:00 PM to be very different from the average number at 3:00 AM, and you will find the average number of cars at 1:00 PM on Monday to be very different from the average at 1:00 PM on Sunday. If Wall Street stock prices were statistically stationary, the Dow Jones averages would be very boring indeed because they would never change by any significant amount.

So it is with brain activity. Every bit of experimental evidence we have says $\partial p(\Xi, t)/\partial t \neq 0$ for brain activity. Nonstationary analysis of adaptation dynamics in general is very, very difficult. It is possible to reach useful conclusions for certain special cases. For example, important and useful results in the case of the LMS algorithm have been presented by Widrow et al. [WIDR5]. Some of these apply to the IAR and *f*-OAR as well. We will consider two limiting cases.

First, suppose $\partial p(\Xi, t)/\partial t$ is small compared to the adaptation rate. In this case, $E\{W\}$ will tend to ***track*** $E\{X\}$, usually with a small lag δt, over time t. One way to say this formally is to write $E\{W, t\} \cong E\{X, t - \delta t\}$ for some small δt. The consequence of this is that over time the weight vectors W_n in the system will tend to ***drift***. That is, the network system will "forget" what it "learned" in the past in favor of what it has been "seeing" lately. In the neural network literature this is commonly called a ***learning instability***. It is probably true that for at least some neuronal structures in the central nervous system this behavior is characteristic and perhaps even serves a useful purpose. (Think about looking up a telephone number and dialing it; do you remember what that number was five minutes later?) But it is also certainly true that some neuronal structures do *not* behave this way; learning instability is an accommodation that does not preserve

assimilation, and we know this situation to be contrary to many observable behaviors during child development.

Next let us suppose $\partial p(\Xi, t)/\partial t$ is large compared to the adaptation rate but that our probability space has the special character that $p(\Xi, t + T) = p(\Xi, t)$ for some T. The probability space in this case is generally said to ***cyclostationary*** and, for small T, is said to be "rapidly" cyclostationary. What we mean specifically by this is that $1/T$ is large compared to the rate at which the weight vectors change. In this case, the changes in the W vectors will fail to follow the time variations in $p(\Xi, t)$ and will instead tend to "sit" at a value in the near vicinity of the time-average "location" of X,

$$\overline{X} = \frac{1}{T}\int_{t}^{t+T} X(t)\,dt\,.$$

The rapid fluctuations in $p(\Xi, t)$ are "seen" by the network system as a kind of "high frequency interference" that the system tends to filter out. If, on the other hand, $p(\Xi, t)$ is cyclostationary but $1/T$ is small with respect to the adaptation rate, the network system will tend to "track" the changes and the W vectors will be "cyclostationary" too. In this case the network system tends to act like a sort of "low pass filter" and $1/T$ is said to be "in the passband" of this "filter." It would not be an abuse of language to say our competitive networks are ***temporal filters of probability distribution functions***.

Although this discussion has focused on the Instars, the same conclusions also apply to the feedforward *f*-OAR adaptation by the *F*-Outstar node. In slightly different form they also apply to the classical Outstar node if this adaptation does not encounter instabilities resulting from the positive feedback issue that can arise if the Outstar makes a significant contribution to the activity of the nodes v to which it connects. (If this instability occurs, that is another and worse matter).

Finally, the classic competitive networks discussed in this chapter lack the mathematical properties of shift- or rotational-invariance. An Instar is said to be shift-invariant if, for example, the set of inputs $[x_1\ x_2\ x_3\ 0\ 0\ 0]^T$, $[0\ x_1\ x_2\ x_3\ 0\ 0]^T$, $[0\ 0\ x_1\ x_2\ x_3\ 0]^T$ and $[0\ 0\ 0\ x_1\ x_2\ x_3]^T$ all produce the same output y. An Instar has rotational-invariance if $[x_1\ x_2 \cdots x_M]^T$, $[x_M\ x_1 \cdots x_{M\text{-}1}]^T$, etc. all produce the same y. To possess shift- or rotational-invariance requires some sort of corresponding characteristic for the Instar's W vector and, generally, the IAR does not produce this kind of symmetry or invariance property. If the Instars in a competitive network lack these properties, then so will the network as a whole.

Shift-invariance and/or rotational-invariance are often very desirable properties for artificial neural networks in applications such as character recognition or image processing. But do

biological networks possessing these properties actually exist in nature? The answer to this probably depends on the scale of structure we are looking at. There is evidence that relatively large network systems – that is, systems comprised of a collection of many interconnected two-layer competitive networks – do exist that exhibit these properties. The evidence implicating this generally comes from psychological rather than psychophysical testing. But there is little evidence to support the hypothesis that these properties exist on the fine scale, and some evidence has been found suggesting that shift- and rotational-*variance* occurs locally at the sub-millimeter scale [BOSK]. This variance on the local scale is accompanied by strong coupling with nearby regions, which could imply that the appearance of shift- or rotational-invariance in psychological testing might be the consequence of much larger-scale system interactions than is modeled by a simple two-layer competitive network. The simple networks discussed in this chapter are not laterally cross-coupled with other competitive networks, and so the modeling of invariance properties belongs to a higher scale modeling problem of networks of network systems.

Some of these issues have been addressed by adaptive resonance theory. In chapter 15 we turn to the examination of the fundamental concepts of ART.