1

Some recent developments in the theory of neural networks

LEON N. COOPER

A question of great interest in neural network theory is the way such a network modifies its synaptic connections. It is in the synapses that memory is believed to be stored: the progression from input to output somehow leads to cognitive behaviour. When our work began more than ten years ago, this point of view was shared by relatively few people. Certainly, Kohonen was one of those who not only shared the attitude, but probably preceded us in advocating it. There had been some early work done on distributed memories by Pribram Grossberg Longuet Higgins and Anderson. If you consider a neural network, there are at least two things you can be concerned with. You can look at the instantaneous behaviour, at the individual spikes, and you can think of the neurons as adjusting themselves over short time periods to what is around them. This has led recently to much work related to Hopfield's model; many people are now working on such relaxation models of neural networks. But we are primarily concerned with the longer term behaviour of neural networks. To a certain extent this too can be formulated as a relaxation process, although it is a relaxation process with a much longer lifetime.

We realized very early, as did many others, that if we could put the proper synaptic strengths at the different junctions, then we would have a machine which, although it might not talk and walk, would begin to do some rather interesting things. Kohonen has shown us some intriguing examples of such behaviour. We were soon confronted with a fundamental problem. Let us assume that you can form a network which will store memory, but which requires the adjustment of very large numbers of synaptic junctions. It is perfectly obvious that this cannot be done

genetically, or at least not 100% genetically, if you believe that experience has anything to do with the content of your memory. So there must be some kind of rule, or set of rules, by which the synaptic strengths change.

Now, the use of the word 'synapse' may be something of a metaphor. The physiologists among us know that a synapse is a very complicated thing, and what we refer to as a synapse is really a logical grouping of large numbers of synapses, i.e. it is really a relation between inputs and outputs, and it is probable that what happens biologically is considerably more complex than any of the simple rules we write down. But the idea was that if we could write down a few rules which captured some of the qualitative properties, perhaps we would be on the right track.

Now it seemed to me that one of the things most lacking in this field, the thing required to convert it from the hand-waving stage to a field which was science as I understand it, was to construct pieces of theory that were well-defined and had a really rigorous structure, obviously highly over-simplified, but that could be brought into correspondence with serious experiment. Now that is not easy to do, and I am still not sure whether we have done it. Nevertheless this has been one of the dominating themes of our work for the last ten years. We chose a preparation, visual cortex, which may be the wrong starting point since it is a very complicated system, as compared to simpler systems like *Aplysia*. However, it is a system where interesting things seem to be happening, where experiments can be done, where there is a rich tradition of at least 20 years of experimentation, and where one has very robust effects. I cannot discuss all of these here, but those familiar with visual cortex and the history of the Hubel–Wiesel cells – the preference of the cells for certain orientations – know that there is a long history of experimentation in this area in which one can change the input–output relations of individual neurons by changing the visual experience of the animal. One reason this seemed so intriguing to us was that it appeared to be a situation in which one could observe changes in the neuron input–output behaviour almost as well as one could in hippocampus or *Aplysia*. I like to call this 'learning on a single-cell level', but am immediately thrown into conflict with some psychologists, who say learning is a much more complicated thing. However, when we learn something, there must be a change in the neural network; I believe that the origin of that change is what happens in individual cells in the network. And so one should be able to relate the large-scale properties to changes in individual cells. It is those changes that occur on a single-cell level that give us learning in a network level.

Experiments have been done in visual cortex that seem to indicate that the response characteristics of visual cortical cells depend on the visual environment of the animal. For example, if an animal is normally reared, the visual cortical cells will be sharply tuned and will be responsive to edges of one orientation, but not to edges of other orientations. This famous result is due to Hubel and Wiesel. It has been known for many years that if you raise animals in deprived environments, for example in the dark, the cells are not sharply tuned, but are broadly tuned. If you raise an animal with one eye open and the other eye closed, the cells will become responsive to the open eye, and sharply tuned to that eye, and they will lose their responsiveness to the closed eye, and so on. Large numbers of repeatable experiments have been done.

The question becomes: can one introduce a set of rules for synaptic change that will explain this behaviour? We have successfully introduced such a set of synaptic rules which involve what seem to be the variables that would provide the kind of learning we want on a network level. Let me describe this first on a single-cell level. Suppose you have a single cell with inputs from the external environment, and we introduce a modification procedure. What would be the necessary modification procedure to reproduce the experimental results? We were able to successfully reproduce what we call the 'classical' experimental results. In addition, we were able to show that there are some rather subtle new effects. Recent experiments do show these subtle effects. If they had been seen before, they had never been really explicitly pointed out or recognized.

One problem with this work is that, in addition to simplifying the synaptic junction, it made one assumption which was clearly not permissible, in that we considered the cortical synapses of the lateral geniculate nucleus to be capable of being both positive and negative. In other words, they could be both excitatory and inhibitory. Our recent work has solved this problem.

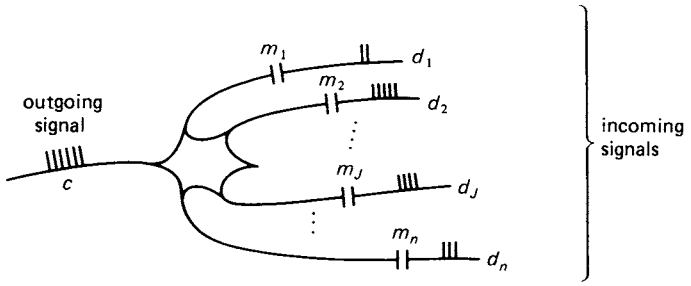
For our purposes, we can think of a set of inputs to a neuron, call them $d_1, \ldots, d_n$. Now the $d_1, \ldots, d_n$ is an n-dimensional vector; think of that n-dimensional vector as being a mapping of what is in the visual space. It is produced via the transduction cells of the retina, and so forth, so that for a particular image on the visual space you have a particular vector in this space. Now these inputs, measured in spikes per second, go through a set of synapses to the cortical cells. The cells integrate the electrical activity, so in effect the output of the cell is a non-linear monotonic function of $m \cdot d$. We are mostly concerned about the roughly linear

region, though some of the other papers in these proceedings will take a different point of view and will try to make it as non-linear as possible. I am not going to insist on the virtue of one rather than the other. The point that we are making is that for the purposes of learning, the linearity is enough (Figs. 1.1 and 1.2).

To summarize, d is the input from the external world, and it changes as the image on the screen changes; m is a set of synaptic strengths that, if it changes at all, changes slowly with time. And it is m that contains the learning of the system. If one has a particular set of ms, with m thought of as a vector, a particular input that is parallel to it could give a large cell output. An input that is orthogonal to m would give zero output, so this set of ms, is, in a certain sense, already a memory because it distinguishes between one set and another set of inputs. Of course, for one cell one gets a rather limited memory, but if there are networks of these cells, all kinds of interesting things could occur.

Now the issue with which we are concerned is precisely how these ms change. Supposing you wish to 'teach' a single cell to recognize a vector. You could have a set of synaptic strengths that give a large response for one input and a very small one for another input. The cell will 'know' something. The question would then be: how do you design a rule so that the cell will learn to recognize a particular input? To correspond to the results in the visual cortex, you want the cell to learn to respond to one pattern and not to respond at all to others when in an environment with

Fig. 1.1. The inputs $d_1, \ldots, d_n$ from axons via *synaptic junctions* $m_1, \ldots, m_n$ produce local depolarizations $m_1d_1, \ldots, m_nd_n$ that are integrated in the cell body to produce a *firing rate*, c. The actual inputs and outputs are rapidly varying functions of time (spikes occur with durations of approximately 10^{-3}). Relevant time intervals for learning and memory are thought to be of the order of 0.5 s. We assume that for learning the relevant inputs and outputs are time averaged frequencies.

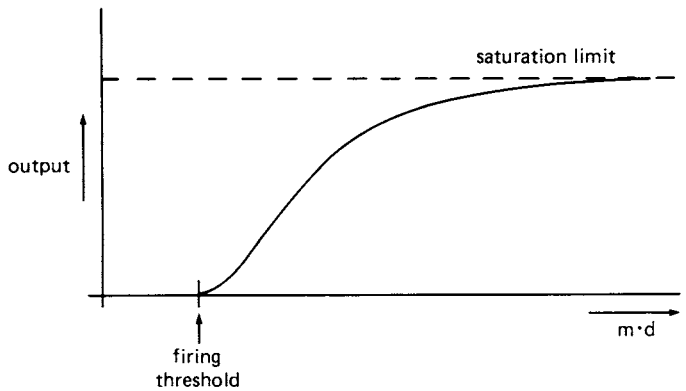


all patterns present. On the other hand, for non-patterned input, a cell's response should not prefer one pattern over another (Fig. 1.3).

Hebb originally proposed that if synaptic junctions change as a product of the output and input, they produce certain interesting correlation properties. It is now generally acknowledged that they cannot change in exactly this way. Still, one question is: is the post-synaptic variable involved? Some experimentalists say no, while every theoretician says yes. Yet there is good experimental evidence, particularly that obtained by Singer and others, which shows this variable must be implicated. If it is implicated, how? There are many other possibilities, particularly those involving what we call 'global' variables, for which there is good evidence.

We introduced a form of modification in which the change in m_j was a product of input d_j and a ϕ function. The properties of ϕ are that if for a given input the output is too low, the synaptic strength decreases in proportion to the input. On the other hand, if above this modification threshold the output is large enough, then the synaptic strength increases proportionally to the input. To explain the existing experimental results we need a negative and a positive region. In addition, to give the system the proper stability properties, it is required that the threshold move back and forth and that, to give it the most beautiful properties of all, it is required that it be a non-linear function of the average output of the cell. We always chose $\bar{c}^2$, but it could actually be $\bar{c}$ to any power that is larger than unity (Fig. 1.4).

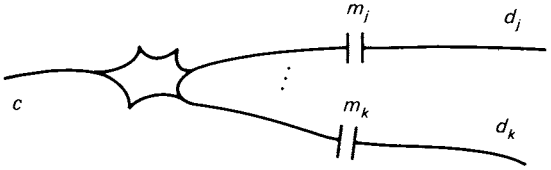
Fig. 1.2. The cell output is a non-linear function of $m \cdot d$. In the linear region, we can write $c = \Sigma_{j=1}^{N} m_j d_j = m \cdot d$ where m_j is a somewhat idealized 'synaptic junction'.



Let me describe the basic properties of the system. Suppose you have a two-dimensional system d^1 and d^2, that is, two inputs, so that the space of inputs can be spanned by these two vectors. You will then have four fixed points. One obtains a set of non-linear coupled differential equations. The non-linearity arises because all the ms are coupled to one another through the cs. If you have only two inputs d^1 and d^2, you look for fixed points in the space. Now the fixed points occur when ϕ is zero. Recall that ϕ is zero when the output is zero, or when the output is at threshold. If you count them up, this gives four fixed points for two dimensions. These four fixed points have different properties. We call m_1^* and m_2^* the selective fixed points. Why selective? It is because if the synapses acquire those strengths, they give you maximum output for one of the inputs and zero output for the other. And it is very easy to see geometrically, because if m_1^* lies perpendicular to d^2, when d^2 comes in you get zero. When d^1 comes in, you get a response. When m_2^* is perpendicular to d^1, you get zero for d^1 and a response for d^2. This is what Kohonen called an optimal mapping; d^1 and d^2 do not have to be orthogonal. In this respect it is a self-organizing optimal mapping. There is another fixed point, which is non-selective because it responds to both d^1 and d^2, and yet another fixed point at zero (in other words the cell does not respond at all). Now the interesting and important question is: which fixed points are stable? The answer is that the only stable fixed points are the selective fixed points. So that wherever you start in the synaptic space, if you keep putting in d^1s and d^2s you eventually end up at the selective fixed points.

Note that in order to get selectivity the synaptic strengths have both positive and negative values, but coming into the cortical cell from the eye, one encounters only excitatory synapses. If we limit synaptic

Fig. 1.3. Neuron learning proceeds through synaptic modification, and an important question is what is the magnitude of $\dot{m}_j$? In general we might write $\dot{m}_j = F(d_j, \ldots, m_j; d_k, \ldots, c; \bar{c}, \ldots; X, Y, Z)$ where the four quantities in parentheses refer to the local instantaneous; quasi-local instantaneous; time averaged; and global contributions, respectively. In the BCM modification, this equation becomes: $\dot{m}_j = \phi(c, \bar{c}; X, Y, Z)\, d_j - \varepsilon m$.

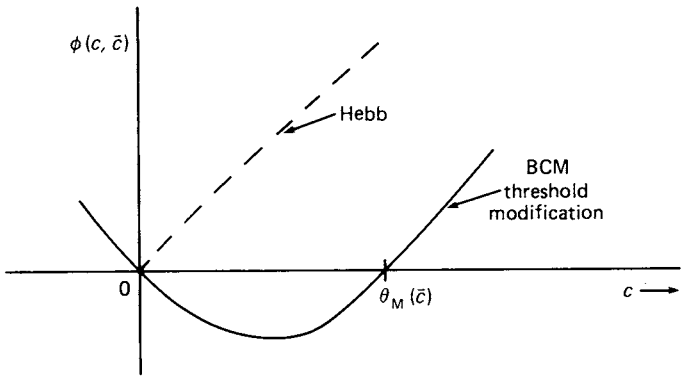


strengths to positive values, we would have partial but not complete selectivity. There are experimental results that show, if you shut off the inhibitory synapses by adding a chemical such as bicuculin, some of the selectivity is lost. This problem will be addressed later.

We can now model various experimental situations. We can have inputs from the left eye and the right eye, and then the output is a summation of left eye and right eye. We can have a normal environment in which we have patterned input to both eyes or, to contrast this, we can have monocular deprivation, patterns to one eye, noise to the other, and then we can run these simulations. We get results that correspond to the classical experimental results. For example, if in normal rearing, with both eyes open, the final stable state is at the selective fixed points in which the cells are binocular and selective, driven by the same pattern. If the eyes are closed, there is no development of selectivity (lid sutured or dark reared) and the cell is binocularly driven. For one eye open, the other eye closed – the very famous paradigm of monocular deprivation – selectivity develops to the open eye, while the closed eye is driven to zero. One instance in which the moving threshold is clearly necessary is in the situation of reverse suture. To get recovery, the threshold must move to a very low value.

In addition, there is a rather subtle connection between occular dominance and selectivity. How is it that, when both eyes are closed, the cell is not necessarily driven to zero, while when one eye is open and the other

Fig. 1.4. The modification threshold, θ_m, is a non-linear function of $\bar{c}$, the average output of the cell. We have used $\theta_m = (1/c_0)(\bar{c})^2$, but the precise form is not critical.



eye is closed, the response of the cell is driven to zero for the closed eye? From our point of view, the reason that it is not driven to zero when both eyes are closed is that the zero fixed point is unstable. But if one eye is closed and the other is open, once the open eye has become selective, the response of the cell is either close to zero or close to threshold. The ϕ function can be expanded close to zero and close to threshold and, depending on whether the input to the open eye is preferred or non-preferred, the appropriate expansion is at one point or the other. Now this expansion eventually results in a differential equation for the synapses between LGN and the closed eye that looks like this: $\dot{x}$ is plus or minus the noise squared times x, depending on whether the input to the open eye is preferred or non-preferred. Non-preferred inputs can only be achieved after the eye has become selective. In other words, before selectivity occurs, there is no driving of the closed eye to zero. And this gives you a correlation between occular dominance and selectivity.

If one looks at some of the original results of Hubel and Wiesel, one can already see this correlation suggested. Recent experiments show a clear correlation between occular dominance and selectivity, and this is precisely what the theory predicts.

This theory has been extended to include the situation in which there are many cells; in other words, where this is input from LGN to excitatory and inhibitory cortical cells and the excitatory and inhibitory cells are connected to one another via intracortical connections. We would like to restrict m, that is to say the inputs between LGN and the cortical cells, to positive values, since these synapses are excitatory.

The output of the jth cell is c_j, and we have an intracortical synapse L_{ij}. We should state that the cell firing rate involves not only being pushed through LGN by the inputs from the right eye and left eye, but also the intracortical connections; that of course is a rather complicated problem which was analysed by myself and a former student, Chris Scofield. We separated excitatory and inhibitory synapses and got some very nice, new predictions. The analysis was complex and we had to rely on extensive computer simulation. This led us to introduce what we call a 'mean field approximation' for the cortical network. Those familiar with theories of magnetism will immediately see where this comes from. The essential idea is that the typical cell gets specific input from LGN and also from large numbers of cortical cells. The other cortical cells are also often pushed by LGN. It is well known that the collaterals can be fairly long, so we replace the effect of many cortical cells on the cell we are watching by an average. In doing that, we simplify things enormously, and finally we

get consistency conditions. The cell i is pushed by the LGN inputs, and then there is a kind of modulatory effect, an average field of the rest of the network. And this average field of the rest of the network is of course also pushed by the LGN inputs.

Very simply stated, previously we have $c = md$, whereas now we have $c = (m - \alpha)d$, and previously we said that m goes to certain fixed points m^*. The fundamental theorem that can be proved is that in the mean field theory, m goes to fixed points that are $m^* + \alpha$.

We can show that all the old fixed points are fixed points here, and that the stability of the fixed points is the same. Thus what was stable before is stable here. Recall that the problem previously was that in the m space one had to find a fixed point, which was inaccessible because of the necessity for negative components. Now, if α is sufficiently inhibitory, this effectively translates the co-ordinates, and all of the values of m^* are now available. So the fixed points will be available with excitatory synapses between LGN and cortical cells if the average inhibition of the network is sufficient.

A question that arises in learning theory is: do all synapses modify in the same way? We do not know the answer, of course, but we have been able to get away with assuming that we have modification between the LGN cortical synapses, and that there is no modification whatsoever among the inhibitory intracortical synapses. This makes the theory very beautiful and very easy to handle.

I would like to summarize some of these ideas. The first point I have already made: the fixed points of the old theory now becomes available, assuming only LGN-cortical excitatory synapses if the network is sufficiently inhibitory. We find that most learning can occur in the LGN cortical synapses. The inhibitory GABAergic cortical–cortical synapses need not modify at all. An experiment was done by Bear and Ebner, at Brown, testing how these GABAergic synapses change under severe conditions of monocular deprivation. They were disappointed to find no such changes. But this was a most welcome result; the fact that we can get away without much inhibitory modification opens the wonderful possibility that the major modification is just for the excitatory LGN cortical synapses. This makes it a much easier problem to treat, mathematically. Some non-modifiable LGN cortical synapses are required. An obvious candidate for these are the synapses onto the inhibitory cells, which go onto shafts rather than spines. It has been seen, in a preliminary experiment by Singer, that the synapses onto the shafts seem to be more resistant, as he put it, than those onto spines.

One of the interesting new results of this theory is that, in binocular deprivation zero is still an unstable fixed point, but the zero is now constructed in the following way: the zero output of the cell comes both from the LGN cortical synapses and from the network synapses which we have taken to be inhibitory. Suppose we suppressed the inhibitory network synapses; then we might expect that cells we could not previously see would suddenly emerge. Such a result has in fact been obtained. Freeman has also shown that an increase in excitability causes cells to appear where they weren't otherwise seen. One of the most interesting results occurs in monocular deprivation. Recall from the previous analysis, that in monocular deprivation the closed eye response goes to zero. And remember that α was previously zero, so that the closed eye result goes to zero only if the closed-eye LGN cortical synapses went to zero. In the mean field theory, the closed-eye response also goes to zero; however that means that the LGN-cortical synapses do not go to zero, but rather, go to α. Therefore, in this theory, we get the monocular deprivation results, but we get them without the LGN–cortical synapses going to zero. So that if inhibition is suppressed, we should get some response for the closed eye. This is in agreement with a result of Sileto and others in which they used bicuculin, which shuts off the inhibitory response, and found an increase in cells responsive to the closed eye. This could be further investigated by post-stimulus histogram, which reveals separate excitatory and inhibitory effects. This makes it possible, even in cells giving an average output of zero, to see excitatory and inhibitory effects, so one can separate the effects of excitatory and inhibitory cells.

In a molecular model we need, among other things, a candidate for the modification threshold, so that the synapses increase above θm, and decrease below θm. Further, θm must vary with the average activity of the cell. Such ideas are being developed actively now by Bear, Ebner and myself at Brown. We are pursuing the idea of using the distinction between the NMDA receptors and the non-NMDA receptors. On the post-synaptic membrane, the NMDA receptors are ones that allow calcium to enter. We are trying to link this with the threshold, θm, and are trying to determine if that threshold varies with the previous experience of the cell.

In summary, what I would like to say is that we propose that we have a theoretical account of the way a little piece of cortex can evolve with experience. There are some dramatic simplifications that would be wonderful, if true. The theory does seem to account for a large variety of